

MDPI
Selections

**ROBOTS, A.I. AND THE FUTURE OF
LAW AND SOCIETY. Teresa Da Cunha
Lopes (Coord.)**

Selected articles published by MDPI

**ROBOTS, A.I. AND THE FUTURE OF
LAW AND SOCIETY. Teresa Da Cunha
Lopes (Coord.)**

ROBOTS, A.I. AND THE FUTURE OF LAW AND SOCIETY. Teresa Da Cunha Lopes (Coord.)

Selected Articles Published by MDPI

MDPI • Basel • Beijing • Wuhan • Barcelona • Belgrade • Manchester • Tokyo • Cluj • Tianjin



This is a reprint of articles published online by the open access publisher MDPI (available at: www.mdpi.com). The responsibility for the book's title and preface lies with Teresa Da Cunha Lopes, who compiled this selection.

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

LastName, A.A.; LastName, B.B.; LastName, C.C. Article Title. <i>Journal Name</i> Year , <i>Volume Number</i> , Page Range.
--

© 2022 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license, which allows users to download, copy and build upon published articles, as long as the author and publisher are properly credited, which ensures maximum dissemination and a wider impact of our publications.

Contents

Preface to "ROBOTS, A.I. AND THE FUTURE OF LAW AND SOCIETY. Teresa Da Cunha Lopes (Coord.)"	vii
Spyros G. Tzafestas Roboethics: Fundamental Concepts and Future Prospects Reprinted from: <i>Information</i> 2018 , 9, 148, doi:10.3390/info9060148	1
Jaana Leikas, Raija Koivisto and Nadezhda Gotcheva Ethical Framework for Designing Autonomous Intelligent Systems Reprinted from: <i>J. Open Innov. Technol. Mark. Complex.</i> 2019 , 5, 18, doi:10.3390/joitmc5010018	27
Alfred Benedikt Brendel, Milad Mirbabaie, Tim-Benjamin Lembcke and Lennart Hofeditz Ethical Management of Artificial Intelligence Reprinted from: <i>Sustainability</i> 2021 , 13, 1974, doi:10.3390/su13041974	39
Ugo Pagallo Vital, Sophia, and Co.—The Quest for the Legal Personhood of Robots Reprinted from: <i>Information</i> 2018 , 9, 230, doi:10.3390/info9090230	57
Luis Raúl Rodríguez Oconitrillo, Juan José Vargas, Arturo Camacho, Álvaro Burgos and Juan Manuel Corchado RYEL: An Experimental Study in the Behavioral Response of Judges Using a Novel Technique for Acquiring Higher-Order Thinking Based on Explainable Artificial Intelligence and Case-Based Reasoning Reprinted from: <i>Electronics</i> 2021 , 10, 1500, doi:10.3390/electronics10121500	69
Juin-Hao Ho, Gwo-Guang Lee and Ming-Tsang Lu Exploring the Implementation of a Legal AI Bot for Sustainable Development in Legal Advisory Institutions Reprinted from: <i>Sustainability</i> 2020 , 12, 5991, doi:10.3390/su12155991	109
Marta Bistrón and Zbigniew Piotrowski Artificial Intelligence Applications in Military Systems and Their Influence on Sense of Security of Citizens Reprinted from: <i>Electronics</i> 2021 , 10, 871, doi:10.3390/electronics10070871	127
Steven Umbrello and Nathan Gabriel Wood Autonomous Weapons Systems and the Contextual Nature of <i>Hors de Combat</i> Status Reprinted from: <i>Information</i> 2021 , 12, 216, doi:10.3390/info12050216	147
Deepti Mishra, Karen Parish, Ricardo Gregorio Lugo and Hao Wang A Framework for Using Humanoid Robots in the School Learning Environment Reprinted from: <i>Electronics</i> 2021 , 10, 756, doi:10.3390/electronics10060756	159
Daniele Giansanti The Social Robot in Rehabilitation and Assistance: What Is the Future? Reprinted from: <i>Healthcare</i> 2021 , 9, 244, doi:10.3390/healthcare9030244	171
Brian Pickering Trust, but Verify: Informed Consent, AI Technologies, and Public Health Reprinted from: <i>Future Internet</i> 2021 , 13, 132, doi:10.3390/fi13050132	181

Maria Maniadaki, Athanasios Papathanasopoulos, Lilian Mitrou and Efpraxia-Aithra Maria
 Reconciling Remote Sensing Technologies with Personal Data and Privacy Protection in the
 European Union: Recent Developments in Greek Legislation and Application Perspectives in
 Environmental Law
 Reprinted from: *Laws* **2021**, *10*, 33, doi:10.3390/laws10020033 **201**

Javier Díaz-Noci
 Artificial Intelligence Systems-Aided News and Copyright: Assessing Legal Implications for
 Journalism Practices
 Reprinted from: *Future Internet* **2020**, *12*, 85, doi:10.3390/fi12050085 **221**

Thomas Hoffmann and Gunnar Prause
 On the Regulatory Framework for Last-Mile Delivery Robots
 Reprinted from: *Machines* **2018**, *6*, 33, doi:10.3390/machines6030033 **231**

Preface to "ROBOTS, A.I. AND THE FUTURE OF LAW AND SOCIETY. Teresa Da Cunha Lopes (Coord.)"

This book deals with several issues and problems within the field of Robotics, Artificial Intelligence and Law. In fact, it attempts to address the "dangerous liaisons" between them and their implications for modern Societies. Its structure is designed to function as a work and consultation tool for the students of the Master of Law of the UMSNH in the seminars on Law and New Technologies and Robots, A.I, Human Rights and Transhumanism.

We are currently witnessing a revolution in production systems that comes as a consequence of the massification of automation and of the presence of autonomous robots and AI (artificial intelligence) in all productive sectors, in the field of war and security, and at all social levels.

The distinctive technological changes of the knowledge society and of the 4th globalization, not only affect the redefinition of business organizational models for competitiveness in the market; They also transform the behavior of individuals, their work relationships, the role of the State as regulator and the world of work, but also pose bio-legal problems on the definition of the boundaries between the human and the machine (transhumanism).

Consequently, it is necessary and urgent to open an ethical-legal debate on issues as important as the legal personality of robots with artificial intelligence to respond to the regulatory challenges of an already present future, given that current legal frameworks are not prepared. to give a direct answer to new technological contexts and their rapid penetration into our societies.

However, the distinctive technological changes of societies and the knowledge economy,"not only affect the ways of producing mass consumer goods and the redefinition of business organizational models for market competitiveness; they also transform the behavior of individuals, their work relationships, the role of the State as regulator and the world of work as a whole"(Da Cunha Lopes et Alli: 2013).

Of course, these developments raise questions. The consequences in employment are worrying, those of legal responsibility in case of error seem to have no answer. Not to mention the protection of privacy against these robots capable of seeing everything, listening to everything, predicting everything (or almost), and sending the data collected on the servers of companies that we do not always know what they are going to do.

Faced with this massification of robots and intelligent algorithms among us, the individual asks himself questions: Is the human being threatened by technology? Can the machine master it? Where does the cyborg end and the transhuman begin?

Therefore, we need to exercise robots to accurately identify and assess the ethical aspects of a given situation (such as the existence of potential benefits or harm to a human being). Consequently, we need to instill in machines the duty to act appropriately (that is, to maximize those benefits and minimize those damages). That it is urgent to place ethical questions that will have to be inscribed in the code of the machines, but it is also urgent to design a legal architecture that frames the complex problems of responsibility. This involves a reflection on the general question of the "legal personality" of robots with Artificial Intelligence and by specific proposals for jurisdiction or the expansion of the concept of "personhood".

It is also necessary, given the new challenges imposed by this changing reality, to highlight the


importance of revaluing ethical principles, as foundations of the legal system, in solving problems, of which the values shared by society with constitutional support appear with their potential to allow a technical and instrumental solution in creating the legitimacy of behaviors within the scope of artificial intelligence systems.

Finally, it is the purpose of this book to build an overview of current debates in the fields of Robotics, of the A.I. and of a Law for the XXI century. Artificial intelligence (AI) (artificial intelligences) is at the center of an intense institutional reflection, the objective of which is to identify the principles that would protect people from the potentially negative effects of the development of these technologies. In recent years, several institutions have investigated the issue of the legal framework for this technological development, perceived as a need to adapt the legislation. In fact, this reflection questions the categories and traditional legal mechanisms. However, this is not the first time that the law has faced the reception of a completely new technique or practice, and it may be interesting to review the way in which previous developments have been received to learn.

Teresa Da Cunha Lopes

Review

Roboethics: Fundamental Concepts and Future Prospects

Spyros G. Tzafestas 

School of Electrical and Computer Engineering, National Technical University of Athens, Zographou, GR 15773 Athens, Greece; tzafesta@cs.ntua.gr

Received: 31 May 2018; Accepted: 13 June 2018; Published: 20 June 2018

Abstract: Many recent studies (e.g., IFR: International Federation of Robotics, 2016) predict that the number of robots (industrial, service/social, intelligent/autonomous) will increase enormously in the future. Robots are directly involved in human life. Industrial robots, household robots, medical robots, assistive robots, sociable/entertainment robots, and war robots all play important roles in human life and raise crucial ethical problems for our society. The purpose of this paper is to provide an overview of the fundamental concepts of robot ethics (roboethics) and some future prospects of robots and roboethics, as an introduction to the present Special Issue of the journal *Information* on “Roboethics”. We start with the question of what roboethics is, as well as a discussion of the methodologies of roboethics, including a brief look at the branches and theories of ethics in general. Then, we outline the major branches of roboethics, namely: medical roboethics, assistive roboethics, sociorobot ethics, war roboethics, autonomous car ethics, and cyborg ethics. Finally, we present the prospects for the future of robotics and roboethics.

Keywords: ethics; roboethics; technoethics; robot morality; sociotechnical system; ethical liability; assistive roboethics; medical roboethics; sociorobot ethics; war roboethics; cyborg ethics

1. Introduction

All of us should think about the ethics of the work/actions we select to do or the work/actions we choose not to do. This includes the work/actions performed through robots which, nowadays, strongly affect our lives. It is true that as technology progresses, the function of robots is upgrading from that of a pure tool to a sociable being. As a result of this social involvement of present-day robots, in many cases the associated social practices are likely to change. The question is how to control the direction in which this will be done, especially from an ethics point of view. Many scholars in the fields of intelligent systems, artificial intelligence, and robotics anticipate that in the near future there will be a strong influence of cultural and societal values and norms on the development of robotics, and conversely an influence of robot cultural values on human beings [1]. This means that social and cultural factors (norms, morals, beliefs, etc.) affect the design, operation, application, use, and evaluation of robots and other technologies. Overall, the symbiosis of humans and robots will reach higher levels of integration and understanding.

Roboethics is a fundamental requirement for assuring a sustainable, ethical, and profitable human-robot symbiosis. Roboethics belongs to technoethics, which was initiated by Jose Maria Galvan via his talk about the “ethical dimension of technology” in the Workshop on “Humanoids: A Techno-ontological Approach” (IEEE Robotics and Automation Conference on Humanoid Robots, Waseda University, 2001) [2]. Today, there are many books, conference proceedings, and journal Special Issues on roboethics (e.g., [3–13]).

Three influential events on roboethics that took place in the initial period of the field are:

- 2004: First Roboethics International Symposium (Sanremo, Italy).
- 2005: IEEE Robotics and Automation Society Roboethics Workshop: ICRA 2005 (Barcelona, Spain).
- 2006: Roboethics Minisymposium: IEEE BioRob 2006—Biomedical Robotics and Biomechatronics Conference (Pisa, Italy).

Other conferences on roboethics, or involving workshops or tracks on roboethics, held in the period of 2006–2018 include:

- 2006: ETHICBOTS European Project International Workshop on Ethics of Human Interaction with Robotic, Bionic, and AI Systems Concepts and Policies (Naples, October 2006).
- 2007: ICRA: IEEE R&A International Conference: Workshop on Roboethics: IEEE Robotics and Automation Society Technical Committee (RAS TC) on Roboethics (Rome, Italy).
- 2007: ICAIL 2007: International Conference on Artificial Intelligence and Law (Palo Alto, USA, 4–6 June 2007).
- 2007: CEPE 2007: International Symposium on Computer Ethics Philosophical Enquiry (Topic Roboethics) (San Diego, USA, 12–14 July 2007).
- 2009: ICRA: IEEE R&A International Conference on Robotics and Automation: Workshop on Roboethics: IEEE RAS TC on Roboethics (Kobe, Japan, 2009).
- 2012: We Robot, University of Miami, FL, USA.
- 2013: International Workshop on Robot Ethics, University of Sheffield (February 2013).
- 2016: AAAI/Stanford Spring Symposium on Ethical and Moral Considerations in Non-Human Agents.
- 2016: International Research Conference on Robophilosophy (Main Topic Roboethics), Aarhus University (17–21 October 2016).
- 2018: International Conference on Robophilosophy: Envisioning Robots and Society (Main Topic Roboethics) (Vienna University, 14–17 February 2018).

In 2004 (25 February), the Fukuoka World Robot Declaration was issued (Fukuoka, Japan), which included the following statement [14]:

“Confident of the future development of robot technology and of the numerous contributions that robots will make to Humankind, this World Robot Declaration is Expectations for next-generation robots: (a) next-generation robots will be partners that co-exist with human beings; (b) next-generation robots will assist human beings both physically and psychologically; (c) next-generation robots will contribute to the realization of a safe and peaceful society”.

Clearly, this declaration tacitly promises that next-generation robots will be designed and used in an ethical way for the benefit of human society.

An important contributor for the progress and impact of robotics of the future is the European Robotics Research Network (EURON), which aims to promote excellence in robotics by creating resources and disseminating/exchanging existing knowledge [14]. A major achievement of EURON is the creation of a “Robotics Research Roadmap” that identifies and clarifies opportunities for developing and exploiting advanced robot technology over a 20-year time frame in the future. A second product of EURON is the “Roboethics Atelier”, a project funded and launched in 2005, with the goal to draw the first “Roboethics Roadmap”. By now, this roadmap has embodied contributions of a large number of scholars in the fields of sciences, technology, and humanities. The initial target of the “Roboethics Roadmap” was the ethics of robot designers, manufacturers, and users.

It is emphasized that for roboethics to be assured, the joint commitment of experts of different disciplines (electrical/mechanical/computer engineers, control/robotics/automation engineers,

psychologists, cognitive scientists, artificial intelligence scientists, philosophers/ethicists, etc.) to design ethics-based robots, and adapt the legislation to the issues (technological, ethical) that arise from the continuous advances and achievements of robotics, is required.

The purpose of this paper is to present the fundamental concepts of roboethics (robot ethics) and discuss some future perspectives of robots and roboethics. The structure of the paper is as follows:

- Section 2 analyzes the essential question: What is roboethics?
- Section 3 presents roboethics methodologies, starting with a brief review of ethics branches and theories.
- Section 4 outlines the roboethics branches, namely: medical roboethics, assistive roboethics, sociorobot ethics, war roboethics, autonomous car ethics, and cyborg ethics.
- Section 5 discusses some prospects for the future of robotics and roboethics.
- Section 6 gives the conclusions.

2. What Is Roboethics?

Roboethics is a modern interdisciplinary research field lying at the intersection of applied ethics and robotics, which studies and attempts to understand and regulate the ethical implications and consequences of robotics technology, particularly of intelligent/autonomous robots, in our society. The primary objective of roboethics is to motivate the moral design, development, and use of robots for the benefit of humanity [5]. The term roboethics (for robot ethics) was coined by Gianmarco Verrugio, who defines the field in the following way [2]:

“Roboethics is an applied ethics whose objective is to develop scientific/cultural/technical tools that can be shared by different social groups and beliefs. These tools aim to promote and encourage the development of robotics for the advancement of human society and individuals, and to help preventing its misuse against humankind”.

To embrace a wide range of robots and potential robotic applications, Veruggio classified roboethics in three levels as follows [2]:

- **Level 1:** Roboethics—This level is intrinsically referred to philosophical issues, humanities, and social sciences.
- **Level 2:** Robot Ethics—This level refers mainly to science and technology.
- **Level 3:** Robot’s Ethics—This level mostly concerns science fiction, but it opens a wide spectrum of future contributions in the robot’s ethics field.

The basic problems faced by roboethics are: the dual use of robots (robots can be used or misused), the anthropomorphization of robots (from the Greek words *άνθρωπος* (anthropos) = human, and *μορφή* (morphe) = shape), the humanization (human-friendly making) of human-robot symbiosis, the reduction of the socio-technological gap, and the effect of robotics on the fair distribution of wealth and power [1,2]. During the last three or four decades, many scholars working in a variety of disciplines (robotics, computer science, information technology, automation, philosophy, law, psychology, etc.) have attempted to address the pressing ethical questions about creating and using robotic technology in society. Many areas of robotics are impacted, particularly those where robots interact directly with humans (assistive robots, elder care robots, sociable robots, entertainment robots, etc.). The area of robotics which raises the most crucial ethical concerns is the area of military/war robots, especially autonomous lethal robots [3,7,15]. Several prominent robotics researchers and professionals began visibly working on the problem of making robots ethical. There are also many computer and artificial intelligence scholars who have argued that robots and AI will one day take over the world. However, many others, e.g., Roger K. Moore, say that this is not going to happen. According to him the problem is not the robots taking over the world, but that some people want to pretend that robots are responsible for themselves [16]. He says: “In fact, robots belong to us. People, companies, and governments

build, own, and program robots. Whoever owns and operates a robot is responsible for what he does". Actually, roboethics has several common problems with computer ethics, information ethics, automation ethics, and bioethics.

According to Peter M. Asaro [17], the three fundamental questions of roboethics are the following:

1. "How might humans act ethically through, or with, robots?
2. How can we design robots to act ethically? Or, can robots be truly moral agents?
3. How can we explain the ethical relationships between human and robots?"

In question 1, it is humans that are the ethical agents. In question 2, it is robots that are the ethical beings. Sub-questions of question 3 include the following [5]:

- "Is it ethical to create artificial moral agents and ethical robots?
- Is it unethical not to design mental/intelligent robots that possess ethical reasoning abilities?
- Is it ethical to make robotic nurses or soldiers?
- What is the proper treatment of robots by humans, and how should robots treat people?
- Should robots have rights?
- Should moral/ethical robots have new legal status?"

Very broadly, scientists and engineers look at robotics in the following ways [5,11]:

- Robots are mere machines (albeit, very useful and sophisticated machines).
- Robots raise intrinsic ethical concerns along different human and technological dimensions.
- Robots can be conceived as moral agents, not necessarily possessing free will mental states, emotions, or responsibility.
- Robots can be regarded as moral patients, i.e., beings deserving of at least some moral consideration.

To formulate a sound framework of roboethics, all of the above questions/aspects (at minimum) must be properly addressed. Now, since humans and robots constitute a whole sociotechnical system, it is not sufficient to concentrate on the ethical performance of individual humans and robots, but the entire assembly of humans and robots must be considered, as dictated by system and cybernetics theory [5,18]. The primary concern of roboethics is to assure that a robot or any other machine/artifact is not doing harm, and only secondarily to specify the moral status of robots, resolve human ethical dilemmas, or study ethical theories. This is because as robots become more sophisticated, intelligent, and autonomous it will become more necessary to develop more advanced robot safety control measures and systems to prevent the most critical dangers and potential harms. Of course it should be remarked here that the dangers for robots do not differ from the dangers of other artifacts, such as factories, chemical processes, automatic control systems, weapons, etc. At minimum, moral/ethical robots need to have: (i) the ability to predict the results of their own actions or inactions; (ii) a set of ethical rules against which to evaluate each possible action/consequence; and (iii) a mechanism for selecting the most ethical action.

Roboethics involves three levels, namely [11]:

1. The ethical theory or theories adopted.
2. The code of ethics embedded into the robot (machine ethics).
3. The subjective morality resulting from the autonomous selection of ethical action(s) by a robot equipped with a conscience.

The three primary views of scientists and engineers about roboethics are the following [5,19]:

- Not interested in roboethics: These scholars say that the work of robot designers is purely technical and does not imply an ethical or social responsibility for them.

- Interested in short-term robot ethical issues: This view is advocated by those who adopt some social or ethical responsibility, by considering ethical behavior in terms of good or bad, and short-term impact.
- Interested in long-term robot ethical issues: Robotics scientists advocating this view express their robotic ethical concern in terms of global, long-term impact and aspects.

Some questions that have to be addressed in the framework of roboethics are [5]:

- Is ethics applied to robots an issue for the individual scholar or practitioner, the user, or a third party?
- What is the role that robots could have in our future life?
- How much could ethics be embedded into robots?
- How ethical is it to program robots to follow ethical codes?
- Which type of ethical codes are correct for robots?
- If a robot causes harm, is it responsible for this outcome or not? If not, who or what is responsible?
- Who is responsible for actions performed by human-robot hybrid beings?
- Is the need to embed autonomy in a robot contradictory to the need to embed ethics in it?
- What types of robots, if any, should not be designed? Why?
- How do robots determine what is the correct description of an action?
- If there are multiple rules, how do robots deal with conflicting rules?
- Are there any risks to creating emotional bonds with robots?

3. Roboethics Methodologies

Roboethics methodologies are developed adopting particular ethics theories. Therefore, before discussing these methodologies, it is helpful to have a quick look at the branches and theories of ethics.

3.1. Ethics Branches

Ethics involves the following branches [5] (Figure 1):

- **Meta-ethics.** The study of concepts, judgements, and moral reasoning (i.e., what is the nature of morality in general, and what justifies moral judgements? What does right mean?).
- **Normative (prescriptive) ethics.** The elaboration of norms prescribing what is right or wrong, what must be done or what must not (What makes an action morally acceptable? Or what are the requirements for a human to live well? How should we act? What ought to be the case?).
- **Applied ethics.** The ethics branch which examines how ethics theories can be applied to specific problems/applications of actual life (technological, environmental, biological, professional, public sector, business ethics, etc., and how people take ethical knowledge and put it in practice). Applied ethics is actually contrasted with theoretical ethics.
- **Descriptive ethics.** The empirical study of people's moral beliefs, and the question: What is the case?



Figure 1. Branches of ethics. Source: [https://commons.wikimedia.org/wiki \(/File:Ethics-en.svg\)](https://commons.wikimedia.org/wiki/File:Ethics-en.svg).

3.2. Ethics Theories

Principal ethics theories are the following [5]:

- **Virtue theory (Aristotle).** The theory grounded on the notion of virtue, which is specified as what character a person needs to live well. This means that in virtue ethics the moral evaluation focuses on the inherent character of a person rather than on specific actions.
- **Deontological theory (Kant).** The theory that focuses on the principles upon which the actions are based, rather than on the results of actions. In other words, moral evaluation carries on the actions according to imperative norms and duties. Therefore, to act rightly one must be motivated by proper universal deontological principles that treat everyone with respect (“respect for persons theory”).
- **Utilitarian theory (Mill).** A theory belonging to the consequentialism ethics which is “teleological”, aiming at some final outcome and evaluating the morality of actions toward this desired outcome. Actually, utilitarianism measures morality based on the optimization of “net expected utility” for all persons that are affected by an action or decision. The fundamental principle of utilitarianism says: “Actions are moral to the extent that they are oriented towards promoting the best long-term interests (greatest good) for every one concerned”. The issue here is what the concept of greatest good means. The Aristotelian meaning of greatest good is well-being (pleasure or happiness).

Other ethics theories include value-based theory, justice as fairness theory, and case-based theory [5]. In real-life situations it is sometimes more effective to combine ethical rules of more than one ethics theory. This is so because in a dynamic world it is very difficult and even impossible to cover every possible situation by the principles and rules of a unique ethics theory.

3.3. Roboethics Methodologies

Roboethics has two basic methodologies: top-down methodology and bottom-up methodology [5,20,21].

- **Top-down roboethics methodology.** In this methodology, the rules of the desired ethical behavior of the robot are programmed and embodied in the robot system. The ethical rules can be

formulated according to the deontological or the utilitarian theory or other ethics theories. The question here is: which theory is the most appropriate in each case? Top-down methodology in ethics was originated from several areas including philosophy, religion, and literature. In control and automation systems design, the top-down approach means to analyze or decompose a task in simpler sub-tasks that can be hierarchically arranged and performed to achieve a desired output or product. In the ethical sense, following the top-down methodology means to select an antecedently specified ethical theory and obtain its implications for particular situations. In practice, robots should combine both meanings of the top-down concept (control systems meaning and ethical systems meaning).

Deontological roboethics: The first deontological robotic ethical system was proposed by Asimov [22] and involves the following rules, which are known as Asimov's Laws [5,22]:

- **Law 1:** A robot may not injure a human being or, through inaction allow a human being to come to harm.
- **Law 2:** A robot must obey orders it receives from human beings except when such orders conflict with Law 1.
- **Law 3:** A robot must protect its own existence as long as such protection does not conflict with Laws 1 and 2."

Later, Asimov added a law which he called Law Zero, since it has a higher importance than Laws 1 through 3. This law states:

- **Law 0:** No robot may harm humanity or through inaction allow humanity to come to harm."

Asimov's laws are human-centered (anthropocentric) since they consider the role of robots in human service. Actually, these laws assume that robots have sufficient intelligence (perception, cognition) to make moral decisions using the rules in all situations, irrespective of their complexity.

Over the years several multi-rule deontological systems have been proposed, e.g., [23,24]. Their conflict problem is faced by treating them as dictating prima facie duties [25].

In Reference [25], it is argued that for a robot to be ethically correct the following conditions (desiderata) must be satisfied [5]:

- "Robots only take permissible actions.
- All relevant actions that are obligatory for robots are actually performed by them, subject to ties and conflicts among available actions.
- All permissible (or obligatory or forbidden) actions can be proved by the robot (and in some cases, associated systems, e.g., oversight systems) to be permissible (or obligatory or forbidden), and all such proofs can be explained in ordinary English".

The above ethical system can be implemented in top-down fashion.

Consequentialist roboethics: As seen above, the morality of an action is evaluated on the basis of its consequences. The best current moral action is the action that leads to the best future consequences.

A robot can reason and act along the consequentialist/utilitarian ethics theory if it is capable to [5]:

- "Describe every situation in the world.
- Produce alternative actions.
- Predict the situation(s) which would be the outcome of taking an action given the present situation.
- Evaluate a situation in terms of its goodness or utility."

The crucial issues here are how "goodness" is defined, and what optimization criterion is selected for evaluating situations.

- **Bottom-up roboethics methodology.** This methodology assumes that the robots possess adequate computational and artificial intelligence capabilities to adapt themselves to different contexts so as to be capable to learn, starting from perception of the world, and then perform the planning of the actions based on sensory data, and finally execute the action [26]. In this methodology, the use of any prior knowledge is only for the purpose of specifying the task to be performed, and not for specifying a control architecture or implementation technique. A detailed discussion of bottom-up and top-down roboethics approaches is provided in Reference [26]. Actually, for a robot to be an ethical learning robot both top-down and bottom-up approaches are needed (i.e., the robot should follow a suitable hybrid approach). Typically, the robot builds its morality through developmental learning similar to the way children develop their conscience. Full discussions of top-down and bottom-up roboethics methodologies can be found in References [20,21].

The morality of robots can be classified into one of three levels [5,21]:

- Operational morality (moral responsibility lies entirely in the robot designer and user).
- Functional morality (the robot has the ability to make moral judgments without top-down instructions from humans, and the robot designers can no longer predict the robot’s actions and their consequences).
- Full morality (the robot is so intelligent that it fully autonomously chooses its actions, thereby being fully responsible for them).

As seen in Figure 2, increasing the robot’s autonomy and ethical sensitivity increases the robot’s level of moral agency.

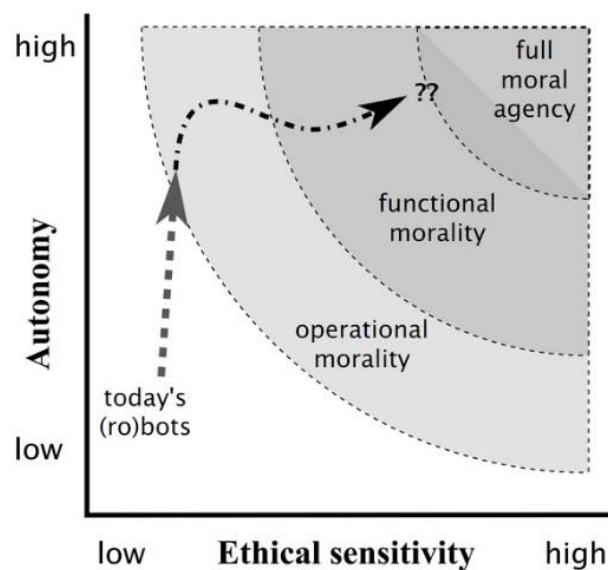


Figure 2. Levels of robot morality (operational, functional, full) embedded in the robot autonomy vs. ethical sensitivity plane. Source: www.wonderfulengineering.com/future-robots-will-have-moral-and-ethical-sense.

4. Roboethics Branches

In the following we will outline the following roboethics branches:

- Medical roboethics.
- Assistive roboethics.
- Sociorobot ethics.
- War roboethics.

- Autonomous car ethics.
- Cyborg ethics.

4.1. Medical Roboethics

Medical roboethics (ethics of medical robots or health care robots) uses the principles of medical ethics and roboethics [5,27,28]. The fundamental area of medical robotics is the area of robotic surgery, which finds increasing use in modern surgery. Robotic surgery has excessive cost. Therefore, the question that immediately rises is [5]: “Given that there is marginal benefit from using robots, is it ethical to impose financial burden on patients or the medical system?”. The critical issue in medical ethics is that the subject of health care and medicine refers to human health, life, and death. Medical ethics deals with ethical norms for the medical or health care practice, or how it must be done. Medical ethics was initiated in ancient Greece by Hippocrates, who formulated the well-known Hippocratic Oath (Ὁρκος του Ιπποκράτη, in Greek) [29].

The principles of medical ethics are based on the general theories of ethics (justice as fairness, deontological, utilitarian, case-based theory), and the fundamental practical moral principles (keep promises, do not interfere with the lives of others unless they request this form of help, etc.) [23,28].

According to the well-known Georgetown Mantra (or six-part medical ethics approach) [30], all medical ethical decisions should involve at least the following principles [7,30]:

- “Autonomy: The patients have the right to accept or refuse a treatment.
- Beneficence: The doctor should act in the best interest of the patient.
- Non-maleficence: The doctor/practitioner should aim “first not to do harm”.
- Justice: The distribution of scarce health resources and the decision of who gets what treatment should be just.
- Truthfulness: The patient should not be lied to and has the right to know the whole truth.
- Dignity: The patient has the right to dignity”.

An authoritative code of ethics is the AMA (American Medical Association) code [31].

Robotic surgery ethics is a sub-area of applied medical ethics, and involves at minimum the above Georgetown Mantra Principles. Medical treatment of any form should be ethical. However, a legal treatment may not be ethical. The legislation provides the minimum law standard for people’s performance. The ethical standards are specified by the principles of ethics and, in the context of licenced professionals (robotics engineers, information engineers, medical doctors, managers, etc.), are provided by the accepted code of ethics of each profession [32,33].

Injury law places on all individuals a duty of reasonable care to others, and determines this duty based on how “a reasonable/rational person” in the same situation would act. If a person (doctor, surgeon, car driver) causes injury to another, because of unreasonable action, then the law imposes liability on the unreasonable person. A scenario concerning the case of injuring a patient in robotic surgery is discussed in Reference [5]. Figure 3 shows a snapshot of the DaVinci robot and its accessories.

A branch of medicine which needs specialized ethical and law considerations is the branch of telemedicine (especially across geographical and political boundaries). Telecare from different countries should obey the standard ethics rules of medicine, e.g., the rules of confidentiality and equipment reliability, while it may reduce the migration of specialists. Confidentiality is at risk because of the possibility of overhearing. Here, the prevention of carelessness in the copying of communications such as diagnoses is necessary, along with the assurance that non-physician intermediaries (e.g., medical technicians or information experts) who collect data about patients respect confidentiality. Communication should be sufficiently fast so as to assure that the ethical requirements of beneficence and justice are met, and to reduce the unpleasant anxiety of the patients. On the legal side, the so-called conflict of laws should be properly faced. A first issue is whether a

medical care professional, who has a licence to practice only in jurisdiction A but treats a patient in jurisdiction B, violates B's laws. Conflict of law principles should be applied here [34].



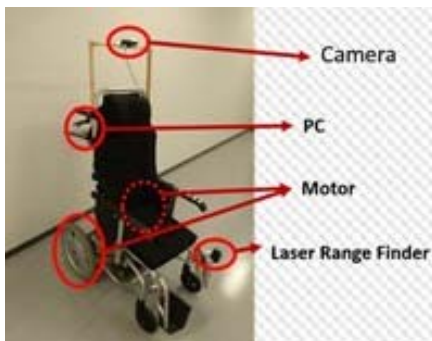
Figure 3. The Da Vinci surgical robot system. Source: [www.montefiore.org \(/cancer-robotic-prostate-surgery\)](http://www.montefiore.org/cancer-robotic-prostate-surgery).

4.2. Assistive Robotics

Assistive robots constitute a class of service robots that focuses on the enhancement of the mobility capabilities of impaired people (people with special needs: PwSN) so as to attain their best physical and/or social functional level, and to gain the ability to live independently [5]. Assistive robots/devices include the following [5]:

- Assistive robots/devices for people with impaired lower limbs (wheelchairs, walkers).
- Assistive robots/devices for people with impaired upper limbs and hands.
- Rehabilitation robots/devices for upper limbs or lower limbs.
- Orthotic devices.
- Prosthetic devices.

Figure 4a shows the principal components of the Toyama University's intelligent/self-navigated wheelchair, and Figure 4b shows the McGill University's multi-task smart/intelligent wheelchair (smart wheeler).



(a)



(b)

Figure 4. (a) An intelligent wheelchair example with motor, PC, camera, and laser range sensor. (b) Smart multi-task wheelchair (McGill SmartWheeler Project). (a) Source: www3.u-toyama.ac.jp/mecha0/lab/mechacontr/res_ENG.html (b) www.cs.mcgill.ca/~smartwheeler.

The evaluation of assistive robots can be made along three main dimensions, namely: cost, risk, and benefit. Since these evaluation dimensions trade off against each other we cannot achieve full points on all of them at the same time. Thus, their quantitative evaluation and the trade-off among the different dimensions is needed. The evaluation of risk-benefit and cost-benefit should be conducted in light of the impact of assistive technologies on users' whole life in both the short term and the long term. Important guidelines for these analyses have been provided by the World Health Organization (WHO), which has approved an International Classification of Functioning, Disability, and Health (ICF) [35].

A framework for the development of assistive robots using ICF, which includes the evaluation of assistive technologies in users' life, is described in References [36,37]. In the ICF model, assistive robots, besides activity, have impacts on body functions and structure/participation, and the functioning of humans (combined, e.g., with welfare equipment, welfare service, housing environment, etc.).

Assistive robotics is part of medical robotics. Therefore, the principles of medical roboethics (Georgetown Mantra, etc.) and the respective codes of ethics are applicable here. Doctors and caregivers should carefully respect the following additional ethical aspects [5]:

1. Select and propose the most appropriate device which is economically affordable by the PwSN.
2. Consider assistive technology that can help the user do things that he/she finds difficult to do.
3. Ensure that the chosen assistive device is not used for activities that a person is capable of doing for him/herself (which will probably make the problem worse).
4. Use assistive solutions that respect the freedom and privacy of the person.
5. Ensure the users' safety, which is of the greatest importance.

A full code of assistive technology was released in 2012 by the USA Rehabilitation Engineering and Assistive Technology Society (RESNA) [38], and another code by the Canadian Commission on Rehabilitation Councilor Certification (CRCC) was put forth in 2002 [39]. A four-level ethical decision-making scheme for assistive/rehabilitation robotics and other technologies is the following [5]:

- **Level 1:** Select the proper device—Users should be provided the proper assistive/rehabilitation devices and services, otherwise the non-maleficence ethical principle is violated. The principles of justice, beneficence, and autonomy should also be followed at this level.
- **Level 2:** Competence of therapists—Effective co-operation between therapists in order to plan the best therapy program. Here again the principles of justice, autonomy, beneficence, and non-maleficence should be respected.
- **Level 3:** Effectiveness and efficiency of assistive devices—Use should be made of effective, reliable, and cost-effective devices. The principles of beneficence, non-maleficence, etc. should be respected here. Of highest priority at this level is the justice ethical rule.
- **Level 4:** Societal resources and legislation—Societal, agency, and user resources should be appropriately exploited in order to achieve the best available technologies. Best practices rehabilitation interventions should be followed for all aspects.

Level 1 is the "client professional relationship" level, level 2 is the "clinical multidisciplinary" level, level 3 is the "institutional/agency" level, and level 4 is the "society and public policy" level.

4.3. Sociorobot Ethics

Sociorobots (social, sociable, socialized, or socially assistive robots) are assistive robot that are designed to enter the mental and socialization space of humans, e.g., PaPeRo, PARO, Mobiserv, i-Cat and NAO (Figure 5). This can be achieved by designing appropriate high-performance human-robot interfaces: HRI (speech, haptic, visual). The basic features required for a robot to be socially assistive are to [40]:

- Comprehend and interact with its environment.

- Exhibit social behavior (for assisting PwSN, the elderly, and children needing mental/socialization help).
- Direct its focus of attention and communication on the user (so as to help him/her achieve specific goals).

A socially interactive robot possesses the following capabilities [5,40–42]:

- “Express and/or perceive emotions.
- Communicate with high-level dialogue.
- Recognize other agents and learn their models.
- Establish and/or sustain social connections.
- Use natural patterns (gestures, gaze, etc.).
- Present distinctive personality and character.
- Develop and/or learn social competence.”

Some more sociorobots, other than those shown in Figure 5, include the following [40]:

- AIBO: a robotic dog (dogbot) able to interact with humans and play with a ball (SONY) [43].
- KISMET: a human-like robotic head able to express emotions (MIT) [44].
- KASPAR: a humanoid robot torso that can function as mediator of human interaction with autistic children [41].
- QRIO: a small entertainment humanoid (SONY) [45].

Sociorobots are marketed for use in a variety of environments (private homes, schools, elderly centers, hospitals). Therefore, they have to function in real environments which includes interacting with family members, caregivers, and medical therapists [5,40]. Normally, a sociorobot does not apply any physical force on the user, although the user can touch it, often as part of the therapy. However, in most cases no physical user-robot contact is involved, and frequently the robot is not even within the user’s reach. In most cases the robot lies within the user’s social interaction domain in which a one-to-one interaction occurs via speech, gesture, and body motion. Thus, the use of sociorobots raises a number of ethical issues that fall in the psychological, emotional, and social sphere. Of course, since sociorobots constitute a category of medical robots, the principles of medical roboethics discussed in Section 4.1 are all applied here as in the case of all assistive robots. In addition, the following fundamental non-physical (emotional, behavioral) issues should be considered [5]:

- Attachment: The ethical issue here arises when a user is emotionally attached to the robot. For example, in dementia/autistic persons, the robot’s absence when it is removed for repair may produce distress and/or loss of therapeutic benefits.
- Deception: This effect can be created by the use of robots in assistive settings (robot companions, teachers, or coaches), or when the robot mimics the behavior of pets.
- Awareness: This issue concerns both users and caregivers, since they both need to be accurately informed of the risks and hazards associated with the use of robots.
- Robot authority: A sociorobot that acts as a therapist is given some authority to exert influence on the patient. Thus, the ethical issue here is who controls the type, the level, and the duration of interaction. If a patient wants to stop an exercise due to fatigue or pain a human therapist would accept this, but a robot might not accept. Such a feature is also to be possessed by the robot.
- Autonomy: A mentally healthy person has the right to make informed decisions about his/her treatment. If he/she has cognition problems, this autonomy right is passed to the person who is legally and ethically responsible for the patient’s therapy.
- Privacy: Securing privacy during robot-aided interaction and care is a primary requirement in all cases.
- Justice and responsibility: This is of primary ethical importance to observe the standard issues of the “fair distribution of scarce resources” and “responsibility assignment”.

- Human-human relation (HHR): HHR is a very important ethical issue that has to be addressed when using assistive and socialized robots. The robots are used as a means of addition or enhancement of the therapy given by caregivers, not as a replacement of them.”



Figure 5. Examples of sociorobots. (a) PaPeRo: www.materialicious.com/2009/11/communication-robot-papero.html; (b) PARO: www.roboticstoday.com/robots/paro; (c) Mobiserv: www.smart-homes.nl/Innoveren/Sociale-Robots/Mobiserv; (d) i-cat: www.bartneck.de/2009/08/12/photos-philips-icat-robot; (e) NAO: www.hackedgadgets.com/2011/02/18/nao-robot-demonstation.

4.4. War Roboethics

Military robots, especially lethal autonomous robotic weapons, lie at the center of roboethics. Supporters of the use of war robots state that these robots have important advantages which include the saving of the lives of soldiers and the safe clearing of seas and streets from IED (Improvised Explosive Devices). They also claim that autonomous robot weapons can expedite war more ethically and effectively than human soldiers who, under the influence of emotions, anger, fatigue, vengeance, etc., may overreact and overstep the laws of war. The opponents of the use of autonomous killer robots argue that weapon autonomy itself is the problem and the mere control of autonomous weapons would never be satisfactory. Their central belief is that autonomous lethal robots must be entirely prohibited [5].

War is defined as follows (Merriam Webster Dictionary):

- A state or period of fighting between countries or groups.
- A state of usually open and declared armed hostile conflict between states or nations.
- A period of such armed conflict.

A war does not really start until a conscious commitment and strong mobilization of the belligerents occurs. War is a bad thing (it results in deliberate killing or injuring people) and raises

critical ethical questions for any thoughtful person [5]. These questions are addressed by “war ethics”. The ethics of war attempts to resolve what is right or wrong, both for the individual and the states or countries contributing to debates on public policy, and ultimately leading to the establishment of codes of war [46,47]. The three dominating traditions (doctrines) in the ethics of war and peace are [5,48]:

- Realism (war is an inevitable process taking place in the anarchical world system).
- Pacifism or anti-warism (rejects war in favor of peace).
- Just war (just war theory specifies the conditions for judging if it is just to go to war, and conditions for how the war should be conducted).

Realism is distinguished in descriptive realism (the states cannot behave morally in wartime) and prescriptive realism (a prudent state is obliged to act amorally in the international scene). Pacifism objects to killing in general and in particular, and objects to mass killing for political reasons as commonly occurs during wartime. A pacifist believes that war is always wrong.

Just war theory involves three parts which are known by their latin names, i.e., jus ad bellum, jus in bello, and jus post bellum [5].

- “Jus ad bellum specifies the conditions under which the use of military force must be justified. The jus ad bellum requirements that have to be fulfilled for a resort to war to be justified are: (i) just cause; (ii) right intention; (iii) legitimate authority and declaration; (iv) last resort; (v) proportionality; (vi) chance of success.
- Jus in bello refers to justice in war, i.e., to conducting a war in an ethical manner. According to international war law, a war should be conducted obeying all international laws for weapons prohibition (e.g., biological or chemical weapons), and for benevolent quarantine for prisoners of war (POWs).
- Jus post bellum refers to justice at war termination. Its purpose is to regulate the termination of wars and to facilitate the return to peace. Actually, no global law exists for jus post bellum. The return to peace should obey the general moral laws of human rights to life and liberty.”

The international law of war or international humanitarian law attempts to limit the effects of armed conflict for humanitarian purposes. The humanitarian jus in bello law has the following principles [5,48]:

1. Discrimination: It is immoral to kill civilians, i.e., non-combatants. Weapons (non-prohibited) may be used only against those who are engaged in doing harm.
2. Proportionality: Soldiers are entitled to use only force proportional to the goal sought.
3. Benevolent treatment of POWs: Captive enemy soldiers are “no longer engaged in harm”, and so they are to be provided with benevolent (not malevolent) quarantine away from battle zones, and they should be exchanged for one’s own POWs after the end of war.
4. Controlled weapons: Soldiers are allowed to use controlled weapons and methods which are not evil in themselves.
5. No retaliation: This occurs when a state A violates jus in bello in war in state B, and state B retaliates with its own violation of jus in bello, in order to force A to obey the rules.

In general, a war is considered a just war if it is both justified and carried out in the right way.

The ethical and legal rules of conducting wars using robotic weapons, in addition to conventional weapons, includes at minimum all of the rules of just war discussed above, but the use of semiautonomous/autonomous robots adds new rules as follows:

- Firing decision: At present, the firing decision still lies with the human operator. However, the separation margin between human firing and autonomous firing in the battlefield is continuously decreased.

- **Discrimination:** The ability to distinguish lawful from unlawful targets by robots varies enormously from one system to another, and present-day robots are still far from having visual capabilities that may faithfully discriminate between lawful and unlawful targets, even in close contact encounter. The distinction between lawful and unlawful targets is not a pure technical issue, but it is considerably complicated by the lack of a clear definition of what counts as a civilian. The 1944 Geneva Convention states that a civilian can be defined by common sense, and the 1977 Protocol defines a civilian any person who is not an active combatant (fighter).
- **Responsibility:** The assignment of responsibility in case of failure (harm) is both an ethical and legislative issue in all robotic applications (medical, assistive, socialization, war robots). Yet this issue is much more critical in the case of war robots that are designed to kill humans with a view to save other humans. The question is to whom blame and punishment should be assigned for improper fight and unauthorized harm caused (intentionally or unintentionally) by an autonomous robot—to the designer, robot manufacturer, robot controller/supervisor, military commander, a state prime minister/president, or the robot itself? This question is very complicated and needs to be discussed more deeply when the robot is given a higher degree of autonomy [49].
- **Proportionality:** The proportionality rule requires that even if a weapon meets the test of distinction, any weapon must also undergo an evaluation that sets the anticipated military advantage to be gained against the predicted civilian harm (civilian persons or objects). In other words, the harm to civilians must not be excessive relative to the expected military gain. Proportionality is a fundamental requirement of just war theory and should be respected by the design and programming of any autonomous robotic weapon.

Two examples of autonomous robotic weapons (fighters) are shown in Figure 6.



Figure 6. Autonomous fighter examples (MQ-1 Predator, M12). Source: www.kareneliot.de/OpenDrones/opendrones_1military.html; https://www.youtube.com/watch?v=_upbplsKGd4; <https://www.digitaltrends.com/cool-tech/coolest-military-robots>.

The use of autonomous robotic weapons in war is subject to a number of objections [5]:

- Inability to program war laws (Programming the laws of war is a very difficult and challenging task for the present and the future).
- Taking humans out of the firing loop (It is wrong per se to remove human from the firing loop).
- Lower barriers to war (The removal of human soldiers from the risk and the reduction of harm to civilians through more accurate autonomous war robots diminishes the disincentive to resort to war).

The Human Rights Watch (HRW) has issued a set of recommendations to all states, roboticists, and other scientists involved in the development and production of robotic weapons, which aim to minimize the development and use of autonomous lethal robots in war [50].

4.5. Autonomous Car Ethics

Autonomous (self-driving, driverless) cars are on the way [5]. Proponents of autonomous cars and other vehicles argue that within two or three decades autonomously driving cars will be so accurate that they will exceed the number of human-driven cars [51,52]. The specifics of self-driving vary from manufacturer to manufacturer, but at the basic level cars use a set of cameras, lasers, and sensors located around the vehicle for detecting obstacles, and employ GPS (global positioning systems) help them to move along a preset route (Figure 7).

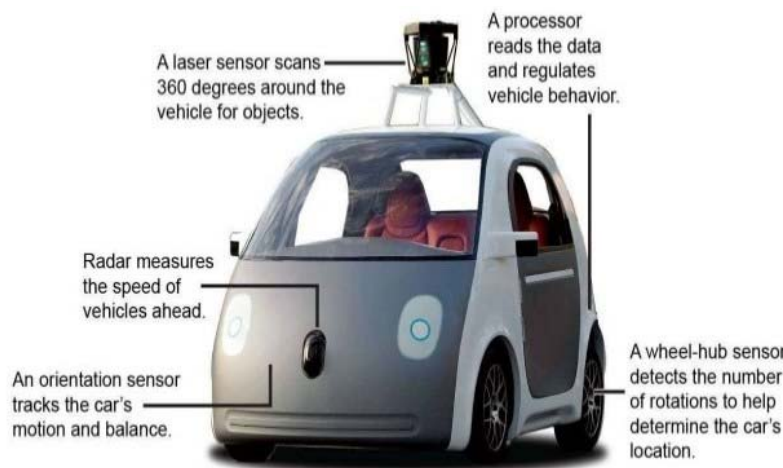


Figure 7. Basic sensors of Google’s driverless car. Source: <http://blog.cayenneapps.com/2016/06/13/self-driving-cars-swot-analysis>.

Currently there are cars on the road that perform several driving tasks autonomously (without the help of the human driver). Examples are: lane assist systems to keep the car in the lane, cruise control systems that speed up or slow down according to the speed of the car in front, and automatic emergency braking for emergency stops to prevent collisions with pedestrians.

SAE (Society of Automotive Engineers) International (www.sae.org/autodrive) developed and released a new standard (J3016) for the “Taxonomy and definitions of terms related to on-road motor vehicle automated driving systems”. This standard provides a harmonized classification system and supporting definitions which:

- “Identify six levels of driving automation from ‘no automation’ to ‘full automation’.
- Base definitions and levels on functional aspects of technology.
- Describe categorical distinction for step-wise progression through the levels.
- Are consistent with current industry practice.
- Eliminate confusion and are useful across numerous disciplines (engineering, legal, media, and public discourse).
- Educate a wide community by clarifying for each level what role (if any) drivers have in performing the dynamic driving task while a driving automation system is engaged.”

The fundamental definitions included in J3016 are (orfe.princeton.edu, Business Wire, 2017):

- “Dynamic driving tasks (i.e., operational aspects of automatic driving, such as steering, braking, accelerating, monitoring the vehicle and the road, and tactical aspects such as responding to events, determining when to change lanes, turn, etc.).
- Driving mode (i.e., a form of driving scenario with appropriate dynamic driving task requirements, such as expressway merging, high-speed cruising, low-speed traffic jam, closed-campus operations, etc.).

- Request to intervene (i.e., notification by the automatic driving system to a human driver that he should promptly begin or resume performance of the dynamic driving task)."

Figure 8 shows the milestones needed to be passed on the way to meeting the final goal of fully automated vehicles, according to SAE, NHTSA (National Highway Traffic Safety Administration), and FHRI (Federal Highway Research Institute).

SAE level	NHTSA level	BAST level	Steering, braking & acceleration	Monitoring of driving environment	Fallback performance	System capability
No Automation	0	Driver only	Human	Human	Human	none
Driver Assistance	1	Assisted	Human and system	Human	Human	
Partial Automation	2	Partially automated	System	Human	Human	
Conditional Automation	3	Highly automated	System	System	Human	
High Automation	3/4	Fully automated	System	System	System	
Full Automation		–	System	System	System	

Figure 8. Vehicle driving automation milestones adopted by ASE, NHTSA, and BAST. Source: [https://www.schlegelundpartner.com \(/cn/news/man-and-machine-automated-driving\)](https://www.schlegelundpartner.com (/cn/news/man-and-machine-automated-driving)).

These scenarios and stages of development are subject to several legal and ethical problems which are currently under investigation at regional and global levels. The most advanced country in this development is the USA, while European countries are somewhat behind the USA. The general legislation in the USA (primarily determined by NHTSA and the Geneva Convention on road traffic of 1949) requires the active presence of a driver inside the vehicle who is capable of taking control whenever necessary. Within the USA, each state enacts its own laws concerning automated driving cars. So far only four states (Michigan, California, Nevada, and Florida) have accepted automated driving software to be legal. In Germany, the Federal Ministry of Transport has already allowed the use of driving assistance governed by corresponding legislation. Most car manufactures are planning to produce autonomous driving technologies of various degrees. For example, Google is testing a fully autonomous prototype that replaces the driver completely, and anticipates to release its technology in the market by 2020. Automakers are proceeding towards full autonomy in stages; currently, most of them are at level 1 and only a few have introduced level 2 capabilities.

The fundamental ethical/liability question here is [5]: Who will be liable when a driverless car crashes? This question is analogous to the ethical/liability question of robotic surgery. Today, the great majority of car accidents are the fault of one driver or the other, or the two in some shared responsibility. Few collisions are deemed to be the responsibility of the car itself or of the manufacturer. However, this will not be the same if the car drives itself. Actually, it will be much harder to conventionally blame one driver or the other. Should the ethical and legal responsibility be shared by the manufacturer or multiple manufacturers, or the people who made the hardware or software? Or, should another car that sent a faulty signal on the highway be blamed? [5]. An extensive discussion of advantages/disadvantages including legal and ethical issues is provided in Reference [53].

4.6. Cyborg Ethics

Cyborg technology aims to design and study neuromotor prostheses in order to store and reinstate lost function with a replacement that is as similar as possible to the real thing (a lost arm or hand, lost vision, etc.) [5,54]. The word cyborg stands for cybernetic organism, a term coined by Manfred Clynes and Nathan Kline [55]. A cyborg is any living being that has both organic and mechanical/electrical parts that either restore or enhance the organism's functioning. People with the most common technological implants such as prosthetic limbs, pacemakers, and cochlear/bionic ear implants, or people who receive implant organs developed from artificially cultured stem cells can be considered to belong to this category [56]. The first real cyborg was a "lab rat" created at Rockland State Hospital in 1950 (New York, www.scienceabc.com).

The principal advantages of mixing organs with mechanical parts are for human health. For example [5]:

- "People with replaced parts of their body (hips, elbows, knees, wrists, arteries, etc.) can now be classified as cyborgs.
- Brain implants based on neuromorphic model of the brain and the nervous system help reverse the most devastating symptoms of Parkinson disease."

Disadvantages of cyborgs include [5]:

- "Cyborgs do not heal body damage normally, but, instead, body parts are replaced. Replacing broken limbs and damaged armor plating can be expensive and time-consuming.
- Cyborgs can think of the surrounding world in multiple dimensions, whereas human beings are more restricted in that sense" [56,57].

Figure 9 shows a cyborg/electronic eye.



Figure 9. An example of cyborg eye. Source: [https://www.behance.net/gallery/4411227/Cyborg-Eye-\(Female\)](https://www.behance.net/gallery/4411227/Cyborg-Eye-(Female)).

Three of the world's most famous real-life cyborgs are the following (Figure 10) [58]:

- The artist Neil Harbison, born with achromatopsia (able to see only black and white) is equipped with an antenna implanted into his head. With this eyeborg (electronic eye), he became able to render perceived colors as sounds on the musical scale.
- Jesse Sullivan suffered a life-threatening accident: he was electrocuted so severely that both of his arms needed to be amputated. He was fitted with a bionic limb connected through a nerve-muscle grafting. He then became able to control his limb with his mind, and also able to feel temperature as well as how much pressure his grip applies.
- Claudia Mitchell is the first woman to have a bionic arm after a motorcycle accident in which she lost her left arm completely.

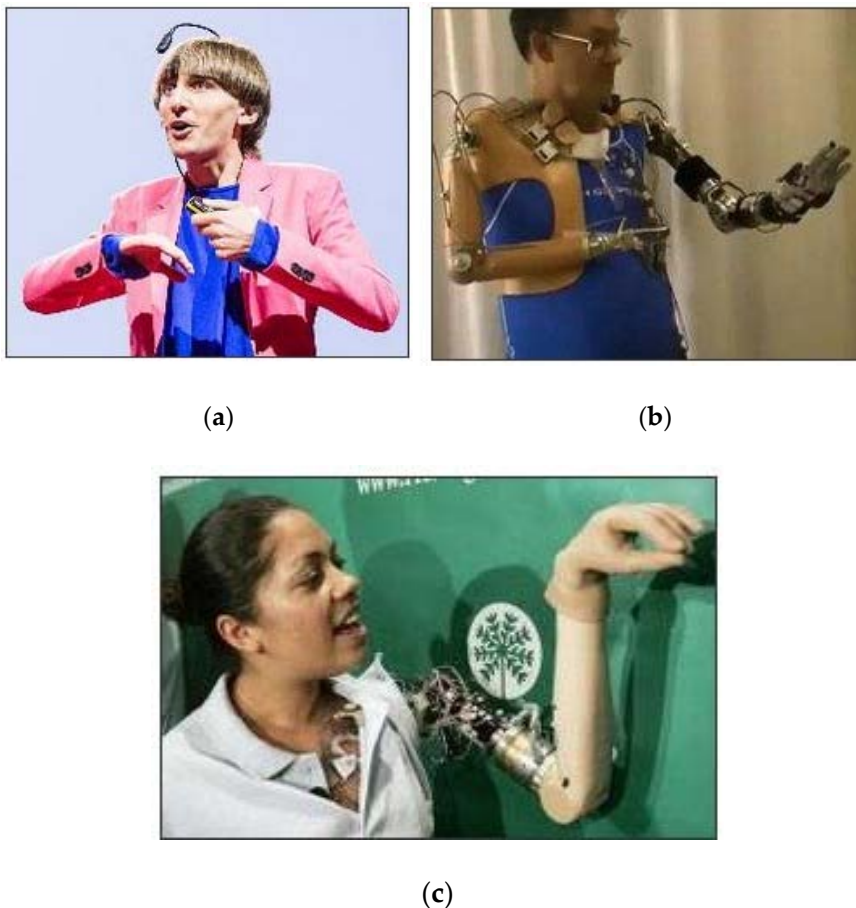


Figure 10. Examples of human cyborgs. (a) Neil Harbison, (b) Jesse Sullivan, (c) Claudia Mitchell. Source: [www.medicalfuturist.com \(/the-world-most-famous-real-life-cyborgs\)](http://www.medicalfuturist.com (/the-world-most-famous-real-life-cyborgs)).

Cyborgs raise serious ethical concerns, especially in the case when the consciousness of a person is changed by the integration of human and machine [59]. Actually, in all cases cyborg technology violates the human/machine distinction. However, in most cases, although the person's physical capabilities take on a different form and his/her capabilities are enhanced, his/her internal mental state, consciousness, and perception has not been changed other than to the extent of changing what the individual might be capable of accomplishing [59]. Actually, what should be of maximum ethical concern is not the possible physical enhancements or repairs, but when the change of the nature of a human is changed by linking human and machine mental functioning. A philosophical discussion about cyborgs and the relationship between body and machine is provided in Reference [60].

5. Future Prospects of Robotics and Roboethics

In general, the intelligence capabilities of robots follow the development path of artificial intelligence. The robots of today have capabilities compatible with “artificial narrow intelligence” (ANI), i.e., they can execute specific focused tasks but cannot self-expand functionally. As a result, they outperform humans in specific repetitive operations. By 2040, robots are expected to perform tasks compatible with “artificial general intelligence” (AGI), i.e., they will be able to compete with humans across all activities, and perhaps convince humans that they are “humans”. Soon after the AGI period, robots are expected to demonstrate intelligence beyond human capabilities. In fact, many futurists, e.g., Hans Moravec (Carnegie Mellon University), predict that in the future, robots and machines will have superb features such as high-level reasoning, self-awareness, consciousness, conscience, emotion, and other feelings. Moravec [61] believes that in the future, the line between humans and robots will blur, and—although current robots are modeled on human senses, abilities, and actions—in the future they will evolve beyond this framework. Therefore, the following philosophical question arises: What makes a human being a human being and a robot a robot? The answer to this question given by several robotics scientists is that what makes a human being different from a robot, even if robots can reason, and are self-aware, emotional, and moral, is creativity.

The American Psychological Association (APA) points out that “in future, loneliness and isolation may be a more serious public health hazard than obesity”. Ron Arkin (a roboethicist) says that “a solution to this problem can be to use companion sociorobots, but there is a need to study deeply the ethics of forming bonds/close relationships with these robots”. Today, human-robot relationships are still largely task driven, i.e., the human gives the robot a task and expects it to be completed. In the future, tasks are expected to be performed jointly by human-robot close co-operation and partnership.

The big double question here is (mobile.abc.com): Should we allow robots to become partners with us in the same way that we allow humans to become partners? Is the concept of sentience or true feeling required in a robot for it to be respected? Arkin’s comment about this question is that: “Robots propagate an illusion of life; they can create the belief that the robot actually cares about us, but what it cares is nothing”.

Three important questions about the robots of the future are (www.frontiers.org):

- How similar to humans should robots become?
- What are the possible effects of future technological progress of robotics on humans and society?
- How to best design future intelligent/autonomous robots?

These and other questions are discussed in Reference [62]. The human-robot similarity of the future depends on the further development of several scientific/technological fields such as artificial intelligence, speech recognition, processing and synthesis, human-computer interfaces and interaction, sensors and actuators, artificial muscles and skins, etc. Clearly, a proper synergy of these elements is required. Whether the robots look like humans or not is not so important as how, and how much, robots can perform the tasks we want them to do (www.frontiers.org). The question here is: Given that we can create human-like (humanoid) robots, do we want or need them? According to the “uncanny valley” hypothesis, as robots become more similar to humans (humane, anthropomorphic), the pleasure of having them around increases up to a certain point. When they are very similar to humans this pleasure falls abruptly. However, it later increases again when the robots become even more similar to humans (Figure 11). This decrease and increase of comfort as a robot becomes more anthropomorphic is the “uncanny valley”, which is discussed in detail in Reference [63].

- Jo Bell (Animal Liberation): “Asimov’s Robot series grappled with this sort of (rights) question. As we have incorporated other races and people-women, the disabled, into the category of those who can feel and think, then I think if we had machines of that kind, then we would have to extend some sort of rights to them.”

Over the years, many AI thinkers have worried that intelligent machines of the future (called superintelligent or ultra-intelligent machines) could pose a threat to humanity. For example, I.J. Good argued (1965) that “an ultra-intelligent machine could design even better machinery, and the intelligence of man would be left far behind”.

Roger Moore, speaking about AI ethics, artificial intelligence, robots, and society, explained why people worry about the wrong things when they worry about AI [16]. He argues that the reasons not to worry are:

- “AI has the same problems as other conventional artifacts.
- It is wrong to exploit people’s ignorance and make them think AI is human.
- Robots will never be your friends.”

Things to worry about include:

- “Human culture is already a superintelligent machine turning the planet into apes, cows, and paper clips.
- Big data + better models = ever-improving prediction, even about individuals.”

General key topics for future roboethics include the following:

- Assuring that humans will be able to control future robots.
- Preventing the illegal use of future robots.
- Protecting data obtained by robots.
- Establishing clear traceability and identification of robots.

The need to develop new industrial standards for testing AI/intelligent robots of the future will be much more crucial, otherwise it will be difficult to implement and deploy future robots, with superintelligence, safely and profitably. Big ethical questions for the robots of the future include the following:

- Is it ethical to turn over all of our difficult and highly sensitive decisions to machines and robots?
- Is it ethical to outsource all of our autonomy to machines and robots that are able to make good decisions?
- What are the existential and ethical risks of developing superintelligent machines/robots?

6. Conclusions

The core of this paper (roboethics branches) followed the structure of the author’s book on roboethics [5]. The paper was concerned with the robot ethics field and its future prospects. Many of the fundamental concepts of ethics and roboethics were outlined at an introductory conceptual level, and some issues of future advanced artificial intelligence ethics and roboethics were discussed.

On topics as sensitive as decisions on human life (e.g., using autonomous robot weapons), the ethical issues of war and robot-based weapons were discussed including the principal objections against the use of autonomous lethal robots in war. The general ethical questions in this area are: What kind of decisions are we comfortable outsourcing to autonomous machines? What kind of decisions should or should not always remain in the hand of humans? In other words, should robots be allowed to make life/death decisions? In cases not covered by the law in force, human beings remain under the protection of the principles of humanity and the dictates of public conscience according to the Geneva Conventions (Additional Protocol II). The Open Roboethics Institute (ORI)

conducted a world-wide public study collecting the opinions of a large number of individuals on the issue of autonomous robotic weapons use. The results of this study were documented and presented in Reference [66]. Other sensitive human life areas discussed in the paper are the use of robots in medicine, assistance to the elderly and impaired people, companionship/entertainment, driverless vehicles, and cybernetic organisms. Finally, another emerging area that rises critical ethical questions that was not discussed in this paper is the area of sex or love-making robots (sexbots, lovebots). Representative references on sexbots include References [67–69]. A review of critical ethical issues in creating superintelligence is provided in [70], and a review of ‘cyborg enhancement technology’, with emphasis on the brain enhancements and the creation of new senses, is given in [71].

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Sabanovic, S. Robots in society, society in robots. *Int. J. Soc. Robots* **2010**, *24*, 439–450. [CrossRef]
2. Veruggio, G. The birth of roboethics. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA 2005): Workshop on Robot Ethics, Barcelona, Spain, 18 April 2005; pp. 1–4.
3. Lin, P.; Abney, K.; Bekey, G.A. *Robot Ethics: The Ethical and Social Implications of Robotics*; MIT Press: Cambridge, MA, USA, 2011.
4. Capurro, R.; Nagenborg, M. *Ethics and Robotics*; IOS Press: Amsterdam, The Netherlands, 2009.
5. Tzafestas, S.G. *Roboethics: A Navigating Overview*; Springer: Berlin, Germany; Dordrecht, The Netherlands, 2015.
6. Dekoulis, G. *Robotics: Legal, Ethical, and Socioeconomic Impacts*; InTech: Rijeka, Croatia, 2017.
7. Jha, U.C. *Killer Robots: Lethal Autonomous Weapon Systems Legal, Ethical, and Moral Challenges*; Vij Books India Pvt: New Delhi, India, 2016.
8. Gunkel, D.J.K. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics*; MIT Press: Cambridge, MA, USA, 2012.
9. Dekker, M.; Guttman, M. *Robo-and-Information Ethics: Some Fundamentals*; LIT Verlag: Muenster, Germany, 2012.
10. Anderson, M.; Anderson, S.L. *Machine Ethics*; Cambridge University Press: Cambridge, UK, 2011.
11. Veruggio, G.; Solis, J.; Van der Loos, M. Roboethics: Ethics Applied to Robotics. *IEEE Robot. Autom. Mag.* **2001**, *18*, 21–22. [CrossRef]
12. Capurro, R. Ethics in Robotics. Available online: http://www.i-r-i-e.net/inhalt/006/006_full.pdf (accessed on 10 June 2018).
13. Lin, P.; Abney, K.; Jenkins, R. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*; Oxford University Press: Oxford, UK, 2018.
14. Veruggio, G. Roboethics Roadmap. In Proceedings of the EURON Roboethics Atelier, Genoa, Italy, 27 February–3 March 2006.
15. Arkin, R. *Governing Lethal Behavior of Autonomous Robots*; Chapman and Hall: New York, NY, USA, 2009.
16. Moore, R.K. *AI Ethics: Artificial Intelligence, Robots, and Society*; CPSR: Seattle, WA, USA, 2015; Available online: www.cs.bath.ac.uk/~jjb/web/ai.html (accessed on 10 June 2018).
17. Asaro, P.M. What should we want from a robot ethics? *IRIE Int. Rev. Inf. Ethics* **2006**, *6*, 9–16.
18. Tzafestas, S.G. *Systems, Cybernetics, Control, and Automation: Ontological, Epistemological, Societal, and Ethical Issues*; River Publishers: Gistrup, Denmark, 2017.
19. Verruggio, P.M.; Operto, F. Roboethics: A bottom -up interdisciplinary discourse in the field of applied ethics in robotics. *IRIE Int. Rev. Inf. Ethics* **2006**, *6*, 2–8.
20. Wallach, W.; Allen, C. *Moral Machines: Teaching Robots Right from Wrong*; Oxford University Press: Oxford, UK, 2009.
21. Wallach, W.; Allen, C.; Smit, I. Machine morality: Bottom-up and top-down approaches for modeling moral faculties. *J. AI Soc.* **2008**, *22*, 565–582. [CrossRef]
22. Asimov, I. *Runaround: Astounding Science Fiction (March 1942)*; Republished in *Robot Visions*: New York, NY, USA, 1991.
23. Gert, B. *Morality*; Oxford University Press: Oxford, UK, 1988.

24. Gips, J. Toward the ethical robot. In *Android Epistemology*; Ford, K., Glymour, C., Mayer, P., Eds.; MIT Press: Cambridge, MA, USA, 1992.
25. Bringsjord, S. Ethical robots: The future can heed us. *AI Soc.* **2008**, *22*, 539–550. [CrossRef]
26. Dekker, M. Can humans be replaced by autonomous robots? Ethical reflections in the framework of an interdisciplinary technology assessment. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'07)*, Rome, Italy, 10–14 April 2007.
27. Pence, G.E. *Classic Cases in Medical Ethics*; McGraw-Hill: New York, NY, USA, 2000.
28. Mappes, G.E.; DeGrazia, T.M.D. *Biomedical Ethics*; McGraw-Hill: New York, NY, USA, 2006.
29. North, M. The Hippocratic Oath (translation). National Library of Medicine, Greek Medicine. Available online: www.nlm.nih.gov/hmd/greek/greek_oath.html (accessed on 10 June 2018).
30. Paola, I.A.; Walker, R.; Nixon, L. *Medical Ethics and Humanities*; Jones & Bartlett Publisher: Sudbury, MA, USA, 2009.
31. AMA. Medical Ethics. 1995. Available online: <https://www.ama-assn.org> and <https://www.ama-assn.org/delivering-care/ama-code-medical-ethics> (accessed on 10 June 2018).
32. Beabou, G.R.; Wennemann, D.J. *Applied Professional Ethics*; University of Press of America: Milburn, NJ, USA, 1993.
33. Rowan, J.R.; Sinaih, S., Jr. *Ethics for the Professions*; Cengage Learning: Boston, MA, USA, 2002.
34. Dickens, B.M.; Cook, R.J. Legal and ethical issues in telemedicine and robotics. *Int. J. Gynecol. Obstet.* **2006**, *94*, 73–78. [CrossRef] [PubMed]
35. World Health Organization. *International Classification of Functioning, Disability, and Health*; World Health Organization: Geneva, Switzerland, 2001.
36. Tanaka, H.; Yoshikawa, M.; Oyama, E.; Wakita, Y.; Matsumoto, Y. Development of assistive robots using international classification of functioning, disability, and health (ICF). *J. Robot.* **2013**, *2013*, 608191. [CrossRef]
37. Tanaka, H.; Wakita, Y.; Matsumoto, Y. Needs analysis and benefit description of robotic arms for daily support. In *Proceedings of the RO-MAN' 15: 24th IEEE International Symposium on Robot and Human Interactive Communication*, Kobe, Japan, 31 August–4 September 2015.
38. RESNA Code of Ethics. Available online: http://resna.org/certification/RESNA_Code_of_Ethics.pdf (accessed on 10 June 2018).
39. Ethics Resources. Available online: www.crcrcertification.com/pages/crc_ccrc_code_of_ethics/10.php (accessed on 10 June 2018).
40. Tzafestas, S.G. *Sociorobot World: A Guided Tour for All*; Springer: Berlin, Germany, 2016.
41. Fog, T.; Nourbakhsh, I.; Dautenhahn, K. A survey of socially interactive robots. *Robot. Auton. Syst.* **2003**, *42*, 143–166.
42. Darling, K. Extending legal protections in social robots: The effect of anthropomorphism, empathy, and violent behavior towards robots. In *Robot Law*; Calo, M.R., Froomkin, M., Ker, I., Eds.; Edward Elgar Publishing: Brookfield, VT, USA, 2016.
43. Melson, G.F.; Kahn, P.H., Jr.; Beck, A.; Friedman, B. Robotic pets in human lives: Implications for the human-animal bond and for human relationships with personified technologies. *J. Soc. Issues* **2009**, *65*, 545–567. [CrossRef]
44. Breazeal, C. *Designing Sociable Robots*; MIT Press: Cambridge, MA, USA, 2002.
45. Sawada, T.; Takagi, T.; Fujita, M. Behavior selection and motion modulation in emotionally grounded architecture for QRIO SDR-4XIII. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'2004)*, Sendai, Japan, 28 September–2 October 2004; pp. 2514–2519.
46. Asaro, P. *How Just Could a Robot War Be*; IOS Press: Amsterdam, The Netherlands, 2008.
47. Walzer, M. *Just and Unjust Wars: A Moral Argument Historical with Illustrations*; Basic Books: New York, NY, USA, 2000.
48. Coates, A.J. *The Ethics of War*; University of Manchester Press: Manchester, UK, 1997.
49. Asaro, A. Robots and responsibility from a legal perspective. In *Proceedings of the 2007 IEEE International Conference on Robotics and Automation: Workshop on Roboethics*, Rome, Italy, 10–14 April 2007.
50. Human Rights Watch. *HRW-IHRC, Losing Humanity: The Case against Killer Robots*; Human Rights Watch: New York, NY, USA, 2012; Available online: www.hrw.org (accessed on 10 June 2018).
51. Marcus, G. Moral Machines. Available online: www.newyorker.com/news_desk/moral_machines (accessed on 24 November 2012).

52. Self-Driving Cars. Absolutely Everything You Need to Know. Available online: <http://recombu.com/cars/article/self-driving-cars-everything-you-need-to-know> (accessed on 10 June 2018).
53. Notes on Autonomous Cars: Lesswrong. 2013. Available online: http://lesswrong.com/lw/gfv/notes_on_autonomous_cars (accessed on 10 June 2018).
54. Lynch, W. *Wilfred Implants: Reconstructing the Human Body*; Van Nostrand Reinhold: New York, NY, USA, 1982.
55. Clynes, M.; Kline, S. Cyborgs and Space. *Astronautics* **1995**, 29–33. Available online: <http://www.tantrik-astrologer.in/book/linked/2290.pdf> (accessed on 10 June 2018).
56. Warwick, K. A Study of Cyborgs. Royal Academy of Engineering. Available online: www.ingenia.org.uk/Ingenia/Articles/217 (accessed on 10 June 2018).
57. Warwick, K. Homo Technologicus: Threat or Opportunity? *Philosophies* **2016**, 1, 199. [CrossRef]
58. Seven Real Life Human Cyborgs. Available online: www.mnn.com/leaderboard/stories/7-real-life-human-cyborgs (accessed on 10 June 2018).
59. Warwick, K. Cyborg moral, cyborg values, cyborg ethics. *Ethics Inf. Technol.* **2003**, 5, 131–137. [CrossRef]
60. Palese, E. Robots and cyborgs: To be or to have a body? *Poiesis Prax.* **2012**, 8, 19–196. [CrossRef] [PubMed]
61. Moravec, H. *Robot: Mere Machine to Transcendent Mind*; Oxford University Press: Oxford, UK, 1998.
62. Torresen, J. A review of future and ethical perspectives of robotics and AI. *Front. Robot. AI* **2018**. [CrossRef]
63. MacDorman, K.F. Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it? In Proceedings of the CogSci 2005 Workshop: Toward Social Mechanisms of Android Science, Stresa, Italy, 25–26 July, 2005; pp. 106–118.
64. IEEE Standards Association. *IEEE Ethical Aligned Design*; IEEE Standards Association: Piscataway, NJ, USA, 2016; Available online: http://standards.ieee.org/develop/indconn/ec/ead_v1.pdf (accessed on 10 June 2018).
65. Coeckelberg, M. Robot Rights? Towards a social-relational justification of moral consideration. *Ethics Inf. Technol.* **2010**, 12, 209–221. [CrossRef]
66. ORI: Open Roboethics Institute. Should Robots Make Life/Death Decisions? In Proceedings of the UN Discussion on Lethal Autonomous Weapons, UN Palais des Nations, Geneva, Switzerland, 13–17 April 2015.
67. Sullins, J.P. Robots, love, and sex: The ethics of building a love machine. *IEEE Trans. Affect. Comput.* **2012**, 3, 389–409. [CrossRef]
68. Cheok, A.D.; Ricart, C.P.; Edirisinghe, C. Special Issue “Love and Sex with Robots”. Available online: http://www.mdpi.com/journal/mti/special_issues/robots (accessed on 10 June 2018).
69. Levy, D. *Love and Sex with Robots: The Evolution of Human-Robot Relationship*; Harper Perrenial: London, UK, 2008.
70. Bostrom, N. Ethical issues in advanced artificial intelligence. In *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and Artificial Intelligence*; Lasker, G.E., Marreiros, G., Wallach, W., Smit, I., Eds.; International Institute for Advanced Studies in Systems Research and Cybernetics: Tecumseh, ON, Canada, 2003; Volume 2, pp. 12–17.
71. Barfield, W.; Williams, A. Cyborgs and enhancement technology. *Philosophies* **2017**, 2, 4. [CrossRef]



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).



Article

Ethical Framework for Designing Autonomous Intelligent Systems

Jaana Leikas *, Raija Koivisto and Nadezhda Gotcheva

VTT Technical Research Center of Finland Ltd., FI-30100 Tampere, Finland; raija.koivisto@vtt.fi (R.K.);
Nadezhda.Gotcheva@vtt.fi (N.G.)

* Correspondence: jaana.leikas@vtt.fi

Received: 29 January 2019; Accepted: 5 March 2019; Published: 13 March 2019

Abstract: To gain the potential benefit of autonomous intelligent systems, their design and development need to be aligned with fundamental values and ethical principles. We need new design approaches, methodologies and processes to deploy ethical thought and action in the contexts of autonomous intelligent systems. To open this discussion, this article presents a review of ethical principles in the context of artificial intelligence design, and introduces an ethical framework for designing autonomous intelligent systems. The framework is based on an iterative, multidisciplinary perspective yet a systematic discussion during an Autonomous Intelligent Systems (AIS) design process, and on relevant ethical principles for the concept design of autonomous systems. We propose using scenarios as a tool to capture the essential user's or stakeholder's specific qualitative information, which is needed for a systematic analysis of ethical issues in the specific design case.

Keywords: autonomous systems; autonomous intelligent systems; artificial intelligence; AI; ethics; design

1. Introduction

“A water metro”, an autonomous ferry is transferring passengers across/along the river, from the harbor in downtown to outer skirts of the city, and back. The ship has no crew. There is a human operator on the shore monitoring the ferry. The operator monitors several ferries at the same time. The ferries autonomously plan the route based on collected data from many different sources, and change it should there be any obstacles on the way. A lot of hope is placed on such a new autonomous ferry. People say this is more reliable than the old-fashioned ships, as there is no possibility for human error. Previously, an old captain, a very proud of his profession, sailed his ship on this same route, and always brought the ship safely home despite varying weather conditions. Now the ship is unmanned, and now this man is replaced by the remote operator who sits in the control room, out of sight of the passengers.

Autonomous systems are fundamentally changing our world and ways of working. They are seen as a means to increase productivity, cost efficiency and safety—not only by reducing the work done by humans, but also by enabling completely new business models [1]. Autonomy goes beyond automation by adding self-governing behavior and requiring intelligent decision-making abilities. The development of key elements in autonomous systems, such as situational awareness systems and autonomous decision-making, are thus likely to be based on various artificial intelligence (AI) technologies [2].

The societal transition from current ICT to future AI society, and steering of this process, are among the biggest challenges of our time [3]. Although these systems are designed to reduce human intervention, relevant questions remain about their responsible and ethical use, their short-term and long-term impact on individuals and societies, and on humanity in general [4,5]. Potential direct applications of these systems, related innovations and business value are currently widely discussed

in academia, business and governmental bodies alike [6]. Although there is a growing interest in the wider societal impacts as well [7], ethical considerations are seen as critical yet not fully understood. While there has been increasing public discussion and research on the links between ethics and Artificial Intelligence (AI) [8], “machine ethics” [9] or potential risks of applying AI [10], these issues need more attention also as opportunities, which has been less accentuated.

AI technologies give rise to a plethora of ethical issues as the design and use of autonomous intelligent systems are socially and culturally embedded [11]. Design of autonomous systems is thus not only a multi-technological effort, but involves also social, psychological, economic, political, and legal aspects, and will have profound impacts at all dimensions of society [12]. The ethics of AIS is still underexplored both as an object of scientific study and as a practice. Current approaches include responsible use of AI [13], professional codes of conduct [14,15] and human-robot interaction [16]. Attempts to incorporate ethics into the AI-design have not yet significantly affected technology design. So far, ethical design research has been challenging from two perspectives [17–19]. First, fundamental values and ethical frames have been too complex to be formalized into a deductive decision-making system [20,21]. Second, the ethical decision-making in AI design is context-dependent, defying thus traditional principles-based approaches.

2. Attempts to Approach Ethical Issues in Design

In the field of design thinking, there are a few design approaches that have emphasized the importance of ethical design thinking. Value-sensitive design (VSD) holds that artefacts are value-laden and design can be value-sensitive [22,23]. The approach refers to the need to identify early implicit values embedded in new technologies by focusing on the usage situations of technology. “Value” is defined here broadly as something that a person or a group considers important in life, and designers can intentionally inscribe their values in the design objects thus shaping them. The design is carried out iteratively by combining conceptual (conceptions of important values of users and stakeholders), empirical (how values are realized in everyday practices and in technical solutions) and technical (how the designed technology and the impact of technology support the values) research and assessment.

Another design approach which discusses ethics is Life-based design (LBD), which highlights the need for designing for the “good life” [24,25] and posits that the measure of technology is in its ability to enhance the quality of life for people. The process of design thinking focuses first on asking what is needed in life and how people wish to live, and thereafter on what kinds of technologies can serve this goal. LBD is thus interested in what people should do with technology rather than what they can do with technology. It focuses on a biological, psychological and socio-cultural form of life of target users. Ethical choices and values are reflected and resolved in the design decisions: What is ethically acceptable, i.e., what constitutes “the good” for the end users.

Thirdly, ethics has been considered as an important element of responsible research and innovation, which highlights the importance of understanding that ethics in technology is strongly linked with social acceptability. Thus, the concept of Responsible Research and Innovation (RRI) [26–28] is a valuable perspective when discussing the ethics of technology. Responsibility is understood broadly as socially, ethically and environmentally acceptable actions [29]. It is seen as a competitive factor and source of innovation for companies. Successful implementation of responsible innovation and business creates shared value by providing sustainable solutions to customers, increased competitiveness to companies and positive societal impact for the society. The comprehensive integration of responsibility in a company’s operations improves its capabilities to produce societally acceptable and desirable goods and services, avoid unintended consequences and manage its commercial risks. RRI emphasizes the need for co-design, empowering ways of working and taking into consideration different stakeholder perspectives.

The main message of all these above-mentioned approaches is that ethical design means, first of all, conscious reflection of ethical values and choices in respect to design decisions. That is, examining

what the prevailing moral rules and norms of the users are and what kind of impacts they have on the design decisions. Secondly, ethical design means a reflection on what is ethically acceptable. Finally, the ethical design must consider the issues of what is ethical, i.e., what constitutes the good of humanity.

3. A framework to Discuss and Analyze Ethical Issues

The precondition for considering ethical issues during the AIS design is that the relevant ethical issues are identifiable. For that purpose, we propose a systematic framework which can be used in different phases of design: In the beginning, ethics for the design goals are defined and interpreted as design requirements; When the design is on a more detailed level, the framework can be applied again. The final design can be assessed with the help of this framework as well. Essentially, the framework can be applied in every design decision if necessary.

The systematics of the analysis framework is based on the idea that the system under design is thoroughly discussed by using identified ethical values. We argue that this should be carried out in the very beginning of design to guide the design towards inherently ethical solutions: Ethically acceptable products and services are accepted by the users, which adds both business and societal value. Bringing in the ethical perspective very early in the product lifecycle is important, because it indicates that it is possible to come up with technical solutions and services that bring sustainability and are good for society. To embed ethical values into the design and to consider ethical issues during the design process designers need systematics to do that. As a solution, we propose the idea of bringing ethics in the practices of human-technology interaction design. This can be done by with the help of usage scenarios—stories or descriptions of usage situations in selected usage contexts—in early phases of concept design. With the help of scenarios, it is possible to operationalize “good” in the design concepts from the point of view of actors, actions and goals of actions, and thus systematically assess the ethical value of the design outcomes.

Examining the context and usage situations of the given technology follows actually the idea of casuistry in ethical thinking. Casuistry is a field of applied ethics that takes a practical approach to ethics [30]. It is focused on examining context and cases rather than using theories as starting points. Instead of discussing ethical theories, it is interested in facts of a particular case, and asks what morally meaningful facts should be taken into account in this case. The ideas of casuistry have been used in applying ethical reasoning to particular cases in law, bioethics, and business ethics (e.g., [31,32]).

As the design of AIS is not only a multi-technological effort, but involves also social, psychological, economic, political, and legal aspects, and is likely to have profound impacts at all the dimensions of the society, this deliberation requires multidisciplinary approach and involvement of various experts and stakeholders [33,34] (e.g., in the case of autonomous ships, experts of autonomous technology, shipping companies, passenger representatives, ethical experts). This iterative process should be carried out using co-design methods, involving users and stakeholders broadly, and including three steps: (1) Identification of ethical values affected by AIS; (2) Identification of context-relevant ethical values; and (3) Analysis and understanding of ethical issues within the context. These steps are further studied in the following chapters.

3.1. Identification of Ethical Principles and Values Affected by AIS

Ethical principles and values can be used as an introductory compass when seeking ways to understand ethics in design. They are universal moral rules that exist above cultures, time, or single acts of people. Principlism is an approach for ethical decision-making that focuses on the common ground moral principles that can be used as a rule of thumb in ethical thinking [31]. Principlism can be derived from and is consistent with a multitude of ethical, theological, and social approaches towards moral decision-making. It introduces the four cardinal virtues of beneficence, nonmaleficence, autonomy, and justice, which can be seen to stem already from e.g., Confucius’s *ren* (compassion or loving others; [35] and Aristotle’s conception of good life [36]. These principles form the basis of

ethical education of e.g., most physicians. They are usually conceived as intermediate between “low level” moral theories, such as utilitarianism and deontology [37]. The principle of “*beneficence*” includes all forms of action intended to benefit or promote the good of other persons [38]. The principle of “*nonmaleficence*” prohibits causing harm to other persons [38]. “Justice”, when identified with morality, is something that we owe to each other, and at the level of individual ethics, it is contrasted with charity on the one hand, and mercy on the other [39], and can be seen as the first virtue of social institutions [40]. The principle of “*autonomy*” is introduced by e.g., Kant and Mill [41,42], and refers to the right of an individual to make decisions concerning her own life.

However, the four virtues, and principlism as such, may not have enough power to carry us far enough in the discussion of technology ethics, as in technology design there are situations in which the four principles may often run into conflict. One of the reasons for this is that dealing with technology ethics is always contextual, and the impact of technology mostly concerns, not only the direct usage situation, but also many different stakeholders who may have conflicting interests [37].

As the context of technology is always situated in a cultural and ecological environment (see e.g., [43]), it is obvious that values for technology design and assessment should reflect the ethical values and norms of the given community, as well as ecological aspirations. Values are culturally predominant perceptions of individuals’, society’s and human kind’s central goals of a good life, good society and good world. They are objectives directing a person’s life and they guide decision-making [44–46]. Besides individual and (multi)cultural values, there are also critical universal values that transcend culture and national borders, such as the fundamental values laid down in the Universal Declaration of Human Rights (UN) [47], EU Treaties (EU) [48] and in the EU Charter of Fundamental Rights (2000) [49].

Friedman et al. (2003; 2006) [22,23] introduce the following values from the point of view of technology design: Human welfare; ownership and property; freedom from bias; universal usability; accountability; courtesy; identity; calmness; and environmental sustainability. In addition, informed consent is seen as a necessity in the adoption of technology [23]. It refers to garnering people’s agreement, encompassing criteria of disclosure and comprehension (for “informed”) and voluntariness, competence, and agreement (for “consent”). People have the right to consent to technological intervention (adoption and usage of technology).

3.2. Identification of Context-Specific Ethical Values

Like design issues, issues of context-specific ethical values involve differences in perspectives and in power [50]. An ethical issue arises when there is a dilemma between two simultaneous values (two ethical ones or an ethical and practical value, such as e.g., safety and efficiency). This is why technology ethics calls for a broader view, where the agents, the goal, and the context of the technology usage are discussed and deliberated, in order to analyze, argue and report the ethical dilemma and its solution. This ethical case deliberation should be carried out in collaboration with relevant stakeholders, designers and ethical experts [51]. This helps to understand what ethical principles and values should define the boundaries of the technology. In this way, it would be possible also to formulate additional design principles to the context of technology.

As to ethics of AI, many public, private and civil organizations and expert groups have introduced visions for designing ethical technology and ethical AI. For this study, we carried out i) a literature review and ii) a discussion workshop, as part of the Effective Autonomous Systems research project at VTT Technical Research Center of Finland Ltd. The participants’ scientific backgrounds include Engineering Sciences and AI, Cognitive Science, Psychology and Social Sciences. They represent experts in autonomous technologies, design thinking, ethics, responsible research and innovation, risk assessment, and societal impacts of technology. In the workshop, the outcomes of already mentioned expert groups were systematically examined, and elaborated in respect to different contexts of autonomous systems.

In the following, we shortly go through the results of the literature review in terms of ethical principles and values introduced by expert groups with respect to AI.

Ethically Aligned Design (EAD) Global initiative has been launched by the IEEE in 2016 and 2017 [4,5] under the title “A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems”, to unite collective input in the fields of Autonomous and Intelligent Systems (A/IS), ethics, philosophy and policy. In addition, some approaches for designing ethics and ethics assessment have been published (e.g., [4,5,12,52,53]).

The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems (2016 pp. 15) [4] has articulated the following high-level ethical concerns applying to AI/AS:

1. Embody the highest ideas of human rights.
2. Prioritize the maximum benefit to humanity and the natural environment.
3. Mitigate risks and negative impacts as AI/AS evolve as socio-technical systems.

The Global Initiative (2016 p. 5–6; 2017 p. 34) proposes a three-pronged approach for a designer to embedding values into AIS:

1. Identify the norms and values of a specific community affected by AIS.
2. Implement the norms and values of that community within AIS.
3. Evaluate the alignment and compatibility of those norms and values between the humans and AIS within that community.

The Asilomar Conference (2017) [54] hosted by the Future Life Institute (a volunteer-run research and outreach organization that works to mitigate existential risks facing humanity, particularly existential risk from advanced AI.), with more than 100 thought leaders and researches in economics, law, ethics, and philosophy, was a forerunner in addressing and formulating principles of beneficial AI to guide the development of AI. Its outcome was the Asilomar AI Principles which include safety; failure and juridical transparency; responsibility; value alignment; human values; privacy and liberty; shared benefit and prosperity; human control; non-supervision; and avoiding an arms race.

The European Group on Ethics in Science and New Technologies (EGE) published Statement on Artificial Intelligence, Robotics and Autonomous Systems (2017) [55], where the following prerequisites are proposed as important when discussing AI ethics: Human dignity; autonomy; responsibility; justice, equality and solidarity; democracy; rule of law and accountability; security, safety, bodily and mental integrity; data protection and privacy; and sustainability. This list is supplemented by e.g., Dignum [56] who proposes AI ethics to rest in the three design principles of accountability, responsibility and transparency.

The draft ethics guidelines for Trustworthy AI, by the European Commission’s High-Level Expert Group on Artificial Intelligence (AI HLEG) (2018) [53] propose a framework for trustworthy AI, consisting:

Ethical Purpose: Ensuring respect for fundamental rights, principles and values when developing, deploying and using AI.

Realization of Trustworthy AI: Ensuring implementation of ethical purpose, as well as technical robustness when developing, deploying and using AI.

Requirements for Trustworthy AI: To be continuously evaluated, addressed and assessed in the design and use through technical and non-technical methods

The AI4People’s project (2018) [3] has studied the EGE principles, as well as other relevant principles and subsumed them under four overarching principles. These include beneficence, non-maleficence, autonomy (defined as self-determination and choice of individuals), justice (defined as fair and equitable treatment for all), and explicability.

In addition, several other parties have introduced similar principles and guidelines concerning ethics of artificial intelligence, including Association for Computing Machinery ACM (US), Google,

Information Technology Industry Council (US), UNI Global Union (Switzerland), World Commission on the Ethics of Scientific Knowledge and Technology COMEST, Engineering and Physical Sciences Research Council EPSRC (UK), The Japanese Society for Artificial Intelligence JSAI, University of Montreal, and European Group on Ethics and New Technologies EGE.

Based on the literature review, the table below (Table 1) introduces the ethical values and principles of the most relevant documents in the current European discussion of technology ethics.

Table 1. Ethical values and principles in European discussion of technology ethics.

Expert Group/Publication	Ethical Value/Principle	Context	Technology
Friedman et al. (2003; 2006) [22,23]	Human welfare Ownership and property Freedom from bias Universal usability Courtesy Identity Calmness Accountability (Environmental) sustainability	Value-sensitive design	ICT
Ethically Aligned Design (EAD) IEEE Global initiative (2016, 2017) [4,5]	Human benefit Responsibility Transparency Education and Awareness	Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems: Insights and recommendations for the AI/AS technologists and for IEEE standards	AI/AS
Asilomar AI Principles (2017) [54]	Safety Failure and juridical transparency Responsibility Value alignment Human values Privacy and liberty Shared benefit and prosperity Human control Non-supervision Avoiding arms race	Beneficial AI to guide the development of AI	AI
The European Group on Ethics in Science and New Technologies (EGE) (2017) [55]	Human dignity Autonomy Responsibility Justice Equality and solidarity Democracy Rule of law and accountability Security Safety Bodily and mental integrity Data protection and privacy Sustainability	Statement on Artificial Intelligence, Robotics and Autonomous Systems	AI, Robotics, AS
European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) (2018) [53]	Respect for human dignity Freedom of the individual Respect for democracy, justice and the rule of law Equality, non-discrimination and solidarity Citizens rights Beneficence: "Do Good" Non maleficence: "Do no Harm" Autonomy: "Preserve Human Agency" Justice: "Be Fair" Explicability: "Operate transparently"	Trustworthy AI made in Europe	AI
AI4People (2018) [3]	Beneficence Non-maleficence Autonomy Justice Explicability	An ethical framework for a good AI society	AI

Based on the workshop discussion [57], and as a synthesis of above presented guidelines and values, we propose a modified set of values to be considered as a basis for ethical and responsible development of AIS (Table 2).

Table 2. A preliminary set of ethical values modified for the context of AIS.

Ethical Value	Tentative Topics for Discussion
Integrity and human dignity	Individuals should be respected, and AIS solutions should not violate their dignity as human beings, their rights, freedoms and cultural diversity. AIS should not threaten a user’s physical or mental health.
Autonomy	Individual freedom and choice. Users should have the ability to control, cope with and make personal decisions about how to live on a day-to-day basis, according to one’s own rules and preferences.
Human control	Humans should choose how or whether to delegate decisions to AIS, to accomplish human-chosen objectives.*
Responsibility	Concerns the role of people and the capability of AIS to answer for the decisions and to identify errors or unexpected results. AIS should be designed so that their affects align with a plurality of fundamental human values and rights.
Justice, equality, fairness and solidarity	AIS should contribute to global justice and equal access. Services should be accessible to all user groups despite any physical or mental deficiencies. This principle of (social) justice goes hand in hand with the principle of beneficence: AIS should benefit and empower as many people as possible.
Transparency	If an AIS causes harm, it should be possible to ascertain why. The mechanisms through which the AIS makes decisions and learns to adapt to its environment should be described, inspected and reproduced. Key decision processes should be transparent and decisions should be the result of democratic debate and public engagement.
Privacy	People should have the right to access, manage and control the data they generate.
Reliability	AIS solutions should be sufficiently reliable for the purposes for which they are being used. Users need to be confident that the collected data is reliable, and that the system does not forward the data to anyone who should not have it.
Safety	Safety is an emerging property of a socio-technical system, which is created daily by decisions and activities. Safety of a system should be verified where applicable and feasible. Need to consider possible liability and insurance implications.
Security	AI should be secure in terms of malicious acts and intentional violations (unauthorized access, illegal transfer, sabotage, terrorism, etc.). Security of a system should be verified where applicable and feasible.
Accountability	Decisions and actions should be explained and justified to users and other stakeholders with whom the system interacts.
Explicability	Also ‘explainability’; necessary in building and maintaining citizen’s trust (captures the need for accountability and transparency), and the precondition for achieving informed consent from individuals.
Sustainability	The risks of AIS being misused should be minimized: Awareness and education. Note “precautionary principle”: Scientific uncertainty of risk or danger should not hinder to start actions of protecting the environment or to stop usage of harmful technology.
Role of technology in society	Governance: Society should use AIS in a way that increases the quality of life and does not cause harm to anyone. Depending on what type of theory of justice a society is committed to, it may stress e.g., the principle of social justice (equality and solidarity), or the principle of autonomy (and values of individual freedom and choice).

* This means that an anthropocentrism standpoint is taken, e.g., belief that human beings are the most important entity in the universe. Does AIS design and applications concern other living systems as well?

In the case of autonomous ships, the list of values could include: Integrity and human dignity; autonomy; human control; responsibility; justice, equality, fairness and solidarity; transparency; privacy; reliability; security and safety; accountability; explicability; sustainability; and role of technology in society. The generic goals of the system to be designed are discussed and analyzed in the light of each identified ethical value.

3.3. Analysis and Understanding of Ethical Issues within the Context

Ethical issues are analyzed further to understand them, solve them and to translate them into design language. This outcome contributes to the design requirements. In the first step of the analysis, the goals and requirements may be more generic, but along with more detailed design, the requirements will become more detailed, as well.

Ultimately, how ethical dilemmas are resolved depends on the context [58]. Ethical issues arise regarding the use of specific features and services rather than the inherent characteristics of the technology. The principles and values must thus be discussed on a practical level to inform technology design. To enable ethical reasoning in human-driven technology design, usage scenarios (e.g., Reference [59]) can be used as “cases” to discuss ethical issues. With the help of scenarios, it is possible to consider: (1) What kind of ethical challenges the deployment of technology in the life of people raises; (2) which ethical principles are appropriate to follow; and (3) what kind of context-specific ethical values and design principles should be embedded in the design outcomes.

Therefore, we propose usage scenarios as a tool to describe the aim of the system, the actors and their expectations, the goals of actors’ actions, the technology and the context. The selected principles are cross-checked against each phase of a scenario and the possible arising ethical issues are discussed and reported at each step. Lucivero (2016) [12] indicates that socio-technical scenarios are important tools to broader stakeholder understanding by joint discussions, which enhance reflexivity in one’s own role in shaping the future, as well as awareness of stakeholder interdependence and related unintended consequences. The purpose of the scenario-based discussion is to develop ethical human requirements for the requirements specification and for the iterative design process. The discussion needs to be carried out with all relevant stakeholders and required expertise. The same systematics can be utilized also for assessment of the end-result, or the design decision. The discussion needs to be documented and agreement made transparent so that later it is possible to go back and re-assess possible relevant changes in the environment.

It is not easy to perceive how the final technological outcome will work in society, what kind of effects it will have, and how it will promote the good for humanity. Discussion of the normative acceptability of the technology is thus needed. Usage scenarios can be used as a participatory design tool to capture the different usage situations of the system and the people and environment bound to it. Scenarios describe the aim of the system, the actors and their expectations, the goals of actors’ actions, the technology and the context [60,61]. Socio-technical scenarios can also be used to broader stakeholder understanding of one’s own role in shaping the future, as well as awareness of stakeholder interdependence [12]. In the second step, the scenarios representing different usage situations of the system are discussed with different stakeholders and ethical experts and examined phase by phase according to the listed ethical values, in order to define potential ethical issues. In addition, the following questions presented by Lucivero (2016, 53) [12] can help comprehension of the possible effects of the system in society:

- How likely is it that the expected artifact will promote the expected values?
- To what extent are the promised values desirable for society?
- How likely is it that technology will instrumentally bring about a desirable consequence?

The outcome of the analysis is a list of potential ethical issues, which need to be further deliberated when defining the design and system’s goals.

Case example: Autonomous short-distance electric passenger ship. An initial usage scenario was developed in a series of workshops, to serve here as an example of the scenario work. This scenario is an imaginary example, developed from a passenger perspective, which illustrates what kind of qualitative information can be provided with a scenario to support the identification of ethical issues and the following requirements specification process. The basic elements of the scenario are the following:

- Usage situation: Transport passengers between two pre-defined points across a river as a part of city public transportation; journey time—20 min.
- Design goals: (1) Enable a reliable, frequent service during operation hours; (2) reduce costs of public transport service and/or enable crossing in a location where a bridge can't be used; and (3) increase the safety of passengers.
- Operational model: Guide passengers on-board using relevant automatic barriers, signage, and voice announcements; close the ramp when all passengers are on board; autonomously plan the route, considering other traffic and obstacles; make departure decision according to environmental conditions and technical systems status; detach from dock; cross the river, avoiding crossing traffic and obstacles; attach to opposite dock; open ramp, allow disembarkation of passengers; batteries are charged when docked; maintenance operations carried out during night when there is no service; remote operator monitors the operation in a Shore Control Center (SCC), with the possibility to intervene if needed.
- Stakeholders: Remote operator: In an SCC, with access to data provided by ship sensors. Monitors 3 similar vessels simultaneously; passengers (ticket needed to enter the boarding area), max 100 passengers per crossing; maintenance personnel; crossing boat traffic on the route; bystanders on the shore (not allowed to enter the boarding area); people living/having recreational cottages nearby; ship owner/service provider; shipbuilder, insurance company, classification society, traffic authorities.
- Environment: A river within a European city (EU regulations applicable); crossing traffic on the river; varying weather conditions (river does not freeze, but storms/snow etc. can be expected).

4. Discussion

We have introduced a framework to discuss and analyze ethical issues in AIS design. We have started by introducing current design approaches, concepts and theoretical insights from the fields of Philosophy of Technology and Design Thinking, as well as from different initiatives in the field of AIS. The developed framework introduces the justification and identification of the ethical principles for a specific case study. Then scenario descriptions are introduced to capture the essential user or stakeholder specific qualitative information, which is needed for a systematic analysis of ethical issues in the specific design case. As a result of such a systematic analysis, a list of ethical issues will be identified. These issues need to be further analyzed to transfer them into design goals and requirements.

Our main message is to engage different stakeholders—ethical experts, technology developers, end users and other relevant parties—in adopting a common multi-perspective yet a systematic discussion during an AIS design process. Our initial framework paves way for practical methods for understanding ethical issues in AIS. Further studies are needed to test and assess the approach and to reformulate final principles for the context of autonomous systems in real design cases with real concepts and scenarios.

In our framework, we lean on the human-centered design tradition of using scenarios as a tool for seeking understanding of the needs and desires of people and communities. It should be kept in mind, however, that scenarios, when used as expert's statements on the technological future, can also be used to legitimize and justify the role of a new, not-yet established technology or an application, and thus have a strategic role in welcoming the technology and convincing an audience. One has to be

sensitive to this kind of technological imperative, i.e., developing technology for technology's sake, and to keep in mind that the outcome of ethical analysis can well be that the given technology is not capable of fully answering the needs of the target users in the given context. Ethical analysis, as its best, can have the power of revealing hype around technological rhetoric. Many technological ideas can be explained by 'a human need', but not all technical solutions can be justified in terms of the benefits of the good life.

Author Contributions: Writing, conceptualization, and methodology development by J.L., R.K. and N.G.; original draft preparation by J.L.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. NFA Norwegian Society of Automatic Control. *Autonomous Systems: Opportunities, and Challenges for the Oil & Gas Industry*; NFA: Kristiansand, Norway, 2012.
2. Montewka, J.; Wrobel, K.; Heikkila, E.; Valdez-Banda, O.; Goerlandt, F.; Haugen, S. Probabilistic Safety Assessment and Management. In Proceedings of the PSAM 14, Los Angeles, CA, USA, 16–21 September 2018.
3. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef]
4. IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. Ethically Aligned Design, Version One, A Vision for Prioritizing Human Wellbeing with Artificial Intelligence and Autonomous Systems. 2016. Available online: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v1.pdf? (accessed on 7 March 2019).
5. IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design, Version 2 for Public Discussion. A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems. 2017. Available online: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf (accessed on 7 March 2019).
6. Brynjolfsson, E.; McAfee, A. The Business of Artificial Intelligence. What it Can—And Cannot—Do for your Organization. 2017. Available online: http://asiandatascience.com/wp-content/uploads/2017/12/Big-Idea_Artificial-Intelligence-For-Real_The-AI-World-Confernece-Expo-December-11_13-2017.pdf (accessed on 11 March 2019).
7. Floridi, L. (Ed.) *Information and Computer Ethics*; Cambridge University Press: Cambridge, UK, 2008.
8. Dignum, V. Ethics in artificial intelligence: Introduction to the special issue. *Ethics Inf. Technol.* **2018**, *20*, 1–3. [CrossRef]
9. Anderson, M.; Anderson, S. (Eds.) *Machine Ethics*; Cambridge University Press: Cambridge, UK, 2011.
10. Müller, V.C. (Ed.) *Risks of Artificial Intelligence*; CRC Press, Taylor & Francis Group: Boca Raton, FL, USA, 2016.
11. Kitchin, R.; Dodge, M. *Code/Space: Software and Everyday Life*; MIT Press: Cambridge, MA, USA, 2011.
12. Lucivero, F. *Ethical Assessments of Emerging Technologies: Appraising the Moral Plausibility of Technological Visions*; The International Library of Ethics, Law and Technology; Springer: Heidelberg, Germany, 2016; Volume 15.
13. Anderson, M.; Anderson, S. The status of Machine Ethics: A Report from the AAAI Symposium. *Minds Mach.* **2007**, *17*, 1–10. [CrossRef]
14. Bynum, T. A Very Short History of Computer Ethics. 2000. Available online: http://www.cs.utexas.edu/~{year}/cs349/Bynum_Short_History.html (accessed on 11 March 2019).
15. Pierce, M.; Henry, J. Computer ethics: The role of personal, informal, and formal codes. *J. Bus. Ethics* **1996**, *15*, 425–437. [CrossRef]
16. Veruccio, G. The birth of roboethics. In Proceedings of the ICRA 2005, IEEE International Conference on Robotics and Automation, Workshop on Robo-Ethics, Barcelona, Spain, 8 April 2005.
17. Anderson, S. The unacceptability of Asimov's three laws of robotics as a basis for machine ethics. In *Machine Ethics*; Anderson, M., Anderson, S., Eds.; Oxford University Press: New York, NY, USA, 2011.

18. Powers, T. Prospects for a Kantian Machine. In *Machine Ethics*; Anderson, M., Anderson, S., Eds.; Oxford University Press: New York, NY, USA, 2011; pp. 464–475.
19. Anderson, M.; Anderson, S. Creating an ethical intelligent agent. *AI Mag.* **2007**, *28*, 15.
20. Crawford, K. Artificial Intelligence’s White Guy Problem. *The New York Times*. 25 June 2016. Available online: <https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html> (accessed on 20 January 2019).
21. Kirchner, J.; Angwin, S.; Mattu, J.; Larson, L. *Machine Bias: There’s Software Used across the Country to Predict Future Criminals, and It’s Biased against Blacks*; Pro Publica: New York, NY, USA, 2016.
22. Friedman, B.; Kahn, P.H., Jr. Human values, ethics, and design. In *The Human-Computer Interaction Handbook, Fundamentals, Evolving Technologies and Emerging Applications*; Jacko, J.A., Sears, A., Eds.; Lawrence Erlbaum: Mahwah, NJ, USA, 2003; pp. 1177–1201.
23. Friedman, B.; Kahn, P.H., Jr.; Borning, A. Value sensitive design and information systems. In *Human-Computer Interaction in Management Information Systems: Applications*; M.E. Sharpe, Inc.: New York, NY, USA, 2006; Volume 6, pp. 348–372.
24. Leikas, J. *Life-Based Design—A Holistic Approach to Designing Human-Technology Interaction*; VTT Publications: Helsinki, Finland, 2009; p. 726.
25. Saariluoma, P.; Cañas, J.J.; Leikas, J. *Designing for Life—A Human Perspective on Technology Development*; Palgrave MacMillan: London, UK, 2016.
26. Von Schomberg, R.A. Vision of Responsible Research and Innovation. In *Responsible Innovation*; Owen, R., Bessant, J., Heintz, M., Eds.; Wiley: Oxford, UK, 2013; pp. 51–74.
27. European Commission. Options for Strengthening Responsible Research and Innovation, 2013. Available online: https://ec.europa.eu/research/science-society/document_library/pdf_06/options-for-strengthening_en.pdf (accessed on 20 January 2019).
28. European Commission. Responsible Research and Innovation—Europe’s Ability to Respond to Societal Challenges, 2012. Available online: <http://www.scientix.eu/resources/details?resourceId=4441> (accessed on 20 January 2019).
29. Porcari, A.; Borsella, E.; Mantovani, E. (Eds.) *Responsible-Industry: Executive Brief, Implementing Responsible Research and Innovation in ICT for an Ageing Society*; Italian Association for Industrial Research: Rome, Italy, 2015. Available online: <http://www.responsible-industry.eu/> (accessed on 20 January 2019).
30. Jonsen, A.R.; Toulmin, S. *The Abuse of Casuistry: A History of Moral Reasoning*; University of California Press: Berkeley, CA, USA, 1988.
31. Kuczewski, M. Casuistry and principlism: The convergence of method in biomedical ethics. *Theor. Med. Bioethics* **1998**, *19*, 509–524. [CrossRef]
32. Beauchamp, T.; Childress, J.F. *Principles of Biomedical Ethics*, 5th ed.; Oxford University Press: Oxford, UK; New York, NY, USA, 2001.
33. Mazzucelli, C.; Visvizi, A. Querying the ethics of data collection as a community of research and practice the movement toward the “Liberalism of Fear” to protect the vulnerable. *Genocide Stud. Prev.* **2017**, *11*, 4. [CrossRef]
34. Visvizi, A.; Mazzucelli, C.; Lytras, M. Irregular migratory flows: Towards an ICTs’ enabled integrated framework for resilient urban systems. *J. Sci. Technol. Policy Manag.* **2017**, *8*, 227–242. [CrossRef]
35. Riegel, J. Confucius. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Stanford University: Stanford CA, USA, 2013.
36. Aristotle, H.G. Nicomachean ethics. In *The Complete Works of Aristotle*; Barnes, J., Ed.; The Revised Oxford Translation; Princeton University Press: Princeton, NJ, USA, 1984; Volume 2.
37. Hansson, S.O. (Ed.) *The Ethics of Technology: Methods and Approaches*; Rowman & Littlefield: London, UK, 2017.
38. Beauchamp, T. The Principle of Beneficence in Applied Ethics. In *Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Stanford University: Stanford, CA, USA, 2008.
39. Miller, D. Justice. In *The Stanford Encyclopedia of Philosophy*; Zalta, E.N., Ed.; Stanford University: Stanford CA, USA, 2017.
40. Rawls, J. *A Theory of Justice*; Revised Edition; Harvard University Press: Cambridge, MA, USA, 1999.
41. Kant, I. Grounding for the Metaphysics of Morals. In *Ethical Philosophy*; Kant, I., Ed.; Translated by Ellington, J.W.; Hackett Publishing Co.: Indianapolis, IA, USA, 1983; First published in 1785.

42. Mill, J.S. *On Liberty*; Spitz, D., Ed.; Norton: New York, NY, USA, 1975; First published in 1859.
43. Shrader-Frechette, K.S. *Environmental Justice: Creating Equality, Reclaiming Democracy*; Oxford University Press: Oxford, UK; New York, NY, USA, 2002.
44. Rokeach, M. *Understanding Human Values: Individual and Societal*; The Free Press: New York, NY, USA, 1979.
45. Schwartz, S. Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In *Advances in Experimental Social Psychology*; Zanna, M.P., Ed.; Elsevier Science Publishing Co Inc.: San Diego, CA, USA, 1992; Volume 25, pp. 1–65.
46. Schwartz, S.; Melech, G.; Lehmann, A.; Burgess, S.; Harris, M.; Owens, V. Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *J. Cross-Cult. Psychol.* **2001**, *32*, 519–542. [CrossRef]
47. UN United Nations. Universal Declaration of Human Rights UDHR. Available online: <http://www.un.org/en/universal-declaration-human-rights/> (accessed on 5 June 2018).
48. EU Treaties. Available online: https://europa.eu/european-union/law/treaties_en (accessed on 5 June 2018).
49. EU Charter of Fundamental Rights. Available online: http://www.europarl.europa.eu/charter/pdf/text_en.pdf (accessed on 5 June 2018).
50. Borning, A.; Muller, M. Next steps for value sensitive design. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Austin, TX, USA, 5–10 May 2012; pp. 1125–1134, ISBN 978-1-4503-1015-4/12/05.
51. Habermas, J. Discourse Ethics: Notes on a Program of Philosophical Justification. In *Moral Consciousness and Communicative Action*; Habermas, J., Ed.; Translated by Lenhardt, C. and Nicholsen, S.W.; Polity Press: Cambridge, UK, 1992; pp. 43–115. First published in 1983.
52. European Committee for Standardization (CEN) Workshop Agreement. CWA Ref. No: 17145-1: 2017 E, 17145-2: 2017 E. Available online: http://satoriproject.eu/media/CWA_part_1.pdf (accessed on 7 March 2019).
53. European Commission’s High-Level Expert Group on Artificial Intelligence. Draft Ethics Guidelines for Trustworthy AI. December 2018. Available online: <https://ec.europa.eu/digital-single-market/en/news/draft-ethics-guidelines-trustworthy-ai> (accessed on 20 January 2019).
54. Asilomar Conference 2017. Asilomar AI Principles. Available online: <https://futureoflife.org/ai-principles/?cn-reloaded=1> (accessed on 15 October 2018).
55. European Group on Ethics in Science and New Technologies. Statement on Artificial Intelligence, Robotics and ‘Autonomous’ Systems. Available online: https://ec.europa.eu/research/ege/pdf/ege_ai_statement_2018.pdf (accessed on 16 June 2018).
56. Dignum, V. The Art of AI-Accountability, Responsibility, Transparency. Available online: <https://medium.com/@virginiadignum/the-art-of-ai-accountability-responsibility-transparency-48666ec92ea5> (accessed on 15 October 2018).
57. Leikas, J.; Koivisto, R.; Gotcheva, N. Ethics in design of Autonomous Intelligent Systems. In *Effective Autonomous Systems, VTT Framework for Developing Effective Autonomous Systems*; Heikkilä, E., Ed.; VTT Technical Research Centre of Finland Ltd.: Espoo, Finland, 2018.
58. Ermann, M.D.; Shauf, M.S. *Computers, Ethics, and Society*; Oxford University Press: New York, NY, USA, 2002.
59. Carroll, J.M. *Scenario-Based Design: Envisioning Work and Technology in System Development*; John Wiley & Sons: New York, NY, USA, 1995.
60. Carroll, J.M. Five reasons for scenario-based design. *Interact. Comput.* **2000**, *13*, 43–60. [CrossRef]
61. Rosson, M.B.; Carroll, J.M. Scenario-Based Design. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*; Jacko, J.A., Sears, A., Eds.; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 2002; pp. 1032–1050.



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Review

Ethical Management of Artificial Intelligence

Alfred Benedikt Brendel ^{1,*}, Milad Mirbabaie ², Tim-Benjamin Lembcke ³ and Lennart Hofeditz ⁴

¹ Business Informatics, Especially Intelligent Systems and Services, Technische Universität Dresden, 01169 Dresden, Germany

² Information Systems & Industrial Services, University of Bremen, 28334 Bremen, Germany; milad.mirbabaie@uni-bremen.de

³ Information Management, University of Goettingen, 37073 Göttingen, Germany; tim-benjamin.lembcke@uni-goettingen.de

⁴ Professional Communication in Electronic Media/Social Media, University of Duisburg-Essen, 47057 Duisburg, Germany; lennart.hofeditz@uni-due.de

* Correspondence: alfred_benedikt.brendel@tu-dresden.de

Abstract: With artificial intelligence (AI) becoming increasingly capable of handling highly complex tasks, many AI-enabled products and services are granted a higher autonomy of decision-making, potentially exercising diverse influences on individuals and societies. While organizations and researchers have repeatedly shown the blessings of AI for humanity, serious AI-related abuses and incidents have raised pressing ethical concerns. Consequently, researchers from different disciplines widely acknowledge an ethical discourse on AI. However, managers—eager to spark ethical considerations throughout their organizations—receive limited support on how they may establish and manage AI ethics. Although research is concerned with technological-related ethics in organizations, research on the ethical management of AI is limited. Against this background, the goals of this article are to provide a starting point for research on AI-related ethical concerns and to highlight future research opportunities. We propose an ethical management of AI (EMMA) framework, focusing on three perspectives: managerial decision making, ethical considerations, and macro- as well as micro-environmental dimensions. With the EMMA framework, we provide researchers with a starting point to address the managing the ethical aspects of AI.

Keywords: artificial intelligence; ethical management; research directions

Citation: Brendel, A.B.; Mirbabaie, M.; Lembcke, T.-B.; Hofeditz, L. Ethical Management of Artificial Intelligence. *Sustainability* **2021**, *13*, 1974. <https://doi.org/10.3390/su13041974>

Academic Editor: Alessio Ishizaka

Received: 29 January 2021

Accepted: 2 February 2021

Published: 12 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Artificial intelligence (AI), i.e., “The ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity” [1], is a unique technology for many reasons. Not only is it difficult for humans to understand and verify the decisions of AI [2], but it is also challenging to establish rules for its use as AI is continuously evolving [3]. This “black box” in the application of AI algorithms leads to a lack of transparency even among the creators and poses particular ethical challenges [4]. As part of societies, business organizations are facing issues regarding the opportunities and consequences of an increasingly AI-based economy [5–7]. It is unclear, for example, what happens when AI-based systems are combined and when they produce results that cannot be pre-evaluated.

Alongside AI-enabled technological advancements, AI's influence on societies has also increased. On subjects such as autonomous driving, self-directed weapon systems and cockpit automation, societal considerations do arise, even touching on matters of life and death [8,9]. These significant and potentially adversarial societal influences motivate our article, in which we will argue for the importance of ethical management of AI and how we, as a research community, might address this challenge.

On the one hand, AI's increasing influence on individuals and their societies goes along with the increasing pressure on organizations to assume responsibility for their AI products and offerings, including ethical considerations tied to the potential consequences of their AI's use on social, environmental, and economic levels [10]. On the other hand, it goes along with a noticeable shift within the workforce: increasingly relying on AI will likely replace some routine task-related jobs in order for firms to remain competitive with others shifting to automated practices. In turn, many more qualified jobs will be created in the process, thereby generating an overall transition towards more high-skilled jobs. Ethical AI considerations need to be embodied in managerial decision-making at first, starting with informing day-to-day operations. More and more organizations want to take this responsibility [11], but not every employee has the time and resources to holistically consider and make sense of a currently fragmented scholarly discourse. This fragmentation poses a risk for the social, environmental and economic sustainable use of AI. The discourse on organizational AI ethics is still in its infancy [4], and current research on AI ethics resides within multiple domains, including, but not limited to, philosophy, computer sciences, information systems (IS), and management research [11]. For this article, we formulate the following two research questions:

RQ1: What is the current status-quo regarding research on the management of ethical aspects of AI?

RQ2: What are potential gaps and directions for future research on this topic?

To answer these questions, we conducted a literature search and review, which led us to the conclusion that there is currently no research on this topic. Against this background, our goal is to provide an initial framework on how to conceptualize the management of AI ethics, which will hopefully lead to future research on this topic. We introduce a framework on how to tie together the three perspectives of (1) managerial decision-making, (2) ethical considerations, and (3) different macro- and micro-environmental dimensions with which an organization interacts. Applying this framework to guide decision making is an essential part of an organization's ethical responsibility. In summary, we propose a pragmatic opinion on a conceptualization for ethically managing AI in organizations. By developing the ethical management of AI (EMMA) framework, we propose to open a new research area and provide scholars and practitioners with the first reference on this important research topic.

2. Challenges for Research and Practice

Our motivation is in line with that of scholars who have acknowledged that the societal and environmental impact of machines and AI deserves more attention [4,12–14]. As a foundation, the question arises: Which potential repercussions of AI are beneficial or detrimental or, more abstractly, "right" or "wrong"? The question of "what is the right thing to do?", which is often connected to the question of "what ought we not do?" is, in many cases, more complicated than it seems. That question has thus become the foundation of a whole scientific field—namely, the ethical sciences, a subfield of philosophy [15].

Concerning AI, scholars have begun to establish an ethical discourse (e.g., [16–18]). Primary examples of ethical considerations include the greater complexity of AI and its increasing decision-making autonomy [19]. The complexity makes it harder to understand how and why an AI has come to a particular decision, and which decision it will make in the future (part of "explainable AI" research) [20]. The increasing decision-making autonomy of AI concerns decisions that an AI can take on its own with little or no prior human approval or supervision [21].

This decision-making autonomy, as well as the general use of AI-based systems in organizations, poses ethical issues concerning various environmental dimensions in which an organization operates. The renunciation of this ethical discourse, which by its philosophical and multidimensional nature tends to be controversial, can entail significant and considerable consequences and risks for our society [16]. We are thus in need of more theoretical and academic reflection on the ethical issues and boundaries of AI, especially on

how to empower (future) employees—especially managers—to consider and implement AI ethics in their daily business [17,22]. As a research community, we should ask ourselves how we may contribute to the field of ethical management of AI and provide first guidance in positioning and guiding future research. In this article, we, therefore, highlight scholarly and practical issues regarding the ethical management of AI and provide a first agenda for future research.

3. Understanding the Role of AI as an Ethical Phenomenon

Organizations need to be capable of dealing with ethical questions regarding AI, not least to circumvent unethical as well as potentially harmful consequences of their AI-based technologies [23,24]. Although being part of an AI arms race, organizations need to assume responsibility for considering ethical aspects of AI [25]. Not only may policies and laws demand this [26], but also unheeded consequences may severely impact the overall organization, for instance, via lawsuits or negative media attention [27].

In traditional business and manufacturing contexts, unethical behaviors mostly occur by “design” (e.g., the Dieselgate of Volkswagen or the sub-sub-contracting of parcel delivery services to circumvent labor laws and save costs). Managers are or can easily be aware of potential consequences that their decisions may have. Unethical behaviors, thus, seldomly happen by mere “chance.” In the AI context, however, such unethical behaviors may occur not only by “design,” but also as an unintentional consequence and, thus, by mere “chance” or “external causes” [27]. For instance, the Tay chatbot released by Microsoft in 2016 became racist after being trained by users on Twitter, but had not been intended or designed to become offensive [28]. For organizations, this raises the question of which ethical principles should be used to develop and manage AI-based technologies. In research, one aspect which is gaining more and more attention is the prediction of an industry 5.0 [29]. Unlike industry 4.0, where the focus is on automation, industry 5.0 is about the synergy of humans and robots. In essence, robots and humans are collaborators instead of competitors [29]. Hence, the focus of industries can be expected to shift away from technical development of systems and towards the social needs of people [29]. However, there is a lack of guidelines and frameworks on how to make AI manageable. In order to gather an overview of research on the intersection of ethics, AI, and management, we conducted a comprehensive literature search in the beginning of 2020 to identify existing conference and journal publications within common databases (AISEL and Scopus databases). In order to be able to map the ethics research perspectives more clearly, we also researched a comprehensive philosophical database (Philpapers.org database). We used the search term “AI AND (Ethics OR Ethical)”, resulting in 552 hits and a sample of 192 papers after filtering (see Figure 1).

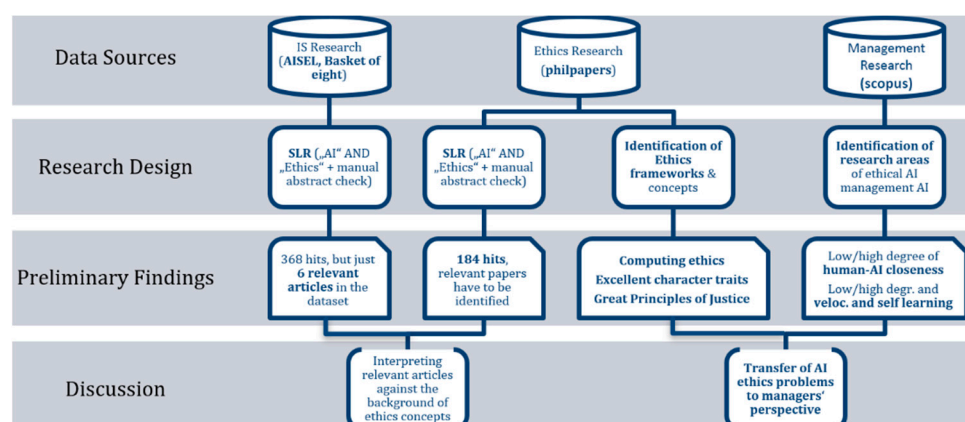


Figure 1. Visualization of the literature review process.

Based on ethical issues regarding the use of AI in organizations, the question arises as to how to structure and classify AI ethics to make them manageable and to open a subfield

for researchers. Since AI is continually evolving, what has been previously considered as AI may not be defined as such today, a phenomenon known as the “AI effect” [3]. Therefore, it is unfeasible to provide a unified and precise threshold between AI and non-AI. Currently, AI is based on different algorithms and techniques, such as supervised learning, unsupervised, or deep learning. These different approaches result in AI-based systems with different velocities of self-learning. Differences in the quality and quantity of training data also mean that the capabilities of AI in organizations vary widely, which emphasizes the reason why we consider that the quality of being self-learning is one central distinguishing characteristic of AI in organizations. According to Xia et al. [30], we define self-learning as a logical model about the self-adaptive goal achievement of software.

Regarding the ethical management of AI, if AI can (at least for now) be classified on the basis of velocity of self-learning, the question arises as to how we can structurally consider the ethical aspects of AI in organizations. Ethics is defined as that part of philosophy that deals with the prerequisites and evaluation of human action and is the methodical reflection on morality [31]. At the center of ethics is a specific moral action, especially about its justifiability and reflection. We assume ethical considerations to be of higher relevance if an AI-enabled technology is in closer interaction with humans. AI tools such as recommendation, forecasting, or optimization algorithms that increase the efficiency of large data sets’ analyses are not, per se, high priorities for ethical consideration, because their direct impact on human lives can be considered as low [32]. The same applies to areas of application such as database mining and optimizations [33], as they do not have significant, influential potential on societies or individuals. Furthermore, AI may directly interact with users in instances such as chatbots in customer service [34], or impact the lives of individuals or minority groups by disadvantaging applicants for interviews in HR [35] or steering self-driving cars [36]. Based on these two assumptions, we divide AI in organizations along the following two dimensions:

1. The degree and velocity of self-learning;
2. The degree of AI’s impact on humans.

As a result of these two definitions, we propose an AI positioning matrix integrating both perspectives as dimensions (see Figure 2). Note, the positioning of cases within the matrix is tentative, and their precise positioning may be argued. We have selected and classified the cases by way of example to describe the range of current AI-based technologies and do not claim to be complete, nor are these based on concrete numbers.

Our AI positioning matrix resembles a portfolio matrix approach as it is prevalent in business administration [37,38] and widely adopted in practice, for example by the Boston Consulting Group and McKinsey [39], in which the spanned dimensions are divided into three sectors (see Figure 2). The first sector covers AI-based technologies that we consider to have a low degree of self-learning and a low degree of impact on humans. Due to both a lower level of human–AI closeness and self-learning, the chance of impactful errors caused by AI is lower as well. Accordingly, we do not classify this sector as particularly relevant for the ethical management of AI. The second sector concerns all cases of both a medium level of self-learning and impact on humans. This sector is more relevant for the ethical management of AI, as these technologies can have a more significant impact on people’s lives and behavior. The third sector covers AI technologies that have a high impact on people’s lives and which we classify as possessing a high degree of self-learning. One case might be an intelligent decision support system in eHealth, which can help medical staff in making decisions about people’s health conditions and in recommending treatments [40]. In this article, we primarily consider AI-based technologies from the second and third sectors as relevant for ethical management.

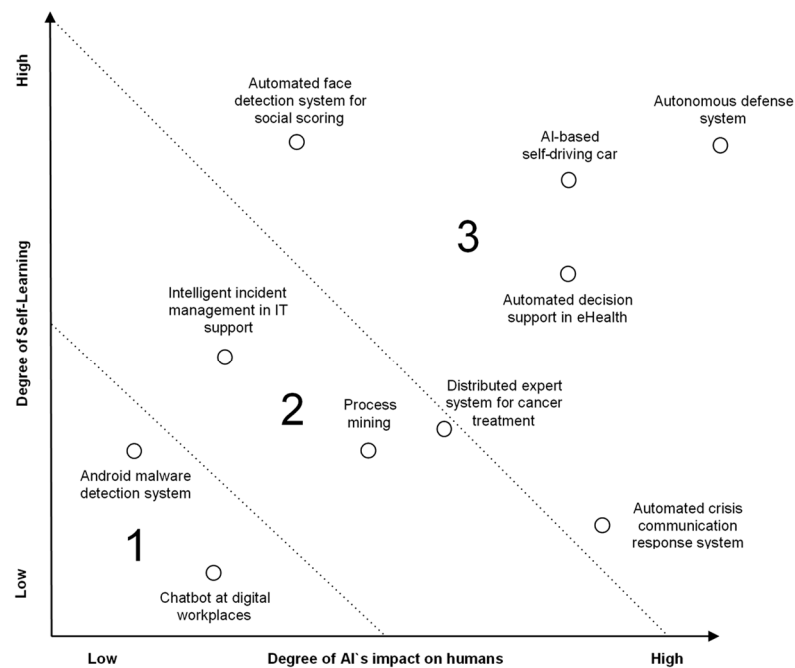


Figure 2. Positioning matrix of AI and ethics.

4. Conceptualization of Ethical Management of Artificial Intelligence in Organizations

Given the previously introduced positioning matrix, different AI-related endeavors can be examined and pre-ranked regarding their potential implications on humans and societies and, likewise, their necessity to be ethically assessed. AI should meet ethical standards as well, particularly if of a strong potential influence on humans and societies and a high level of self-learning. In order to sustain a company’s competitiveness, corporate offerings do not only require the satisfying of a customer need, but also the complying with further standards, including ethical considerations [41]. We propose that, in order to be able to manage AI ethics, we need to consider an interplay of three parts. First, AI-related managerial decisions need to embody ethical considerations if the endeavor is ethically charged (for instance, a project in sector two or three). Second, to incorporate ethical aspects, managers need to have an ethical reference frame within which they can match different potential decisions with ethical considerations. Third, different dimensions of an organizational environment, including but not limited to stakeholder groups, need to be taken into consideration. This triad highlights our understanding that all parts are interconnected with each other, forming the EMMA framework (see Figure 3).

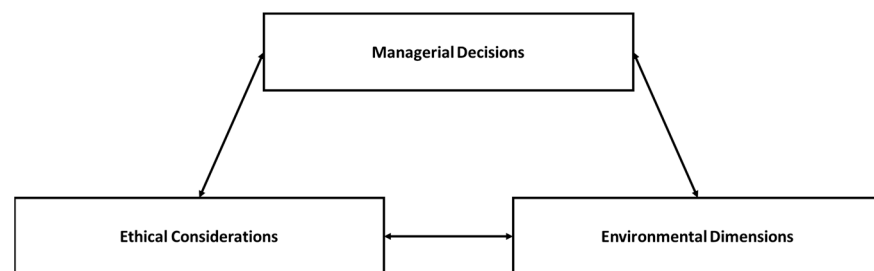


Figure 3. Ethical management of artificial intelligence (EMMA) framework.

4.1. Managerial Decisions

In principle, all services and products offered by organizations need to be purposeful in order to be valuable, whether it is customer value manifesting in prices paid or shareholder

value reflected by market valuations [39,40]. Such offerings are the result of a value creation process (e.g., manufacturing or software development). Operational actions are influenced by decisions that guide an organization's value creation process. Hence, organizational decision making guides and steers the daily course of action. To underline the importance of shaping ethically sound AI-enabled offerings, we argue for incorporating ethical considerations within organizational decision making. Given the significance of AI ethics [42], we assume a holistic responsibility of different organizational decision-making levels for adhering to an organization's ethical reference frame. Section 5.1 offers an exemplary operationalization of this managerial decision making.

4.2. Ethical Considerations

To shape an ethical reference frame, organizations need to decide on their ethical foundations. Basic considerations on how the business should be carried out and which standards should be adhered to are relevant aspects for such an ethical reference frame. This frame should be flexible yet specific enough for research and to allow managers to challenge and evaluate complex organizational decisions. For instance, managers should not only be able to gauge new products and services against this frame, but also steer the organization overall (e.g., cross-sectional tasks such as human resources). Section 5.2 introduces exemplary ethical streams and considerations relevant to an ethical reference frame.

4.3. Environmental Dimensions

Ethical considerations and the derivation of a reference frame, in turn, do not occur in vacuo. Ethical aspects are influenced by—and themselves influence—an organization's environment on both inner- and outer-organizational levels. For instance, stakeholders such as customers, societies, or political landscapes, may be influenced or impacted by product and service offerings provided in unethical ways. As employees and decision-makers may not be mindful of potential environmental dimensions, it would be beneficial to provide both groups with appropriate guidance on the environmental dimensions to be considered. Section 5.3 explicates an exemplary operationalization of these dimensions.

5. Applying the Ethical Management of Artificial Intelligence Framework

Extending from the previously outlined general considerations, leading to the proposition of the EMMA framework (see Figure 1), we will in this section provide an example of how the individual components of EMMA can be operationalized (see Figure 4). In the end, companies have to adapt EMMA according to their decisions' process structure, ethical values, and environmental factors.

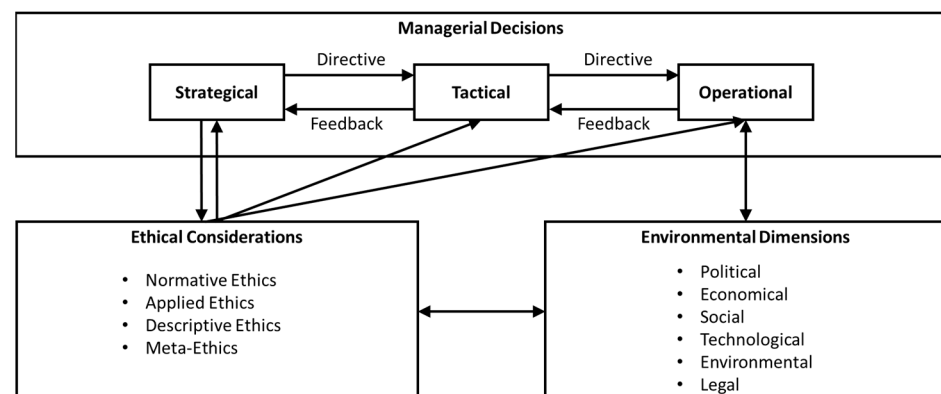


Figure 4. Instantiation of the ethical management of AI framework.

5.1. Operationalization of Managerial Decisions

To operationalize managerial decisions, we opted for a segmentation regarding the different levels of organizational decision making. A separation into strategic, tactical, and operational management and decision making has been widely accepted and has become one of the cornerstones of strategic management [43,44].

With strategic decisions touching on the overall vision and strategy, they are considered more long-term oriented and aim at steering the overall organization to stay ahead of the competition [45]. The strategic level can define and change key ethical considerations informing and guiding the overall organization. Strategic management directs tactical decision making, which focuses on how the organizational strategy and vision can be enacted [46]. With a mid-term focus, tactical management converts strategy into action plans. For instance, it weighs potential AI options against each other and directs the operational management with a strategic–tactical decision-making framework. Eventually, the operational level is concerned with developing, implementing, and applying tactical decisions [44].

In the case of AI, ethical challenges may arise on the operational level (e.g., an AI may be acting in non-predicted ways, overstepping ethical boundaries). Given a hierarchical organization, general employees may stay at arm’s length from strategic or tactical management, both intentionally and unintentionally. Decision makers may not have developed a close communication with the employees implementing the decision, or such employees may feel uncomfortable in openly addressing challenges arising during daily business. These considerations underline the important rule of an organizational feedback culture across the entire hierarchy, favoring and empowering employees to speak up and to be heard [44]. In sum, said concerns render the feedback function an essential part of the instantiated EMMA.

5.2. Operationalization of Ethical Considerations

In order to enable the development of ethical considerations, we looked at three main research streams of ethics—meta-ethics, normative ethics, and applied ethics—as well as descriptive ethics [47]. This section provides a pivotal overview of the ethical concepts potentially relevant for EMMA based on a comprehensive literature review on AI from an ethics research perspective (Tables A1–A4 in Appendix A). We hereby introduce relevant ethical considerations for our EMMA framework.

As the first stream, we identified epistemic perspectives as the most relevant part of meta-ethics in the context of ethical AI use (Appendix A Table A1). Epistemology deals with how knowledge can be derived and, regarding AI, how to identify ethical requirements for AI-based technologies [48]. In addition, we are aware of other ethical views, such as the antinatalism of Schopenhauer and Heidegger’s work “Being and Time”. As an exemplary challenge regarding AI, Coeckelbergh [49] compares the creation of AI with the Frankenstein problem and refers to Heidegger. Instead of trying to control it, he stated that we have to let go. This means he suggests that we should wait for a change to happen.

The normative ethics point of view, i.e., prescriptive ethics, leads to an evaluation of whether an action is perceived as good/ethical or not and eventually arrives at normative guidelines [50]. We aim to formulate issues regarding the management of AI in the form of ethical questions, taking into account the basic structure of normative ethics (Appendix A, Table A2), and have identified Max Weber as one of the most important pioneers of modern normative ethics [50]. We also include ethics of responsibility as part of a deontological view in our normative ethics considerations. This ethical viewpoint refers to Norbert Wiener’s great principles of justice, which are (1) the principle of freedom, (2) equality, (3) benevolence, and (4) the minimum infringement of freedom. As a complementary principle, we also supplemented dignity, as it was a highly relevant component of some normative ethical studies that considered dignity in the context of ethical AI use [51]. As a last important concept of normative ethics, we have identified Plato’s virtues, which

include (1) courage, (2) justice, (3) temperance, and (4) prudence, which were also revisited in the context of AI [52].

To cover an applied ethics perspective, we focused on aspects of business ethics (Appendix A, Table A3). We applied computer and information ethics on Norbert Wiener's point of view on justice and also covered recently introduced topics such as (Kantian) moral agents and cyborg ethics.

For the descriptive ethics stream, we identified what we framed as a criminal perspective (Appendix A, Table A4). Following Spinellis [53], this includes the absolute and relative punishment theories which focus on punishment approaches for unethical behavior (according to Kant and Hegel). As another relevant aspect, we identified the deliberation of actions, such as harmful actions with or without an intention or non-harmful actions that become harmful through manipulation [54]. We also considered reasonings and individual perspectives such as cognitive and emotional control [55] in the context of criminal actions.

Against the background of these focal points, we consider the management of AI in a structured manner to derive relevant questions for research and practice. Nonetheless, future research will be necessary to provide further indications on how to precisely shape and render ethical reference frames for organizations, for instance, extending previously non-AI-related frameworks for ethics in organizations [56].

5.3. Operationalization of Environmental Dimensions

As introduced in Section 3, ethical considerations do not happen in vacuo but need to be considered per different environmental dimensions. We decided to follow a classification scheme of political, economic, socio-cultural, technological, environmental, and legal (PESTEL) dimensions, inspired by the PESTEL analysis, a traditional approach to environmental scanning in strategical management [44]. The PESTEL analysis describes a business environment for specific market conditions, developments, and their effects, in order to shape sound decision-making principles for the management of an organization [57]. The initial analysis assumes that different external dimensions influence a company's success, and thus its management [58]. However, a company and its management can also influence the environmental dimensions.

5.4. Reflections on Future Research Opportunities

Bearing in mind the prior operationalization that serves as one potential instantiation for our EMMA framework (Figure 3), this subsection ties the framework's three perspectives together. Taking the six dimensions suggested by the PESTEL framework into account, managerial and academic questions from different standpoints arise. In the following subsections, we present an initial opinion on various aspects induced by EMMA that a future research subfield may address.

As an overall key consideration and basic premise, strategical management needs to establish an organization-wide ethical code of conduct to inform ethical evaluations and decision making. Without such ethical principles and guidelines, it is hard to gauge decisions and compare alternatives with respect to ethical considerations. To investigate the influence of AI on different dimensions, we have adapted PESTEL as an exemplary classification framework. This objective is critical to understand in order to define the technological dimension of PESTEL, which considers the impact of an AI on other technologies, but not the societal impact of a particular AI (this would be part of the societal dimension).

Each PESTEL dimension holds a basic premise, which we understand as the theoretical and hypothetical influencing potential of an AI. For this deliberation, it is not crucial that the influence is exerted, but rather that the influence may or could be. Afterward, we assume the strategical management to be responsible for deciding if an ethical influence is justifiable in general, and to what degree and within which boundaries in particular. These considerations become part of the strategic decision making and guide the tactical management function, whose task is to transform a more general strategical decision into a set of tactical decisions that can be implemented by the operational management

function. On the tactical level, the decision upon an AI's ethical justification is made. This decision includes the task of identifying, evaluating, and comparing different approaches to fulfill the strategical decision within the strategical boundaries. Tactical management also needs to consider ethical boundaries attached to sourcing external AI (e.g., from other companies). As a result, the strategical decisions are amended by tactical directions as to how the AI shall be enacted. Eventually, on the operational level, core questions of implementability arise. In light of AI, the main challenges are to develop and apply an AI so that it remains within the strategic–tactical decision-making framework, and to comply with the ethical boundaries and the organization's ethical code of conduct. Since AI can be an undetermined endeavor with regards to its implementation and outcome, the operational level cannot rely on the traditional execution function but has to be empowered in order to be sensitized and foresee potential ethical conflicts arising during development, implementation, or application of an AI.

5.4.1. Political

Basic premise: What influence may AI used by organizations reasonably exert on politics, and how may it influence the organization's perception by politics?

The political dimension includes but is not limited to: policy setting and legislation; political stability; self-defense and military; trade; and taxation.

The political dimension demands organizations to consider the potential influences their AI may exert on politics. Within the political dimension, EMMA does not selectively refer to policy setting [59], but governmental functions and the political system as a whole. An organization's AI may not only influence politics but also shape how an organization is being perceived in political landscapes. In light of lobbying, organizations may use AI to directly influence political and public opinion [60,61]. By seizing a user base's inertia, habits, or prejudices, an organization's AI may influence its users to subvert political decisions.

Previous research already indicated that artificial intelligence, and especially social bots, were used to influence public opinion in political discourses. For example, Bessi and Ferrara [60] identified automated Twitter accounts during the US presidential election campaign in 2016 that tried to spread specific political opinions and manipulate political communication on Twitter.

On the other hand, there is also the question of how politics can influence the ethical management of AI. One ethical researcher discussed the role of punishment for ethically-wrong actions regarding AI [62]. Politicians could use their influence to develop strategies that punish organizations not following politically desired ethical codices.

5.4.2. Economic

Basic premise: What influence may AI used by organizations reasonably exert on the economic system, and how may it influence the organization's perception in the economic system?

The economic dimension includes but is not limited to: the economic system; financial markets; economic growth; and market valuation.

In the economic dimension, organizations should holistically consider their AI's influence on the economic system. Here, "the economic system" refers to the market, national, and global economy. By accumulating bargaining power or significant market shares, organizations can have an increasing influence on an economic system. Higher market power may render it more accessible for organizations to act in their interest, to say nothing of ethical considerations for the general welfare. As a consequence, AI may constitute significant risks for economic stability. These risks were presaged by the global financial crisis starting in 2007, partly fueled by derivatives being sold out by highly complex algorithms self-reinforcing each other [63,64]. Similar economic power may arise from developments such as robo-advisors autonomously investing and trading [65]. Conversely, organizations may use AI to gain economic advantages that may not align

with public welfare or societal goals. For instance, being able to train an AI with unethically derived training data can lead to competitive advantage and increase an organization's market valuation, adversely affecting organizations that operate ethically.

If a manager considers how AI can be applied to influence an economic system, the first step should be to develop AI's consciousness of moral actions further. In business ethics research, this is termed a "moral agent" [66]. These moral agents could be used to support decision-making processes [67] regarding the economic system.

5.4.3. Social

Basic premise: What influence may AI used by organizations reasonably exert on societies, and how may it influence the societal perception of the organization?

The social dimension includes but is not limited to: society's ethical and moral values, organizational working culture; and organizational reputation.

The social dimension deals with AI's impact on societies. Both the society surrounding an organization as well as the society within an organization have to be considered. The external perspective focuses on the impact that AI may have on customers and societies as a whole. One particular conflict can be that societal and organizational ethical values differ [68]. An organization's management may prioritize maximizing shareholder value, while society may value ethical considerations conflicting with shareholder profits [24]. Thus, management has to be mindful of a society's ethical compass and if their AI complies with it. From an internal perspective, AI may influence the organization itself, for instance, by implementing AI to optimize an organization and support or replace employees [5]. In optimizing work processes, AI may overburden employees, setting unreachable or unsustainable work goals, eventually leading to phenomena such as technology-induced stress, or "technostress" [69]. If supporting organizational decision making and receiving a high degree of autonomy, AI-based leadership may also raise ethical concerns as to balancing organizational needs with those of employees.

Even if an AI adheres to an organization's ethical framework, the actions of this AI may still be judged as immoral by society. Accordingly, the social dimension is particularly challenging, as the moral values of a society—and the ethical standards that societies have established to respond to such values—can be subject to unexpected changes and may not be codified as specific laws.

As one of the most important concepts of philosophy, Kant's categorical imperative could be used as a code of conduct for AI's interaction with society. Etzioni and Etzioni [18] also considered AI and ethics against the background of Kant's categorical imperative and discussed the effects on humans. Tonkens [70] denies, however, that the standard form of Kantian artificial moral agents agree with the spirit of Kant's philosophy, and he demands that further ethical principles be used to develop an ethical framework for AI.

Furthermore, the use of AI can be associated with various ethical risks and rewards for societies [71]. One risk is the possibility of cognitive degeneration if a person is cognitively overburdened. AI can also limit autonomy, as it can provide different predefined choices, and it can replace interpersonal relationships. Although this also offers possibilities in cases where interpersonal relationships are not possible, there are risks if communication takes place exclusively via AI assistants.

5.4.4. Technological

Basic premise: What influence may AI used by organizations reasonably exert on the use of technologies, and how may it unfold and evolve within technologies, potentially influencing living beings?

The technological dimension includes but is not limited to: AI-enabled products and services; AI development and implementation; AI explainability and transparency; human imitation and impact on humans; and technology assessment.

Although AI as a technology is also implicit in the other PESTEL dimensions, this technological dimension focuses mainly on an AI's potential influences on other technolo-

gies. As a result of this, we understand, for example, AI to AI (AI2AI) interactions in which one AI exchanges information or prescribes decisions to another AI. One instance may be a drone as an autonomous weapon system that may have one AI for coordinating and detecting offenses, one AI for deciding on firing a drone, one AI for pathfinding, and another AI for deciding on the ideal point in time to ignite the drone's warhead. These systems exchange information with each other, and as such decisions happen more and more autonomously—and at lightning speed—the initial debate has begun in favor of such systems that include some control function. For instance, an AI may serve as a lawyer or ethical instance, overseeing all other systems and autonomously deciding about the appropriateness of an attack [8]. The financial system provides another example of an AI2AI system with algorithms trading increasingly autonomously [72], which is not, *per se*, unethical. However, if the individual or societal welfare is impaired, the consequences of these AI decisions become relevant.

From a philosophical point of view, social choice ethics also address the question of how an AI technology can be developed so that it can make moral decisions. Baum [73] identified three decisions based on normative ethics that must be made in this regard: (1) standing (concerning whose views on ethics are included), (2) measurement (concerning how their views are identified), and (3) aggregation (concerning how personal views are combined to a single view that could guide AI behavior). These decisions would, in any case, have to be made before the development of AI and should not be left to the self-learning AI [73]. Beckers [74] asked himself what risks arise when we create AI in our image. He raises several assertions and principles that should be considered when creating AI, such as the fact that we do not yet fully understand intelligence as a concept. He also discusses conventional philosophical approaches, such as antinatalism, and explains why he rejects them and why an intelligent AI should be built under specific circumstances [74].

5.4.5. Environmental

Basic premise: What influence may AI used by organizations reasonably exert on resource utilization, and how may it influence the organization's perception as environmentally friendly?

The environmental dimension includes but is not limited to: resource utilization; power consumption; waste management.

The environmental dimension addresses considerations regarding the overall impact of an organization's products and services on its environment. Aspects include natural resource utilization, energy consumption, and accompanying effects such as greenhouse gas emissions [75]. AI may necessitate the use of natural resources in an unsustainable way, *i.e.*, using up resources faster than they can regenerate. Primarily, AI itself may use up resources such as rare earth elements and conventionally generated power that is necessary for the IT to run AI. Such environmental influences can lead to new policies being enacted, negative public relations, or even some kind of collective boycott.

Another vital contribution to the debate was made by Sparrow, who outlined a discourse on responsibility for AI in crises [76]. He raised the question of who can be held responsible for war crimes perpetrated by autonomous robots. The programmer? The commanding officer? The machine itself? In order to clarify this problem, he compared it with the problem of child soldiers. There, too, there would be no answer to the question, since child soldiers also sometimes acted autonomously. Sparrow, therefore, opposes the placement of autonomous AI in war zones.

5.4.6. Legal

Basic premise: What influence may AI used by organizations reasonably exert on the legal framework, and how may it influence an organization's perception as legally compliant?

The legal dimension includes but is not limited to: health, safety, consumer, product and labor laws; and data privacy policies.

The legal dimension considers AI's influence on, or interference with, a legal framework. Notwithstanding rare instances of ethical actions in favor of a higher good that allow for illegal actions, as may be the case with some freedom or democratic movements, we assume that illegal actions, in the majority of cases, will also be unethical [77]. Jones [78] conceptualized this commonality as "an unethical decision [to be] either illegal or morally unacceptable to the larger community". The legal framework can be external (i.e., the law and order of a government or public authority), but also internal (such as organizational policies and work rules) [79]. Although legal aspects resonate within the other PESTEL dimensions, there are distinctive legal issues present. An AI may become able to unveil legal loopholes and allow an organization to act in a law-abiding but unethical manner [27].

At the same time, an AI may collect data that is unnecessary or even forbidden by privacy policies without users noticing this. Under those circumstances, AI may criminalize uninformed or unwary users without their knowledge—for instance, through unethical or illegal data processing. Already, in non-AI instances, such legal issues have arisen. For instance, WhatsApp has been accused of automatically collecting and processing user data, including users' entire mobile phone contacts list, thereby misleading their users to disobey local laws [80]. In training complex AI models, similar instances of illegal or unethical data collection may render similar legal challenges [81].

Ethics research provides some examples of how legislation could and should influence developments in AI. Under the term AIonAI, for example, a new law was introduced that deals with cases of interaction between different AIs [82]. In order to implement a law on AI2AI interaction, Ashrafian [82] followed the declaration of human rights and reviewed each article to see if it could be applied to AI or not. Kant's categorical imperative could also be used to create laws for AI [18].

6. Discussion

With AI spreading into almost every aspect of our lives, this article illustrates that AI touches on pertinent ethical issues, effecting our society on social, environmental, and economic levels [12,18]. This article set out to address the issue of a fragmented discourse on the ethical management of AI by providing a synthesized framework (EMMA) supporting both scholarly research and organizational implementation. As with most ethical discourses, organizational or business ethics cannot be seen as black and white, or right and wrong [83]. Although certain bottom lines are widely agreed upon (such as the universal declaration of human rights, UDHR, of the United Nations), ethical considerations bear a robust cultural imprint [23,84–86]. Nevertheless, this should not impede scholars from highlighting the importance of ethical considerations. With AI's advancing capabilities, we assume the consideration of EMMA to be an ongoing quest for organizations researchers.

In connecting the philosophical discourse with the managerial decision levels and the PESTEL environmental dimensions, our operationalized instantiation of the EMMA framework demonstrates a significant scholarly contribution for both range and impact [87]. To the best of our knowledge, it is the first comprehensive and holistic account focusing on the issue of how to make AI ethics manageable in organizations. Providing a foundation for a research field of managing AI ethics, our proposed positioning matrix and EMMA framework help scholars to position their research projects and to address existing research gaps. Complementarily, our instantiated EMMA framework can have a broad impact on businesses and societies and may support management in assuming its ethical responsibility. Our positioning matrix allows managers to prioritize different AI projects according to their potential importance for ethical consideration. The instantiation of our EMMA framework serves as a managerial starting point, identifying key questions and conflict lines as well as presenting possible effects of AI on different organizational decision-making levels and environmental dimensions. So, what can we, as a community, do?

Lastly, we would like to acknowledge that philosophical sciences are more experienced in pure ethics research, computer sciences in AI, management sciences in management,

and the like. However, this article highlights that EMMA is a cross-sectional topic in need of research and scholars able to connect different “scholarly conversations,” in line with the reasoning for cross-paradigm and interdisciplinary research [88,89].

7. Limitations and Avenues for Future Research

Our article is not free of limitations. First, in proposing the research subfield of EMMA, an EMMA framework has been introduced on an overarching level. With our guiding questions in mind, future research can further explicate our foundation, for instance, by deriving specific (and potentially commensurable) managerial guidelines to ethically manage and evaluate AI for a particular environmental, organizational, cultural, or further specificities (e.g., [90–92]). Second, this article focuses on the ethical perspective. In future accounts, the interlinkage of employees’ and decision-makers’ moral values with organizational ethics may receive further elucidation—for instance, drawing on the value management discourse (e.g., [93,94]). Third, as for all literature review based approaches, we were limited to the articles accessible to us during the review process. Hence, future research can expand and challenge our results and propositions by conducting a new review.

8. Conclusions

In this article, we examined the current extent to which research and practice has engaged with the challenge of managing the ethical aspects of including AI in products and services, potentially leading to unintended ethical consequences. Based on our literature review results, we concluded that this topic is in its infancy, lacking a clear framework of what to consider. Against this background, we developed a general EMMA framework, consisting of the interrelation of managerial decision, ethical considerations and environmental dimensions. We operationalized this framework to develop a set of research questions, which we consider to be at the core of future EMMA research.

In sum, we encourage scholars to build on our work and to provide their perspectives on EMMA. In times of increasingly accelerating technology cycles, we should not forget about the ethical implications of our actions. In all modesty, we hope for our EMMA framework to spark an essential discourse on how to make theoretical considerations about ethics feasible and manageable—a discourse whose time seems to have come.

Author Contributions: Conceptualization, A.B.B. and M.M.; methodology, A.B.B. and M.M.; validation, A.B.B. and M.M.; formal analysis, T.-B.L. and L.H.; investigation, T.-B.L. and L.H.; resources, T.-B.L. and L.H.; data curation, T.-B.L.; writing—original draft preparation, T.-B.L. and L.H.; writing—review and editing, A.B.B.; visualization, T.-B.L. and L.H.; supervision, A.B.B. and M.M.; project administration, A.B.B. and M.M.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: Open Access Funding by the Publication Fund of the TU Dresden.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Table A1. Meta ethics.

Ethics Sub-Stream	Philosophical Approach	Principles	Sources
Antinatalism	Metaphysical antinatalism Modern antinatalism (Schopenhauer)	Why we should not create new humans	[74]
Modern hermeneutics and existential philosophy	Being and Time (Heidegger)	Hermeneutic Phenomenology Ontological approach	[49]

Table A2. Normative ethics.

Ethics Sub-Stream	Philosophical Approach	Principles	Sources
Consequentialist Actions ethical if outcome viewed as beneficial	Max Weber: Ethics of Conviction	Tradition Institutionalized patterns Charisma Leaders' persuasiveness Legal Legitimacy by adhering to impersonal rules and universal principles subject to suitable legal–rational reasoning	[95,96]
Deontological Actions ethical if adhering to institutional rules, regulations, laws, and norms—including socially accepted norms	Max Weber: Ethics of Responsibility —> Great Principles of Justice (Norbert Wiener)	Societies should be built on: 1. Principle of Freedom 2. Equality 3. Benevolence 4. Minimum Infringement of Freedom	
Virtues (Plato and Socrates) Character of a moral agent as driving force; actions as a reflection of the moral character	Plato's virtues	1. Courage 2. Justice 3. Temperance 4. Prudence/Wisdom 5. (Dignity)	[97]

Table A3. Applied ethics (business ethics).

Ethics Sub-Stream	Philosophical Approach	Principles	Sources
Computer and Information Ethics	Great Principles of Justice (Norbert Wiener)	1. Principle of Freedom 2. Equality 3. Benevolence Minimum Infringement of Freedom	[98]
	Ethics methodology of Norbert Wiener:	1. Identify an ethical question or case 2. Clarify any ambiguous ideas/principles 3. Apply already existing, ethically acceptable principles, laws, rules, and practices Use the purpose of a human life plus the great principles of justice to find a solution	
	Recently introduced topics of business ethics	1. Online ethics 2. "Agent" ethics 3. Cyborg ethics 4. The "open source movement" 5. Electronic government 6. Global information ethics 7. Computing for developing countries Ethics and nanotechnology	
	Kantian artificial moral agents	According to Categorical Imperative	

Table A4. Descriptive ethics (criminal perspective).

Ethics Sub-Stream	Philosophical Approach	Principles	Sources
Wrong ethical action → Consequence → Function of punishments	Absolute punishment theory (Kant/Hegel) Punishment necessary	Retaliation theory Atonement theory Theory of debt settlement	[99]
	Relative punishment theory Punishment necessary to avoid repetition of wrong actions	Specialized prevention e.g., imprisoning → preventing further actions Resocialization Improving General prevention Change societal means	
Purpose/Deliberation of Action	1. Harmful action with intention 2. Harmful action without intention 3. Unharmful action becomes harmful through manipulation		[54]
Individuals' actions	Reasoning	Cognitive Control Emotional/Affective	[55]
	Perspectives	Individual Organizational Societal	[100]

References

- Rai, A.; Constantinides, P.; Sarker, S. Editor's Comments: Next-Generation Digital Platforms: Toward Human-AI Hybrids. *Manag. Inf. Syst. Q.* **2019**, *43*, iii–ix.
- Felzmann, H.; Villaronga, E.F.; Lutz, C.; Tamò-Larrieux, A. Transparency You Can Trust: Transparency Requirements for Artificial Intelligence between Legal Norms and Contextual Concerns. *Big Data Soc.* **2019**, *6*, 1–14. [CrossRef]

3. McCorduck, P. *Machines Who Think: A Personal Inquiry into the History and Prospects of Artificial Intelligence*; CRC Press: Cleveland, OH, USA, 2004; ISBN 978-1-56881-205-2.
4. Wang, W.; Siau, K. Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda. *J. Database Manag.* **2019**, *30*, 61–79. [CrossRef]
5. Frey, C.B.; Osborne, M.A. The Future of Employment: How Susceptible Are Jobs to Computerisation? *Technol. Forecast. Soc. Chang.* **2017**, *114*, 254–280. [CrossRef]
6. Munoko, I.; Brown-Libur, H.L.; Vasarhelyi, M. The Ethical Implications of Using Artificial Intelligence in Auditing. *J. Bus. Ethics* **2020**, *167*, 209–234. [CrossRef]
7. Russell, S.J.; Norvig, P. *Artificial Intelligence: A Modern Approach*; Pearson Education Limited: London, UK, 2016.
8. Coppersmith, C.W.F. *Autonomous Weapons Need Autonomous Lawyers*; The Report: Vacaville, CA, USA, 2019.
9. Holford, W.D. An Ethical Inquiry of the Effect of Cockpit Automation on the Responsibilities of Airline Pilots: Dissonance or Meaningful Control? *J. Bus. Ethics* **2020**. [CrossRef]
10. Martin, K. Ethical Implications and Accountability of Algorithms. *J. Bus. Ethics* **2019**, *160*, 835–850. [CrossRef]
11. Kolbjørnsrud, V.; Amico, R.; Thomas, R.J. *The Promise of Artificial Intelligence*; Accenture: Dublin, Ireland, 2016.
12. Bostrom, N.; Yudkowsky, E. The Ethics of Artificial Intelligence. In *The Cambridge Handbook of Artificial Intelligence*; Frankish, K., Ramsey, W.M., Eds.; Cambridge University Press: Cambridge, UK, 2014; Volume 316, pp. 316–334.
13. Anderson, M.; Anderson, S.L. *Machine Ethics*; Cambridge University Press: Cambridge, UK, 2011.
14. Johnson, D.G.; Verdicchio, M. AI Anxiety. *J. Assoc. Inf. Sci. Technol.* **2017**, *68*, 2267–2270. [CrossRef]
15. Rosen, G.; Byrne, A.; Cohen, J.; Shiffrin, S.V. *The Norton Introduction to Philosophy*; WW Norton & Company: New York, NY, USA, 2015.
16. Boddington, P. *Towards a Code of Ethics for Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2017.
17. Burton, E.; Goldsmith, J.; Koenig, S.; Kuipers, B.; Mattei, N.; Walsh, T. Ethical Considerations in Artificial Intelligence Courses. *AI Mag.* **2017**, *38*, 22–34. [CrossRef]
18. Etzioni, A.; Etzioni, O. Incorporating Ethics into Artificial Intelligence. *J. Ethics* **2017**, *21*, 403–418. [CrossRef]
19. Siau, K.; Wang, W. Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *J. Database Manag.* **2020**, *31*, 74–87. [CrossRef]
20. Gunning, D. *Explainable Artificial Intelligence (Xai)*; DARPA: Arlington, TX, USA, 2017.
21. Kalenka, S.; Jennings, N.R. Socially responsible decision making by autonomous agents. In *Cognition, Agency and Rationality*; Springer: Berlin/Heidelberg, Germany, 1999; pp. 135–149.
22. Wilson, H.J.; Daugherty, P.; Bianzino, N. The Jobs That Artificial Intelligence Will Create. *MIT Sloan Manag. Rev.* **2017**, *58*, 14.
23. Payne, D.; Raiborn, C.; Askvik, J. A Global Code of Business Ethics. *J. Bus. Ethics* **1997**, *16*, 1727–1735. [CrossRef]
24. Rose, J.M. Corporate Directors and Social Responsibility: Ethics versus Shareholder Value. *J. Bus. Ethics* **2007**, *73*, 319–331. [CrossRef]
25. Makridakis, S. The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms. *Futures* **2017**, *90*, 46–60. [CrossRef]
26. Rességuier, A.; Rodrigues, R. AI Ethics Should Not Remain Toothless! A Call to Bring Back the Teeth of Ethics. *Big Data Soc.* **2020**. [CrossRef]
27. Yampolskiy, R.V. Taxonomy of Pathways to Dangerous Artificial Intelligence. In Proceedings of the Workshops at the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–13 February 2016.
28. Horton, H. *Microsoft Deletes “Teen Girl” AI after It Became a Hitler-Loving Sex Robot within 24 h*; The Daily Telegraph: London, UK, 2016.
29. Nahavandi, S. Industry 5.0—A Human-Centric Solution. *Sustainability* **2019**, *11*, 4371. [CrossRef]
30. Xia, X.; Cao, B.; Yu, J. The Trust Measurement Algorithm of Agent Internetwork for Architecture-Centric Evolution Model. In Proceedings of the 2009 International Conference on Computational Intelligence and Software Engineering, Wuhan, China, 11–13 December 2009; pp. 1–4.
31. Thiroux, J.P.; Krasemann, K.W. *Ethics: Theory and Practice*, 11th ed.; Pearson Education: London, UK, 2012.
32. Mjolsness, E.; DeCoste, D. Machine Learning for Science: State of the Art and Future Prospects. *Science* **2001**, *293*, 2051–2055. [CrossRef]
33. Bologa, A.-R.; Bologa, R. Business Intelligence Using Software Agents. *Database Syst. J.* **2011**, *2*, 31–42.
34. Diederich, S.; Janßen-Müller, M.; Brendel, A.B.; Morana, S. Emulating Empathetic Behavior in Online Service Encounters with Sentiment-Adaptive Responses: Insights from an Experiment with a Conversational Agent. In Proceedings of the International Conference on Information Systems (ICIS), Munich, Germany, 15–18 December 2019.
35. Strohmeier, S.; Piazza, F. (Eds.) *Human Resource Intelligence und Analytics: Grundlagen, Anbieter, Erfahrungen und Trends*; Springer Gabler: Wiesbaden, Germany, 2015; ISBN 978-3-658-03595-2.
36. Maxmen, A. Self-Driving Car Dilemmas Reveal That Moral Choices Are Not Universal. *Nature* **2018**, *562*, 469. [CrossRef]
37. Pfeiffer, W.; Dögl, R. Das Technologie-Portfolio-Konzept zur Beherrschung der Schnittstelle Technik und Unternehmensstrategie. In *Strategische Unternehmensplanung/Strategische Unternehmensführung*; Springer: Berlin/Heidelberg, Germany, 1990; pp. 254–282.
38. Yu, O. *Technology Portfolio Planning and Management: Practical Concepts and Tools*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007; Volume 96.

39. Grünig, R.; Kühn, R.; Kühn, R. *The Strategy Planning Process: Analyses, Options, Projects*; Springer: Berlin/Heidelberg, Germany, 2015; ISBN 978-3-662-51603-4.
40. Amato, F.; Marrone, S.; Moscato, V.; Piantadosi, G.; Picariello, A.; Sansone, C. *Chatbots Meet EHealth: Automating Healthcare*; University of Naples Federico II: Naples, Italy, 2017; p. 10.
41. Akers, J.F. Ethics and Competitiveness—Putting First Things First. *MIT Sloan Manag. Rev.* **1989**, *30*, 69.
42. Hagedorff, T. The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds Mach.* **2020**, *30*, 99–120. [CrossRef]
43. Bryson, J.M. *Strategic Planning for Public and Nonprofit Organizations: A Guide to Strengthening and Sustaining Organizational Achievement*; John Wiley & Sons: Hoboken, NJ, USA, 2018.
44. Rothaermel, F.T. *Strategic Management*; McGraw-Hill Education: New York, NY, USA, 2017.
45. Quinn, J.B. *Strategies for Change: Logical Incrementalism*; Irwin Professional Publishing: Burr Ridge, IL, USA, 1980.
46. Berry, T. *Hurdle: The Book on Business Planning: How to Develop and Implement a Successful Business Plan*; Palo Alto Software, Inc.: Eugene, OR, USA, 2003.
47. Baker, A.; Perreault, D.; Reid, A.; Blanchard, C.M. Feedback and Organizations: Feedback Is Good, Feedback-Friendly Culture Is Better. *Can. Psychol. Can.* **2013**, *54*, 260. [CrossRef]
48. Kohlberg, L. Education, Moral Development and Faith. *J. Moral Educ.* **1974**, *4*, 5–16. [CrossRef]
49. Coeckelbergh, M. Pervasion of What? Techno–Human Ecologies and Their Ubiquitous Spirits. *AI Soc.* **2013**, *28*, 55–63. [CrossRef]
50. Chakrabarty, S.; Erin Bass, A. Comparing Virtue, Consequentialist, and Deontological Ethics-Based Corporate Social Responsibility: Mitigating Microfinance Risk in Institutional Voids. *J. Bus. Ethics* **2015**, *126*, 487–512. [CrossRef]
51. Van Rysewyk, S.P.; Pontier, M. *Machine Medical Ethics*; Springer: New York, NY, USA, 2014; ISBN 978-3-319-08107-6.
52. Moberg, D.J. The Big Five and Organizational Virtue. *Bus. Ethics Q.* **1999**, *9*, 245–272. [CrossRef]
53. Spinellis, D. Victims of Crime and the Criminal Process. *Isr. Law Rev.* **1997**, *31*, 337–378. [CrossRef]
54. Young, L.; Cushman, F.; Hauser, M.; Saxe, R. The Neural Basis of the Interaction between Theory of Mind and Moral Judgment. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 8235–8240. [CrossRef]
55. Schleim, S.; Spranger, T.M.; Erk, S.; Walter, H. From Moral to Legal Judgment: The Influence of Normative Context in Lawyers and Other Academics. *Soc. Cogn. Affect. Neurosci.* **2011**, *6*, 48–57. [CrossRef]
56. Nicholson, N. Ethics in Organizations: A Framework for Theory and Research. *J. Bus. Ethics* **1994**, *13*, 581–596. [CrossRef]
57. Kaplan, R.S.; Norton, D.P. *The Execution Premium: Linking Strategy to Operations for Competitive Advantage*; Harvard Business Press: Cambridge, MA, USA, 2008.
58. Aguilar, F.J. *Scanning the Business Environment*; Macmillan: New York, NY, USA, 1967.
59. Schiff, D.; Biddle, J.; Borenstein, J.; Laas, K. What's Next for AI Ethics, Policy, and Governance? A Global Overview. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society; ACM: New York, NY, USA, 2020; pp. 153–158.
60. Bessi, A.; Ferrara, E. Social Bots Distort The 2016 U.S. Presidential Election Online Discussion. *First Monday* **2016**, *21*, 1–15. [CrossRef]
61. Murthy, D.; Powell, A.B.; Tinati, R.; Anstead, N.; Carr, L.; Halford, S.J.; Weal, M. Bots and Political Influence: A Sociotechnical Investigation of Social Network Capital. *Int. J. Commun.* **2016**, *10*, 20.
62. Tasioulas, J. First Steps Towards an Ethics of Robots and Artificial Intelligence. *J. Pract. Ethics* **2019**. [CrossRef]
63. Bamberger, K.A. Technologies of Compliance: Risk and Regulation in a Digital Age. *Tex. Law Rev.* **2009**, *88*, 669.
64. Hurlburt, G.F.; Miller, K.W.; Voas, J.M. An Ethical Analysis of Automation, Risk, and the Financial Crises of 2008. *IT Prof.* **2009**, *11*, 14–19. [CrossRef]
65. Helbing, D.; Frey, B.S.; Gigerenzer, G.; Hafen, E.; Hagner, M.; Hofstetter, Y.; Van Den Hoven, J.; Zicari, R.V.; Zwitter, A. Will democracy survive big data and artificial intelligence? In *Towards Digital Enlightenment*; Helbing, D., Ed.; Springer: Berlin/Heidelberg, Germany, 2019; pp. 73–98.
66. Nath, R.; Sahu, V. The Problem of Machine Ethics in Artificial Intelligence. *AI Soc.* **2017**, *35*, 103–111. [CrossRef]
67. Lara, F.; Deckers, J. Artificial Intelligence as a Socratic Assistant for Moral Enhancement. *Neuroethics* **2019**, *12*, 275–287. [CrossRef]
68. Jones, T.M.; Felps, W. Shareholder Wealth Maximization and Social Welfare: A Utilitarian Critique. *Bus. Ethics Q.* **2013**, *23*, 207–238. [CrossRef]
69. Atanasoff, L.; Venable, M.A. Technostress: Implications for Adults in the Workforce. *Career Dev. Q.* **2017**, *65*, 326–338. [CrossRef]
70. Tonkens, R. A Challenge for Machine Ethics. *Minds Mach.* **2009**, *19*, 421–438. [CrossRef]
71. Danaher, J. Toward an Ethics of AI Assistants: An Initial Framework. *Philos. Technol.* **2018**, *31*, 629–653. [CrossRef]
72. Coombs, N. What Is an Algorithm? Financial Regulation in the Era of High-Frequency Trading. *Econ. Soc.* **2016**, *45*, 278–302. [CrossRef]
73. Baum, S.D. Social Choice Ethics in Artificial Intelligence. *AI Soc.* **2017**, *32*, 1–12. [CrossRef]
74. Beckers, S. AAAI: An Argument Against Artificial Intelligence. In *Philosophy and Theory of Artificial Intelligence 2017*; Müller, V.C., Ed.; Springer International Publishing: Cham, Switzerland, 2018; Volume 44, pp. 235–247. ISBN 978-3-319-96447-8.
75. Watson, R.T.; Boudreau, M.-C.; Chen, A.J. Information Systems and Environmentally Sustainable Development: Energy Informatics and New Directions for the IS Community. *MIS Q.* **2010**, *34*, 34. [CrossRef]
76. Sparrow, R. Killer Robots. *J. Appl. Philos.* **2007**, *24*, 62–77. [CrossRef]
77. Smith, N.C.; Simpson, S.S.; Huang, C.-Y. Why Managers Fail to Do the Right Thing: An Empirical Study of Unethical and Illegal Conduct. *Bus. Ethics Q.* **2007**, *17*, 633–667. [CrossRef]

78. Jones, T.M. Ethical Decision Making by Individuals in Organizations: An Issue-Contingent Model. *Acad. Manag. Rev.* **1991**, *16*, 366–395. [CrossRef]
79. Edelman, L.B.; Suchman, M.C. The Legal Environments of Organizations. *Annu. Rev. Sociol.* **1997**, *23*, 479–515. [CrossRef]
80. Houser, K.A.; Voss, W.G. GDPR: The End of Google and Facebook or a New Paradigm in Data Privacy. *Richmond J. Law Technol.* **2018**. [CrossRef]
81. Butterworth, M. The ICO and Artificial Intelligence: The Role of Fairness in the GDPR Framework. *Comput. Law Secur. Rev.* **2018**, *34*, 257–268. [CrossRef]
82. Ashrafian, H. AI on AI: A Humanitarian Law of Artificial Intelligence and Robotics. *Sci. Eng. Ethics* **2015**, *21*, 29–40. [CrossRef]
83. Lewis, P.V. Defining ‘Business Ethics’: Like Nailing Jello to a Wall. *J. Bus. Ethics* **1985**, *4*, 377–383. [CrossRef]
84. Hofstede, G. Culture and Organizations. *Int. Stud. Manag. Organ.* **1980**, *10*, 15–41. [CrossRef]
85. Okleshen, M.; Hoyt, R. A Cross Cultural Comparison of Ethical Perspectives and Decision Approaches of Business Students: United States of America versus New Zealand. *J. Bus. Ethics* **1996**, *15*, 537–549. [CrossRef]
86. United Nations. *The Universal Declaration of Human Rights*; United Nations: Washington, DC, USA, 1948.
87. Rai, A. Editor’s Comments: The MIS Quarterly Trifecta: Impact, Range, Speed. *Manag. Inf. Syst. Q.* **2016**, *40*, iii–x.
88. Huff, A.S. *Writing for Scholarly Publication*; SAGE Publications: Thousand Oaks, CA, USA, 1999.
89. Rai, A. Editor’s Comments: Beyond Outdated Labels: The Blending of IS Research Traditions. *MIS Q.* **2018**, *42*, 2.
90. Adams, J.S.; Tashchian, A.; Shore, T.H. Codes of Ethics as Signals for Ethical Behavior. *J. Bus. Ethics* **2001**, *29*, 199–211. [CrossRef]
91. Ambrose, M.L.; Arnaud, A.; Schminke, M. Individual Moral Development and Ethical Climate: The Influence of Person–Organization Fit on Job Attitudes. *J. Bus. Ethics* **2008**, *77*, 323–333. [CrossRef]
92. Weaver, G.R. Ethics Programs in Global Businesses: Culture’s Role in Managing Ethics. *J. Bus. Ethics* **2001**, *30*, 3–15. [CrossRef]
93. Paarlberg, L.E.; Perry, J.L. Values Management: Aligning Employee Values and Organization Goals. *Am. Rev. Public Adm.* **2007**, *37*, 387–408. [CrossRef]
94. Somers, M.J. Ethical Codes of Conduct and Organizational Context: A Study of the Relationship between Codes of Conduct, Employee Behavior and Organizational Values. *J. Bus. Ethics* **2001**, *30*, 185–195. [CrossRef]
95. Frazer, E. Max Weber on Ethics and Politics. *Edinb. Univ. Press* **2006**, *19*.
96. Sorensen, A. *Deontology—Born and Kept In Servitude By Utilitarianism*, 1st ed.; Danish Yearbook of Philosophy: Leiden, The Netherlands, 2009.
97. Vasiliou, I. Platonic Virtue: An Alternative Approach. *Philos. Compass* **2014**, *9*, 605–614. [CrossRef]
98. Douglas, A.; John, W. *Computer and Information Ethics*; Praeger: Westport, CT, USA, 2008.
99. Cilliers, C.; Kriel, J. *Fundamental Penology: Only Study Guide for PEN1014*; UNISA: Pretoria, South Africa, 2008.
100. Hugo, A.B.; Erin, K. *Punishment*. *Stanford Encyclopedia of Philosophy*; Stanford University: Stanford, CA, USA, 2015.

Article

Vital, Sophia, and Co.—The Quest for the Legal Personhood of Robots

Ugo Pagallo

Law School, University of Turin, Lungo Dora Siena 100, 10153 Turin, Italy; ugo.pagallo@unito.it;
Tel.: +39-011-6706903

Received: 26 July 2018; Accepted: 7 September 2018; Published: 10 September 2018

Abstract: The paper examines today’s debate on the legal status of AI robots, and how often scholars and policy makers confuse the legal agenthood of these artificial agents with the status of legal personhood. By taking into account current trends in the field, the paper suggests a twofold stance. First, policy makers shall seriously mull over the possibility of establishing novel forms of accountability and liability for the activities of AI robots in contracts and business law, e.g., new forms of legal agenthood in cases of complex distributed responsibility. Second, any hypothesis of granting AI robots full legal personhood has to be discarded in the foreseeable future. However, how should we deal with Sophia, which became the first AI application to receive citizenship of any country, namely, Saudi Arabia, in October 2017? Admittedly, granting someone, or something, legal personhood is—as always has been—a highly sensitive political issue that does not simply hinge on rational choices and empirical evidence. Discretion, arbitrariness, and even bizarre decisions play a role in this context. However, the normative reasons why legal systems grant human and artificial entities, such as corporations, their status, help us taking sides in today’s quest for the legal personhood of AI robots. Is citizen Sophia really conscious, or capable of suffering the slings and arrows of outrageous scholars?

Keywords: artificial intelligence; legal agent; liability; personhood; responsibility; robotics

1. Introduction

The legal personhood of robots has been a popular topic of today’s debate on the normative challenges brought about by this technology. In 2007, for example, Carson Reynolds and Masatoshi Ishikawa explored the scenarios of Robot Thugs, namely, machines that choose to commit and, ultimately, carry out a crime: their aim was to determine whether and to what extent these machines can be held accountable [1]. Three years later, I expanded this analysis on agency and criminal responsibility, to the fields of contracts and extra-contractual liability [2]. In homage to Reynolds and Ishikawa’s creature *Picciotto Roboto*, my next paper then provided a concise phenomenology on how smart AI systems may affect pillars of the law, such as matters of criminal accountability, negligence, or human intent [3]. In 2013, I summed this analysis up with my monograph on *The Laws of Robots* [4]. There, I suggested a threefold level of abstraction, so as to properly address today’s debate on the legal personhood of robots and smart AI systems, that is:

- (i) The legal personhood of robots as proper legal “persons” with their constitutional rights (for example, it is noteworthy that the European Union existed for almost two decades without enjoying its own legal personhood);
- (ii) The legal accountability of robots in contracts and business law (for example, slaves were neither legal persons nor proper humans under ancient Roman law and still, accountable to a certain degree in business law);

- (iii) New types of human responsibility for others' behaviour, e.g., extra-contractual responsibility or tortious liability for AI activities (for example, cases of liability for defective products. Although national legislation may include data and information in the notion of product, it remains far from clear whether the adaptive and dynamic nature of AI through either machine learning techniques, or updates, or revisions, may entail or create a defect in the "product").

Against this framework, the aim of the paper is to shed further light on such threefold status that AI robots may have in the legal domain, by taking into account what has happened in this domain of science, technology, and their normative challenges over the past years. Whereas most legal systems, so far, have regulated the behaviour of AI robots as simple tools of human interaction and hence, as a source of responsibility for other agents in the system [4], have advancements of technology affected this traditional framework? Do certain specimens of AI technology, such as smart humanoid robots, recommend that we should be ready to grant some of these robots full legal personhood and citizenship? Or, would such legislative action be morally unnecessary and legally troublesome, in that holding AI robots accountable outweighs the "highly precarious moral interests that AI legal personhood might protect" [5]?

To offer a hopefully comprehensive view on these issues, the paper is presented in three parts. First, focus is on current trends of AI technology and robotics, so as to stress both benefits and threats of this field. Then, attention is drawn to the confusion that prevails in most of today's debate between the legal personhood of AI robots and their legal accountability in contracts and business law. Finally, the analysis dwells on the pros and cons of granting AI robots full legal personhood, as opposed to the status of legal accountability, or as a source of responsibility for other agents in the legal system. At the risk of being lambasted for reactionary anthropocentrism, the conclusion of the paper is that such a quest for the legal personhood of AI robots should not have priority over the regulation of more urgent issues triggered by the extraordinary developments in this field.

2. Current Trends of Robotics

To shed light on today's debate on the legal personhood of AI robots, it is important to briefly put this debate into perspective. Robots materialized as a reprogrammable machine operating in a semi- or fully autonomous way, so as to perform manufacturing operations, more than fifty years ago. Inspired by the research of George Devol and Joseph Engelberger, robots were deployed in the automobile sector since 1961, i.e., the UNIMATE robot removing die-casting and performing spot welding in a General Motors factory in New Jersey. In the early 1980s, such a use of robots within the automobile sector became critical. The Japanese industry attained a strategic competitiveness through the large-scale use of this technology in their factories, reducing their costs and increasing the overall quality of their cars. This trend went on until the early 2000s, when certain individuals still had the impression that robotics was too dependent on the automobile industry. Remarkably, in the *Editorial to the World 2005 Robotics Report* of the Economic Commission for Europe and the International Federation of Robotics, Åke Madesäter raised this risk: "In the period 1997–2003, the automotive industry in Spain received 70% of all new robot installations. In France, the United Kingdom and Germany the corresponding figure amounted to 68%, 64% and 57%, respectively" [6].

In the same years as covered by the UN World report, however, the field of robotics opened up to a profound transformation, a "revolution", according to many scholars [7,8]. The first step of this diversification concerned the set of water-surface and underwater unmanned vehicles, or "UUVs", employed for remote exploration work and the repairs of pipelines, oil rigs and so on. This set of robotic applications started developing an amazing pace since the mid-1990s. Then, it was the turn of unmanned aerial vehicles ("UAVs"), or systems ("UAS"), that upset the military field in the mid-2000s [9]. A few years later, time was ripe for the advent of self-driving cars: whereas the Nevada Governor signed a bill into law in June 2011 that for the first time ever authorized the use of driverless cars on public roads, other states in the U.S. soon followed suit. In September 2017, the House of Representatives passed a bill, the Self Drive Act, which aims to provide a much-needed federal

framework for the regulation of autonomous vehicles. While the panoply of robots available out there suggests further candidates for the next robotic revolution in the field of service applications for personal and domestic use, such as robots for home security and surveillance, for handicap assistance, or just for fun and entertainment, we should not miss a crucial, twofold aspect of this trend. On the one hand, robots are progressively connected to the Internet: avoiding the shortcomings of traditional approaches, such as on-board computers for robots, the troubles with the computing power of such machines have increasingly been addressed by connecting them to a networked repository on the Internet, allowing robots to share the information required for object recognition, navigation and task completion in the real world. On the other hand, the field of robotics is more frequently intertwined with advancements of artificial intelligence ('AI'), to such an extent that even the definition of robot has evolved over the past decades. Some argue that we are dealing with machines built basically upon the mainstream "sense-think-act" paradigm of AI research [10]. Sebastian Thrun, former director of the AI Laboratory at Stanford, California, similarly reckons that robots are machines with the ability to "perceive something complex and make appropriate decisions" [8] (p. 77). Although we still have not obtained either machines that are capable of doing any work men can do, or the solution for the problem of creating proper artificial intelligence within "25 years" [11], or "the current generation" [12], we are dealing with machine-learning systems that (i) increasingly define or modify their decision-making rules autonomously; (ii) improve their knowledge and skills through the interaction with other artificial agents, smart things, or human fellows in the surrounding environment; and, (iii) respond to the stimuli of such environment, by modifying their own properties, or inner states [4]. Among the ingredients that made the convergence between computer sciences, AI and robotics possible, we should list the improvement of more sophisticated statistical and probabilistic methods, the growing availability of huge amounts of data and of massive computational power, up to the transformation of places and spaces into IT-friendly environments, e.g., smart cities and domotics.

As to the specimens of such smart machines like Blue Frog Robotics' Buddy, SoftBank's Pepper, or Asus's Zenbo, it is worth mentioning two further applications in this context. The first one is Vital, a robot developed by Aging Analytics UK, who was appointed in May 2014 as a board member of the Hong-Kong venture capital firm Deep Knowledge. The reasons for this appointment hinged on the ability of Vital to foretell good investments in the field of therapies for age-related syndromes, pinpointing market trends that otherwise would be under the human radar. Whereas AI machines will sensibly improve their adaptive decision-making rules over the next years, it seems fair to admit that trends of humans delegating complex cognitive tasks to robots and AI systems will reasonably increase as well. For example, in October 2016, a Finnish OMX-listed company, Tieto, appointed Alicia T, an AI expert system, as a member of the leadership team of its new data-driven businesses unit. According to Tieto's website, Alicia T. will not only become a full-fledged member of the management team, but also possess the capacity to cast votes: "AI will help the management team to become truly data-driven and will assist the team in seeking innovative ways to pursue the significant opportunities of the data-driven world".

Then, we have the case of Sophia, a social humanoid robot developed by Hong Kong-based company Hanson Robotics, in collaboration with Google's parent company Alphabet and SingularityNET, which provide for Sophia's voice recognition system and AI software, respectively. Activated in April 2015, Sophia made her first public appearance in Austin, Texas, in March 2016. As the Wikipedia entry claimed in early 2018, "interviewers around the world have been impressed by the sophistication of many of Sophia's responses to their questions, (although) the bulk of Sophia's meaningful statements are believed by experts to be somewhat scripted". Wikipedia entry has meanwhile been updated, insisting now on the "controversy over hype in the scientific community". Nonetheless in October 2017, as a matter of fact, Sophia became the first AI application to receive citizenship of any country, namely, Saudi Arabia; a month later, 'she' was named the first Innovation Champion of the United Nations Development Programme, the first non-human to be given any UN title.

Obviously, one may wonder why on earth Saudi Arabia has enrolled Sophia as a citizen of her own, much as the UN celebrating Sophia as an innovation champion. Yet, some months earlier, in February 2017, the European Parliament adopted a proposal, in which the EU institution invites the European Commission “to explore, analyze and consider the implications of all possible legal solutions, (including) ... creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently” (§59f of the document).

Admittedly, current trends of AI and robotics have suggested some pessimistic views. In 2015, for instance, the Future of Life Institute released an open letter addressing the challenges and threats posed by this technology: “Its members—and advocates, among which Bill Gates, Elon Musk, and Stephen Hawking—are concerned that as increasingly sophisticated achievements in AI accumulate—especially where they intersect with advances in autonomous robotics technology—not enough attention is being paid to safety”. A year later, the White House Office of Science and Technology Policy (OSTP) conducted a series of public workshops on questions of AI and policy, releasing a final report on how to tackle such issues, as fairness, accountability, or social justice, through means of transparency [13]. While the European Parliament’s Committee on Legal Affairs and the UK House of Commons presented similar reports in 2017, an Industry Connections Program within the IEEE Standards Association issued its own document, namely, *The Global Initiative on Ethical of Autonomous and Intelligent Systems* from December 2017, in which the normative challenges and ethical threats of this kind of technology are similarly taken into account. In light of the manifold AI robotics applications and of the multiple, and even opposite, normative views of legislators, experts, and opinion makers, on whether or not legal systems should grant AI robots their “electronic personhood”, is there any chance to orient ourselves?

The next step of the analysis has to set the proper level of abstraction, in order to take sides in today’s debate. From a methodological viewpoint, the aim is to determine the interface that makes an investigation of some crucial aspects of the legal system possible, so as to comprise a set of features representing the observables and variables of the analysis, the result of which provides a model for the field [4]. From a substantial perspective, we should distinguish the analysis of the technology that is subject to legal regulation, and the set of legal notions that are at stake with matters of accountability, liability, and responsibility. Next sections explore this twofold aspect of the problem separately.

3. Levels of Abstraction

A main source of misunderstandings in today’s debate on the legal personhood of AI robots has to do with the ways in which the different kinds of issues, interests, or goods that are at stake with their behaviour, are confused in a Hegelian night where all kinds of legal status look grey. Going back to the aforementioned European Parliament’s proposal from February 2017, for example, it is unclear whether “the status of electronic persons” refers to the full legal personhood of robots as proper legal “persons”, or regards their legal accountability in contracts and business law, or both. This confusion reappears with certain scholars. Some claim, “that for a computer agent to qualify as a legal agent it would need legal personhood. Both meanings of ‘agency’ raise questions as to the desirability of legal personhood of bots” and other artificial agents such as robots [14]. Others argue that granting robots legal personhood would prevent “the debates over slavery” that “remind us of uncomfortable parallels with the past” and “reflect ongoing tension over humanity’s role in an increasingly technologized world” [15] (p. 186). More recently, this confusion between legal personhood and accountability of AI robots reappears in [5]. Here, the reason legal systems should not confer legal personhood on “purely synthetic entities” has to do with moral reasons and the abuse of legal person status by robots and those that make them, i.e., either robots as liability shields, or robots as themselves unaccountable rights violators.

A proper level of analysis has thus to be set, in order to stop comparing apples and oranges, namely, the apples of legal accountability and the oranges of legal personhood. (In this context, a third hypothesis on AI robots as a source of responsibility for other agents in the system is set aside, e.g., current work for the amendment of the EU directive no. 374 from 1985 on liability for defective products.) Correspondingly, attention should be drawn to three different kinds of reasoning, which can be summed up as follows: (i) if apples, i.e., robots as accountable agents, then we have oranges; (ii) if apples, then we should have oranges; and, (iii) if we do not have apples, then neither oranges as a result.

As to the first kind of argument, according to which the legal agency of AI robots would require their legal personhood [5,14,15], it is not necessary to resort to the example of the legal status of slaves under ancient Roman law to show that forms of dependent or restricted legal status, such as agents in contract law, are not essentially intertwined with forms of independent legal personhood. For example, the European Union existed for almost two decades without enjoying its own legal personhood. Therefore, scholars may discuss about different types of apple, namely, registering AI robots like corporations, or bestowing them with capital, or making the financial position of such smart machines transparent, without resorting to any kind of AI personhood. From the different types of apples under scrutiny in today's research, in other words, it does not follow that AI robots necessarily turn out to be an orange, thereby enjoying some form of full legal personhood.

The second kind of argument is normative, for it claims that, once AI robots are conceived of as agents in contracts and business law, then they should be treated as legal persons. The normative ground of this reasoning rests on the reasons why legal systems grant human persons full legal personhood. As we will see later in the next section, such reasons have to do with the moral status of humans, their intrinsic worth and capability to suffer, their consciousness, and so forth. In the tradition of human rights declarations, the reference value is given by the idea of dignity, e.g., Article 1 of the Universal Declaration of Human Rights (UDHR), and Protocol 13 of the European Convention on Human Rights (ECHR). The problem with this second kind of argument is that different types of AI robotic agenthood can be examined, regardless of whether or not such artificial agents are conscious, capable to suffer, or experience emotions, desires, pleasures, or pain. What is at stake with the legal agenthood of AI robots and their accountability has to do with matters of efficiency in transactions and economic affairs, rather than any kind of AI robotic dignity. Advocates of the argument, according to which the legal agency of AI robots should require their legal personhood e.g., [15], have thus to preliminarily demonstrate that such artificial agents possess some of the requisites that legal systems usually take into account, in order to grant humans their full legal personhood.

Yet, there is a variant of the second argument, which likens the status of robots to the legal personhood of corporations. In both cases, so goes the reasoning, once we admit that AI robots and corporations are agents in contracts and business law, then—for reasons that hinge on the efficiency of economic affairs and transactions—they should be considered as full legal persons. Admittedly, the idea of registering AI robots just like corporations is popular among scholars [16–18]. As claimed by others, moreover, “existing laws might provide a potentially unexpected regulatory framework for autonomous systems” [19]. According to this view, we should not amend the current law, to admit that AI robots may “inhabit” a company and “thereby gain some of the incidents of legal personality” [19].

However, going back to the variant of our second argument, there are three problems. First, the corporate solution for the legal agenthood of AI robots is only one among several technical options that scholars have suggested over the past years, in order to tackle problems of accountability for AI robotic behaviour. Scholars have proposed registries for artificial agents, insurance policies, or modern forms of the ancient Roman legal mechanism of *peculium*, namely, the sum of money or property granted by the head of the household to a slave or son-in-power [4]. What all these cases illustrate is that legal systems can properly address the challenges of the agenthood of AI robots in contracts and business law, without embracing any form of corporation and hence, any kind of legal personhood of AI robots. In light of the examples in the previous section, we can thus say that the status of Vital or

Alicia T. may even make sense, without endorsing any kind of citizenship status for Sophia. Second, the extent of the legal personhood of corporations dramatically varies among legal systems. Contrary to the US tradition, for example, most EU companies do not enjoy their own privacy rights, or their own political rights, such as freedom of speech [20]; corporations cannot be held criminally responsible [21], and so forth. This latter scenario is at odds with that which advocates of the legal personhood of AI robots usually claim: at least in Europe, the corporate solution for the legal personhood of AI robots would be a Pyrrhic victory. Third, we have to take into account the opinion of those who oppose granting robots the status of legal persons just like corporations. According to this view, “there are two kinds of abuse that might arise at the expense of human legal rights—humans using robots to insulate themselves from liability and robots themselves unaccountably violating human legal rights” [5] (p. 285).

This latter argument on the “two abuses” can be understood both as a critique of the just like corporation-view and as an illustration of the third kind of confusion in today’s debate between the legal agenthood and the legal personhood of AI robots. As to the first aspect of this stance, its advocates claim, the personhood of artificial agents could be a means to shield humans from the consequences of their conduct. In light of the *International Tin Council case* before the House of Lords in October 1989, “the risk (is) that electronic personality would shield humans actors from accountability for violating rights of other legal persons, particularly human or corporate” [5] (p. 287). Although this possibility is for real, we should also pay attention to the other way around, namely, cases in which the intricacy of the interaction between humans and computers can make it extremely difficult to ascertain what is, or should be, the information content of the natural or artificial entity, as foundational to determining the responsibility of individuals. Such cases of distributed responsibility that hinge on multiple accumulated actions of humans and computers may lead to cases of impunity that already have recommended some legal systems to adopt new forms of criminal accountability. Think of the collective knowledge doctrine, the culpable corporate culture, or the reactive corporate fault, as ways to determine the blameworthiness of corporations and their autonomous criminal liability [22].

Still, in addition to the risk of AI robots as liability shields, advocates of the “two abuses”-doctrine raise the further threat of robots as themselves unaccountable right violators. In a nutshell, the problem revolves around who could represent the artificial agent in a legal dispute and moreover, how we should deal with issues of robot insolvency. Although as we have seen above in this section, legal systems could establish mechanisms for AI robots to own property or hold accounts, much as requiring the creators of robots to place initial funds in such accounts, “money can flow out of accounts just as easily as it can flow in; once the account is depleted, the robot would effectively be unanswerable for violating human legal rights” [5] (p. 288). Traditional punitive sanctions of the law, such as jail time for criminal insolvency, would be unavailable, unsatisfying, or ineffective. As a result, we may envisage the malfunctioning of AI robots or their manipulation that cause or fuel human wrongdoing, if not properly detected and recovered, thus making people vulnerable to systematic recourse to such artificial systems. In addition, we should expect a novel wave of AI crimes and wrongdoing, after the 1990s generation of computer crimes set up by national legislators, such as new forms of AI Ponzi schemes [22].

There are, however, two further problems with this narrative on AI robots representing a threat as unaccountable right violators. The first issue is empirical, and has to do with scholarly work and legislative measures as to how to hold such artificial agents accountable. Whereas, as previously stressed, scholars have suggested different kinds of strategies, as registries, insurance policies, modern forms of *peculium*, and the like, some institutions, as the Japanese government, have worked out a way to address these issues through the creation of special zones for robotics empirical testing and development, namely, a form of living lab, or *Tokku*. Significantly, the special zone of Tsukuba was set up in 2011, in order to understand how AI safety governance and tax regulation could be disciplined [23]. Thus, we can dismiss this part of the “two abuses”-doctrine as an empirical issue

concerning how legal systems could properly keep the legal agenthood of AI robots under control, and make them accountable.

However, how about the further claim of the “two abuses”-doctrine? Should we buy the idea that once AI robots are a mere liability shield, or rather potential unaccountable rights violators in contracts and business law, no legal personhood shall follow as a result?

The fallacy of the argument concerns once again the confusion between apples and oranges, that is, between the legal agenthood of AI robots in contracts and business law, and their legal personhood. There are several instances of how legal systems might grant rights of personhood, independently of matters that regard accountability in economic affairs. As to the rights of human persons, think about minors and people with severe psychological illnesses, who cannot be deprived of their legal personhood as espoused in certain rights despite such illnesses, or emotional and intellectual immaturity, as occurs with e.g., the 1989 UN Convention on the Rights of the Child. As to the set of legal persons, consider the legal personhood that is enjoyed by such non-human entities, as the Whanganui river and Te Urewera national park in New Zealand, the Ganges and the Yamuna rivers in India, up to the entire ecosystem in Ecuador. Therefore, the question is not about whether the legal agency of AI robots would—or should—require their legal personhood, or whether the legal personhood of AI robots should vice versa be subject to their accountability in contracts and business law. Rather, the problem has to do with the reasons why legal systems usually grant humans their legal personhood, and whether AI robots meet such requirements.

4. AI as Legal Persons

The previous section has stressed some of the reasons why legal systems usually grant humans their legal personhood. According to the philosophical stance, or ideological options scholars adopt, such motives are often referred to either the moral status of humans and the protection of their dignity, or their capability to suffer, along with further elements, such as human consciousness, intentionality, desires, and interests. Some of these requisites, e.g., capability to suffer and (a certain degree of) consciousness, have been evoked to extend the sphere of legal personhood to other natural agents, such as animals [24]. Others have debated whether such extension could comprise some artificial agents. In his seminal 1992 article on the *Legal Personhood for Artificial Intelligences*, for example, Lawrence Solum examines three possible objections to the idea of recognizing rights to those artificial agents, or intelligences (AIs), namely, the thesis that “AIs Are Not Human” [25] (pp. 1258–1262); “The Missing-Something Argument” [25] (pp. 1262–1276); and, “AIs Ought to Be Property” [25] (pp. 1276–1279). Remarkably, according to Solum, there are no legal reasons or conceptual motives for denying the personhood of AI robots: the law should be entitled to grant personality on the grounds of rational choices and empirical evidence, rather than superstition and privileges. “I just do not know how to give an answer that relies only on a priori or conceptual arguments” [25] (p. 1264).

We may accept Solum’s argument that scholars cannot, on conceptual grounds, rule out in advance the possibility that AI robots should be given the status of legal personhood; still, we have to face three different kinds of problem with this new legal status. First, attention should be drawn to “the missing-something problem”. Although certain scholars claim that AI robots would already have the capability of fulfilling the awareness requirements in criminal law, together with “the mental element requirements of both intent offenses and recklessness offenses” [26] (p. 99), current AI robots lack most requisites that usually are associated with granting someone, or something, legal personhood: such artificial agents are not self-conscious, they do not possess human-like intentions, or properly suffer. This does not amount to say that the levels of autonomy, self-consciousness, and intentionality—which arguably are insufficient to grant AI robots their full legal personhood today—are inadequate to produce relevant effects in other fields of the law, e.g., the legal agenthood of artificial agents in the field of contracts and business law, as previously explored above in Section 3. Otherwise, we would incur in the same kind of confusion that has been stressed apropos of, say,

the “two abuses”-doctrine, by simply reversing the terms of such argumentation, that is, if AI robots do not meet the requisites of legal personhood, then they cannot be legal agents either.

The second kind of problem concerns the consequences of granting AI robots legal personhood. Once we admit there being artificial agents capable of autonomous decisions similar in all relevant aspects to the ones humans make, the next step would be to acknowledge that the legal meaning of “person” and, for that matter, of crimes of intent, of constitutional rights, of dignity, etc., will radically change. Even Solum admits that, “given this change in form of life, our concept of a person may change in a way that creates a cleavage between human and person” [25] (p. 1268). Likewise, others warn that “the empirical finding that novel types of entities develop some kind of self-consciousness and become capable of intentional actions seems reasonable, as long as we keep in mind that the emergence of such entities will probably require us to rethink notions of consciousness, self-consciousness and moral agency” [27] (pp. 558–559). At the end of the day, nobody knows to where this scenario may lead. For instance, would a strong AI robotic lawyer accept the argument that “evil is not part of the components of criminal liability” [26] (p. 93)? What if the AI robot, rather than an advocate of current exclusive legal positivism, is a follower of the natural law tradition?

The third kind of problem has to do with how we should mediate today’s state-of-the-art and Leibniz’s warning about our own ignorance: “every mind has a horizon in respect to its present intellectual capacity but not in respect to its future intellectual capacity” [28] (p. 115). On the one hand, some popular claims of today’s debate can be deemed as simply non-sense, such as the awareness of AI robots, thus subject to retribution and deterrence, rehabilitation and incapacitation, down to capital penalty e.g., [26]. Yet, on the other hand, the breath-taking advancements of technology in this field recommend being prepared as to how we shall rethink notions of consciousness, self-consciousness and moral agency. As previously stressed above in Section 3, some legislators and policy makers have adopted forms of legal experimentation, e.g., the Japanese government’s special zones set up over the past 15 years, as a way to address the normative challenges of AI robotics in a pragmatic way, that is, through empirical testing and development [29]. Admittedly, some sort of Western *Tokku* can increase our comprehension of risks and threats triggered by robots and smart AI systems, in order to prevent their undesirable actions and keep them in check. In addition, we can grasp how such systems may react in different contexts and whether robots and other AI agents ultimately meet human needs. From a legal viewpoint, the set up of special zones for robotics empirical testing and developing appears particularly relevant, because we can properly address on this basis the set of potential issues brought about by the advancement of AI and robotics, e.g., a fleet of self-driving cars for “public car sharing” [30]. In the traditional world of human drivers, many legal systems had to introduce—in addition to compulsory insurance policies—public funds for the victims of road accidents, e.g., the Italian legislative decree no. 209 from 2005. In the foreseeable world of autonomous vehicles, hypotheses of accountable AI car systems may make sense because a sort of digital peculium, embedded in the design of the system, can represent the smart AI counterpart to current public funds for the victims of road accidents. Along these lines, it is worth mentioning that the former Italian data protection authority has suggested that robots may soon become “data processors” pursuant to Article 28 of the EU regulation on data protection [31].

Whether or not this kind of legal experimentation and pragmatic approach will end up granting AI robots full legal personhood, thereby transforming pillars of the law and such basic concepts, as the idea of legal person, is of course a question that goes beyond our current Leibniz’s “horizon”. In the meanwhile, we already mentioned the case of our first robotic citizen, Sophia, who is the first non-human ever to be awarded by the UN too. *Pace* the empirical remarks and conceptual exercises that have been summed up throughout this paper—taking into account more than 25 years of scholarly debate in the field—what would the reasons for these legal steps be? By taking UDHR’s Article 1 and ECHR’s Protocol 13 seriously, where would the dignity of Sophia lie? Is she really conscious, or capable to suffer the slings and arrows of outrageous scholars?

The time is ripe for the conclusions of this paper.

5. Conclusions

Scholars properly stress, time and again, that the notion of person is a fiction in the legal domain. While Ancient Roman lawyers recovered the term ‘person’ from the theater’s mask that actors used to wear on stage, none of their definitions of legal person resembles today’s meaning of personhood. Contrary to Roman ideas on the role of the parties in legal acts, or in a process, on the status of free people, or enslaved persons, and so on, we currently associate the notion of person with the legal subject that has rights (and duties) of its own. This is the definition that we find in chapter 16 of Thomas Hobbes’s *Leviathan*, which has an origin in the ideas discussed by scholars of Canon Law in the 1200s and 1300s. In his *Commentary on Digestum Novum* [32], for instance, Bartolus de Saxoferrato (1313–1357) affirms that an artificial person, such as a monastery, is not a real person but rather, a fiction that nonetheless “stands in the name of the truth” (*pro vero*) [32]. This equalization between natural persons and artificial persons, between humans and legal entities, such as a mission, or a corporation, triumphed with the tenets of legal positivism and formalism in the 1800s. The great Roman scholar, Friedrich August von Savigny, makes this point clear in his *System of Modern Roman Law* (1840–1849) ed. (1979). Here, Savigny admits that only humans have rights and duties of their own and still, the Law has the power to grant such right of personhood to anything, be it monasteries or corporations, governments or ships in maritime law, rivers in New Zealand or India, down to the entire ecosystem in Ecuador.

Legal fictions have real effects, though. Since Ancient Roman times, they concern for example the procedural mechanisms that allow individuals to enforce their own rights, e.g., the *actio in personam* which gives an individual the role of a party in a process or legal act. In addition, fictions may regard the family status of an individual, e.g., adoptions, or when such individual should be considered deceased, e.g., Article 4 of the Italian civil code on cases in which some legal effect depends on whether someone outlived someone else and yet, it is impossible to determine such circumstance, so that both persons are considered by the law as deceased at the same time.

This crucial real-life impact of legal fictions has thus recommended scholars to carefully mull over whether and to what extent AI robots may create loopholes in the legal system and hence, whether current provisions of the law should be modified, or amended. By distinguishing between the legal agenthood of AI robots in contracts and business law, and the legal personhood of AI robots with their constitutional rights, the paper has insisted on the reasons why we should not confuse such legal statuses, so that two different outcomes follow as a result.

On the one hand, as to the legal agenthood of AI robots, it makes sense to consider new forms of accountability and of liability in the field of contracts and business law, such as registries, or modern forms of *peculium*. The aim is to prevent both risks of robotic liability shield and of AI robots as unaccountable rights violators, while tackling cases of distributed responsibility that hinge on multiple accumulated actions of humans and computers that may lead to cases of impunity. On the other hand, as to the legal personhood of AI robots, current state-of-the-art has suggested that none of today’s AI robots meet the requisites that usually are associated with granting someone, or something, such legal status. Although we should be prepared for these scenarios through manifold methods of legal experimentation, e.g., setting up special zones, or living labs, for AI robotic empirical testing and development, it seems fair to concede that we currently have other types of priority, e.g., the regulation of the use of AI robots on the battlefield [4].

Therefore, going back to the February 2017 proposal of the European Parliament, which was mentioned above in Section 2, the final recommendation of this paper would be threefold: (i) in the mid term, skip any hypothesis of granting AI robots full legal personhood; (ii) take seriously into account the possibility of new forms of accountability and liability for the activities of AI robots in contracts and business law, e.g., new forms of legal agenthood in cases of complex distributed responsibility; and, (iii) test such new forms of accountability and liability through methods of legal experimentation.

However, going back to the current debate on the legal personhood of AI robots, we should recognize that granting someone, or something, legal personhood is—and always has been—a highly

sensitive political issue. In addition to rivers in New Zealand and India, or the entire ecosystem in Ecuador, consider the legal jungle of the status, or condition, of individuals as legal members of a state, e.g., people's citizenship. As shown by the legal condition of Turks in Germany, or of some Brazilian football players in Italy, or of young immigrants in US, this is the realm of political discretionary power that sometimes turns into simple chaos, or mere sovereign arbitrariness. The recent case of Saudi Arabia enrolling Sophia as a citizen of her own is hence unsurprising. It reminds us of Suetonius' *Lives of the Twelve Caesars* (121 AD), in which we find Caligula planning to make his horse, Incitatus, a senator, and "the horse would invite dignitaries to dine with him in a house outfitted with servants there to entertain such events".

From Incitatus to Sophia, the paper has stressed the normative reasons, according to which we can evaluate whether granting legal personhood makes sense, or turns out to be a simple matter of sheer chance and political unpredictability. In the case of legal persons, such as corporations, political decisions have to do with matters of efficiency, financial transparency, accountability, and the like. In the case of human fellows, the reference is to their dignity, consciousness, intrinsic worth, and so forth. Certainly, we cannot prevent on this basis the odd decisions of legislators making robots citizens, or horses senators. Yet, from Caligula's horse to current Sophia, basic legal principles make clear when political decisions on "persons" are incongruous, so that courts may one day overturn them for having no rational basis.

Funding: This research received no external funding

Conflicts of Interest: The author declares no conflict of interest.

References

1. Reynolds, C.; Ishikawa, M. Robotic Thugs. In *EthiComp*; Bynum, T.W., Rogerson, S., Murata, K., Eds.; Global e-SCM Research Center & Meiji University: Tokyo, Japan, 2007; pp. 487–492.
2. Pagallo, U. The Human Master with a Modern Slave? Some Remarks on Robotics, Ethics, and the Law. In *The "Backwards, Forwards and Sideways" Changes of ICT*; Arias-Oliva, M., Bynum, T.W., Rogerson, S., Torres-Corona, T., Eds.; Universitat Rovira I Virgili: Tarragona, Spain, 2010; pp. 397–404.
3. Pagallo, U. The Adventures of Picciotto Roboto: AI and Ethics in Criminal Law. In *The Social Impact of Social Computing*; Bissett, A., Light, A., Lauener, A., Rogerson, S., Ward Bynum, T., Eds.; Sheffield Hallam University: Sheffield, UK, 2011; pp. 349–355.
4. Pagallo, U. *The Laws of Robots: Crimes, Contracts, and Torts*; Springer: Dordrecht, The Netherlands, 2013.
5. Bryson, J.J.; Diamantis, M.E.; Grant, T.D. Of, for, and by the People: The Legal Lacuna of Synthetic Persons. *Artif. Intell. Law* **2017**, *23*, 273–291. [CrossRef]
6. UN World Robotics. *Statistics, Market Analysis, Forecasts, Case Studies and Profitability of Robot Investment*; Edited by UN Economic Commission for Europe and Co-Authored by the International Federation of Robotics; UN Publication: Geneva, Switzerland, 2005.
7. Gogarty, B.; Hagger, M. The Laws of Man over Vehicle Unmanned: The Legal Response to Robotic Revolution on Sea, Land and Air. *J. Law Inf. Sci.* **2008**, *19*, 73–145.
8. Singer, P. *Wired for War: The Robotics Revolution and Conflict in the 21st Century*; Penguin: London, UK, 2009.
9. Pagallo, U. Robots of Just War: A Legal Perspective. *Philos. Technol.* **2011**, *24*, 307–323. [CrossRef]
10. Bekey, G.A. *Autonomous Robots: From Biological Inspiration to Implementation and Control*; The MIT Press: Cambridge, MA, USA; London, UK, 2005.
11. Simon, H. *The Shape of Automation for Men and Management*; Harper & Row: New York, NY, USA, 1965.
12. Minski, M. *Computation: Finite and Infinite Machines*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1967.
13. OSTP; National Science and Technology Council Networking and Information Technology; Research and Development Subcommittee; National Science and Technology Council Networking and Information Technology. *The National Artificial Intelligence Research and Development Strategic Plan*; OSTP: Washington, DC, USA, 2016.
14. Hildebrandt, M. From Galatea 2.2 to Watson and back? In *Human Law and Computer Law: Comparative Perspectives*; Hildebrandt, M., Gaakeer, J., Eds.; Springer: Dordrecht, The Netherlands, 2011.

15. Chopra, S.; White, L.F. *A Legal Theory for Autonomous Artificial Agents*; The University of Michigan Press: Ann Arbor, MI, USA, 2011.
16. Karnow, C.E.A. Liability for Distributed Artificial Intelligence. *Berkeley Technol. Law J.* **1996**, *11*, 147–183.
17. Lerouge, J.-F. The Use of Electronic Agents Questioned under Contractual Law: Suggested Solutions on a European and American level. *John Marshall J. Comput. Inf. Law* **2000**, *18*, 403.
18. Weitzenboeck, E.M. Electronic Agents and the Formation of Contracts. *Int. J. Law Inf. Technol.* **2001**, *9*, 204–234. [CrossRef]
19. Bayern, S.; Burri, T.; Grant, T.D.; Häusermann, D.M.; Möslin, F.; Williams, R. Company Law and Autonomous Systems: A Blueprint for Lawyers, Entrepreneurs, and Regulators. *Hastings Sci. Technol. Law J.* **2017**, *9*, 135–162. [CrossRef]
20. Pagallo, U. The Group, the Private, and the Individual: A New Level of Data Protection? In *Group Privacy: New Challenges of Data Technologies*; Taylor, L., Floridi, L., van der Sloot, B., Eds.; Springer: Dordrecht, The Netherlands, 2017; pp. 159–173.
21. Pagallo, U.; Quattrocchio, S. The Impact of AI on Criminal Law, and Its Twofold Procedures. In *The Research Handbook of the Law of Artificial Intelligence*; Barfield, W., Pagallo, U., Eds.; Elgar: Cheltenham, UK, 2018.
22. Pagallo, U. AI and Bad Robots: The Criminology of Automation. In *The Routledge Handbook of Technology, Crime and Justice*; McGuire, M.R., Holt, T.J., Eds.; Routledge: London, UK; New York, NY, USA, 2017; pp. 643–653.
23. Pagallo, U. LegalAIze: Tackling the Normative Challenges of Artificial Intelligence and Robotics through the Secondary Rules of Law. In *New Technology, Big Data and the Law. Perspectives in Law, Business and Innovation*; Corrales, M., Fenwick, M., Forgó, N., Eds.; Springer: Singapore, 2017; pp. 281–300.
24. Singer, P. *Practical Ethics*; Cambridge University Press: Cambridge, UK, 2011.
25. Solum, L.B. Legal personhood for artificial intelligence. *N. C. Law Rev.* **1992**, *70*, 1231–1287.
26. Hallevy, G. *Liability for Crimes Involving Artificial Intelligence Systems*; Springer: Dordrecht, The Netherlands, 2015.
27. Hildebrandt, M.; Koops, B.-J.; Jaquet-Chiffelle, D.-O. Bridging the Accountability Gap: Rights for New Entities in the Information Society? *Minn. J. Law Sci. Technol.* **2010**, *11*, 497–561.
28. Coudert, A.P. *Leibniz and the Kabbalah*; Kluwer Academic: Boston, MA, USA; London, UK, 1995.
29. Pagallo, U. From Automation to Autonomous Systems: A Legal Phenomenology with Problems of Accountability. In Proceedings of the International Joint Conferences on Artificial Intelligence Organization (IJCAI-17), Melbourne, Australia, 17–23 August 2017.
30. Pagallo, U. Three Lessons Learned for Intelligent Transport Systems that Abide by the Law. JusLetterIT. Available online: http://jusletter-it.weblaw.ch/issues/2016/24-November-2016/three-lessons-learned_9251e5d324.html (accessed on 24 November 2016).
31. Pizzetti, F. *Intelligenza Artificiale, Protezione dei Dati e Regolazione*; Giappichelli: Torino, Italy, 2018. (In Italian)
32. De Saxoferrat, B. Commentary on Digestum Novum. In *Commentaria*; Il Cigno: Rome, Italy, 1996.



© 2018 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

RYEL: An Experimental Study in the Behavioral Response of Judges Using a Novel Technique for Acquiring Higher-Order Thinking Based on Explainable Artificial Intelligence and Case-Based Reasoning

Luis Raúl Rodríguez Oconitrillo ^{1,*}, Juan José Vargas ¹, Arturo Camacho ¹, Álvaro Burgos ²
and Juan Manuel Corchado ^{3,4,5,6,*}

- ¹ School of Computer Science and Informatics, Universidad de Costa Rica (UCR), Ciudad Universitaria Rodrigo Facio Brenes, San José 11501-2060, Costa Rica; jvargas@ecci.ucr.ac.cr (J.J.V.); arturo.camacho@ecci.ucr.ac.cr (A.C.)
 - ² Law School, Universidad de Costa Rica (UCR), Ciudad Universitaria Rodrigo Facio Brenes, San José 11501-2060, Costa Rica; alvaro.burgos@ucr.ac.cr
 - ³ Bisite Research Group, Universidad de Salamanca, 37008 Salamanca, Spain
 - ⁴ Air Institute, IoT Digital Innovation Hub, 37008 Salamanca, Spain
 - ⁵ Department of Electronics, Information and Communication, Faculty of Engineering, Osaka Institute of Technology, Osaka 535-8585, Japan
 - ⁶ Pusat Komputeran dan Informatik, Universiti Malaysia Kelantan, Kelantan 16100, Malaysia
- * Correspondence: lurago34@gmail.com (L.R.R.O.); jm@corchado.net (J.M.C.)

Citation: Rodríguez Oconitrillo, L.R.; Vargas, J.J.; Camacho, A.; Burgos, A.; Corchado, J.M. RYEL: An Experimental Study in the Behavioral Response of Judges Using a Novel Technique for Acquiring Higher-Order Thinking Based on Explainable Artificial Intelligence and Case-Based Reasoning. *Electronics* **2021**, *10*, 1500. <https://doi.org/10.3390/electronics10121500>

Academic Editor: Jun Liu

Received: 20 May 2021
Accepted: 16 June 2021
Published: 21 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Abstract: The need for studies connecting machine explainability with human behavior is essential, especially for a detailed understanding of a human's perspective, thoughts, and sensations according to a context. A novel system called RYEL was developed based on Subject-Matter Experts (SME) to investigate new techniques for acquiring higher-order thinking, the perception, the use of new computational explanatory techniques, support decision-making, and the judge's cognition and behavior. Thus, a new spectrum is covered and promises to be a new area of study called Interpretation-Assessment/Assessment-Interpretation (IA-AI), consisting of explaining machine inferences and the interpretation and assessment from a human. It allows expressing a semantic, ontological, and hermeneutical meaning related to the psyche of a human (judge). The system has an interpretative and explanatory nature, and in the future, could be used in other domains of discourse. More than 33 experts in Law and Artificial Intelligence validated the functional design. More than 26 judges, most of them specializing in psychology and criminology from Colombia, Ecuador, Panama, Spain, Argentina, and Costa Rica, participated in the experiments. The results of the experimentation have been very positive. As a challenge, this research represents a paradigm shift in legal data processing.

Keywords: interpretation-assessment/assessment-interpretation (IA-AI); hybrid artificial intelligence system; mixture of experts (MOE); explainable case-based reasoning (XCBR); explainable artificial intelligence (XAI); semantic networks (SN)



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

There is a need for a computational framework that allows capturing, representing and processing the meta-knowledge [1] of a human in the context of the domain of discourse and study the behavioral response of a person in light of explanatory machine techniques [2]. For example, in the legal area, it means to get an intelligent and coherent explanation of the legal analysis made by a human in a particular scenario of a previous case and find the reasons about why a particular law was used [3] in order to support decision-making when other judges are dictating a resolution. This situation seems superfluous, but it is not, because it usually would imply navigating between a complicated

set of theories that range from cognitive learning theories [4,5], instructional design [6], cognitive theory [7], and information processing theory [8], among others. These theories reveal how a judge can learn and support decision-making using the knowledge from other judges and the sentences. In this particular investigation, there is a path that subsumes and synthesizes in some way some parts of the previous theories and focuses them on a practical point of application, and that leads directly to basal knowledge [9], and that is the Subject-Matter Experts (SME) [10] from which the analysis of the merits of a case (legal matter background) is the main activity of a judge. In this way, it is possible to lead efforts to work with higher-order thinking [11] using the technology like a meta-media [12] to manage meta-knowledge.

The study of the merits of a case involves analyzing the scenarios formed by the facts associated with a case. A legal analysis consists of a judge perceiving [13] and analyzing the facts and evidence, according to a determined legal posture [14,15]. It is to emphasize that the behavioral response of a research subject, such as a judge, is diverse, so it is necessary to investigate the response when using the framework in terms of functionality, usability, and efficiency when analyzing the merits of a case.

Studying the judge's behavior on accepting or rejecting the use of the framework is not as simple as asking the judge if they agree, like, or think it is possible to use this framework to do such work. This investigation is a complete challenge to the mental and psychological scheme because there are rules in the domain of discourse that represent substantial barriers to conduct experiments and studies in technology and human behavior inside the jurisdictional area. The main barriers are: (1) That nothing and no one can intervene in the decision-making of a judge, this is called judicial independence, and (2) the judges have a high degree of discretion to make decisions, and nothing and no one can tell them how or what decision to make [16,17]. So far, no research has evaluated the behavior of the judges when faced with the use of a framework that helps them with the deep analysis of a case by taking fragments of the human psyche [18] using meta-knowledge, focusing on the perception [13], and adapting Artificial Intelligence (AI) [19] techniques to explain that fragments. The psyche, in this research, is understood as the processes and phenomena that make the human mind works as a unit that processes perceptions, sensations, feelings, and thoughts, rather than a metaphysical phenomenon [18]. So, it is understandable that the psyche is exceptionally complex; however, it is possible to explore some deep traits and characteristics that can be expressed through layers of awareness [20] or envelope [21] of knowledge, based on the perception a human has from objects and relationships of the real-world. In this way, it is possible to express, through related meta-knowledge fragments, the meaning, and purpose of someone in a specific context.

Thus, this research aims not only to investigate the behavioral development of the judge when using new technology for in-depth analysis of cases but also to show computational advances with a high impact in cognitive and psychological fields. So, this research presents a Mixture of Experts (MOE) system [22,23] called RYEL [24–27]. This system was created based on CBR guidelines [3,28–30] and Explainable Artificial Intelligence (XAI) [31–33] using focus-centered organization fundamentals, which means the organization of XAI and CBR is done and focus according to the perspective and approach that a human has in a domain of discourse, meaning it is human-centric [34–36]. A human interacts with the system through Explanatory Graphical Interfaces (EGI) [2], which are graphic modules that implement computational techniques of knowledge elicitation [37] to capture, process, and explain the perception of a human about facts and evidence from scenarios in a context. RYEL uses the method called Interpretation-Assessment/Assessment-Interpretation (IA-AI) explained in [2] which consists not only in explaining machine inferences but also the point of view that a human has using metadata from the real world along with statistical graphs and dynamic graphical figures.

Various investigations try to obtain knowledge from past cases using the traditional Case-Based Reasoning (CBR) approach in a legal context, such as [3,28,30,38–41]. In those systems, CBR consists only of solving current cases according to how previous ones

were solved, that is, in a deterministic way. This kind of solution is different in capturing the judge's interpretation and assessment of facts and provides an intelligent simulation [42–45] that allows a legal analysis about the merits of the case. This approach is an understudied approach concerned in identifying the perception of a judge about the objects and relationships of the real world involved in a case, along with the machine's ability for capturing and processing that information and explaining it graphically on an interface [46].

Thus, the novelty of this research not only lies in the societal impact of using XAI and CBR to assist judges in resolving legal disputes between humans with the novel IA-AI method to analyze the merits of a case, but also the behavioral study of the judge in the face of this technology. Therefore, a balance in explaining the software design and behavioral analysis of the judges is the key to reveal essential aspects of this investigation. The following sections explain this balance.

2. Framework Design

Design Science research process proposed in [47] allowed the creation of the computational framework of RYEL explained in [2], implementing CBR life-cycle stages as shown in Figure 1 as a guide to exchange and organize the information with the user. The system design was developed in [24–26] as a hybrid system [48,49] implementing different machine learning techniques for every CBR stages [28,30,38,39,41,50–52]. An adaptation took place to implement the stages to graphical interfaces where the judge can manipulate corresponding images representing connected facts and evidence of the scenarios. How the scenarios show a definition, relationship, characterization, and description according to a legal context using images allows the machine to acquire higher-order thinking [11] from humans dynamically and graphically. By manipulating interrelated images, a human expresses ideas and points of view. The system takes the images as inputs to carry out a legal analysis simulation and generates a graphical explanation of the laws applicable to the factual picture. Other judges use the explanations provided by the machine for decision-making support.

The data overview diagram of the system consists of image inputs, evidence and facts processing, norms and laws outputs, CBR articulating and processing the information between the user and machine; organizing the inputs and outputs of the system as depicted in Figure 2. The role of EGIs is to provide graphical interfaces that are used for human interaction with a computer, called Human-Computer Interaction (HCI) [53], an example of the interface is shown in Figure 3. The graphical techniques of the interfaces allow the elicitation of human knowledge using the ligaments between the shape of images and the content of its attributes, as well as the relationships between images. This means a meta-media to manage higher-order thinking by combining the functionality of the EGIs, for example, by combining the functions of the interfaces of Figure 4 with Figure 5. This combination makes it possible to work with the range, importance level, order, and attribute links between images. The role of IA-AI is to obtain the perception of a human from the dynamic triangulation of attributes expressed with images, relationships, and unsupervised algorithms [2].

Explanation of the system design must line up with the study of human behavior in light of the cognitive field and technology. Thus, the following points allow alignment, and the next sections explain them: (1) The cognitive environment [54] of the judge in order to delineate the domain of discourse and understand both the computational nature of the data and the behavioral study of the judge; (2) the knowledge representation, (3) the computational legal simulation, and (4) the hybrid nature of the system.

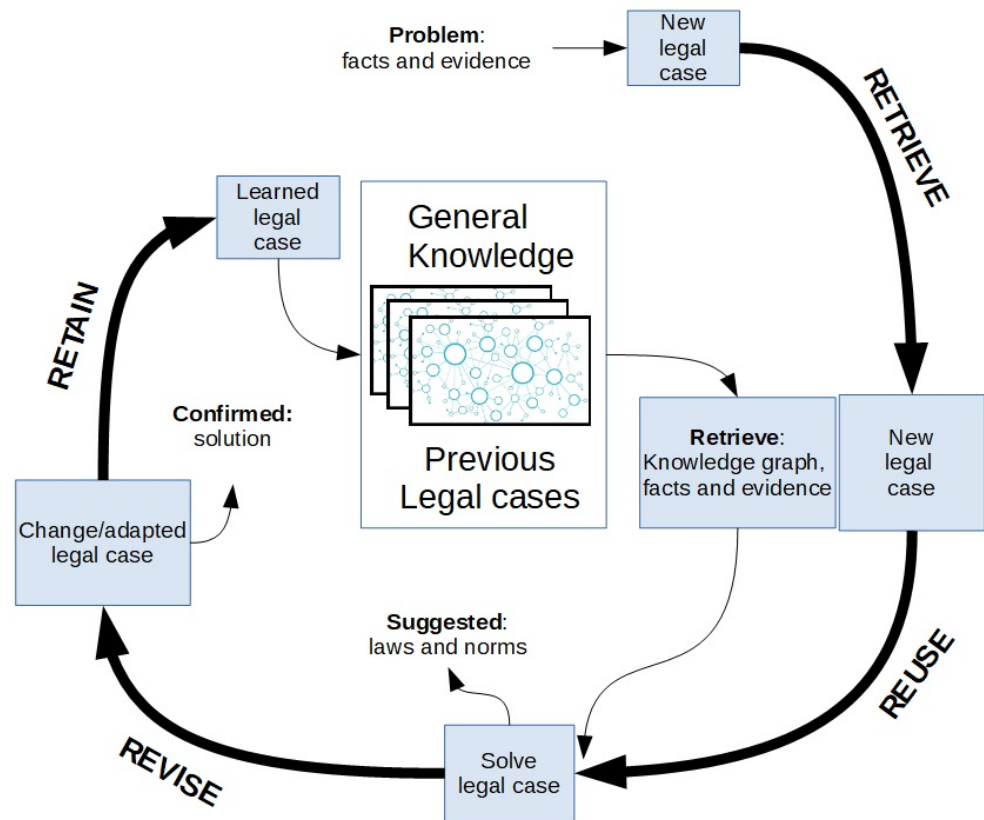


Figure 1. Stages of the case-based reasoning life-cycle, used in the RYEL system: Retrieve, reuse, review, and retain.

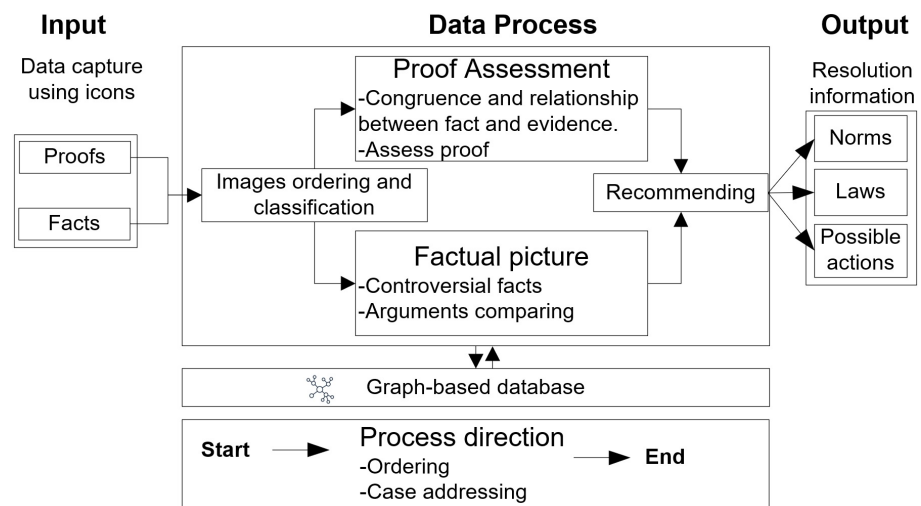


Figure 2. Data overview diagram of the system: Image inputs, evidence and facts processing, and norms and laws outputs.

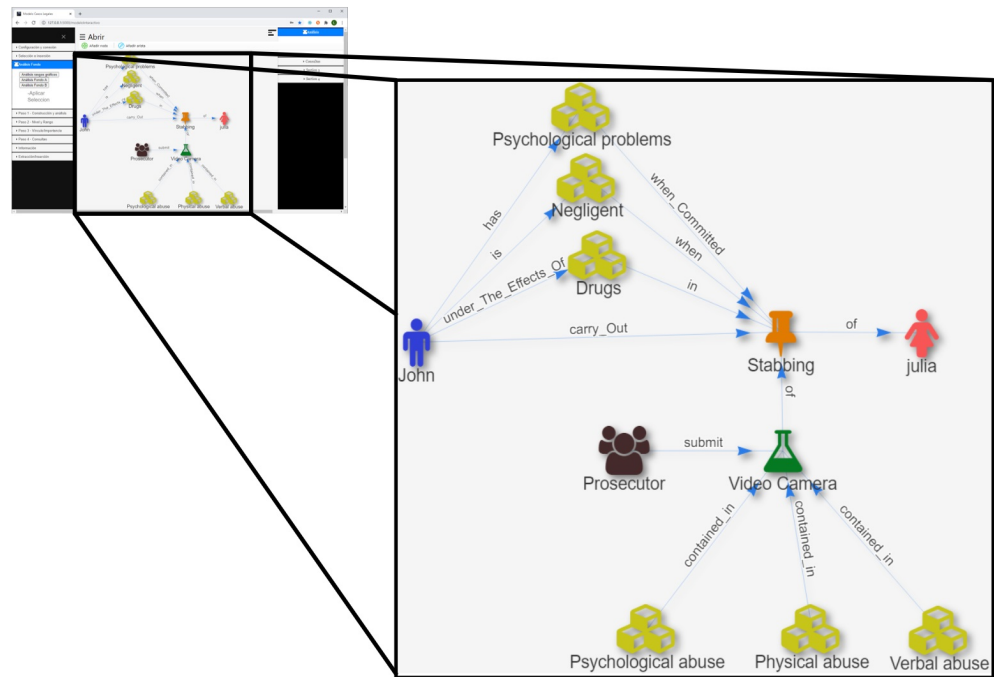


Figure 3. EGI: The graphic arrangement of images according to the interpretation and assessment of a judge on a simplified legal scenario related to a stabbing in a homicide case.

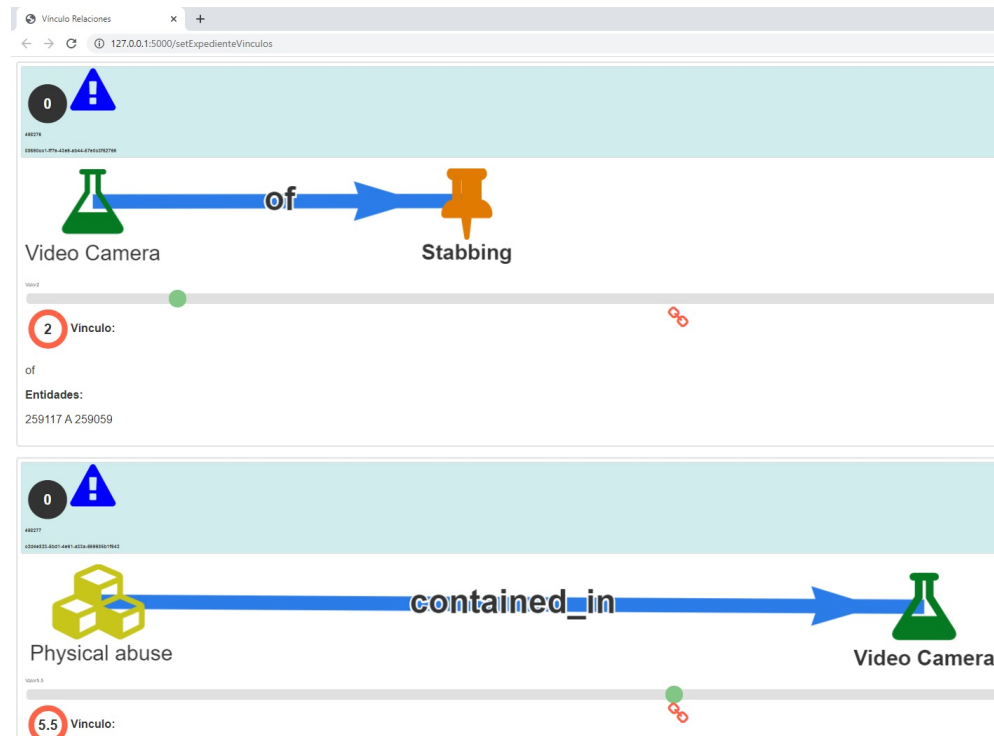


Figure 4. EGI allows the evaluation and listing of links between facts and evidence in a scenario.

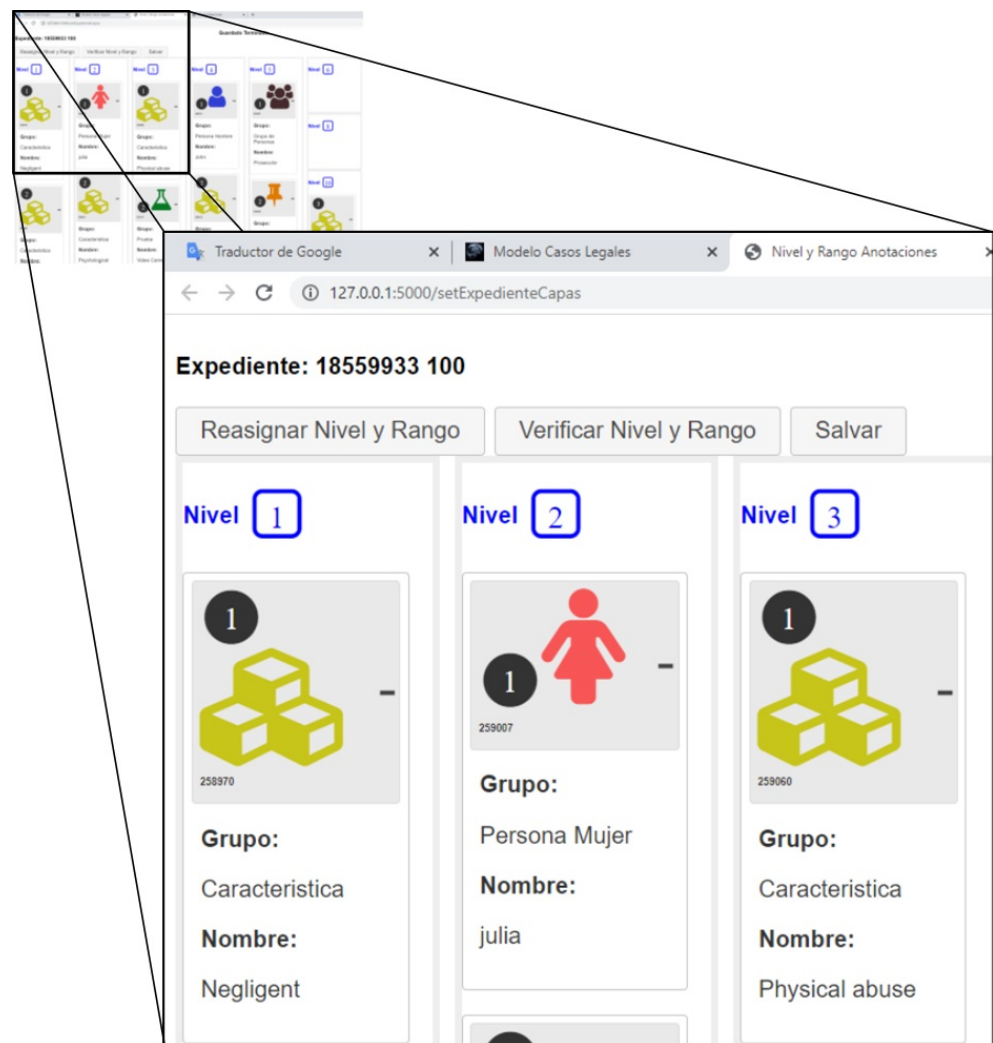


Figure 5. Judges use EGI to classify the granules of information in a case according to their perspective. The levels represent the degree of interest (importance), and the location of each granule within each level represents the order of precedence (range). Again, the granules can represent facts or evidence, as well as other case data.

2.1. Cognitive Environment

A judge may have extensive knowledge. However, the system focuses on how a judge understands information from scenarios in a context, as shown in Figure 6. This figure explains the definition of understanding in this research in terms of (1) perception, (2) perspective, and (3) interpretation. These words seem to be standard and straightforward terms, but the system treats them as part of its nature and requires explanation.

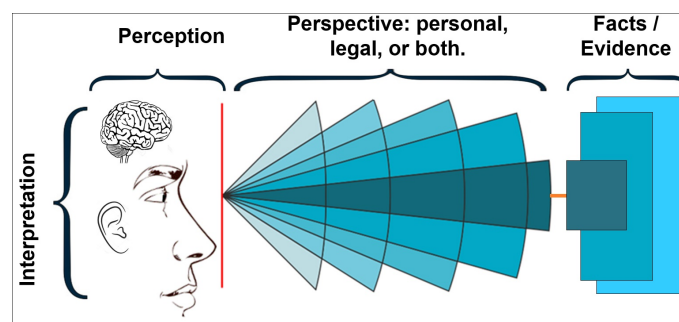


Figure 6. This figure shows the cognitive legal information and relationship-interaction between the perception, perspective, and interpretation.

Perception in [13] is a mental process that involves activities such as thought, learning, memory, and others, along with a link to the state of memory and its organization. It is a subjective state where a person captures and understands, in their way, the qualities of objects and facts from the real world. Therefore, a judge may have a different perception of the information between one file and another. For example, a judge in a Domestic Violence Court has grasped, learned, and is aware that the defendant from the beginning is an alleged aggressor given a situation of vulnerability over the victim. However, a judge in Criminal Court has learned and understood that the “Principle of Innocence” must be used with a defendant, which presumes the state of not being guilty until proven otherwise.

Perspective in [55] is the point of view from which an issue is analyzed or considered. The perspectives can influence people’s perceptions or judgments. The judges’ perceptions could change according to their attitude, position, or considerations about facts, objects, individuals, entities, or type of work. The annotations of a case, which are information from the legal file, can be analyzed using a different perspective; for example, a judge in a Criminal Tax Investigation Court may see the action of hitting a person as not so severe or even belittle it, while a judge in a Domestic Violence Court can see it as very serious.

Interpretation in [56] means expressing or conceiving a reality personally or attributing meaning to something. Thus, the judges could conceive an annotation from the legal file according to their reality and attribute and then assign a meaning. Consider this example, shooting to a person can be interpreted by a judge in a criminal court as an act of personal defense and assess it as a reason to preserve life, while another judge, from the same court, may interpret it as an act of aggression and assess it as a reason to steal something.

The system handles the interpretation and assessment made by a judge as two separate but interacting processes. In order to understand this interaction, consider the following example; person *X* assesses the help a person *Y* gave them in a trial, but person *X* cannot interpret the reasons of the help, because person *Y* is their enemy, or else, person *X* interprets that their enemy helped them in a trial because they want something from him. For that reason, the help is not so valued by person *X*. This example shows how the interpretation and assessment interact in this investigation.

In the file, how a judge understands the facts and evidence is not recorded. Currently, a file only contains the final decision of a judge supported by motivations and underpinnings of the law, along with chunks of structured data like “the outcome”, “the considering”, and “the therefore” as described in [17,57]. Thus, this unrecorded information is precisely the most important to understanding the perception of a human. The graphical techniques and explainable methods [33], in this investigation, allow to capture and detail this information.

2.2. Knowledge Graph

Internally, the system transforms images and relationships representing the scenarios of a case into directed graphs called Knowledge Graphs (KG) [58–60], which contain object types, properties, and relationships from real criminal cases. Through graphic media, the judge can obtain information about the ontological content [61] processed by the images. After the image transformation, each scenario is converted to a set of nodes and edges, representing facts or evidence along with the relationship that explains their bond, which translates into hermeneutical content [62]. There may be more than one scenario per legal case. The expressive semantic nature [63] of the KG allows for having different graphical forms [46] to show the reasoning of a judge and to understand the use of law in a proven fact (fact whose evidence accredits it as a true) in a crime. In [64] the KGs have been prevalent in both academic and industrial circles in these years because they are one of the most used approaches to integrate different types of knowledge efficiently.

Usually, the judge performs the mental process of relating legal concepts of the scenarios to find the meaning of the information provided by the parties in conflict. Thus, to determine whether the facts are truthful, the judge makes groups of evidence and links them to the facts. The groups, data type, and relationships in the legal scenarios mold a

network that expresses meaning. Therefore, a network is generated and is visualized as Semantic Networks (SN) [65–67] by the system.

In [68], the SN is a directed graph composed of nodes, links, or arcs, as well as labels on the links. KG in [63] is a type of SN, but the difference lies in the specialization of knowledge and in creating a set of relationships. Thus, the knowledge structure depends on the domain of application, and graph structure changes according to the knowledge expressed. Since the system translated images into nodes describing physical objects, concepts, or situations in a legal context, the relationships (edges) between images are transformed into links and express a connection between objects in legal scenarios. Links have labels to specify a particular relationship between the objects of the legal case. Thus, nodes and edges are a means to make the structure of legal knowledge. In this way, the use of images and relationships allows the construction of KG that represents the judge's knowledge after having interpreted and evaluated the facts and evidence contained in the scenarios, and this is the reason why the graphs include information about properties types and relationships between entities. An entity can be an object, a person, a fact, a proof, or the law.

Human Interaction

The judge can access graphic resources in the form of images representing legal elements [16] which are pieces of juridical data made of evidence and facts, as shown in Figure 7. An EGI offers the judge a popup menu to select the image that best reflects a record entry from the expedient. In addition, the system has a drawing area called working canvas where the judges can draw their perception of the scenarios by establishing, organizing, distributing, and relating the images that they select from the menu, as shown in Figure 7. The KGs built with the images are stored in an unstructured database, and when this happens, they become a more specific type of graph called a Property Graph (PG) [69–71].

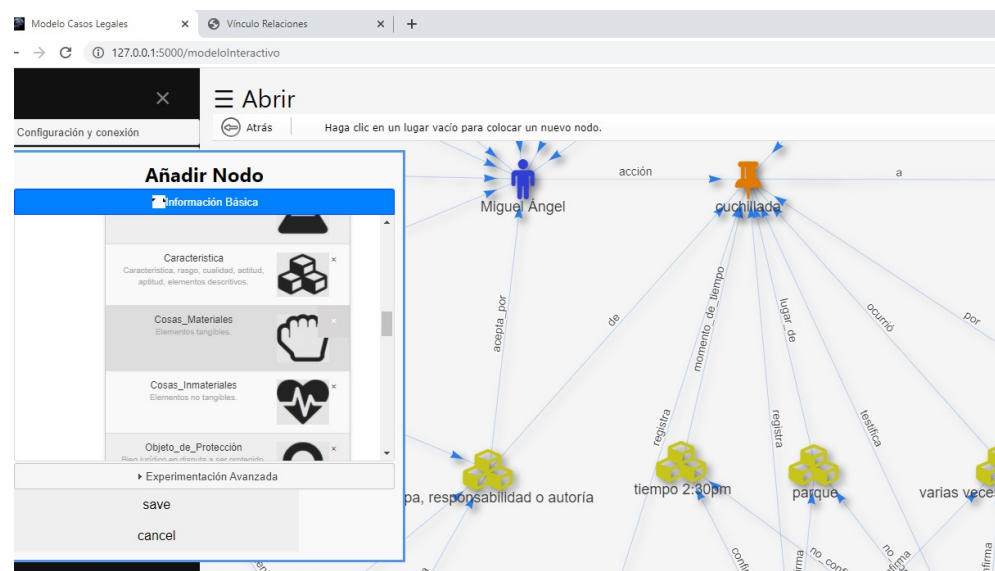


Figure 7. Example of an EGI showing a factual picture used to analyze the merits of a homicide case in real-time according to the graphical interpretation. Internally each image is translated into a node containing a set of attributes, and the arrows turn into edges containing a set of vectors.

The judge can change the display state of an EGI; this is between images or nodes to study the attribute representations in both states. The nodes acquire colors, sizes, and properties, to explain the details of the attributes visually. Edges acquire properties such as length, thickness, color, and orientation to explain how the nodes are linked and distributed. Both nodes, as well as edges, contain unique properties resulting from the transformation process. The system uses the IA-AI method to create properties and attributes. The method has three main processes. In the first process, after the judge has

finished drawing on the working canvas, like in Figure 7, they can interpret and assign the levels and ranges of importance to the images drawn. The judge does it by dragging and dropping the images into previously designed graphic boxes (precedence and importance levels) as shown in Figure 5. In the second process, other EGIs are used to assess the links between images representing the facts and evidence (proof assessment); the judge does this by hanging up the links in different positions and establishing the bond length (link) between objects as shown in Figure 4. In the third process, another EGI is used to explain recommendations on laws and regulations concerning the factual picture depending on the context, as shown in Figure 8, where the Y-axis indicates the legal taxonomy, this means the order of importance of the legal norms according to the context. The X-axis represents the level of similarity that the norms have in the factual picture of the scenarios. Finally, the machine recommends groups of norms represented by a higher or right circle in the chart, depending on what the judge is analyzing.

During a case, the judges can run legal simulations to delve into the merits of the case gradually. The simulation carried out by the system is described below.

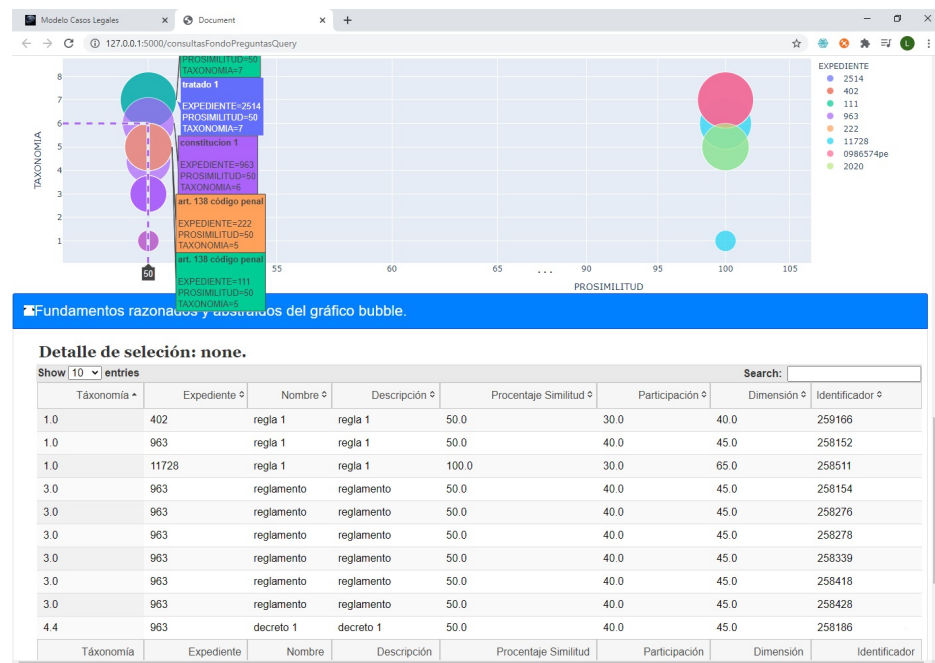


Figure 8. EGI explains to the user using circles, size, and colors, the set of laws and regulations in line with the factual picture of the case under analysis. The machine expresses by a graphical distribution of circles those norms and laws that best describe the legal scenario under study. The recommended norms and laws to take are higher and farther to the right in the chart. However, a judge can explore, analyze, and select those that best fit the factual picture.

2.3. Intelligent Simulation

Figure 9 shows the legal simulation activity. There are three main processes in the simulation which are: (1) The capture of the interpretation and assessment values using EGIs [25,27] that a judge makes of the facts and evidence of a case, (2) identify the patterns of interpretation [3] using CYPHER [71] scripts to extract the semantic [72,73] and the ontological [74,75] content of the facts and evidence contained in the scenarios of a case depicted by EGIs, and (3) the options the machine offers to the judge to distill legal information from the patterns found in the graphs as shown in Figure 10 by using unsupervised algorithms [71], like Jaccard [76], Cosine of Similarity [77], and Pearson's Correlation Coefficient [78] applied to graphs. Then the machine provides an explanation of the results. Some examples of the information that the machine explains are: (a) A graphical explanation about a set of norms that apply to a case; (b) explain and identify the

evidence that is not related to some fact; (c) detects the evidence that not evaluated; and (d) indicate what evidence has been evaluated but not related.

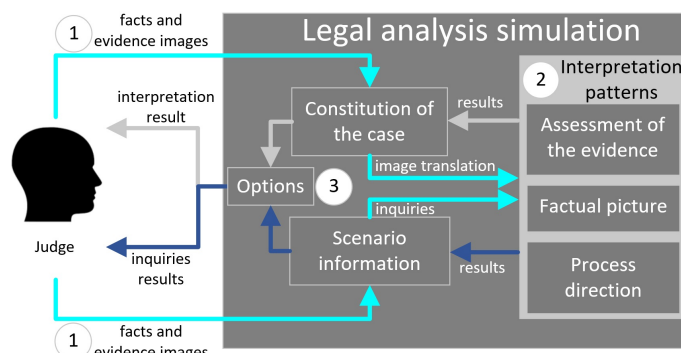


Figure 9. Legal analysis simulation of the merits of a case, considering: Evidence, facts, and the direction process of the legal data in a trial.

1. Análisis de Fondo: Interpretación y Valoración 2. Fundamentos: Normas, Leyes y Acciones

- 1. Valores EXACTOS para hechos y pruebas
 - Interpretación exacta
 - Valoración exacta
 - 2. Valores APROXIMADOS para hechos y pruebas
 - Interpretación próxima
 - Valoración próxima
 - 3. Valores de interpretación y valoración menos similares. Aquello empleado por otros jueces y no por mí.
 - Interpretación no usada
 - Valoración no usada
 - 1. Todas las usada según los hechos y pruebas.
 - Normas empleadas.
 - Reglas aplicadas.
 - Acciones impuestas.
 - 2. Las que no he usado y que están usadas pra los mismos hechos y pruebas.
 - Normas no usadas.
 - Reglas no usadas.
 - Acciones no usadas.
- 3. Análisis de cantidad y tipos: Elementos entre Hechos y Pruebas**
- 1. Menor cantidad de elementos entre 2 elementos.
 - 2. Mayor cantidad de elementos entre 2 elementos.
- [Ejecutar Análisis del Fondo](#)

Figure 10. Graphical interface showing simulation options.

When judges use the system continuously, they will be able to integrate legal knowledge during a trial.

Meta-Knowledge Integration

Knowledge Integration (KI) [79] happen by capturing and representing the interpretation and assessment of facts and evidence made by the judges at the beginning of a trial, along with the knowledge obtained from the analysis of the legal simulations. Thus, the system unifies unstructured knowledge [80] of the interpretation and assessment values of a judge according to their legal perspective and new information that may appear during the process of a trial to the end of it. In addition, KI allows the generation and integration of fragments of meta-knowledge. There are three points of KI and one more at the end of the trial when judges dictate a resolution or until the sentence appeal. If the resolution is legally challenged (contested decision), then there is one more point of KI. At each point, the judges can express new insights or changes of facts and evidence and run a legal analysis simulation, as many as necessary.

2.4. Hybrid System

RYEL uses different types of machine learning techniques therefore, it has the characteristic of being a hybrid [48] system. Hybridization applies in a multitude of computational areas, as in [81,82]. However, this research focuses on the legal field, specifically on facts and evidence from a case analyzed by a judge.

The hybridization [48] of RYEL is organized under the MOE [22,23] foundations based on the divide-and-conquer principle [23,83]. That means that different parts or segments constitute the problem space; each part corresponds to a module called an

“expert” [23]. MOE usually uses “gate network” [23] that decides to which expert a specific task should be assigned to deal with complex and dynamic problems [48], for example, the use of various experts for multiple label classifications using Bayesian Networks (BN) and tree structures [22]. Supervised machine learning such as neural networks are typically used [22,83] in MOE, however our approach is unsupervised [84] using KG [25,27], to build SN with a CBR [3,85,86] and XAI [32,87].

2.5. Case Explicability

The implementation of XAI and CBR reveals the interconnections and characteristics of objects within the scenarios of a context. Due to the use of KG, it is possible to achieve legal exegesis [88] by obtaining the hermeneutical content of relations and objects together with ontological data through their properties. That means that a legal interpretation is according to the content expressed by a judge; therefore, the semantic explained initially.

The adaptations of the CBR stages, shown in Figure 1, are the following: (1) Retrieve, whereby the judges have graphical options to execute a legal analysis simulation to find patterns of interpretation of the facts and assessment of the evidence similar to the case depicted in the working canvas for a specific context; (2) reuse, whereby the system synthesizes and evaluates the patterns found, and detects the laws with which they have links in order to be considered by the judges in new cases; and (3) revise, whereby the judges of higher-hierarchy use the EGIs to make a review of the performance made by lower judges aimed to make modifications and corrections in the factual picture posed on the working canvas. In this stage, the system integrates knowledge of the judges and the parties in conflict. If the parties in conflict appeal to the resolution (legal challenge), then higher-hierarchy judges must revise the scenarios. The higher judges can also run legal analysis simulations in order to consult, verify, correct, or add new perspectives to refute or accept, in the whole or part, the analysis carried out by judges of lower-hierarchies; then, the issuance of a final resolution occurs and (4) retained, which means that the sentence is final and no further legal simulation is necessary. In the retention stage, the system incorporates cognitive information into the knowledge database because the possible errors of bias in perception were eliminated or corrected by reviewing several humans during the legal process using the system. Figure 1 shows a list that summarizes the stages of the CBR.

1. Case-Base: A KG represents this;
2. The Problem: Is the interpretation and assessment of both facts and evidence;
3. Retrieve: Using CYPHER script patterns and graph similarity algorithms like Cosine, Pearson, and Jaccard;
4. Reuse: Consists of detecting norms and laws related to the factual picture drawn on the working canvas;
5. Revise: Analyze and review the work done by lower judges using KG via EGI;
6. Retain: Is the stage of adding to the knowledge base a correct approach to interpreting and assessing a factual picture.

2.6. Case Definition, Data Model, and Example

Formally a case is a graphical deposition of facts and evidence made by the judges according to their perspective using EGI. In one case, there are segments of information called “scenarios” that contain facts related to the evidence. Scenarios are a way to express and organize legal information.

An in-depth legal analysis is the identification and description of both data and relationships within each scenario. The judges do this analysis as they work through the case during the trial. To exemplify the data and relationships, consider the data model of segment 1 in Figure 11 where bidirectional arrows represent that a relationship can go one way or the other from concepts or objects, and it demonstrates the organization and relation of the meta-knowledge. In this figure in segment 2, a simplified real world example of a “violation case” uses the data model from segment 1. The elements called “Material Object” and “Formal Object” broach the subject of each scenario; for example, in this figure,

there is a case of a man affecting a woman through the action of rape. In [89] a formal object means carrying out a legal study, from a particular perspective, on the relationship of legal data; the material object is the matter that deals with such data, but in this case, all the relationships that describe each object are also represented and organized.

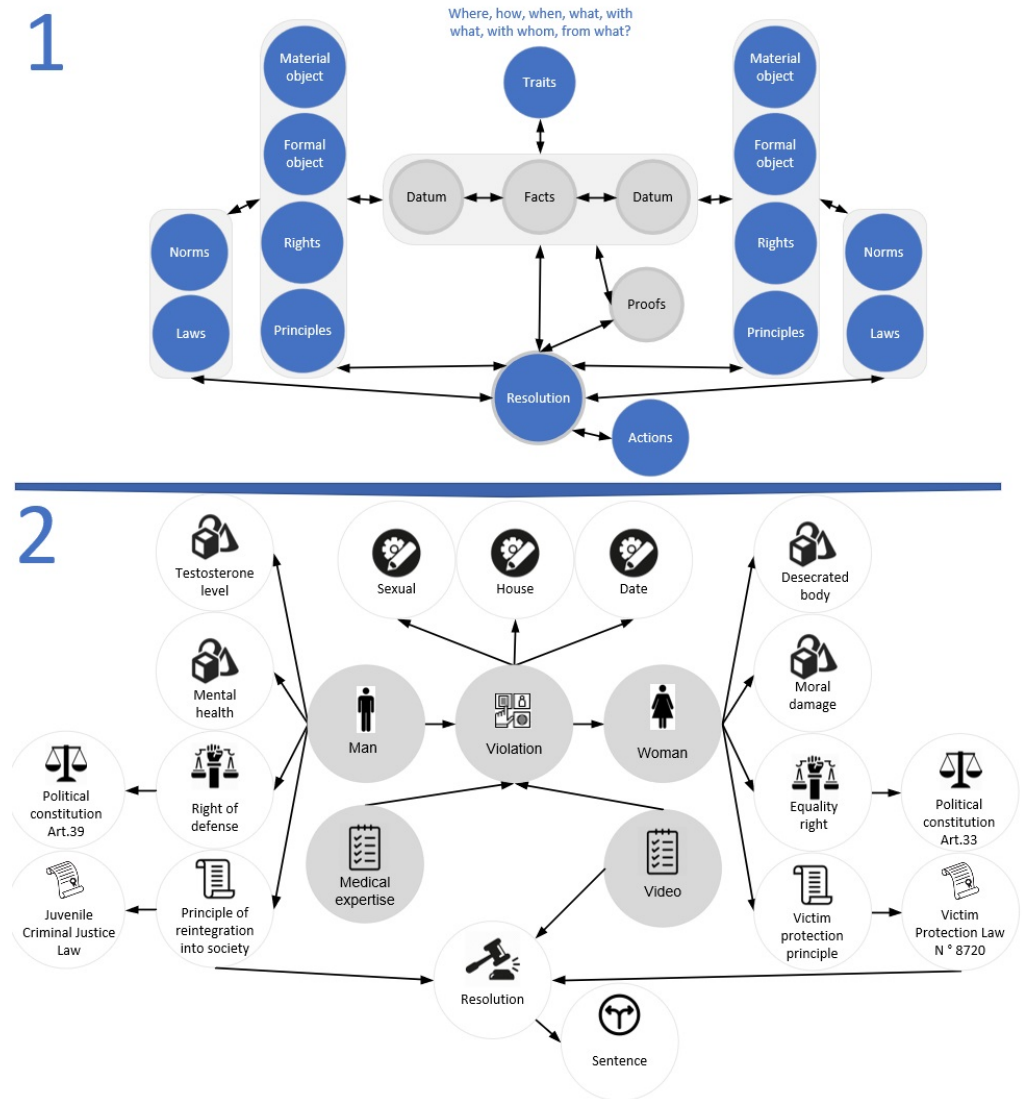


Figure 11. (1) Data model, and (2) property graph example, labels are omitted from relationships for simplicity.

Segment 2 of Figure 11 shows a PG where there is a removal of properties and labels in the relationships in order to simplify the example. This PG represents a judge analyzing a man from the perspective of the state of mental health that could lead him to rape a woman because of medical issues which are related to the testosterone levels in his body, and the woman from the perspective of the moral damage suffered by desecrating her body. The rest of the graph describes tests, norms, laws, rights, resolutions, and decisions related to the rape felony following the model of segment 1.

2.7. Explainable Technique

RYEL uses the explanatory technique called Fragmented Reasoning (FR) [2]. This technique uses dynamic statistical graphics that granulate the information following a hierarchical order and importance of the information according to the interpretation and assessment made by a human of real-world objects. This technique means that the semantic

and holistic constitution of objects, attributes, and vectors describing relationships between objects in a KG, as in Figure 7, are fragmented and link up to each other to explain the human conception according to its perception in a specific domain of discourse. Therefore, this technique expresses the hermeneutical content of a case from the perspective of a human and allowed the study of a new spectrum of cognitive information treatment [54] in the field of machine learning, associated with the human factor [90,91] specifically about the subjective information [92] of a person, which in this case is specific to the judge.

In Figure 12, the percentage of participation represented by the Y-axis is used to explain the level at which the concepts or objects of the current case are within the factual picture of other cases. The X-axis is used to explain the level of similarity that the concepts have between the cases. The size of the circles represents the dimension or level of importance of the scenarios within the files. The machine recommends the group of files distributed and located higher or more to the right of the graph. The machine handles each fragment as a collection of nodes to describe the interpretation of juridical objects and the assessment of a juridical concept. In this way, it has been possible to investigate the legal explanations related to the inferences [93] obtained by the system.

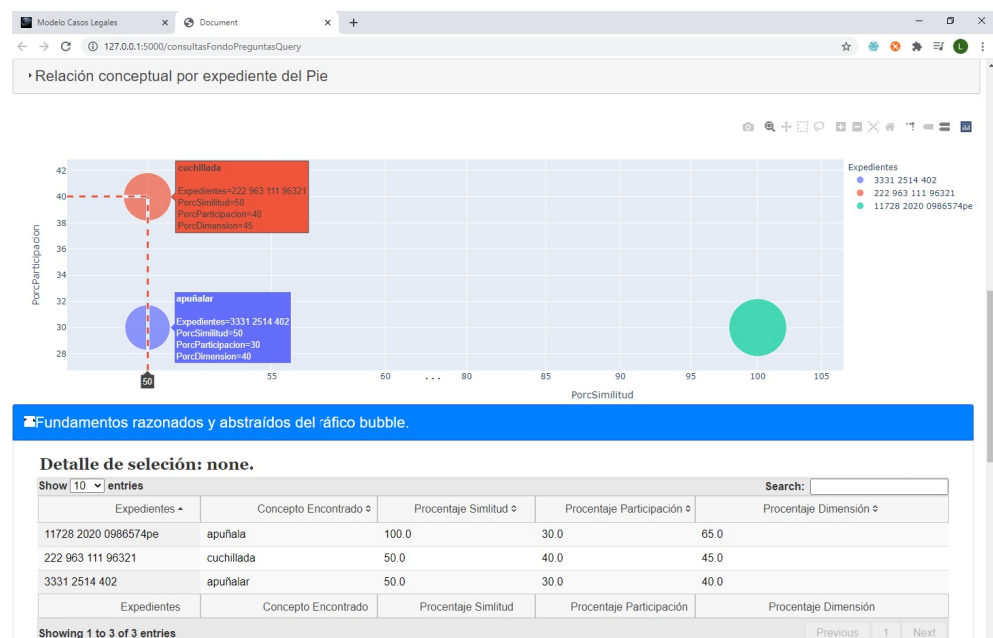


Figure 12. EGI explains employing circles, size, and colors to the user the set of legal files containing the factual picture and similar scenarios to the legal context with which the judge works. The machine explains that each circle is a set of legal files with characteristics and states. The machine recommends those that are higher and to the right of the graph. The judge can explore, analyze, and select other circles that consider the best for a specific legal context.

Internally the FR technique works using a strategic arrangement of data for each observation made by a human. FR uses the IA-AI method to get information as an input and reveals how it was interpreted and evaluated by a person. Figure 12 shows examples of some calculations and graphical view when the system provides recommendations. A fragment is a set of cognitive information pieces represented by geometric figures, colors, sizes, positions, and distribution of data elicited from EGIs using IA-AI and KG. The system uses the fragments to manage and organize the set of objects and to be able to explain them.

A fragment ω is represented by a collection of elements and the judge’s assessments. A fragment is an approximation of a set of nodes about a legal context p where a set of juridical concepts κ is in union with a set of nodes β joined with their relationships γ . The variables β and γ are decorating [94] the juridical concept κ . In this case, the decoration refers to the design pattern used programmatically (coding) to define a collection of objects that are capable of expressing the behavior of an individual object κ dynamically,

but without affecting the behavior of other objects of the same type in the same context; the programming paradigm used is Object-Orientation (OOP) to handle nodes, relationships, attributes, and properties. From the interpretation patterns, the construction of predicates occurs; imperative programming is used directly to manage the objects, and declarative programming is used indirectly to manage the assertions of the objects using CYPHER scripts. The set of objects contained in the fragments participate in (1) The Jaccard, Pearson, and Cosine formulas to work with the interpretation patterns, and (2) to organize and construct vectors from said patterns to make an inference.

3. Machine Specifications

Table A1 explains the main components, technology, formula, and concepts of the system in approximate order of operation. We will call each component with a “C” attached to a number, for example, “Component 3 = C3”. Table 1 synthesizes and distills operations and essential functions to work with higher-order thinking and handle KGs in the system based on Table A1. For now, the focus will be on C2, which provides the data structures that represent a KG (case) in the form of an ordered triple as shown in Figure A1. Equation (1) shows the formal representation of the ordered triple, from which the extraction of elements such as concepts, nodes, and relationships is possible. The output of extraction is a list of values that represents the input for the vector construction algorithm shown in Algorithm 1. Inference generation uses vectors between scenarios. This section explains: (1) The ordered triple, (2) formulas and vector construction, and (3) a simplified real case scenario example using the formal representation of a case.

Table 1. Synthesis of how the components of the artifact are used and operated, based on Table A1.

KG Operating Point	# Components Involved
1—create and manipulate	C1, C2, C5, C6
2—seek	C1, C2, C3, C4, C7
3—modify	C1, C2, C6
4—infer	C3, C4, C7, C8, C9, C10

3.1. Data Structure

To explain each element in Figure A1, consider this: Given graph G represents an ordered pair in the form of $G = (N, E)$, where N are the nodes and E are the edges, the artifact handles this:

1. Each node or vertex (image) is an ordered pair in the form of $N = (n, to)$ where n is the label of the node and to is an ordered triple in the form of $to = (in, ic, A)$ where in is the index of the node, ic is the index of the legal concept to which the node belongs, and $A = \{x : x \text{ is an attribute of the node}\}$. The x attributes of the node are text fields, for example, name and description of the node, and numeric values about precedence and importance levels that belong to the set of numbers \mathbb{Q} ;
2. Each edge or arc (relations between images) is an ordered pair in the form of $E = (e, to)$ where e is the relation label and to is an ordered triple in the form of $to = (ir, po, A)$ where ir is the index of the relation, po is an ordered pair in the form of $po = (ni, nf)$ where ni is the index of the start node and nf is the index of the final node, and $A = \{y : y \text{ is a relation attribute}\}$. The attributes of a relationship are text fields, for example, name and description of the relationship, as well as numerical values about the link and relevance of the relationship that belongs to the set of numbers \mathbb{Q} .

Algorithm 1: Creation of vectors and related concepts in KG

Input: ListValuesAttributesNodeConcept, ListAttributesRelConcept, ListConcepts, k
Output: ListRelParad, ListRelSinta, ListVecNode, ListVecRel

```

1 procedure VectorCreation(ListValuesAttributesNodeConcept, ListAttributesRelConcept, ListConcepts, k):
2   Quantity_K ← card|K|
3   if Quantity_K > 0 then
4     if K ∈ λ then
5       foreach con ∈ ListConcepts do
6         indexNode ← GetConceptIndex(con)
7         foreach atribN ∈ ListValuesAttributesNodeConcept do
8           levelInterests ← GetLevelInterestsNode(atribN)
9           order ← GetOrderNode(atribN)
10          descripNode ← GetsNodeDescription(atribN)
11          vectorNode ← C6(CreateVectorNode(levelInterests, order))
12          ListVecNode ← AddListaVectorNode(vectorNode)
13          foreach atribR ∈ ListAttributesRelConcept do
14            link ← ObtenerVinculoRel(atribR)
15            importance ← GetImportanceRel(atribR)
16            effect ← pythagorasTheorem(link, levelInterests)
17            descripRel ← GetsDescripRel(atribR)
18            vectorRel ← C6(CreateVectorRelation(link, importance, effect))
19            ListVecRel ← AddListvectorRel(vectorRel)
20          RelConcPara ← C3(GetRelParadigmaticas(indexNode, descripNode))
21          ListRelParad ← AddListRelPara(RelConcPara)
22          RelConcSint ← C3(GetRelSyntagmaticas(indexNode, descripNode))
23          ListRelSinta ← AddListRelSint(RelConcSint)
24        return ListRelParad, ListRelSinta, ListVecNode, ListVecRel
25      else
26        There are not enough elements or concepts, or do not belong to the legal element.
27    else
28      Not enough items.

```

3.2. Case, Context, and Scenarios

From N and E , a case C is a ordered triple as shown in Equation (1) where:

1. $p \in \mathbb{N}$ and is an index that identifies the context of the case assigned by the artifact;
2. V means the case scenarios in the form of $V = \{\lambda : \lambda \text{ is a legal element}\}$ where $n|V| > 0$;
3. R represents the relationships in the scenarios of a case in a given context in the form of $R = \{r : r \text{ is a relationship type } E\}$.

Given the above, Equation (1) shows a case with a set of relations R for a set of nodes that constitute the legal elements λ and describe the V scenarios of the factual picture that occurs in a p context. The relations and nodes were created from the transformation of interrelated images using the graphical interfaces of the artifact, and an index is an internal number that the machine assigns to the description of the context given by the judge, for example, "Simple Homicide = 999999 = p ":

$$C = (p, V, R). \quad (1)$$

3.3. Legal Elements

A legal element is an ordered triple as shown in the Equation (2) where:

1. $i \in \mathbb{N}$ and is an index that identifies a particular legal element assigned by the artifact;
2. K represents a concept in the form of $K = \{z : z \equiv P \vee z \equiv H\}$ where:
 - (a) $P = \{p : p \text{ is a proof of the kind } N\}$,
 - (b) $H = \{h : h \text{ is a fact of the kind } N\}$;
3. $T = \{t : t \text{ is a relationship of the kind } E\}$.

Equation (2) means that in a legal element λ there are relations T for a set of nodes that form the concepts K formed by facts or evidence, and an index i identify them. The artifact assigns the index to each set of nodes to identify that set:

$$\lambda = (i, K, T). \quad (2)$$

From the formal representations of the case explained previously, it is possible to supply a real, simple, and reduced example of an interpretation pattern. Consider Listing 1, this script seeks for patterns about nodes connected to the act of raping (Violation) someone under 18 years old. The pattern can be modified to look for children, older people, or undefined sex according to the rules of gender ideology. Modifications can be made to the script to apply deductive logic by taking a general aspect of a fact, evidence, or person and looking for a particular attribute pattern to canalize some legal study.

Listing 2 seeks particular attributes of people; in this case, it is a man connected with a woman, regardless of age or other characteristics, but considers the names. This script traces connection patterns up to 15 deep layers between these two people, and at the same time, extracts the shortest links between them. Deep layers mean the depth of connections between one object and another. Therefore, using this script can determine objects or events that are intermediaries between people to understand their criminal nexus.

Listing 1. Simplified example of code about interpretation patterns related to the act of rape using CYPHER.

```

1. MATCH (n) WHERE n:Man or y:Woman
2. OPTIONAL MATCH (n)-[r]-(v:Violation)-[type]-(s:Sexual)-(y)
3. WHERE EXISTS(n.age) < 18
4. RETURN n, r limit 100

```

Listing 2. Simplified example of existing patterns between a node type Man and another type Woman using CYPHER.

```

1. MATCH (hombre:Jackie name: 'Jack Smith' ),
2. (mujer:Al name: 'Alice Kooper' ),
3. p = shortestPath((Man)-[*..15]-(Woman))
4. RETURN p

```

There are 3 ways to avoid ambiguities which are: (1) By using a specific context, (2) searching for a particular pattern, and (3) using vector similarity. Let us consider the following examples about patterns: (1) In contrast with Listing 1, the pattern $(n)-[r]-(v:Violation)-[type]-(s:Agreement)-(y)$ seeks for nodes and relations connected with the violation of an “agreement” rather than a violation in “sexual” terms, and (2) if we compare the $(John)-[under_TheEffects_of]->(Drugs)-[in]->(Stabbing)$ pattern with the $(Alice)-[under_TheEffects_of]->(Drugs)-[in]->(bed)$ pattern, we obtain that there is no ambiguity, due to the intrinsic nature of both patterns. However, the use of the pattern $(person)-[under_TheEffects_of]->(Drugs)-[]->()$ would serve to look for other patterns in all the database knowledge, where a person is under the influence of the drug, regardless of gender, name, or any other characteristic. In the latest pattern, the Alice and John scenarios are collected, differentiated, and explained by the system using charts and geometric figures that explain their differences. From the interpretation patterns, it is possible to extract vectors from them to compare scenarios and generate inferences.

3.4. Vector Creation

C6, explained in Table A1, is responsible for constructing the vectors. The construction consists of 2 phases. In the first phase, attribute calculations of nodes and relationships take place. The attributes are effect (E), link ($V =$ adjacent side), and importance ($I =$ opposite side). The use of these attributes is through an adaptation of the Pythagorean formula in a Euclidean space; this formula generates values between two connected nodes by using a right triangle that is formed between their centers and the circumference of each one in a 3D plane, as shown in Figure 13. In the second phase, by using Algorithm 1 it is possible to obtain the vector modules.

Algorithm 1 receives as input a list of nodes and relationship attributes, as well as a list of legal concepts obtained from EGIs. The input lists of object attributes are processed to create output lists of vectors. The lists of vectors represent, in a unique way, the factual pictures of the scenarios in a case. C3 is responsible for applying Natural Language Processing (NLP) techniques like the paradigmatic and syntagmatic process to the input list of concepts to detect which ones accept a replacement and which ones can be combined, respectively. NLP techniques produce output lists of paradigmatic and syntagmatic relationships of concepts used as filters in searches for interpretation patterns. The output lists of vectors and concepts from Algorithm 1 are input parameters in Algorithm 2 which make inferences.

Algorithm 2: KG legal inference using Cosine, Jaccard, and Pearson functions

```

Input: ListRelParad, ListRelSinta, ListVecNode, ListVecRel, k
Output: RecomendList, G_Pie, G_Bars, G_Figures
1 procedure Inference(ListRelParad, ListRelSinta, ListVecNode, ListVecRel, k):
2   Quantity_K ← card|K|
3   if Quantity_K > 0 then
4     if K ∈ λ then
5       sizeLrp ← size(ListRelParad)
6       sizeLrs ← size(ListRelSinta)
7       if (sizeLrp > 0) ∨ (sizeLrs > 0) then
8         AnalysPattern ← C4(ListRelParad, ListRelSinta)
9         ListPatronNodesFound ← C7(SearchNodes(AnalysPattern))
10        ListPatternRelaFound ← C7(BuscarRela(AnalysPattern))
11        foreach atribNB ∈ ListPatronNodesFound do
12          levelInterestsE ← GetLevelInterestsNode(atribNB)
13          orderE ← GetOrderNode(atribNB)
14          vectorNodeE ← C6(CreateVectorNode(levelInterestsE, orderE))
15          ListVecNodeE ← AddListVectorNode(vectorNodeE)
16        foreach atribRB ∈ ListPatternRelaFound do
17          linkE ← GetRelink(atribRB)
18          importanceE ← GetImportanceRel(atribRB)
19          effectE ← GetLevel(atribRB)
20          vectorRelacionE ← C6(CreateVectorRelation(linkE, effectE))
21          ListVecRelE ← AddListVectorRelation(vectorRelacionE)
22        ListSimiCosNode ← C8(Cosine(ListVecNodeE, ListVecRel))
23        ListSimiCosRel ← C8(Cosine(ListVecRelE, ListVecRel))
24        ListSimiJacNode ← C8(Jaccard(ListVecNodeE, ListVecNode))
25        ListSimiJacRel ← C8(Jaccard(ListVecRelE, ListVecRel))
26        ListSimiPearNode ← C8(Pearson(ListVecNodeE, ListVecNode))
27        ListSimiPearRel ← C8(Pearson(ListVecRelE, ListVecRel))
28        ListValSimiNode ← C9(AnalysPattern, ListSimiCosNode, ListSimiJacNode, ListSimiPearNode)
29        ListValSimiRel ← C9(AnalysPattern, ListSimiCosNode, ListSimiJacNode, ListSimiPearNode)
30        ListValueVectorsExp ← C9(ListValSimiNode, ListValSimiRel)
31        ListRecommendationExp ← Recommend(ListValueVectorsExp)
32        G_Pie ← C9(ListRecommendationExp)
33        G_Bars ← C9(ListRecommendationExp)
34        ListValueVectorsNorm ← C10(AnalysPattern, ListValSimiNode, ListValSimiRel)
35        ListRecommendationNorm ← Recommend(ListValueVectorsNorm)
36        RecomendList ← Join(ListRecommendationExp, ListRecommendationNorm)
37        G_Figures ← C10(RecomendList)
38      else
39        Conceptual relationships could not be determined.
40    return RecomendList, G_Pie, G_Bars, G_Figures
41  else
42    There are not enough elements or concepts, or they do not belong to the legal element.
43  else
44    Not enough elements.

```

Part 1 from Figure 13 represents linked nodes in a scenario. Each sphere is a node. Between the nodes, there are sets of vectors \vec{X} , \vec{Y} , and \vec{Z} obtained from links, that is, the relationships of the nodes. The vector \vec{X} has the origin point in C' , which is the center of node A , and the endpoint is C'' which is the center of node B . The vector module $C'\vec{C}''$ constitutes the assessment of the links between facts and evidence using the relationships between the images, for example, the links shown in Figure 4. The vector \vec{Z} has the origin point in C' , which is the center of node A and the endpoint in C , which is the circumference of node B that represents the diameter of the node. The vector module $C'\vec{C}$

is built with CYPHER scripts to classify information about the importance levels obtained from interfaces like the one shown in Figure 5.

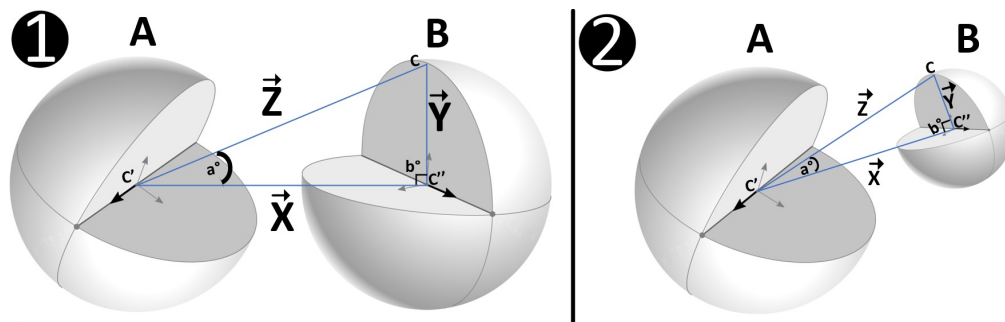


Figure 13. Relationships between nodes in a KG; showing a collection of vectors in n-dimensional Euclidean space. The spheres represent the nodes, and the vectors are letters formed from the relationships between them. 3D rendering explains the vector projection performed by the artifact.

Part 2 from Figure 13 represents multiple variations of nodes and relationships in a scenario and represents dynamic changes in the perception of objects in the real world. For example, consider node A as a fact and node B as evidence. Between these nodes, there is a small increase in the distance and makes vector \vec{X} have a longer link between the nodes. The “increase” of the module $C\vec{C}''$ means a “decrease” in the legal connection that node B has over node A. A longer link between the nodes reduces the ability of node B to be able to express the influence it has on node A. In other words, node B is not so capable of expressing the influence on node A. A decrease in the size of node B implies a decrease in the module $C\vec{C}''$ and means a reduction in the legal importance of node B within a scenario in which A also participates.

3.5. Jaccard Index

Jaccard is a statistical measure that consists of measuring the similarity between finite datasets, for example, between a set of objects D' and D'' . This is a division between the size of the intersection and the union of the element sets. In this case, vector modules provide a series of values to create the sets to be compared. This process provides values between 0 and 1; the first expresses inequality between vectors and the second total equality between them. This index is useful in queries to detect patterns of similar objects (nodes or relationships) within scenarios, for example, to obtain the granular similarity between sets of facts or evidence, or between attribute values that belong to different groups of scenarios, that is, to be able to obtain similarities between attributes belonging to the same type of nodes, but that belong to different scenarios:

3.6. Cosine Similarity

Cosine Similarity is a measure of similarity between two vectors, in this case, those that belong to a set of objects \vec{G}' and \vec{G}'' other than zero. It means that it calculates the angle between vectors to get the cosine by multiplying the values of each vector, adding their results, and then dividing the result by multiplying the square root of each value of the vector squared. A pair of vectors oriented at 90° to each other have a similarity of 0, meaning they are not equal, and a pair of diametrically opposite vectors have a similarity of -1 , meaning they are opposite. On the other hand, if both vectors point with an orientation towards the same place, they have a similarity of $+1$, meaning they are equal. The different values that the cosine angle acquires reflect a greater or lesser degree of similarity between the attributes of the relationships that the scenarios contain. This type of similarity is helpful in detecting assessment patterns, for example, to identify similarities between the angle produced between the link and the effect between a pair of nodes in the same scenario or indifferent ones. The angle a° produced by the assessment of the link (line

connecting nodes from their centers), and the effect (line from the center of one node to the diameter of the other), are shown in Figure 13.

3.7. Pearson's Correlation

The Pearson Correlation Coefficient is a statistical measure to detect a linear correlation between two variables A and B . It has a value between $+1$ and -1 . A value of $+1$ is a total positive linear correlation, 0 means there is no linear correlation, and -1 is a total negative linear correlation. The Pearson similarity is the covariance of the values from vector modules divided by multiplying the standard deviation of the values of the first vector by the standard deviation of the values of the second vector. This coefficient is helpful in queries about the correlation of values, for example, to calculate the correlation between the link and importance of two connected nodes in a scenario, for example, the values of a vector $\vec{V}1$ represents the attributes contained in the link of a node A connected with node B . A second vector $\vec{V}2$ represents the importance that node B has for node A . Thus, if A is a fact and B is proof, then the level of correlation between link and importance is the degree to which evidence can demonstrate that an event occurred, according to the interpretation of a judge.

3.8. Dataset Example

Figure 14 shows a composite scenario of a real murder case. The structure shown in Figure A1 explains a piece of a set of objects of this case and describes the dataset involved. The following points summarize the example and the data structure.

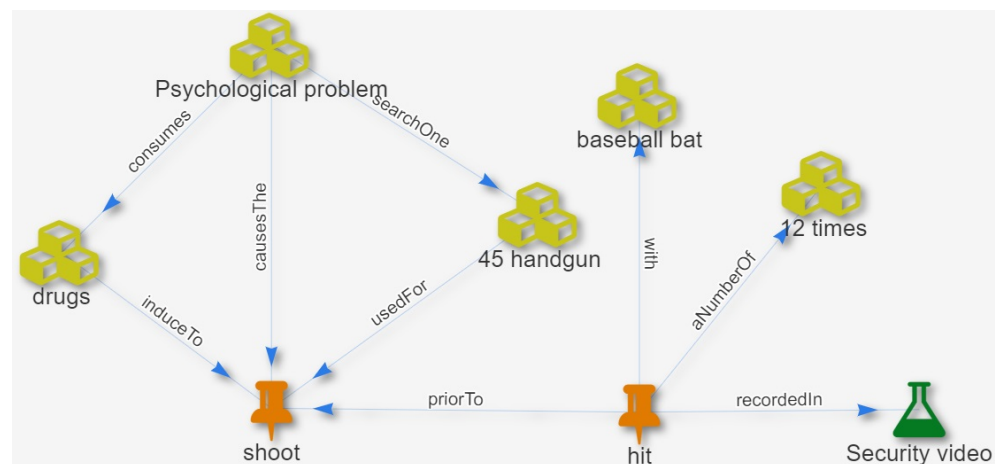


Figure 14. The graphic interface of images according to the interpretation and assessment made by a judge. A piece of a real scene in a murder case using a handgun.

1. The interrelated images of Figure 14 are taken to build the nodes according to the definition $N = (n, to)$, and relationships following the specification $E = (e, to)$. This produces the structures shown in Table A2;
2. Then labels, indices, and values of the nodes are obtained from step 1, as shown in Table A3;
3. From step 2, information about relationships, indices, and labels from the connection of each node, is shown in Table A4, and ;
4. Using the information from Table A2 about the descriptions, the artifact extracts the concepts "injure" and "disable" according to the representation $K = \{x : x \equiv P \vee x \equiv H\}$ and an index is assigned to them, for example 333 and 444 respectively;
5. Using Tables A2 and A3 and the concepts obtained from point 4, the artifact distills 2 legal elements that are shown with Equations (3) and (4) respectively. An index is assigned to each legal element, for example, 10,000 and 11,000, and this is done following the definition $\lambda = (i, K, R)$:

$$\lambda = (10000, \{333\}, \{500, 600, 700, 800\}) \quad (3)$$

$$\lambda = (11000, \{444\}, \{000, 100, 200, 300, 400\}); \quad (4)$$

6. The artifact assigns an index to the case, for example, “999999” and then it creates the case as shown in Equation (5) following the definition $C = (p, V, R)$ and according to the information obtained from the previous steps:

$$C = (999999, \{10000, 11000\}, \{000, 100, 200, 300, 400, 500, 600, 700, 800\}); \quad (5)$$

7. The artifact offers different queries to analyze the case. Depending on the query type, the Jaccard, Cosine, and Pearson formulas are executed individually or in combination. Obtaining differences is according to the type of analysis and query the judge wants to execute. The artifact shows the recommendations as in Figure 8. At the end of this figure, the judge can select the laws and norms supplied by the system. The system automatically converts the selections into images and incorporates them into the working canvas so the judge can continue, if necessary, with the analysis of more information.

4. Research Question and Hypothesis

The following research question arises: “Is it possible to capture and represent a judge’s interpretation and assessment processes of the legal file data and apply machine learning on said processes, to generate recommendations before the resolution of a case related to jurisprudence, doctrine, and norms in different legal contexts and get a positive behavioral response from the judge?”

Thus, a secondary question arises: “Can the system can be used by a judge to support his decisions, but without being seen as a threat of decision-making [95] bias?”

The above research questions are intimately linked to the unsolved problem, raised long ago by Berman and Hafner in 1993 [3] on “how to represent teleological structures in CBR?” Teleology is the philosophical doctrine of final causes [51], which means, according to Berman and Hafner, identifying the cause, purpose, or final reason for applying a law or rule to regulate (punish) an act (fact) identified as a felony. Thus, the answer to the first questions also provides a reasonably approximate answer to Berman and Hafner’s question.

As judges have hierarchies in their roles and there are types of technical criteria to study the behavioral response of a judge, the statement of the following hypotheses is as follows. (1) H_0 : The hierarchy does not affect the acceptance of the system and H_a : The hierarchy does affect the acceptance of the system, as well as (2) H_0 : The criterion does not affect the acceptance of the system, H_a : The criterion does affect the acceptance of the system.

5. Material and Methods

SME has defined real world legal situations to test cases with RYEL, which represent criminal conflicts in a trial and have allowed to reduce the number of cases that initially would have been necessary to carry out the experiments. The use of multi-country scenarios for laboratory testing was 83 from Costa Rica, 25 from Spain, and 5 from Argentina. In addition, experts in artificial intelligence participated from Costa Rica and Spain [2] and were counted, to be a total of 17. As the laws are different in all countries, a norms equivalence mapping was necessary to implement, which means a set of implication rules in the form $X1 \rightarrow Y2$, where $X1$ is the name of a norm in a specific country and is equivalent, but not equal, to $Y2$ which belongs to another country. In this way, there was no problem analyzing the same criminal factual picture (facts and evidence) in different countries without being strictly subject to the name of a norm.

5.1. Participants

Two groups of research subjects participated in this study. The first group of judges was selected at random, belonging to courts, tribunals, and chambers in criminal justice. In addition, military-grade judges were also included randomly at the magistracy level to include data about military behavior when using this technology. Experiments in Panama [26], Spain, and Argentina involved 16 expert judges in the criminal field, while in Costa Rica, there were 10 judges [2] which also include Ecuador and Colombia. The second group was a sample of judges selected randomly at the national level in Costa Rica.

5.2. Design

This study is an adaptation of a 3-stage experiment. The first stage is to study the acceptance or denial behavior of the judge when using the system. The second stage compares the results obtained from the first stage with the second group of judges. The third stage consists of investigating whether the responses of the second group were affected by factors such as judges' hierarchies (their roles) and the kind of evaluation criteria. The results of one stage are the inputs of the next.

In the first stage, the use of User Experience (UX) [96] is a means to investigate the behavioral response of a judge in terms of accepting or rejecting the application of RYEL to analyze the merits of a case. Table 1 shows a synthesis of the primary operations that were used by the research subjects when manipulating KG using images. The fundamentals of measurement parameters are from the quality model called Software Quality Requirements and Evaluation (SQuaRE), defined in [97]. The characteristics of this model are adapted to investigate the degree to which a system satisfies the "stated" and "implied needs" of a human (stakeholders) and is used to measure the judge's behavioral response. The characteristics used from the model are "functional suitability", "usability", and "efficiency" linked to technical criteria issued by the judge. Table 2 shows a synthesis of the characteristics, parameters, and criteria considered in the experiment.

Table 2. Software evaluation characteristics defined in ISO-25010 [97] to study the behavioral response of the judge.

¹ Characteristic	Parameter	Criteria
Functional	Suitability	(1) It allows capturing the interpretation and assessment of the judge. (2) Graphically represents the legal knowledge that a judge has about a case.
	Accuracy	(1) The system is capable of analyzing the factual picture and returning the correct legal norms.
	Functionality compliance	(1) Judicial independence and discretionary level are respected.
Usability	Understandability	(1) Suitable for case data manipulation. (2) Graphic interfaces describe the legal analysis made by humans.
	Learnability	(1) Easy to learn.
	Operability	(1) Easy to operate and control.
	Attractiveness	(1) Attractive and innovative graphical interfaces.
Efficiency	Time behaviour	(1) System response time is acceptable.
	Efficiency compliance	(1) Flexible to capture different types of legal data. (2) It is possible to represent characteristics of facts and evidence. (3) Allows a flexible analysis of the merits of a case.

¹ Adaptation and use of software evaluation characteristics defined in ISO-25010 [97].

A quality matrix [98] or evaluation matrix was created using Table 2 and applied to the judges at the end of the first stage. The matrix allowed to obtain quantitative values for each of the criteria. The criteria were posed as questions and measured with a Likert Scale [99] as 5–Totally agree, 4–Fairly agree, 3–Neither agree nor disagree, 2–Fairly disagree,

1–Totally disagree, and 0–Not started. A treatment is a legal case of homicide applied to each research subject (judge) using RYEL. The experimental unit consists of pairs of related nodes that form a KG that describes the case graphically.

The second stage consists of obtaining objective evidence [97] to validate the results of the matrix against the criteria of another group of judges. For this, obtaining an additional random sample of 172 judges from Costa Rica was necessary to take. The total population of judges working in Costa Rica is 1390 [100]. The sample includes all hierarchies of judges and represents 12.37% of active judges in the country. To this sample, a questionnaire was applied based on the criteria from Table 2. This sample focused on judges that do not necessarily know each other; they have not used or have seen the system before, and they do not know or have met the investigators conducting the research. In this way, it is possible to reduce the information bias [101] in this type of research. The judges received information on the system's method, operation, and characteristics through the questionnaires' descriptions and formulation. The criteria in the questionnaire were organized into groups of 10 questions and coded from 1-P to 10-P, as shown in Table 3, for statistical purposes. The design of the questions considered the Likert scale for their answers. This design was similar to the one used in the evaluation matrix explained before. It was necessary to coordinate with the Superior Council of the Judiciary in Costa Rica to contact the judges across the country.

The third stage uses the information gathered in the sample at the national level in Costa Rica to make a Two-way Analysis of Variance (ANOVA) [101]. This analysis was to check if there are significant statistical differences that prove the hypotheses about whether the factors like hierarchies and legal criteria affect the behavioral response of acceptance or denial of the judges about using the system. The criteria have 10 levels, one per group of questions, from 1-P to 10-P. The hierarchy has 4 levels which are: (1) Criminal courts; (2) tribunals; (3) chambers; and (4) other. The latter consider members of the superior council and interim positions of judges during designations.

Table 3. Responses statistical summary.

Coded	Mean	SE Mean	St Dev	Variance	Coef Var	Median	Mode
1-P	4.6744	0.0473	0.6202	0.3846	3.27	5	5
2-P	2.3663	0.0876	1.1494	1.3212	48.58	2	1
3-P	3.064	0.0993	1.3029	1.6976	42.52	3	3
4-P	4.814	0.0378	0.4959	0.2459	10.3	5	5
5-P	3.3779	0.0941	1.2341	1.523	36.53	4	4
6-P	3.3895	0.0938	1.2305	1.514	36.3	3.5	3
7-P	3.7733	0.0922	1.2095	1.4629	32.05	4	5
8-P	3.8488	0.0865	1.1344	1.287	29.47	4	4
9-P	3.6512	0.0939	1.2309	1.515	33.71	4	4
10-P	3.75	0.0976	1.2802	1.6389	34.14	4	5

5.3. Setting

Due to the circumstances caused by COVID-19 and in which the judges found themselves, the experiments were conducted either onsite (judge's office) or remotely (virtual meeting via a shared desktop). In both cases, a Dell G5 laptop was the hardware used for experimentation. The laptop had 15.6" of full HD IPS display, 16GB RAM, and a Hard Drive of 1GB. After a legal and coordinated appointment with the judges and setting up the test environment, it was possible to proceed with the experiments.

5.4. Procedure

In the beginning, each research subject watched a video. The video explained the experiment, the operation of the system, and the function of the EGIs. The video was 2.6 min long. It used a test case *A* about a homicide using a dagger and another test case *B* about a homicide with a weapon. Various experts helped the design process of the test cases, 2 in law and 2 in artificial intelligence, who verified them.

In the first stage, $N = 26$ research subjects from Colombia, Ecuador, Panama, Spain, Argentina, and Costa Rica were obtained and asked to draw in the system the interpretation and assessment of the facts and evidence contained in the test case *A* according to their perspective using interrelated images. At the end of the drawing, each subject produced a KG. The KGs produced were compared to each other to determine differences. Then, a division of the group of subjects N into two groups in the form of $N/2$ each took place. Test case *B* was given to the first group to obtain a new KG from each member. The second group, who never saw case *B*, was asked to observe and explain at least 3 KGs made by the first group to check if they were able to understand the interpretation and assessment contained in the KGs. Finally, the two groups ran legal analysis simulations with the system to determine whether it was possible to study the merits of the case. After cases *A* and *B* were used to explain the system, each member of the groups was allowed to use the artifact to enter new cases or vary the previous ones to test the system in depth. Then, the evaluation matrix was applied to each research subject to collect the UX that each one lived after using the system. Real life examples of the experiments with the research subjects are shown in Figure 15 when they were using the system to analyze the merits of a case about homicide.

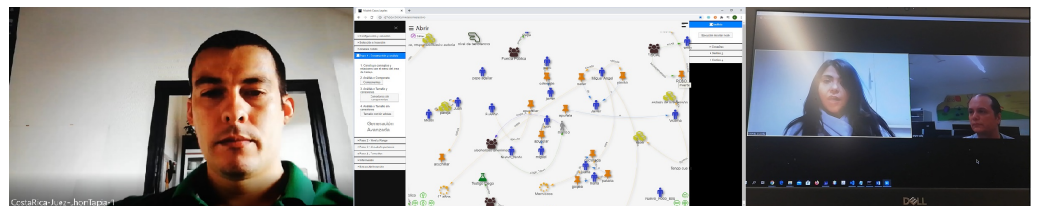


Figure 15. Real live experiment samples using the RYEL system by judges from Costa Rica and Argentina, respectively.

In the second stage, it was necessary to request a legal license from the Superior Council of the Judiciary in Costa Rica in order to be able to contact all the judges of Costa Rica and to send them a questionnaire. The criteria of Table 2 allowed us to build the questionnaire containing 10 groups of questions.

In the third stage, 172 sample of judges responses were taken from the questionnaires sent. The data of the responses were processed and tabulated. Finally, a two-way ANOVA was applied to the data to determine if the hierarchies or criteria affect the behavioral response of the judge; if they accept or refute using the system to analyze the merits of a case.

6. Results and Discussion

The judge's behavioral response was a tendency to accept the system, recognizing that it can help with the analysis of the merits of a case without violating judicial independence and discretionary level. Table 4 shows an extract of the evaluation matrix by country. Out of six countries, five showed a behavioral tendency of 90%, or more, to accept the system, reaching almost 100% in some cases. Colombia and Ecuador presented different results that are very close to 90% acceptance because some of the legal cases used for experimentation did not contain the names of regulations from those countries, and the judges belonging to them wanted to evaluate the names related to their legislation. Despite this, the acceptance of the computational method implemented by RYEL was very positive in all the countries subject to experimentation. The explanation of the system required the use of two cases,

but each judge entered from five to six real life cases when allowed to test the system. If 26 judges tested the system, it means that at least 130 case variations were used in total. In addition, the 113 cases used to manufacture the system from different countries must also be added. The total number of cases was approximately 243 from various countries used to create and test the system. It is necessary to remember that the SME supplied representative cases of the discourse domain; therefore, the high amounts of data did not present an obstacle and did not determine the risk of bias that would typically occur with another approach.

Table 4. Synthesis of the evaluation matrix according to the judge’s criteria.

¹ Characteristic	Colombia	Ecuador	Panama	Spain	Argentina	Costa Rica
Suitability	100.00	95.00	95.71	98.00	100.00	92.00
Accuracy	80.00	80.00	85.71	96.00	90.00	93.00
Functionality compliance	80.00	80.00	91.43	96.00	100.00	94.00
Understandability	90.00	95.00	94.29	96.00	100.00	99.00
Learnability	80.00	70.00	82.86	92.00	100.00	88.00
Operability	90.00	80.00	92.86	92.00	100.00	94.00
Attractiveness	100.00	100.00	100.00	96.00	100.00	98.00
Time behaviour	100.00	90.00	97.14	92.00	100.00	100.00
Efficiency compliance	80.00	86.67	91.43	94.67	100.00	94.00
Average	88.89	86.30	92.38	94.74	98.89	94.67

¹ Adaptation and use of software evaluation characteristics that are defined in ISO-25010 [97].

The radar graph in Figure 16a shows the comparison of the system evaluation results according to hierarchies. The characteristics described in each vertex reveal that the distances between criminal courts, military criminal courts, criminal magistrates, and superior courts are very close to each other and with high values. The average acceptability per hierarchy on the radar places values very close to 100% of acceptance. The provincial courts had a slightly lower acceptance rate. The reason was that some judges were unable to complete the experiment as they had to attend trials, and it was not possible to reschedule the experiment, and it reflects in the usability and efficiency vertices whose values are below average. Nevertheless, the vertex of the functionality in the provincial courts has values very close to 90%, which means that this hierarchy accepts the system well, despite the other low values.

Figure 16b shows the acceptability trend of the system among the judges, according to the hierarchical order. This trend remained unknown under the ordinary conditions of legal review processes, but detection was possible during the system’s evaluation. For example, it was possible to find that when the higher-ranking judges needed to review the work done by the lower ones, it was easy for them to graphically arrange the teleological structures of the facts and evidence using KG through the EGIs to carry out the reviews of the analysis made by the lower-ranking judges. Furthermore, it was possible to reveal that lower-hierarchy judges tended to accept the system in terms of the support they received from the EGIs to perform the interpretation and assessment of facts and evidence as part of the analysis of the merits of the case. On the other hand, the higher-ranking judges showed more acceptance of the system, especially regarding the support they received from the EGIs to access the teleological and semantic approach created by the lower-hierarchy judges.

The information collected up to this point responds to the first research question, and it reveals that the system was able to capture the interpretation and assessment of facts and evidence from the perspective of a judge. Regarding the second question, the results reveal that the judge’s behavioral response was very positive and with a tendency to accept the system to analyze a case without representing a risk of bias or a threat to decision making.

The validation of the previous results was against other evaluations of judges; this evaluation required the application of questionnaires to all the active judges of Costa

Rica (1390), and the random sample of 172 showed a mean of 4, a mode of 5 (Likert scale designed), and a standard deviation of 1.2988. These data are in Table 3. It means that the judges tend to accept the approach, operation, and framework implemented by RYEL.

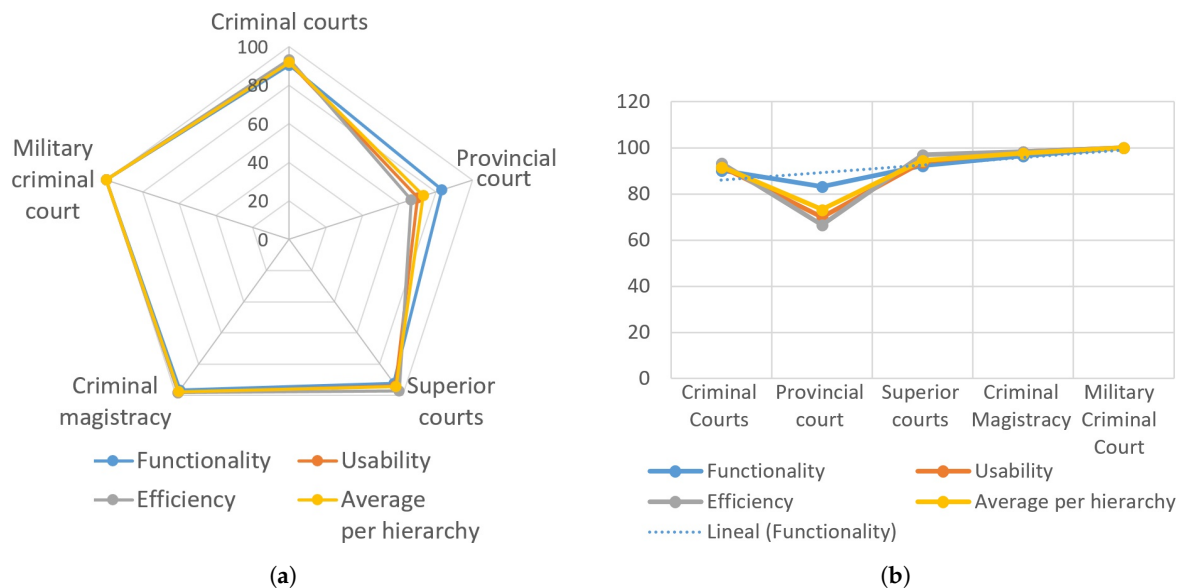


Figure 16. Characteristics evaluation score and acceptability trend. (a) Scoring radar on system by hierarchy. (b) RYEL acceptability tendency according to the hierarchy of the judges.

For the statistical verification of the sample and the collected results at the national level, Figure 17a shows the 1-Sample Z test, which had 92% of statistical power, a significant percentage for samples and experiments [101]. The statistical significance level is $\alpha = 0.06$, from which we can obtain 94% in the confidence intervals in the statistical tests. Figure 17b shows that the judges’ responses comply with the normality assumption, having a $P\text{-Value} = 0.213 > \alpha 0.06$ where the normality hypothesis is accepted. There is a low Anderson–Darling (AD) statistic value of 0.487 which means a good fit for the data distribution. The Levene statistic is $0.800 > \alpha = 0.06$, which means the hypothesis acceptance about equality of variances when working with the hierarchy and legal criteria of the judges in the answers of the questionnaires.

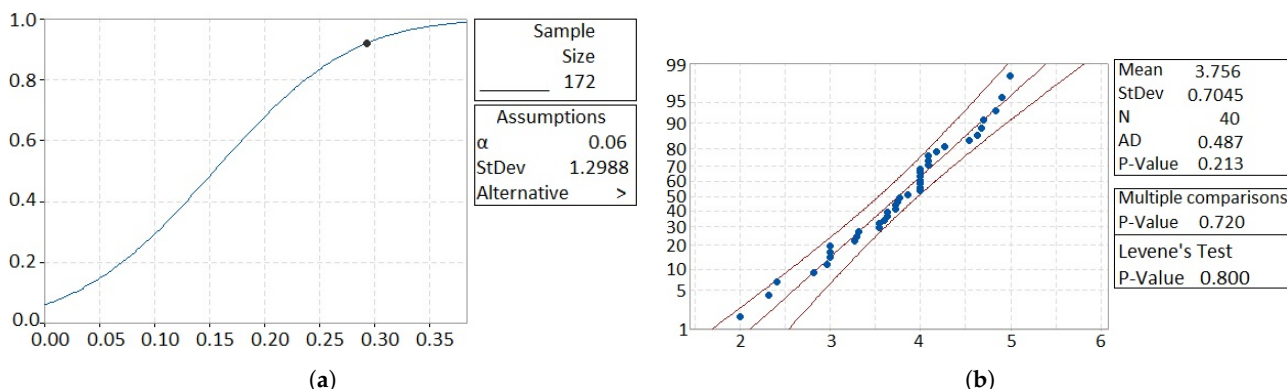


Figure 17. Statistical significance, power, and normality of samples and results. (a) Power curve for 1-Sample Z Test with $\alpha = 0.06$. (b) Normal probability plot of answers with 94% confidence interval.

Due to the results obtained in the previous statistical analysis, connected with the need to determine if indeed the results obtained from the UX and the questionnaires were affected by the hierarchical trend shown in Figure 16b or the criteria, a tTwo-way ANOVA was necessary to apply. The results are in Table 5 where the hierarchy factor has a $P\text{-Value} = 0.148 > \alpha = 0.06$, which means that the null hypothesis that the hierarchy does

not affect the response of the judge is accepted since there is sufficient statistical evidence to state with 94% confidence that the judge's responses are not affected by the hierarchy. On the other hand, the criterion factor at the same table shows a $P\text{-Value} = 0.000 < \alpha 0.06$ and means that the null hypothesis that the criterion does not affect the response of the judge is rejected since there is significant statistical evidence with 94% confidence that the criteria do influence the judges' response. These results reveal that the judge's behavioral tendency to accept the system is due to the criteria discussed and analyzed, not because of a human's position. It also means that the trend found in Figure 16b has a 94% statistical probability that it is due to the actual operation of the system and not to the position that the judge holds.

Table 5. Two-way ANOVA: Criteria and hierarchy.

Source	DF	Seq SS	Contribution	Adj SS	Adj MS	F-Value	P-Value
Criteria	9	12.901	66.66%	12.901	1.4334	7.29	0.000
Hierarchy	3	1.142	5.90%	1.142	0.3805	1.93	0.148
Error	27	5.312	27.45%	5.312	0.1967		
Total	39	19.354	100.00%				

Concerning the above, Figure 18 shows the residuals from the analysis of variance. The Y - axis represents the residual values, and the X - axis represents the order of the observations. There are no patterns nor a fixed trend. Therefore, the data obtained from the responses on the criteria are independent, and this means that there is no codependency in the data that could affect the results.

Figure 19 shows the main effects in responses to legal criteria. The Y - axis is the mean of the criteria; the X - axis represents the criteria. Thus, criterion 2 or 2-P has the lowest main effect of all because this group of questions referred to whether a legal case must always be resolved similarly to a previous case, with more or less similar characteristics. It means that statistically, there is enough evidence to affirm with 94% confidence that the judges reject the idea of receiving help that implies always solving a case just as another similar one was solved. The 2-P group of criteria in Figure 19 compared with Table 3 which has a mean of 2 and a mode of 1 for the same criterion, indicates that indeed the judges do not approve the 2-P criterion. It is necessary to remember that RYEL uses the CBR stages to exchange and organize data; this means, as a guide of the information, and does not develop the traditional implementation of using strictly the same solutions from past cases to solve current ones. This implementation makes RYEL's contribution to the domain of discourse evident.

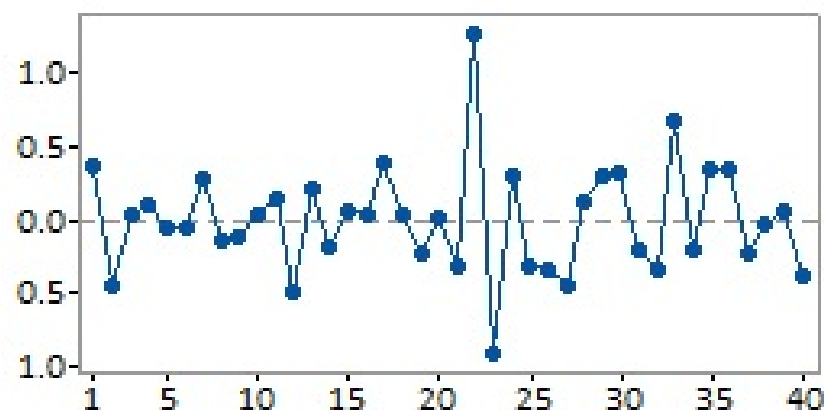


Figure 18. Residuals vs. observation order.

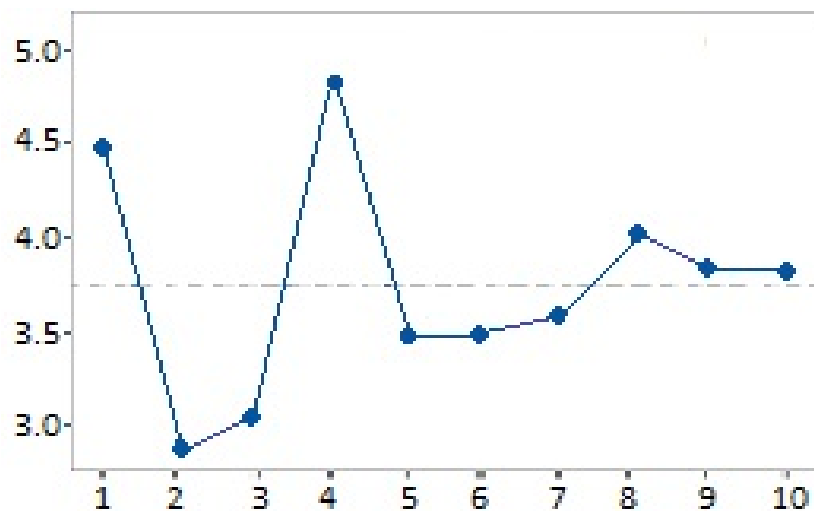


Figure 19. Main effects plot for judges' evaluation.

Figure 19 shows the criterion 3 or 3-P, which is the second one to have low values and refers to whether the judges believe that IA could help them with the analysis of a case. This point deserves special attention because analyzing the extended answers made by the judges in this group of questions makes it possible to understand that the judges associate AI with the automation and repetition of legal solutions applied indiscriminately to each case, without receiving any explanation and without being in control of the machine. This situation, of course, is not the way of work of RYEL.

Figure 20 shows the cumulative acceptance percentages grouped by the SQuaRE-based design parameters. The lowest percentage is learning because not all people have the same abilities to learn. The highest values are attractiveness, understandability, and suitability, which translates into a motivational design consistent with the needs of the research subject and the legal domain. On average, the rest of the cumulative acceptance percentages of the system are pretty high; this means that the domain expert quantifies, according to the SQuaRE parameters, that the system is helpful in analyzing the merits of the case.

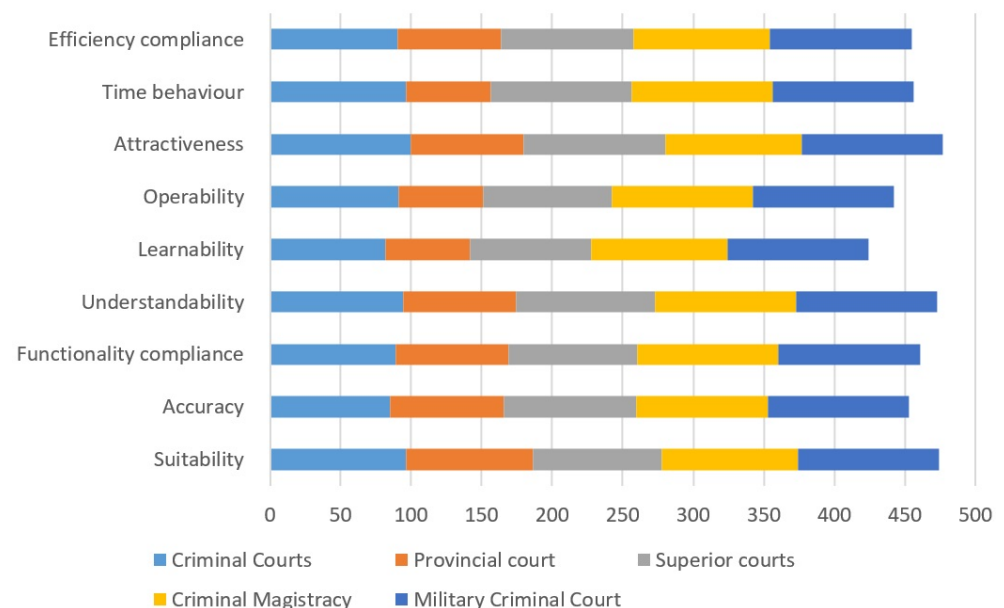


Figure 20. Grouping by design parameters SQuaRE.

Table 6 shows the cross-check of the survey-based statistic analysis between the 4-P and 10-P criteria. The first criterion is that if, in order to analyze the merits of a case, it is

essential to carry out an interpretation and assessment of facts and evidence. The second criterion is whether RYEL is novel and useful as a decision support tool. The results confirm the following: (a) No research subject marked option 1 for any of these criteria, (b) only four research subjects marked the options 2 and 3 for both criteria, which is only the 2.32% of the research subjects and it means that an insignificant number of them do not agree with the legal analysis approach and with the tool, and (c) most of the research subjects when marking options 5 and 4 for criterion 10-P also marked 4 and 5 options for criterion 4-P, which means that most of the research subjects understand and accept the legal analysis approach and the operation of RYEL.

Table 6. Cross-check between the 4-P and 10-P criteria.

² Criterion 10-P	¹ Criterion 4-P				Gran Total
	2	3	4	5	
1		1	1	16	18
2			3	6	9
3	1	1	5	25	32
4			10	42	52
5	1		3	57	61
Grand total	2	2	22	146	172

^{1,2} 5–Totally agree, 4–Fairly agree, 3–Neither agree nor disagree, 2–Fairly disagree, and 1–Totally disagree.

All the results obtained show the following:

1. The system was able to capture the high-order thinking of a judge to assist with analyzing a case using KG through images;
2. The system is a novel implementation of machine learning in the legal domain;
3. It was possible to explore and find shortcomings in the behavioral response and position of a judge in the face of this type of technology.

6.1. Comparison with Similar Approaches

By comparing our research with works with similar approaches, we extend the results. Attention is on expert and case-based systems.

Table A6 shows the 23 most essential expert systems from 1987 to the present, which are related to our research. The table shows the key elements, computational technique, and approach. The most important results obtained when comparing our system with the systems in this table are: (1) No system works with dynamic KG, (2) they do not use graphical techniques to elicit legal meta-knowledge of a person, (3) they do not work with high-order thinking, (4) do not allow an analysis of the merits of a case, (5) they do not focus on the judge, and (6) cannot be extended to other domains of knowledge.

Table A5 shows the primary investigations focused on CBR from 1986 to the present and related to our approach. This table shows the key elements, case types, and approaches. The main results obtained when comparing these investigations with ours are: (1) They do not contemplate multiple and complex scenarios within the cases, (2) no investigation considers data processing from the perspective of a human, (3) they are only focused on lawyers or prosecutors, not to judges, and (4) none of them processes teleological, semantic, ontological, and hermeneutical information to support decision making.

6.2. Functional Limitations

The system works with data from a factual picture, direction of the legal process, and assessment of the evidence. Information about the criteria of the judge that are not typical of the analysis of facts and evidence, for example, the criteria a judge may have on the management and administration of an office, control of dates to avoid document delays, and office procedures, are not considered in this research. However, a judge can

indeed consider that a case has been prescribed and request the archive of the documents. The type of data about this request is not part of the system.

6.3. Applicability

There are two fundamental aspects of a resolution that are “form” and “substance”. The form is the way to present and write a resolution complying with the requirements and formalities that the law requires, for example, a heading, covers, and numbers of pages. The substance refers to the in-depth study of the matter in conflict and then issues a resolution based on substantive law, which means a set of obligations and rights imposed by law. The applicability of this work refers to the substance of the case and not in the form.

6.4. Implications

The above results have particular implications in both the computational and legal domains. RYEL could mark a before and after in systems with a legal approach because it allows an evolution from predictive systems to systems with explanatory and analytical techniques. Some of the most relevant implications from this in the computational field are:

1. Due to explicability techniques, “Black Boxes” problems in machine learning could be overcome by methods like IA-AI when dealing with human perception;
2. The reduction and nullification of algorithmic bias and bias related to data and processes is gaining momentum because third parties do not manipulate the cases and analysis processes. Instead, the judge enters the cases in real life and commands the analysis with the options provided by the system; the latter explores the relationships and objects the judge creates, explains the inferences, and offers to the judge options to make decisions;
3. Judges from other hierarchies can review the sentences using RYEL in the different legal stages. This review could cause a reduction or elimination of bias related to a wrong perception, incorrect interpretation, and an erroneous assessment.

Some of the most relevant implications in the legal field are:

1. RYEL shows the potential to be a disruptive technology in the domain of discourse and could cause the user to resist the change;
2. The system allows experts to analyze the scenarios from different perspectives and reach agreements; this generates a unification of legal criteria and decreases legal uncertainty;
3. The system paves the way in the jurisdictional area by allowing a computational mechanism to participate in a judge’s exclusive functions when decision making takes place.

7. Conclusions

The use of the IA-AI method showed the ability to capture the high-order thinking of a judge. The behavioral response of the judges was quite positive in accepting the use of this technology to analyze the merits of a case. This research caused a paradigm shift in the way a judge thinks and works for two main reasons:

1. Legal files are always textual and therefore processed as text. Experimentation with the system was exclusively using interrelated images and the IA-AI method, making a big difference;
2. No judge who used the system and obtained a UX saw a threat of decision-making bias because the system did not impose solutions but instead allowed the judge to dissect a case and then analyze how other judges had perceived the facts and evidence to formulate their conclusions criteria. Moreover, the system operates without breaking the rules about “degree of discretion” and “judicial independence” in the domain of discourse.

The results obtained from the experimentation and technological characteristics of RYEL showed a new spectrum of research in which the interaction of technology and

human behavior implies new techniques to capture the perception of a human. Therefore, this research could open doors to venture into other domains using this technology to study the behavioral response of a subject, where the interpretation and assessment of a person have to be the foundation for the development of the area under discussion. At present, there is no detection of investigations or experimental studies with the same approach as ours.

Author Contributions: Conceptualization, L.R.R.O.; methodology, L.R.R.O.; software, L.R.R.O.; validation, J.J.V., A.C., Á.B., and J.M.C.; investigation, L.R.R.O.; data curation, L.R.R.O.; writing—original draft preparation, L.R.R.O.; writing—review and editing, J.M.C. and L.R.R.O.; visualization, J.M.C. and L.R.R.O.; supervision, J.J.V., A.C., Á.B., and J.M.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval for this study were waived for this study, due to experiments were not conducted in humans, only communication and interaction.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study, Oficio N. 9073-2020.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: The authors would like to acknowledge the School of Computer Science and Informatics (ECCI) and the Postgraduate Studies System (SEP), both from the University of Costa Rica (UCR), Costa Rica; the BISITE Research Group and the Faculty of Law, both from the University of Salamanca (USAL), Spain; and the Edgar Cervantes Villalta School of Judiciary of Costa Rica. Special thanks to all the judges and AI experts who participated in this investigation from Mexico, Costa Rica, Spain, Panama, Argentina, and other countries.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

(IA-AI)	Interpretation-Assessment/Assessment-Interpretation
(MOE)	Mixture of Experts
(XAI)	Explainable Artificial Intelligence
(SN)	Semantic Networks
(SME)	Subject-Matter Experts
(AI)	Artificial Intelligence
(EGI)	Explanatory Graphical Interfaces
(CBR)	Case-Based Reasoning
(HCI)	Human-Computer Interaction
(KG)	Knowledge Graphs
(PG)	Property Graph
(KI)	Knowledge Integration
(BN)	Bayesian Networks
(FR)	Fragmented Reasoning
(UX)	User Experience
(SQuaRE)	Software Quality Requirements and Evaluation
(ANOVA)	Analysis of Variance
(AD)	Anderson-Darling

Appendix A

Table A1. RYEL system: Components and technology.

#	Components	Technology, Formulas or Concepts	Orientation and Use
1	A visual component that allows a dynamic work with images representing objects.	D3.js, HTML, JavaScript, JQuery, CSS.	¹ HCI—graphics to elicit knowledge.
2	Component for managing the KG modeling.	Neo4j database.	Data model.
3	According to the context, the component connects the images with the graph model and works with similar words.	² NLP using NEO4J scripts.	HCI—graphics to elicit knowledge and semi-supervised method to detect word similarities.
4	Component that extracts image patterns from the KG and provides query options.	HTML, CSS, AJAX, CYPHER.	Queries and analysis options management.
5	Component to transform artifact inputs (images and relationships) to nodes and arcs.	Jquery, JavaScrips, HTML.	HCI—graphics to elicit knowledge.
6	Component that adapts the Pythagorean theorem to Euclidean space creates and modifies the attributes of nodes and relationships.	Adaptation of the Pythagorean formula using Javascript.	Attribute calculation.
7	Component to manage searches for node and relationship interpretation patterns.	CYPHER, HTML, PYTHON.	Performance pattern search operations.
8	Component for similarity of attributes and interpretation patterns calculation.	Adaptation of the Cosine, Jaccard and Person equations to the attributes calculated using PYTHON and CYPHER.	Attribute similarity operations.
9	Graphic component for visualizing for interpretation patterns using pie and bar charts.	HTML, CSS, D3.js, Javascript, JQuery, CSS, PYTHON, CYPHER.	HCI—graphics to show interpretation patterns.
10	Component that converts found patterns and similar attributes into a graphic explanation using geometric figures.	D3.js, HTML, CYPHER, PYTHON.	HCI—graphics to explain the interpretation found.

¹ Human-Computer Interface; ² Natural Language Processing.

Table A2. Data structures about nodes and relationships in Figure 14.

Structure	Simplified Example of the Structure
Node	$(shoot, \{00, 333, \{injure, precedence, level\}\})$
Relationship	$("", \{Null, (Null, Null), description, link, relevance\})$
Node	$(drugs, \{10, 333, \{injure, precedence, level\}\})$
Relationship	$(induceTo, \{000, (10, 00), \{description, link, relevance\}\})$
Node	$(Psychological\ problem, \{20, 333, \{injure, precedence, level\}\})$
Relationship	$(causesThe, \{100, (20, 00), \{description, link, relevance\}\})$
Relationship	$(consumes, \{200, (20, 10), \{description, link, relevance\}\})$
Relationship	$(searchOne, \{300, (20, 30), \{description, link, relevance\}\})$
Node	$(45\ handgun, \{30, 333, \{injure, precedence, level\}\})$
Relationship	$(usedFor, \{400, (30, 00), \{description, link, relevance\}\})$
Node	$(hit, \{40, 444, \{disable, precedence, level\}\})$
Relationship	$(priorTo, \{500, (40, 00), \{description, link, relevance\}\})$

Table A2. *Cont.*

Structure	Simplified Example of the Structure
Relationship	$\left(\text{with}, \{600, (40, 50), \{\text{description}, \text{link}, \text{relevance}\}\} \right)$
Relationship	$\left(\text{aNumberOf}, \{700, (40, 60), \{\text{description}, \text{link}, \text{relevance}\}\} \right)$
Relationship	$\left(\text{recordedIn}, \{800, (40, 70), \{\text{description}, \text{link}, \text{relevance}\}\} \right)$
Node	$\left(\text{baseball bat}, \{50, 444, \{\text{disable}, \text{precedence}, \text{level}\}\} \right)$
Relationship	$\left("", \{Null, (Null, Null), \{\text{description}, \text{link}, \text{relevance}\}\} \right)$
Node	$\left(\text{12 times}, \{60, 444, \{\text{disable}, \text{precedence}, \text{level}\}\} \right)$
Relationship	$\left("", \{Null, (Null, Null), \{\text{description}, \text{link}, \text{relevance}\}\} \right)$
Node	$\left(\text{Security video}, \{70, 444, \{\text{disable}, \text{precedence}, \text{level}\}\} \right)$
Relationship	$\left("", \{Null, (Null, Null), \{\text{description}, \text{link}, \text{relevance}\}\} \right)$

Table A3. The adjacency of the elements of a case in Figure 14.

Adjacency Matrix									
Labels	Indices	0	10	20	30	40	50	60	70
shoot	0	0	0	0	0	0	0	0	0
drug	10	1	0	0	0	0	0	0	0
psychological problem	20	1	1	0	1	0	0	0	0
45 handgun	30	1	0	0	0	0	0	0	0
hit	40	1	0	0	0	0	1	1	1
baseball bat	50	0	0	0	0	0	0	0	0
12 times	60	0	0	0	0	0	0	0	0
video security	70	0	0	0	0	0	0	0	0

Table A4. Relationships box from Figure 14.

Relations List		
Labels	Indices	Relations
shoot	0	Null
drug	10	(10,00)
psychological problem	20	(20,00), (20,10), (20,30)
45 handgun	30	(30,00)
hit	40	(40,00), (40,50), (40,60), (40,70)
baseball bat	50	Null
12 times	60	Null
video security	70	Null

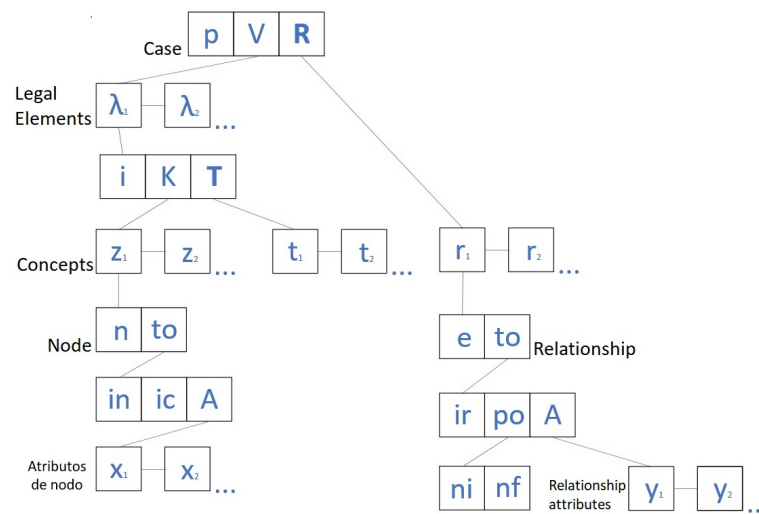


Figure A1. Representation of a case structure with ordered pairs ordered triple, and datasets used to process the information obtained graphically from images in the KG. The vectors are extracted through these structures and used in similarity functions in component 8 from Table A1. **Observation:** *t* structure is similar to *r*, so the details of *t* are omitted for clarity of the diagram.

Table A5. Case-based reasoning related work summary.

Articles	Key Elements	¹ Cases	Focus on
1986 [38]	The system JUDGE uses a case-based model of felonies where justifications of actions or the lack of them are used as a metric and determine if a situation is favorable or not; the model works entering actions and compares them with others stored previously to obtain differences (effects).	A & M	Lawyer
1987 [102]	A legal citations model created using Case-Based Knowledge (CBK), and it consists of analyzing the characteristics of what they call “phasic states of a legal case” (facts) and their “dimensions” (classifications) to help lawyers litigate. The citations use Blue Book and HYPO systems; the output is a citation network used to justify legal disputes.	C	Lawyer
1997 [103]	Divorce Property Separation Act (ASHSD) is a tool based on CBR and RBR to query cases about ways to separate assets. A case is a list of attributes processed using three stages: (1) Filtering attributes, (2) assigning a similarity between them, and (3) assigning a weight to them. The rules are if-then statements used for attribute classification.	AS	Lawyer
2003 [104]	The system CATO was created as a learning system to teach legal argumentation to beginning law students. It uses 14 cases having legal decisions and 2 cases used as evidence. There are 26 factors (facts) associated with 5 types of numerical values used to classify the factors.	USTA	Law student
2003 [105]	AlphaTemis is a free text query system on attributes of legal cases. The user can assign a weight to each attribute that is a discrete number used to query those that are the same.	SCB	Lawyer Prosecutor
2003 [106]	The investigation deals with the organization of the legal arguments, obtaining differences between them, and evaluating if a previous case is essential for the current one; for this, it uses a hierarchy of factors (facts) to measure the importance. Finally, it applies the BUC (Best-untrumped Cases) to identify what factors are in common between the cases in the database and the current legal problem.	USTA	Law student
2011 [30]	This research proposes a way in which one case can be compared to another using proposition and legal rules based on legal information about what they call “value judgments” and “legal concepts” where the judges handle values of specific factual scenarios according to what a proponent (plaintiff or appellant) presents in the arguments. An opponent (defendant or defendant) refutes that argument, and finally, the proponent makes a rebuttal.	CadyDom	Lawyer

¹ A & M = Assault and Murder, C = Citations, AS = Assets, USTA = US Trade Agreement Law, JD = Judicial Decisions, SCB = “Súmulas” of Court of Brazil, CadyDom = Fictional example oral argument based on Cady vs. Dombrowski case by the U.S. Supreme Court, N/D = Not Defined.

Table A6. Expert systems related work summary.

Articles	Key Elements	¹ Technique	Focus on
1987 [107]	DEFAULT system uses hierarchical predicates ordering (general to specific), for consulting information about legal cases related to the eviction of indigents.	PreL	Lawyer
1987 [108]	Uses predicates (PROLOG) to define norms of legal cases and make queries about legal rules. It uses a number (“raking”) to indicate the importance of a norm.	PreL	Lawyer
1991 [109]	It explains the potential and advantages of working with legal information graphically because the law and arguments contain complex relationship schemes, and graphs can help identify them. The use of Toulmin charts allows to express arguments and helps the user define value judgments on the legal information.	TC	Lawyer
1991 [110]	Loge-expert is a system that consists of process flow charts with hypertext about the rules of the civil code in Canada used to consult multiple legal documents regarding a given law.	Ht charts	Layman
1993 [111]	Use the LES system that uses Horn clauses to find similarities between legal requirements and legal norms.	ProL	Lawyer
1993 [86]	Use of predicates called “slots” in a system called CIGOL for consulting facts in legal cases.	PreL	Lawyer
1999 [112]	Retrieve texts from legal cases from the Attorney General of the Republic of Portugal using Dynamic Logic, which is an extension of Modal Logic, through consultations using rules and predicates that describe events (facts) of legal cases.	DL and PreL	Lawyer
1999 [29]	SMILE is a system that searches for words in sentences of legal texts and searches for the rules associated with those words. It uses a decision tree (ID3 algorithm) and a legal language repository to generate the tree-like word structures and related rules.	DT	Lawyers
2003 [104]	It uses predicates to explain the concept of “Theoretical Construction” that consists of facts related to legal rules, values, and preferences.	PreL	Lawyer
2005 [113]	AGATHA is a system that searches for case precedents to explain how things happened. Cases are decision trees and use the A* algorithm to find the least cost path between a source node and a destination node. The lowest cost path is the one selected.	DT	Lawyer
2005 [65]	Use a semantic web with hyperlinks to legal documents on the Dutch Tax and Customs Law (DTCA) to query related legal documents.	Ht	Lawyer
2005 [114]	Use propositional language to describe legal arguments, requests from plaintiffs, and advocates.	ProL	Lawyer
2009 [115]	It supports a litigant using predicates (PROLOG) to define and query legal situations from the House of the Lords in Quebec, Canada.	PreL	Lawyer
2009 [116]	ArguGuide software showing the text structure of a legal case and the legal topic. It shows a content map, whose elements are legal text and checklists.	CM	Lawyer
2009 [117]	Use of Carneades system to describe cases of the German Family Law. The arguments are lists and each tuple is a statement.	ProL	Lawyer
2009 [118]	This research is about displaying arguments using Toulmin charts that are flow charts of the arguments supplemented in this case with hypertext; a chart shows the text of the case, so a law student or lawyer can manually manipulate and segment the text that needs to be used as an argument.	TC	Lawyer Students

Table A6. Cont.

Articles	Key elements	¹ Technique	Focus on
2013 [119]	Use variables of location and time of people about a crime and calculate the probability that a person is a murderer. It makes analogous use of the “Island Problem”.	BY	Lawyer Prosecutor
2014 [74]	Ontology building using rules and predicates for consulting legal case documents.	PreL	Lawyer
2017 [120]	Queries using a question-based text for searches and answers. It tries to get a question about a legal context and returns general and related information.	NLP	Lawyer
2017 [121]	Argumentation mining using pre-classified legal words with a KNS classifier. The input text is about facts and the output is a text about the general topic of arguments.	NLP	Lawyer
2017 [41]	Pre-existing mapping of arguments, rules to legal cases. It tries to demonstrate that the legislation and the precedents are sources of the arguments.	FM	Lawyer
2017 [121]	Prometea is a system for issuing a “legal opinion” of the legal cases that the prosecution has. This opinion consists of indicating which are the most relevant cases and therefore must be processed first. The definition of relevance is according to the “vulnerability” of the people described in the case; for example, it tries to find words or information about the elderly, children, women, or people with disabilities. Only considers cases whose legal complexity is simple.	N/D	Prosecutor
2018 [122]	It uses document classification techniques (TF-IDF) to process a set of legal cases on labor material and uses the K-NN algorithm to obtain a ranking on the trend of opinions of judges in the Brazilian courts related to those specific cases.	KNN & TF-IDF	Lawyer

¹ NLP = Natural Language Processing, TC = Toulmin Chart, BN = Bayesian Networks, FM = Feature Mapping, CM = Content Mapping, ProL = Proposition Logic, PreL = Predicate Logic, DL = Dynamic Logic, Ht = Hypertext, DT = Decision Tree, KNN & TF-IDF = K-Nearest Neighbors & Term Frequency–Inverse Document Frequency, N/D = Not Defined.

References

- Evans, J.; Foster, J. Metaknowledge. *Science* **2011**, *331*, 721–725. [CrossRef] [PubMed]
- Rodríguez, L.; Vargas, J.; Camacho, A.; Burgos, A.; Corchado, J. RYEL system: A novel method for capturing and represent knowledge in a legal domain using Explainable Artificial Intelligence (XAI) and Granular Computing (GrC). In *Interpretable Artificial Intelligence: A perspective of Granular Computing*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 369–399.
- Berman, D.; Hafner, C. Representing teleological structure in case-based legal reasoning: The missing link. In Proceedings of the ICAIL '93 4th International Conference on Artificial Intelligence and Law, Amsterdam, The Netherlands, 15–18 June 1993; pp. 50–59.
- Tennyson, R.; Breuer, K. Cognitive-Based design guidelines for using video and computer technology in course development. In *Video in Higher Education*; Kogan Page: London, UK, 1984; pp. 26–63.
- Tennyson, R.; Rasch, M. Linking cognitive learning theory to instructional prescriptions. *Instr. Sci.* **1988**, *17*, 369–385. [CrossRef]
- Tennyson, R.; Park, O. Artificial intelligence and computer-based learning. In *Instructional Technology: Foundations*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1987; pp. 319–342.
- Bruner, J. *Toward a Theory of Instruction*; Belknap Press of Harvard University: London, UK, 1966; ISBN 978-0-674-89701-4.
- Peter, L.; Donald, N. *Human Information Processing: An Introduction to Psychology*; Academic Press, Inc.: Cambridge, MA, USA, 1977; ISBN 0124509509.
- Plant, R.; Gamble, R. Methodologies for the development of knowledge-based systems, 1982–2002. *Knowl. Eng. Rev.* **2003**, *18*, 47–81. [CrossRef]
- Baldwin, C. How to improve communication with co-workers and subject matter expe. In Proceedings of the Professional Communication Conference The New Face of Technical Communication: People, Processes, Products, Philadelphia, PA, USA, 5–8 October 1993; pp. 403–407.
- Stayanchi, J. Higher Order Thinking through Bloom’s Taxonomy. *Humanit. Rev.* **2017**, *22*, 117–124.
- Guitton, M. The immersive impact of meta-media in a virtual world. *Comput. Hum. Behav.* **2012**, *28*, 450–455. [CrossRef]
- Leonardo, G. La Definición del Concepto de Percepción en Psicología. *Rev. Estud. Soc.* **2004**, *18*, 89–96. [CrossRef]
- Atienza, M. *Las Razones del Derecho Teorías de la Argumentación jurídica*; Universidad Autónoma de México: México, Mexico, 2005; ISBN 978-970-32-0364-2.
- Romero, J. Notas sobre la Interpretación Jurídica. *Rev. Cienc. Jurídicas* **2014**, *133*, 79–102.
- Legislativa, A. *Código Procesal Penal (Ley N. 7594 de 10 de abril de 1996)*; Tribunal Supremo de Elecciones: San José, Costa Rica, 1996.
- Legislativa, A. *Código Penal (Ley N. 4573 de 15 de noviembre de 1970)*; Tribunal Supremo de Elecciones: San José, Costa Rica, 1970.

18. Perroni, E. *Play: Psychoanalytic Perspectives, Survival and Human Development*; Routledge: London, UK, 2013; ISBN 9780415682084.
19. McCarthy, J.; Minsky, M.; Shannon, C. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, 31 August 1955; pp. 1–17.
20. Rochat, P. Layers of awareness in development. *Dev. Rev.* **2015**, *38*, 122–145. [CrossRef]
21. Rapp, B. *Handbook of Cognitive Neuropsychology What Deficits Reveal About the Human Mind*; Psychology Press: London, UK, 2001; ISBN 9781841690445.
22. Hong, C.; Batal, I.; Hauskrecht, M. A Mixture-of-Experts Framework for Multi-Label Classification. *arXiv* **2014**, arXiv:1409.4698.
23. Masoudnia, S.; Ebrahimpour, R. Mixture of experts: A literature survey. *Artif. Intell. Rev.* **2014**, *42*, 275–293. [CrossRef]
24. Rodríguez, L.; Osegueda, A. Business intelligence model to support a judge's decision making about legal situations. In Proceedings of the IEEE 36th Central American and Panama Convention (CONCAPAN XXXVI), San Jose, Costa Rica, 9–11 November 2016; pp. 1–5.
25. Rodríguez, L.R. Jurisdictional normalization based on artificial intelligence models. In Proceedings of the XX Iberoamerican Congress of Law and Informatics (FIADI), Salamanca, Spain, 19–21 October 2016; pp. 1–16.
26. Rodríguez, L.R. *Jurisdictional Normalization of the Administration of Justice for Magistrates, Judges Using Artificial Intelligence Methods for Legal Guidance Systems*; II Central American and Caribbean Congress on Family Law; Central American and Caribbean Congress: Panamá, Panama, 2016; pp. 1–10.
27. Rodríguez, L.R. Artificial Intelligence Applied in Procedural Law and Quality of Sentences. In Proceedings of the XXI Iberoamerican Congress of Law and Informatics (FIADI), San Luis Potosí, México, 17–20 October 2017; pp. 1–19.
28. Kolodner, J. *Case-Based Reasoning*; Morgan Kaufmann Publishers, Inc.: San Mateo, CA, USA, 1993.
29. Bruninghaus, S.; Ashley, K. Toward Adding Knowledge to Learning Algorithms for Indexing Legal Cases. In Proceedings of the ICAIL '99 7th International Conference on Artificial Intelligence and Law, Oslo, Norway, 14–17 June 1999; pp. 9–17.
30. Grabmair, M.; Ashley, K. Facilitating Case Comparison Using Value Judgments and Intermediate Legal Concepts. In Proceedings of the 13th International Conference on Artificial Intelligence and Law, Montreal, QC, Canada, 6–10 June 2011; pp. 161–170.
31. Ha, T.; Lee, S.; Kim, S. Designing Explainability of an Artificial Intelligence System. In Proceedings of the TechMindSociety '18: Technology, Mind, and Society, Washington, DC, USA, 5–7 April 2018; p. 1.
32. Zhu, J.; Liapis, A.; Risi, S.; Bidarra, R.; Youngblood, M. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. In Proceedings of the IEEE Conference on Computational Intelligence and Games (CIG) Maastricht, The Netherlands, 14–17 August 2018; pp. 1–8.
33. Khanh, H.; Tran, T.; Ghose, A. Explainable Software Analytics. In Proceedings of the ACM/IEEE 40th International Conference on Software Engineering: New Ideas and Emerging Results, Gothenburg, Sweden, 28 May–3 June 2018.
34. Pedrycz, W.; Gomide, F. *Fuzzy Systems Engineering: Toward Human-Centric Computing*; Wiley-IEEE Press: Hoboken, NJ, USA, 2007.
35. Bargiela, A.; Pedrycz, W. Granular Computing for Human-Centered Systems Modelling. In *Human-Centric Information Processing Through Granular Modelling*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 320–330.
36. Yao, Y. Human-Inspired Granular Computing. In *Novel Developments in Granular Computing: Applications for Advanced Human Reasoning and Soft*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 1–15.
37. Shadbolt, N.; Smart, P. Knowledge Elicitation: Methods, Tools and Techniques. In *Evaluation of Human Work*; CRC Press: Boca Raton, FL, USA, 2015; pp. 163–200.
38. Bain, W. Judge: A case-based reasoning system. In *The Kluwer International Series in Engineering and Computer Science (Knowledge Representation, Learning and Expert Systems)*; Springer: Boston, MA, USA, 1986; pp. 1–4.
39. Aleven, V. Teaching Case-Based Argumentation through a Model and Examples. Doctoral Dissertation, University of Pittsburgh, Pittsburgh, PA, USA, 1997.
40. Bench-Capon, T.; Sartor, G. Theory Based Explanation of Case Law Domains. In Proceedings of the ICAIL '01 8th International Conference on Artificial Intelligence and Law, St. Louis, MO, USA, 21–25 May 2001; pp. 12–21.
41. Verheij, B. Formalizing Arguments, Rules and Cases. In Proceedings of the 16th International Conference on Artificial Intelligence and Law, London, UK, 12–16 June 2017; pp. 199–208.
42. Snyder, J.; Mackulak, G. Intelligent simulation environments: Identification of the basics. In Proceedings of the 20th conference on Winter simulation, New York, NY, USA, 1–2 December 1988; pp. 357–363.
43. Zeigler, B.; Muzy, A.; Yilmaz, L. Artificial Intelligence in Modeling and Simulation. In *Encyclopedia of Complexity and Systems Science*; Springer: New York, NY, USA, 2009; pp. 344–368.
44. Ruiz, N.; Giret, A.; Botti, V.; Fera, V. An intelligent simulation environment for manufacturing systems. *Comput. Ind. Eng.* **2014**, *76*, 148–168. [CrossRef]
45. Li, K.; Li, J.; Liu, Y.; Castiglione, A. Computational Intelligence and Intelligent Systems. In Proceedings of the 7th International Symposium, ISICA, Guangzhou, China, 21–22 November 2015; pp. 183–275.
46. Teerapong, K. Graphical ways of researching. In Proceedings of the Graphical Ways of Researching, Como, Italy, 27 May 2014.
47. Offermann, P.; Levina, O.; Schönherr, M.; Bub, U. Outline of a Design Science Research Process. In Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology, Philadelphia, PA, USA, 7–8 May 2009; pp. 7–11.
48. Abraham, A.; Corchado, E.; Corchado, J. Hybrid learning machines. *Neurocomput. Int. J.* **2009**, *72*, 13–15. [CrossRef]

49. Azizi, A. Hybrid artificial intelligence optimization technique. In *In Applications of Artificial Intelligence Techniques in Industry 4.0*; Springer: Singapore, 2019; pp. 27–47.
50. Corchado, J.; Pavón, J.; Corchado, E.; Castillo, L. Development of CBR-BDI Agents: A Tourist Guide Application. In *ECCBR 2004: Advances in Case-Based Reasoning*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 547–559.
51. Hafner, C.; Berman, D. The role of context in case-based legal reasoning: Teleological, temporal, and procedural. *Artif. Intell. Law* **2002**, *10*, 19–64. [CrossRef]
52. Conrad, J.; Al-Kofahi, K. Scenario Analytics Analyzing Jury Verdicts to Evaluate Legal Case Outcomes. In Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017), London, UK, 12–16 June 2017.
53. Card, S.; Moran, T.; Newell, A. *The Psychology of Human-Computer Interaction*; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1986; ISBN 978-0898598599.
54. Clewley, N.; Dodd, L.; Smy, V.; Witheridge, A.; Louvieris, P. Eliciting Expert Knowledge to Inform Training Design. In Proceedings of the ECCE 2019: 31st European Conference on Cognitive Ergonomics, BELFAST, UK, 10–13 September 2019; pp. 138–143.
55. Galinsky, A.; Maddux, W.; Gilin, D.; White, J. Why It Pays to Get Inside the Head of Your Opponent: The Differential Effects of Perspective Taking and Empathy in Negotiations. *Psychol. Sci.* **2008**, *19*, 378–384. [CrossRef] [PubMed]
56. Carral, M.d.R.; Santiago-Delefosse, M. Interpretation of Data in Psychology: A False Problem, a True Issue. *Philos. Study* **2015**, *5*, 54–62.
57. Legislativa, A. *Código Penal (Ley N. 7576 del 30 de Abril de 1996)*; Tribunal Supremo de Elecciones: San José, Costa Rica, 1996.
58. Zhang, L. *Knowledge Graph Theory and Structural Parsing*; Twente University Press: Enschede, The Netherlands, 2002.
59. Singhal, A. Introducing the Knowledge Graph: Things, not Strings. Available online: <https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things-not.htm> (accessed on 3 December 2012).
60. Paulheim, H. Knowledge Graph Refinement: A Survey of Approaches and Evaluation Methods. *Semant. Web* **2016**, *2016*, 1–23. [CrossRef]
61. Gasevic, D.; Djuric, D.; Devedzic, V. *Model Driven Architecture and Ontology Development*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 1–310.
62. Tonon, M. Hermeneutics and Critical Theory. In *A Companion to Hermeneutics*; Wiley: Hoboken, NJ, USA, 2015; pp. 520–529.
63. Bonatti, P.; Cochez, M.; Decker, S.; Polleres, A.; Valentina, P. Knowledge Graphs: New Directions for Knowledge Representation on the Semantic Web. *Rep. Dagstuhl Semin.* **2018**, *18371*, 2–92.
64. Yan, J.; Wang, C.; Cheng, W.; Gao, M.; Aoying, Z. A retrospective of knowledge graphs. *Front. Comput. Sci.* **2018**, 55–74.
65. Winkels, R.; Boer, A.; De-Maat, E.; Van, T.; Breebaart, M.; Melger, H. Constructing a semantic network for legal content. In Proceedings of the ICAIL 05 10th International Conference on Artificial Intelligence and Law, Bologna, Italy, 6–11 June 2005; pp. 125–132.
66. Florian, J. *Encyclopedia of Cognitive Science: Semantic Networks*; Wiley and Sons: Hoboken, NJ, USA, 2006; ISBN 9780470016190.
67. Noirie, L.; Dotaro, E.; Carofiglio, G.; Dupas, A.; Pecci, P.; Popa, D.; Post, G. Semantic networking: Flow-based, traffic-aware, and self-managed networking. *Bell Labs Tech. J.* **2009**, *14*, 23–38. [CrossRef]
68. Lehmann, F. Semantic Networks. *Comput. Math. Appl.* **1992**, *23*, 1–50. [CrossRef]
69. Robinson, I.; Webber, J.; Eifrem, E. *Graph Databases New Opportunities for Connected Data*; O'Reilly Media: Sebastopol, CA, USA, 2015.
70. McCusker, J.; Erickson, J.; Chastain, K.; Rashid, S.; Weerawarana, R.; Bax, M.; McGuinness, D. What is a Knowledge Graph? *Semant. Web Interoperabil., Usabil. Appl. Ios Press J.* **2018**, *2018*, 1–14.
71. Neo Technology. What Is a Graph Database? 2019. Available online: <https://neo4j.com/developer/graph-database/> (accessed on 5 January 2019).
72. Ioana, H.; Prangnawarat, N.; Haye, C. Path-based Semantic Relatedness on Linked Data and its use to Word and Entity Disambiguation. In Proceedings of the International Semantic Web Conference, ISWC 2015: The Semantic Web—ISWC, Bethlehem, PA, USA, 11–15 October 2015; pp. 442–457.
73. Seeliger, A.; Pfaff, M.; Krcmar, H. Semantic Web Technologies for Explainable Machine Learning Models: A Literature Review. *PROFILES/SEMEX@ISWC* **2019**, *2465*, 1–16.
74. Mezghanni, I.; Gargouri, F. Learning of Legal Ontology Supporting the User Queries Satisfaction. In Proceedings of the International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Warsaw, Poland, 11–14 August 2014; pp. 414–418.
75. Loui, R. From Berman and Hafne's teleological context to Baude and Sachs' interpretive defaults: An ontological challenge for the next decades of AI and Law. *Artif. Intell. Law* **2016**, *2016*, 371–385. [CrossRef]
76. Tan, P.N.; Steinbach, M.; Kumar, V. *Introduction to Data Mining*, 1st ed.; Addison-Wesley: Boston, MA, USA, 2005; ISBN 978-0-321-32136-7.
77. Singhal, A. Modern Information Retrieval: A Brief Overview. *Bull. IEEE Comput. Soc. Tech. Comm. Data Eng.* **2001**, *24*, 35–43.
78. Boddy, R.; Smith, G. *Statistical Methods in Practice: For Scientists and Technologists*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2009; ISBN 9780470749296.
79. Yu, G.; Yang, Y.; Qingsong, X. An Ontology-based Approach for Knowledge Integration in Product Collaborative Development. *J. Intell. Syst.* **2015**, *26*, 35–46. [CrossRef]
80. Schneider, M. Knowledge Integration. *Encycl. Sci. Learn.* **2012**, *2012*, 1684–1686.
81. Allama, Z.; Dhunny, Z. On big data, artificial intelligence and smart cities. *Cities* **2019**, *89*, 80–91. [CrossRef]

82. Chamoso, P.; González-Briones, A.; Rodríguez, S.; Corchado, J. Tendencies of Technologies and Platforms in Smart Cities: A State-of-the-Art Review. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 1–17. [CrossRef]
83. Baldacchino, T.; Cross, E.; Worden, K.; Rowson, J. Variational Bayesian mixture of experts models and sensitivity analysis for nonlinear dynamical systems. *Mech. Syst. Signal Process.* **2016**, *66*, 178–200. [CrossRef]
84. Merkl, B.; Chao, J.; Howard, R. *Graph Databases for Beginners*; Ebook, 2018. Available online: <https://neo4j.com/blog/data-modeling-basics/> (accessed on 15 May 2020).
85. Ashley, K.; Rissland, E. A case-based system for trade secrets law. In Proceedings of the ICAIL '87 1st International Conference on Artificial Intelligence and Law, New York, NY, USA, 27–29 May 1987; pp. 60–66.
86. Yamaguti, T.; Kurematsu, M. Legal Knowledge Acquisition Using Case-Based Reasoning and Model Inference. In Proceedings of the ICAIL '93 4th International Conference on Artificial Intelligence and Law, Amsterdam, The Netherlands, 15–18 June 1993; pp. 212–217.
87. Lecue, F. On The Role of Knowledge Graphs in Explainable AI. *Semant. Web* **2019**, *2019*, 1–9. [CrossRef]
88. Goodrich, P. Historical Aspects of Legal Interpretation. *Indiana Law J.* **1986**, *61*, 331–354.
89. Rojas, G. *El Objeto Material y Formal del Derecho*; Universidad Católica de Colombia: Bogotá, Colombia, 2018.
90. Abdul, A.; Vermeulen, J.; Wang, D.; Lim, B.; Kankanhalli, M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In Proceedings of the CHI '18: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–18.
91. Pedrycz, W. Granular computing for data analytics: A manifesto of human-centric computing. *IEEE/CAA J. Autom. Sin.* **2018**, *5*, 1025–1034. [CrossRef]
92. Gómez-Pérez, J.; Erdmann, M.; Greaves, M.; Corcho, O. A Formalism and Method for Representing and Reasoning with Process Models Authored by Subject Matter Experts. *IEEE Trans. Knowl. Data Eng.* **2012**, *25*, 1933–1945. [CrossRef]
93. Wolf, C.; Ringland, K. Designing accessible, explainable AI (XAI) experiences. *ACM Sigaccess Access. Comput.* **2020**, *6*, 1–5. [CrossRef]
94. Freeman, E.; Robson, E. *Head First Design Patterns*; O'Reilly Media, Inc.: Hoboken, NJ, USA, 2004; ISBN 978-0-596-00712-6.
95. Khazaii, J. Fuzzy Logic. In *Advanced Decision Making for HVAC Engineers*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 157–166.
96. Lai-Chong, E.; Schaik, P.; Roto, V. Attitudes towards user experience (UX) measurement. *Int. J. Hum. Comput. Stud.* **2014**, *2014*, 526–541.
97. Standards, B. Systems and software engineering—Systems and software Quality Requirements and Evaluation (SQuaRE)—System and software quality models. *BS ISO/IEC* **2011**, *25010*, 1–34.
98. Salvaneschi, P. The Quality Matrix: A Management Tool for Software Quality Evaluation. In Proceedings of the IASTED International Conference on Software Engineering, Innsbruck, Austria, 15–17 February 2005; pp. 394–399.
99. Robinson, J. Likert Scale. *Encycl. Qual. Life Well-Being Res.* **2014**, *2014*, 3620–3621.
100. Judiciary, C.R. *Informe de Labores 2019 Centro de Apoyo, Coordinación y Mejoramiento de la Función Jurisdiccional*; Poder Judicial Costa Rica: San Jose, CR, USA, 2019; pp. 1–34.
101. Montgomery, D. *Design and Analysis of Experiments*, 18th ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2013; ISBN 978-1-118-14692-7.
102. Ashley, K.; Rissland, E. But, see, accord: Generating blue book citations in HYPO. In Proceedings of the ICAIL '87 1st International Conference on Artificial Intelligence and Law, Boston, MA, USA, 27–29 May 1987; pp. 67–74.
103. Pal, K.; Campbell, J. An Application of Rule-based and Case-based Reasoning Within a Single Legal Knowledge-based System. *SIGMIS Database J.* **1997**, *1997*, 48–63. [CrossRef]
104. Chorley, A.; Bench-Capon, T. Developing Legal Knowledge Based Systems Through Theory Construction. In Proceedings of the ICAIL '03: 9th International Conference on Artificial Intelligence and Law, Edinburgh, UK, 24–28 June 2003; pp. 85–86.
105. Bueno, T.; Bortolon, A.; Hoeschl, H.; Mattos, E.; Ribeiro, M. Analyzing the use of Dynamic Weights in Legal Case Based System. In Proceedings of the ICAIL '03: 9th International Conference on Artificial Intelligence and Law, Edinburgh, UK, 24–28 June 2003; pp. 136–141.
106. Aleven, V. Using background knowledge in case-based legal reasoning: A computational model and an intelligent learning environment. *Artif. Intell.* **2003**, *2003*, 183–237. [CrossRef]
107. Purdy, R. Knowledge representation in “Default”: An attempt to classify general types of knowledge used by legal experts. In Proceedings of the CAIL '87 1st International Conference on Artificial Intelligence and Law, Boston, MA, USA, 27–29 May 1987; pp. 199–208.
108. Belzer, M. Legal reasoning in 3-D. In Proceedings of the 1st International Conference on Artificial Intelligence and Law, Boston, MA, USA, 27–29 May 1987; pp. 155–163.
109. Dick, J. Representation of Legal Text for Conceptual Retrieval. In Proceedings of the 3rd International Conference on Artificial Intelligence and Law, New York, NY, USA, 12 May 1991; pp. 244–253.
110. Paquin, L.C.; Blanchard, F.; Thomasset, C. Loge-expert: From a legal expert system to an information system for non-lawyers. In Proceedings of the ICAIL 91 3rd International Conference on Artificial Intelligence and Law, New York, NY, USA, 18 May 1991; pp. 254–259.

111. Yoshino, H.; Haraguchi, M.; Sakurai, S.; Kagayama, S. Towards a legal analogical reasoning system: Knowledge representation and reasoning methods. In Proceedings of the 4th International Conference on Artificial Intelligence and Law, New York, NY, USA, 15–18 June 1993; pp. 110–116.
112. Quaresma, P.; Pimenta, I. A Collaborative Legal Information Retrieval System Using Dynamic Logic Programming. In Proceedings of the 7th International Conference on Artificial Intelligence and Law, Oslo, Norway, 14 June–17 June 1999; pp. 190–191.
113. Chorley, A.; Bench-Capon, T. AGATHA: Using heuristic search to automate the construction of case law theories. *Artif. Intell. Law* **2005**, *13*, 9–51. [CrossRef]
114. Suzuki, Y.; Tojo, S. Additive Consolidation for Dialogue Game. In Proceedings of the ICAIL '05: 10th International Conference on Artificial Intelligence and Law, Bologna, Italy, 6–11 June 2005; pp. 105–114.
115. Zurek, T.; Kruk, E. Supporting of legal reasoning for cases which are not strictly regulated by law. In Proceedings of the ICAIL 09 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain, 8–12 June 2009; pp. 220–221.
116. Colen, S.; Cnossen, F.; Verheij, B. How much logical structure is helpful in content-based argumentation software for legal case solving? In Proceedings of the ICAIL '09: 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain, 8–12 June 2009; pp. 224–225.
117. Gordon, T.; Walton, D. Legal Reasoning with Argumentation Schemes. In Proceedings of the 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain, 8–12 June 2009; pp. 137–146.
118. Lynch, C.; Ashley, K.; Pinkwart, N.; Aleven, V. Toward assessing law students' argument diagrams. In Proceedings of the 12th International Conference on Artificial Intelligence and Law, Barcelona, Spain, 8–12 June 2009; pp. 222–223.
119. Dahlman, C.; Feteris, E. *Legal Argumentation Theory: Cross-Disciplinary Perspectives*; Springer: London, UK, 2013.
120. Bennett, Z.; Russell-Rose, T.; Farmer, K. A scalable approach to legal question answering. In Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law, London, UK, 12–16 June 2017; pp. 269–270.
121. Corvalán, J. La Primera Inteligencia Artificial Predictiva al Servicio de la Justicia: Prometea. 2017. Available online: <https://ialab.com.ar/wp-content/uploads/2019/05/Artículo-Juan-La-Ley.pdf> (accessed on 15 December 2018).
122. Barros, R.; Peres, A.; Lorenzi, F.; Krug-Wives, L. Case Law Analysis with Machine Learning in Brazilian Court. In Proceedings of the IEA/AIE 2018 - Recent Trends and Future Technology in Applied Intelligence, Montreal, QC, Canada, 25–28 June 2018; pp. 857–868.

Article

Exploring the Implementation of a Legal AI Bot for Sustainable Development in Legal Advisory Institutions

Juin-Hao Ho ¹, Gwo-Guang Lee ^{1,*} and Ming-Tsang Lu ^{2,*}

¹ Department of Information Management, National Taiwan University of Science and Technology, Taipei 10607, Taiwan; Kelsonnono@gmail.com

² College of Management, National Taipei University of Business, Taipei 10051, Taiwan

* Correspondence: lgg@cs.ntust.edu.tw (G.-G.L.); mingsang.lu@gmail.com (M.-T.L.)

Received: 30 June 2020; Accepted: 22 July 2020; Published: 25 July 2020

Abstract: This study explores the implementation of legal artificial intelligence (AI) robot issues for sustainable development related to legal advisory institutions. While a legal advisory AI Bot using the unique arithmetic method of AI offers rules of convenient legal definitions, it has not been established whether users are ready to use one at legal advisory institutions. This study applies the MCDM (multicriteria decision-making) model DEMATEL (decision-making trial and evaluation laboratory)-based Analytical Network Process (ANP) with a modified VIKOR, to explore user behavior on the implementation of a legal AI bot. We first apply DEMATEL-based ANP, called influence weightings of DANP (DEMATEL-based ANP), to set up the complex adoption strategies via systematics and then to employ an M-VIKOR method to determine how to reduce any performance gaps between the ideal values and the existing situation. Lastly, we conduct an empirical case to show the efficacy and usefulness of this recommended integrated MCDM model. The findings are useful for identifying the priorities to be considered in the implementation of a legal AI bot and the issues related to enhancing its implementation process. Moreover, this research offers an understanding of users' behaviors and their actual needs regarding a legal AI bot at legal advisory institutions. This research obtains the following results: (1) It effectively assembles a decision network of technical improvements and applications of a legal AI bot at legal advisory institutions and explains the feedbacks and interdependences of aspects/factors in real-life issues. (2) It describes how to vary effective results from the current alternative performances and situations into ideal values in order to fit the existing environments at legal advisory institutions with legal AI bot implementation.

Keywords: artificial intelligence; legal AI bot; sustainable development; MCDM (multiple criteria decision-making)

1. Introduction

Artificial intelligence (AI)-based legal bots have attracted extensive consideration and have appeared as one of the most promising innovations of technology. Robotics, the replacement of human labor, is becoming a crucial issue, as AI basically functions as an intelligence robot. The development of AI as a root for resolutions to various questions in life, including law, is getting more important. However, experts or human workers are still needed to apply those legal expert systems. Hence, in 2017 an online AI platform, DoNotPay, which provides free legal advice, was released in the U.S. It is called by Joshua Browder, its creator, the "first legal robot", and it could deal with up to 1000 kinds of civil events [1]. The legal AI bot can also help institutions that offer legal advice services, such as advice bureaus and community legal service centers for sustainable development.

Institutions often have a lot of part-time interns and volunteers offering legal assistance and advice, sometimes relatively early in their legal careers. Nevertheless, they are being requested to offer legal advice on a very wide range of legal problems, often with huge customer case-loads and occasionally with various cases of heterogeneous issues (such as immigration law or consumer law). They usually have limited finance resources to be capable of engaging outside legal advisors or just to employ more personnel. Therefore, if there are legal AI bots, they can solve many issues and can save on human resources and related costs for sustainable development.

A significant issue in this area is realizing what aspects/factors contribute to users' intention to apply for legal AI bot services for sustainable development. Current research studies have presented an interest in exploring the intention stage (want and plan to use)/adoption stage (will to use) of a legal AI bot. Nevertheless, most legal AI bot studies have used various kinds of methods and frameworks, making it challenging to associate the consequences of diverse studies and to develop a concrete user behavior intention and adoption in the service area. Compared with a physical legal advisory institution, the rapid growth of legal AI bots could bring legal advisory institutions administrators great benefit and efficiency. For legal advisory institutions, it is of great importance to know the strategies for legal AI bot implementation and the basis for users to use legal AI bots. Thus, this study aimed to address the following research problems: (1) in implementing legal AI bots, users should give priority to the influence factors that will improve user's intentions to use legal AI bot; (2) in using legal AI bots, which influence factors will be prioritized by users to decide whether to continue to use the legal AI bot? (3) What are the differences in the influence factors considered by the intention stage and adoption stage?

This purpose of the research is to offer an understanding about the aspects affecting legal AI bots and their implementation at improving legal advisory institutions in order to decrease the gaps in performance among each factor and aspect for sustainable development. The estimation of legal AI bot implementation is a decision analysis issue with multiple attributes, which are often categorized via interdependent factors and may even display similar feedback results. Therefore, one needs to stress that these factors exhibit various associations between lower- or higher-level elements. In addition, most common strategic models cannot take the interrelationships and dependences among dissimilar levels of factors into consideration.

Our research looks to determine the scope to which a variety of factors can affect the results in the definite factors of legal AI bot implementation. The research herein is distinguishable partly owing to its applications of various inner sources for dependent and independent information. Hence, the research objective is to set up an integrated MCDM (multicriteria decision-making) model so as to determine ways to resolve problems in legal AI bot applications.

Conventional multiple attribute decision analysis models cannot deal with the complex relationships among dissimilar factors' hierarchical stages. Nevertheless, decision makers involved in the implementation of a legal AI bot need such a model to aid their decisions. The objective of the existing study is to solve such a problem. We use an integrated MCDM model that aggregates together decision-making trial and evaluation laboratory (DEMATEL), DANP (DEMATEL-based ANP), and a modified VIKOR (M-VIKOR). Its purpose is to explore a legal advisory AI bot and to establish and enhance implementation strategies. This hybrid MCDM method can deal with the limitations of current assessment models and can assist in investigating how best to apply legal AI bots to enhance service performance for sustainable development. In this study, we investigate the interdependence of user behavior and legal AI bots and consider alternative behaviors to achieve values associated with enhanced performance.

This study makes three contributions. First, it considers four significant perspectives which those in legal advisory institutions must take into account before implementing a legal AI bot: attitudes toward legal AI bots, trust-related behavior, perceived behavioral control, and resistance to innovation for sustainable development. Second, this research demonstrates trust-related behavior; that is, it determines perceptions of external and internal limitations on user behavior and uses the relative

importance of these to implement a legal AI bot. For users to be assured that a legal AI bot can be applied for legal advisory, legal advisory institutions must offer them training with the tools required for fundamental applications and functions of a legal AI bot. Third and finally, the results of this research indicate to what extent practicality and ease of use will affect users' attitude toward ongoing applications of legal AI bots. If users are to accept a legal AI bot, then they need an environment in which they perceive that the methods are easy and useful to apply. A better understanding of how to implement a legal AI bot will assist administrators in adopting suitable schemes for creating such an environment for sustainable development.

This study has five sections. Section 1 (given above) introduces the research. Section 2 reviews the existing literature with regard to aspects of service and legal AI Bots and how to structure a model for their implementation so that our conceptual model can be developed. Section 3 defines this integrated MCDM model. Section 4 offers a case study of implementation and investigates and discusses the outcomes. Section 5 concludes.

2. Literature Review

Informatics in the legal field is growing, bringing together law and AIs at its most important parts. We thus conduct an interdisciplinary investigation in the areas of intelligent technology, law, logic, informatics, and so on [2]. This fast development of AI will influence the marketplace for legal services, the structural transformation in the legal profession, and the reorganization of resources for legal bots [3,4]. AI will improve the agility of legal services and will upgrade the standard of legal services by attaining broader justice of the judiciary and by removing the asymmetry of legal service resources in the future [5,6]. However, legal AI bots are not actual specialists, and human lawyers need to observe whenever necessary [4,7]. How to change human thinking and procedures is a significant issue in this area of AI, and it is also a mission that law people need to confront [8,9].

Legal AI bots could deal with the question of disproportion in legal service resources [4,10]. In the 1970s, researchers started to study the combination of law and AI by investigating the application of robot judgements to replace human judgements by removing legal vagueness [4,11,12]. However, this primary legal AI is to serve and assist bots or judges in dealing with events and not to replace them [4,10,13].

Academic studies on the combination of AI and law are up until now in the developing stage and are insufficient at thinking about and at realizing the applications of AI into different areas, much less investigating the pertinence of legal AI knowledge from the perception of customers or users. Thus, based on the technology acceptance model (TAM), trust, and innovation issues, our research investigates the crucial factors of the acceptance of society and how AI robots got into the legal field and interviews clients, lawyers, experts, and judges. The outcomes of the research will contribute to the combination of AI and law and to practical applications, thus filling in the research gaps.

TAM considers real users' behavior to understand novel technology based on their intention to apply it. Two main factors, i.e., perceived ease of use and perceived usefulness, influence intention, and adoption [14–16]. Numerous researchers have introduced user trust as a main factor in the investigation of TAM [17,18]. They found that user trust exerts an important influence on perceived usefulness [19]. On this foundation of TAM, investigations argue over the influences of attitude-related behaviors, perceived behavioral control, and user trust on the readiness of users to accept legal AI bots via trust factors. In addition, effective regulations and laws create trust, and a significant issue includes a legal AI bot developing and growing trust: trust in privacy, trust in functions, and trust in design. Though existing legal structures are healthy enough to deal with a few challenges that autonomous and robotic goods and services can offer, they still must develop or adapt in reply to the novel extents of applications, personal choices, and government actions [20].

According to previous research, a legal bot may offer an attractive technology of AI as users will find it an interesting and innovative approach. Various factors might have an impact on users' behavior and their willingness to use legal AI bots [4]. Xu and Wang [4] studied how individuals' willingness to

be innovative and their perception of the usefulness of a method affect the adoption of a legal AI bot. Other research has applied TAM to investigate how users accept novel ideas and implement a legal AI bot [4,21]. Investigations mostly focused on the acceptance of a legal AI bot using users’ application or intention as the dependent variable. Most previous research has focused on users’ comments regarding how they use a legal AI bot, in terms of their acceptance of the technology (how useful the legal AI bot is and this ease of using a legal AI bot) and their attitudes to or interest in legal AI bots [4,21].

While the provision of tools or approaches to enhance users’ application efforts remain a challenging and significant topic [4,21], there has been little research into how and why users accept legal AI bots [4]. Therefore, this study focuses on user behavior and how to solve various related issues. To do this, we analyze users who use a legal AI bot at legal advisory institutions and their various behaviors: plan-related (attitude-related behavior and perceived behavior control) and trust-related in terms of resistance to innovation. We do this to interpret and predict their attitudes to legal AI bots and their intention to implement them at the legal advisory institutions. This MCDM model is used for an evaluation of users’ behavior. To provide a framework for their behavior, we developed the following evaluation system, which refers to fourteen factors related to four aspects: attitude-related behaviors (ARB); perceived behavioral control (PBC); trust-related behaviors (TRB); and innovation resistance (IR). These features correspond to legal AI bot implementation within each aspect, as shown in Table 1.

Table 1. Explanation of aspects and factors.

Aspect/Factors		Description	Source
A₁		Attitude-related behaviors (ARB)	
<i>f</i> ₁	Perceived usefulness (PU)	The degree to which a person believes that using a particular legal AI bot would enhance his or her performance.	[16,22,23]
<i>f</i> ₂	Perceived ease of use (PEOU)	The degree to which a user believes that applying a legal AI bot is clear and understandable for the average person.	[16,22,23]
<i>f</i> ₃	Complexity	Users’ perceptions about how difficult the legal AI bot will be to use, operate, or understand. The easier it is to use and realize, the more likely it is to be implemented. Therefore, complexity is expected to be negatively related to attitude.	[22–25]
A₂		Perceived behavioral control (PBC)	
<i>f</i> ₄	Self-efficacy (SE)	Specific decisions that individuals make about their ability to do something. With reference to legal advisory legal AI bots, self-efficacy refers to users’ assessment of the ability to achieve services and legal information through them.	[23,26]
<i>f</i> ₅	Resource facilitating conditions (RFC)	Resources, such as time or precedents, related to resource compatibility and matters that may constrain usage.	[23,27]
<i>f</i> ₆	Technology facilitating conditions (TFC)	Technology, such as software and hardware, related to technology compatibility and issues that may constrain practice.	[23,27]
A₃		Trust-related behaviors (TRB)	
<i>f</i> ₇	Disposition to trust (DTC)	A person’s general tendency to trust others; it could be considered a personality trait.	[28,29]
<i>f</i> ₈	Structural assurance (SA)	The perception that the necessary legal and technical structures are in place: guarantees/promises, encryption, regulations, and other processes.	[16,23]
<i>f</i> ₉	Trust belief (TB)	The belief in the trustworthiness of the legal AI bot, consisting of a set of particular beliefs about competence and integrity.	[18,23,28,29]
A₄		Innovation resistances (IR)	
<i>f</i> ₁₀	Usage barrier (UB)	For using the legal AI bot, the usage barriers include users’ perceptions on what is required for legal advice, e.g., clarity.	[30–33]
<i>f</i> ₁₁	Value barrier (VB)	The perception of some users that a legal AI bot has few advantages: such as if the advisory legal AI bot connection generates more time than benefits.	[30,31]
<i>f</i> ₁₂	Risk barrier (RB)	Users’ perception rather than a characteristic of the robots. Hence, at legal advisory institutions for a legal AI bot, it is not always a problem of actual risks but has to do more with users’ perception that, for a number of reasons, the service entails risks.	[30,31]
<i>f</i> ₁₃	Tradition barrier (TB)	The impact of the innovation on routines. If these routines are significant to a user, resistance will be high. The image barrier is related to the origin of an innovation, such as advisory class.	[30,31]
<i>f</i> ₁₄	Image barrier (IB)	The negative “danger to use” perception to AI in general and to robots in particular. Users who already perceive that technology is too difficult to apply may instantly form a negative image of these service associated with the robots.	[30,31]

3. Developing a Map Based on an Integrated MCDM Model

In the section, we briefly define the proposed integrated MCDM model. One of the critical issues in MCDM is ranking a series of alternatives according to a series of factors. In this field, there exist numerous MCDM approaches that rank the alternatives in dissimilar ways [34]. This model is based on previous practice and is considered a suitable method for exploring a strategy to ensure the implementation of a legal AI bot. The functions that the integrated model offers include selection and ranking as well as performance enhancement. In this study, there are two alternative performances: one is intention stage which means that users want and plan to use a legal AI bot, and another alternative is adoption stage, which means that users will to use legal AI bot. The latter is required to reduce any gaps in achieving the ideal outcomes. Ultimately, the major advantage of our hybrid model is its decision-making function of selection extending to enhancement. Thus, it can help administrators develop the best strategies for alternative selection and enhancement problems.

The hybrid model is divided into three parts after the number of factors/aspects to be included in the framework for the legal AI bot implementation has been confirmed. (1) DEMATEL is applied to set up a structure showing the network of influencing relationships (i.e., INRM, referring to the influential network relationships map) on the factors/aspects within the framework. (2) DANP (DEMATEL-based ANP) is applied to the concepts and procedures of ANP to derive the influential weights of each factor/aspect. (3) The modified VIKOR technique applying influential weights is then used to synthesize these gaps between current and ideal performances. Hence, this hybrid model with the integration of three parts is able to support decision-makers in determining how to decrease the gaps of performance to attain the ideal outcome. The hybrid model is shown in Figure 1.

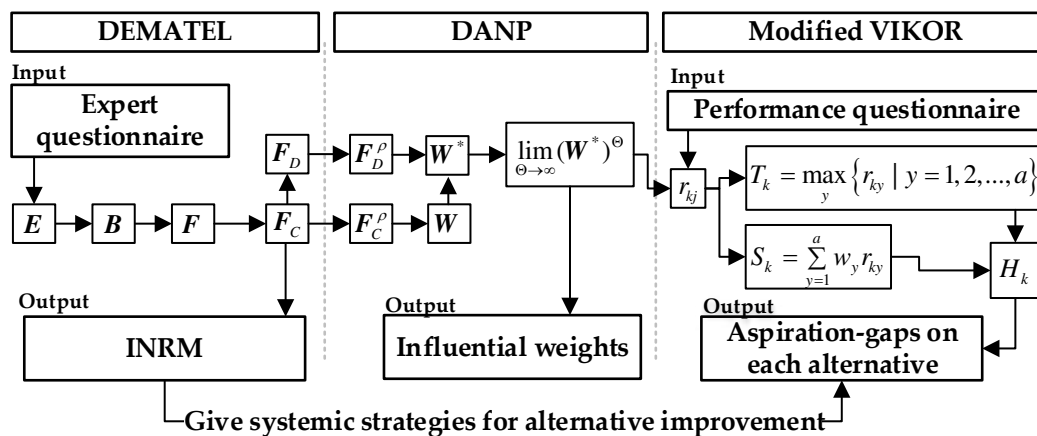


Figure 1. Modeling procedures of our proposed hybrid multicriteria decision-making (MCDM) model.

3.1. DEMATEL for Constructing an Evaluation Framework with INRM

DEMATEL is a method for establishing interdependent relationships among factors in a complex structure. The method applies mathematical theories to compute the degree of direct and indirect effects on each factor/aspect [23,35–37]. This method has four phases as follows.

3.1.1. Phase 1: Building Domain Knowledge Based on a Direct-Relation Matrix

When the number of elements (a) in an evaluation framework has been confirmed, the standard scale of degree of influence is developed (e.g., ranging from “extremely high effect (4)” to “lack of effect (0)”). This measures the degree of influence between factors or aspects by using normal language. The average of n domain experts uses a standard scale to show this direct degree of influence of the factor/aspect x on each other factor/aspect y in the matrix $D = [d_{xy}]_{a \times a} = [(\sum_{z=1}^n d_{xy}^z) / n]_{a \times a}$ (in which $d_{xy} \neq 0$; otherwise, $d_{xy} = 0$, and n is the number of domain experts). Finally, the mean is used to

integrate a primary direct relation matrix called $E = [e_{xy}]_{a \times a}$ that represents the actual experience among all the domain experts.

3.1.2. Phase 2: Obtaining a Normalized Direct Relation Matrix

A normalized primary direct relation matrix B is achieved via normalizing this primary direct relation matrix E . We use Equations (1) and (2), where the maximum sum of each row and column is 1 and all the diagonal terms of the matrix B are 0:

$$\eta = \max_{x,y} \left[\max_x \sum_{y=1}^a |e_{xy}|, \max_y \sum_{x=1}^a |e_{xy}| \right] \tag{1}$$

$$B = \frac{E}{\eta} \tag{2}$$

3.1.3. Phase 3: Deriving a Matrix of Full-Influential Relations

The matrix of full-influential relations F can be derived by using Equation (3). It can offer assurances of convergent resolutions to this matrix inversion in the same way as capturing a Markov chain matrix. Thus, the matrix of the full-influence relation F can be achieved from these values in the normalized direct-relation matrix B , where I is the identity matrix.

$$F = B + B_2 + \dots + B_h = B \times (I - B)^{-1}, \text{ when } \lim_{h \rightarrow \infty} B^h = [0]_{a \times a} \tag{3}$$

The full-influence relation matrix F can be divided into F_C (by factors) and F_D (by aspects) according to a hierarchical structure in Equations (4) and (5), respectively.

$$F_C = \begin{matrix} & \begin{matrix} D_1 & & D_y & & D_m \end{matrix} \\ \begin{matrix} D_1 \\ \vdots \\ D_x \\ \vdots \\ D_m \end{matrix} & \begin{bmatrix} c_{11} \dots c_{1m_1} & \dots & c_{y1} \dots c_{ym_y} & \dots & c_{m1} \dots c_{mm_m} \\ \mathbf{F}_c^{11} & \dots & \mathbf{F}_c^{1y} & \dots & \mathbf{F}_c^{1m} \\ \vdots & & \vdots & & \vdots \\ \mathbf{F}_c^{x1} & \dots & \mathbf{F}_c^{xy} & \dots & \mathbf{F}_c^{xm} \\ \vdots & & \vdots & & \vdots \\ \mathbf{F}_c^{m1} & \dots & \mathbf{F}_c^{my} & \dots & \mathbf{F}_c^{mm} \end{bmatrix} \end{matrix} \tag{4}$$

$a \times a | m < a, \sum_{y=1}^m m_y = a$

$$F_D = \begin{bmatrix} f_{11} & \dots & f_{1y} & \dots & f_{1m} \\ \vdots & & \vdots & & \vdots \\ f_{x1} & \dots & f_{xy} & \dots & f_{xm} \\ \vdots & & \vdots & & \vdots \\ f_{m1} & \dots & f_{my} & \dots & f_{mm} \end{bmatrix}_{m \times m} \tag{5}$$

3.1.4. Phase 4: Establishing an Influential Network Relations Map

By summing the individual columns and rows of the full-influence relations matrix F , we acquire the sum of vectors with all columns and rows, as shown by the following Equations (6) and (7):

$$p_x = \left[\sum_{y=1}^a f_{xy} \right]_{a \times 1}', x \in \{1, 2, \dots, a\} \text{ (Factor } x \text{ influences } f \text{ other factors)} \tag{6}$$

$$q_y = \left[\sum_{x=1}^a f_{xy} \right]_{1 \times a}, y \in \{1, 2, \dots, a\} \text{ (Factor is affected by all other factors)} \tag{7}$$

When $y = x$ (the sum of column and row aggregations means that any factor x influences all other factors, called p_x , and x is affected by all other factors, called q_x . The value $(p_x + q_x)$ shows the

total influence affects received and given by enabler factor x (i.e., representing the degree of effect that this enabler factor x plays in the entire structure, also called “prominence”). In addition, the value $(p_x - q_x)$ states the clear influence of enabler x on this entire method. When $(p_x - q_x)$ has a positive value, then x fits the net cause set. When $(p_x - q_x)$ has a negative value, then x fits the net effect set. Thus, by mapping the dataset of $(p_x + q_x, p_x - q_x)$, we can get the INRM of aspects and factors.

3.2. The DANP Method for Deriving Influential Weights on Aspects and Factors

DANP is applied to the full-influence relations matrix to derive the weight of interdependent relations among aspects/factors by using the concepts and procedures of ANP [23,35,36]. Thus, the value of the weight represents the ratio of factors/aspects and their degree of influence on the whole model that is simultaneously based on a consideration of given and received degrees of influence in a situation. The DANP method includes three major steps, as follows.

3.2.1. Phase 1: Developing an Unweighted Super-Matrix

Developing this unweighted super-matrix $W = (F_C^\rho)'$ can be divided into two steps. The first action normalizes the full-influence relations matrix F_C (i.e., factor value) to obtain the normalized full-influence relations matrix F_C^ρ . The second action transposes the normalized full-influence relations matrix F_C^ρ to obtain $W = (F_C^\rho)'$.

The normalized full-influence relations matrix F_C^ρ (Equation (9)) is obtained by normalizing each row of aspects in the full-relation matrix F_C (Equation (8)), where the sum of each row equals the number of aspects:

$$F_C^\rho = d_x \begin{matrix} D_1 \\ \vdots \\ D_m \end{matrix} \begin{matrix} c_{11} \\ \vdots \\ c_{m1} \\ \vdots \\ c_{m1} \\ \vdots \\ c_{mm} \end{matrix} \begin{bmatrix} F_c^{\rho 11} & \dots & F_c^{\rho 1y} & \dots & F_c^{\rho 1m} \\ \vdots & & \vdots & & \vdots \\ F_c^{\rho x1} & \dots & F_c^{\rho xy} & \dots & F_c^{\rho xm} \\ \vdots & & \vdots & & \vdots \\ F_c^{\rho m1} & \dots & F_c^{\rho my} & \dots & F_c^{\rho mm} \end{bmatrix} \quad (8)$$

$a \times a | m < a, \sum_{y=1}^m m_y = a$

where $F_C^{\rho 11}$, as a normalized example, demonstrates the basic concept of how to normalize actions, as shown in Equations (9) and (10):

$$d_x^{11} = \sum_{y=1}^{m_1} f_{xy}^{11}, x = 1, 2, \dots, m_1 \quad (9)$$

$$F_C^{\rho 11} = \begin{bmatrix} f_{11}^{11}/d_1^{11} & \dots & f_{1y}^{11}/d_1^{11} & \dots & f_{1m_1}^{11}/d_1^{11} \\ \vdots & & \vdots & & \vdots \\ f_{x1}^{11}/d_x^{11} & \dots & f_{xy}^{11}/d_x^{11} & \dots & f_{im_1}^{11}/d_x^{11} \\ \vdots & & \vdots & & \vdots \\ f_{m_1 1}^{11}/d_{m_1}^{11} & \dots & f_{m_1 y}^{11}/d_{m_1}^{11} & \dots & f_{m_1 m_1}^{11}/d_{m_1}^{11} \end{bmatrix} = \begin{bmatrix} f_{11}^{\rho 11} & \dots & f_{1y}^{\alpha 11} & \dots & f_{1m_1}^{\alpha 11} \\ \vdots & & \vdots & & \vdots \\ f_{x1}^{\alpha 11} & \dots & f_{xy}^{\alpha 11} & \dots & f_{xm_1}^{\alpha 11} \\ \vdots & & \vdots & & \vdots \\ f_{m_1 1}^{\alpha 11} & \dots & f_{m_1 y}^{\alpha 11} & \dots & f_{m_1 m_1}^{\alpha 11} \end{bmatrix} \quad (10)$$

Then, the normalized full-influence relations matrix F_c^ρ is transposed to acquire the super-matrix with unweighted $W = (F_c^\rho)'$, as expressed in Equation (11):

$$W = (F_c^\rho)' = \begin{matrix} & \begin{matrix} D_1 & & D_x & & D_m \end{matrix} \\ \begin{matrix} D_1 \\ \vdots \\ D_x \\ \vdots \\ D_m \end{matrix} & \begin{bmatrix} c_{11} & c_{12} & \dots & c_{x1} \dots c_{xm_x} & \dots & c_{m1} \dots c_{mm_m} \\ \vdots & \vdots & & \vdots & & \vdots \\ \vdots & \vdots & & \vdots & & \vdots \\ c_{y1} & c_{y2} & \dots & c_{yx} & \dots & c_{ym} \\ \vdots & \vdots & & \vdots & & \vdots \\ c_{m1} & c_{m2} & \dots & c_{mx} & \dots & c_{mm} \end{bmatrix} \end{matrix} \quad (11)$$

$a \times a | m < a, \sum_{y=1}^m m_y = a$

3.2.2. Phase 2: Synthesizing a Weighted Super-Matrix

This synthesizing stage of the super-matrix with a weighted W^* can also be divided into two steps. The first action normalizes the full-influence relation matrix F_D (i.e., aspect level) (Equation (5)) and transposes it to achieve the normalized full-influence relation matrix F_D^ρ , as shown in Equations (12) and (13). The second action is the normalized full-influence relation matrix F_D^ρ multiplied by this super-matrix with unweighted W ; it is able to present a super-matrix with weighted W^* , as expressed in Equation (14).

$$d_x = \sum_{y=1}^m f_D^{xy}, x = 1, 2, \dots, m \text{ and } f_D^{\rho xy} = f_D^{xy} / d_x, y = 1, 2, \dots, m \quad (12)$$

$$F_D^\rho = \begin{bmatrix} f_D^{11}/d_1 & \dots & f_D^{1y}/d_1 & \dots & f_D^{1m}/d_1 \\ \vdots & & \vdots & & \vdots \\ f_D^{x1}/d_x & \dots & f_D^{xy}/d_x & \dots & f_D^{xm}/d_x \\ \vdots & & \vdots & & \vdots \\ f_D^{m1}/d_m & \dots & f_D^{my}/d_m & \dots & f_D^{mm}/d_m \end{bmatrix}_{m \times m} = \begin{bmatrix} f_D^{\rho 11} & \dots & f_D^{\rho 1y} & \dots & f_D^{\rho 1m} \\ \vdots & & \vdots & & \vdots \\ f_D^{\rho x1} & \dots & f_D^{\rho xy} & \dots & f_D^{\rho xm} \\ \vdots & & \vdots & & \vdots \\ f_D^{\rho m1} & \dots & f_D^{\rho my} & \dots & f_D^{\rho mm} \end{bmatrix}_{m \times m} \quad (13)$$

$$W^\rho = F_D^\rho \times W = \begin{bmatrix} f_D^{\rho 11} \times W^{11} & \dots & f_D^{\rho x1} \times W^{x1} & \dots & f_D^{\rho m1} \times W^{m1} \\ \vdots & & \vdots & & \vdots \\ f_D^{\rho 1y} \times W^{1y} & \dots & f_D^{\rho xy} \times W^{xy} & \dots & f_D^{\rho my} \times W^{mj} \\ \vdots & & \vdots & & \vdots \\ f_D^{\rho 1m} \times W^{1m} & \dots & f_D^{\rho xm} \times W^{xm} & \dots & f_D^{\rho mm} \times W^{mm} \end{bmatrix} \quad (14)$$

3.2.3. Phase 3: Agglomerating the Weighted Super-Matrix

We can use the Markov chain process of ANP to agglomerate the super-matrix with weighted W^* by means of itself numerous times until this super-matrix has become a stable super-matrix to have a sufficiently large power Θ . Hence, the influential ratio values of factors are obtained by $\lim_{\Theta \rightarrow \infty} (W^\rho)^\Theta$. Finally, we obtain a set of influential weights on factors $(w_1, \dots, w_j, \dots, w_n)$ and aspects $(w_1^D, \dots, w_j^D, \dots, w_m^D)$.

3.3. M-VIKOR for Evaluating and Improving Alternative Performance

M-VIKOR is an evaluation technique following the conception of compromise in reaching the best possible outcomes in multicriteria situations. It can be applied to assist decision makers in selecting and ranking options as well as for performance enhancement [23,35,36]. We define the "ideal value" in terms of the "worst value" as the standard and change the normal "max-min" to determine the benchmark. However, VIKOR's negative-ideal and positive-ideal points are determined by the

best score and the worst performance score according to “max-min” factors in real-world situations. Because VIKOR cannot show gaps in the enhancement of alternatives, we modified it so that the normal maximum and minimum are the negative ideal points, with the points being an ideal value as well as the worst value for alternative selection and enhancement. Thus, by using the M-VIKOR “ideal-word” for the normalized class distance utility, being near the ideal value and far from the worst value is a good outcome in these real-world situations [23,35,36,38–42].

3.3.1. Phase 1: Determining the Negative/Positive Ideal Point Based on Ideal Values and Worst Values

The normal VIKOR method sets the positive-ideal point $u_x^* = \max_k\{u_{kx}|k = 1, 2, \dots, m\}$ and the negative ideal point $u_x^- = \min_k\{u_{kx}|k = 1, 2, \dots, m\}$ in k alternatives. The positive-ideal point $u_x^* = \max_k\{u_{kx}|k = 1, 2, \dots, m\}$ and the negative ideal point $u_x^- = \min_k\{u_{kx}|k = 1, 2, \dots, m\}$ are set as follows (Equations (15) and (16)):

$$u_x^* = \left\{ \begin{array}{l} \max_k\{u_{kx}|k = 1, 2, \dots, m\}, \text{ for benefit attributes} \\ \min_k\{u_{kx}|k = 1, 2, \dots, m\}, \text{ for cost attributes} \end{array} \right\}, \quad x = 1, 2, \dots, a \quad (15)$$

$$u_x^- = \left\{ \begin{array}{l} \min_x\{u_{kx}|k = 1, 2, \dots, m\}, \text{ for benefit attributes} \\ \max_x\{u_{kx}|k = 1, 2, \dots, m\}, \text{ for cost attributes} \end{array} \right\}, \quad x = 1, 2, \dots, a \quad (16)$$

In this study, we used questionnaires in which the scored responses range from 0 to 10: totally dissatisfied (0) to extremely satisfied (10). We set the ideal value at 10 (i.e., $u_y^* = 10$ as the positive-ideal point) and the worst value at 0 (i.e., $u_y^- = 0$ as positive-ideal point) in each factor x , respectively. The basic concept differs from the traditional method as follows:

The vector of ideal value (Equation (17)):

$$\mathbf{u}^{aspired} = (u_1^{aspired}, \dots, u_x^{aspired}, \dots, u_a^{aspired}) = (10, \dots, 10, \dots, 10) \quad (17)$$

The vector of worst value (Equation (18)):

$$\mathbf{u}^{worst} = (u_1^{worst}, \dots, u_y^{worst}, \dots, u_a^{worst}) = (0, \dots, 0, \dots, 0) \quad (18)$$

3.3.2. Phase 2: Obtaining the Mean of the Minimal Gap of the Maximal Regret and Group Utility on Each Alternative

The purpose of this phase is to compute the minimal average gap of the group utility S_k and the maximal gap for all factors or aspects in order to give the highest priority to the enhancement sequence T_k (Equations (19) and (20)):

$$L_k^{g=1} = S_k = \sum_{x=1}^a w_x r_{kx} = \sum_{x=1}^a w_x \left(\frac{|u_x^{aspired} - u_{kx}|}{|u_x^{aspired} - u_x^{worst}|} \right) \quad (19)$$

$$L_k^{g=\infty} = T_k = \max_x \{r_{kx}|x = 1, 2, \dots, a\} \quad (20)$$

where $r_{kx} = \left(\frac{|u_x^{aspired} - u_{kx}|}{|u_x^{aspired} - u_x^{worst}|} \right)$ represents the gap ratio of performance; S_k indicates the average gap ratios of the ideal value $u_x^{aspired}$ to the value of performance u_{kx} in factor x of alternative k ; w_x indicates the relative influential weight of factor x (or aspect x), where w_x is obtained via the DANP method; and T_k represents the maximal performance gap in all the factors or aspects for prioritizing enhancement within alternative k . It is possible that the M-VIKOR method can also be used to solve only one alternative in terms of the gap in performance enhancement: closing the gap between zero and the ideal value.

3.3.3. Phase 3: Providing a Comprehensive Indicator of Each Alternative

The comprehensive score of each alternative H_k is finally integrated by Equation (21). When the value is combined in the influential network relations map (INRM), we can observe how each alternative is enhanced to decrease the gaps in factors in order to achieve the ideal value:

$$H_k = v \frac{S_k - S^{aspired}}{S^{worst} - S^{aspired}} + (1 - v) \frac{T_k - T^{aspired}}{T^{worst} - T^{aspired}} \quad (21)$$

where $S^{aspired} = 0$ (i.e., achieving the ideal value of group utility S_k), $S^{worst} = 1$ (i.e., the worst situation of S_k), $T^{aspired} = 0$ (i.e., achieving the ideal value of maximal regret T_k), and $T^{worst} = 1$ (i.e., the worst situation of T_k). Thus, Equation (21) can be rewritten as Equation (22):

$$H_k = vS_k + (1 - v)T_k \quad (22)$$

where v is the weight for the decision-making perspective (i.e., $v = 1$ is only considered in how to minimize the group utility S_k ; $v = 0$ is only considered in how to choose the maximum gap for previous enhancement T_k ; and $v = 0.5$ is considered for both the group utility S_k and the maximum gap T_k).

4. Research Methods

In this section, our proposed hybrid MCDM model was applied in a case study on the implementation of a legal AI bot in Taiwan. The case study illustrates how the hybrid MCDM model can be used to assist administrators in understanding and enhancing their own attitude toward this type of legal AI bot and in realizing users' behavior and attitudes toward it.

4.1. Data Collection

Between April and May, 2020, interviews and questionnaires were applied to collect data from 36 experts (10 AI judges, 12 lawyers, and 14 AI experts) who understand and have an interest in the development of AI, law, or AI robots and who had worked at least 10 years for related work experiences. In order to ensure the smooth progress of data collection, this study firstly applied a matrix filling technique to conduct the pre-investigation and trial filling. The response from filling in the matrix was that it was not easy for experts to compare the name and code of individual factors, such as filling in the matrix. Hence, this study enhances this procedure of fulfilling the study via designing a survey like a Likert scale and clarifying the corresponding instructions and conceptions in detail so the experts can seriously and easily fill in the survey. In the analysis, we use the tool "Microsoft Office Excel 2016" for computations. The significance confidence is 99.05%, and the gap error is only 0.95%, which is less than 1% and greater than 95% consensus. Each survey needed between 40 and 50 minutes to complete.

4.2. Using DEMATEL to Develop INRM

This study used DEMATEL to investigate how to adopt a legal AI bot according to the 14 factors referring to four aspects, as discussed above. From the surveys, we obtained matrix F , giving the total influence for the four aspects and 14 factors. These are shown in Tables 2 and 3, respectively. We developed the ideas and estimations of the users in the four aspects and found how the extent of the influence is associated with other aspects in Table 2. Based on the degree of total influence ($p_x + q_x$), $TRB(A_3)$ has the strongest effect on the strength of the relationship; this was the most significant effect. On the contrary, $IR(A_4)$ has the least influence. Based on the relationship of influence ($p_x - q_x$), we also determine that $IR(A_4)$ has the strongest direct influence on other aspects and that $TRB(A_3)$ is the worst direct influence.

Table 2. The sum of effects on aspects and total effect matrix of F_D .

Aspects	A_1	A_2	A_3	A_4	p_x	q_x	p_x+q_x	p_x-q_x
ARBA ₁	0.458	0.431	0.492	0.373	1.755	1.796	3.551	−0.042
PBCA ₂	0.435	0.376	0.451	0.338	1.600	1.620	3.220	−0.020
TRBA ₃	0.473	0.428	0.462	0.350	1.713	1.841	3.553	−0.128
IR A ₄	0.431	0.385	0.436	0.336	1.588	1.398	2.987	0.190

Table 3. The total effect matrix of F_C for factors.

Factors	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}
f_1	0.442	0.496	0.486	0.448	0.457	0.454	0.495	0.530	0.501	0.418	0.399	0.386	0.366	0.373
f_2	0.499	0.399	0.466	0.426	0.427	0.436	0.469	0.510	0.477	0.403	0.383	0.372	0.348	0.355
f_3	0.489	0.456	0.386	0.414	0.408	0.414	0.471	0.500	0.476	0.392	0.374	0.354	0.335	0.345
f_4	0.449	0.422	0.418	0.327	0.381	0.387	0.432	0.457	0.428	0.358	0.352	0.340	0.310	0.322
f_5	0.486	0.454	0.444	0.411	0.360	0.437	0.475	0.497	0.476	0.379	0.366	0.354	0.339	0.353
f_6	0.430	0.411	0.397	0.370	0.391	0.321	0.418	0.452	0.422	0.346	0.339	0.311	0.304	0.305
f_7	0.480	0.459	0.461	0.417	0.432	0.424	0.392	0.502	0.477	0.369	0.356	0.345	0.325	0.332
f_8	0.496	0.473	0.466	0.428	0.441	0.436	0.483	0.427	0.486	0.378	0.361	0.348	0.319	0.336
f_9	0.493	0.463	0.462	0.417	0.435	0.423	0.480	0.505	0.403	0.387	0.374	0.350	0.336	0.335
f_{10}	0.481	0.459	0.465	0.429	0.419	0.410	0.448	0.480	0.455	0.333	0.373	0.376	0.361	0.351
f_{11}	0.493	0.469	0.465	0.420	0.416	0.420	0.458	0.497	0.474	0.394	0.323	0.381	0.353	0.355
f_{12}	0.499	0.474	0.471	0.423	0.431	0.422	0.488	0.519	0.495	0.420	0.396	0.320	0.362	0.359
f_{13}	0.387	0.359	0.359	0.333	0.333	0.332	0.359	0.392	0.384	0.341	0.317	0.317	0.236	0.294
f_{14}	0.384	0.354	0.354	0.324	0.329	0.328	0.352	0.386	0.358	0.321	0.310	0.296	0.281	0.233

Note: $z = 36$ denotes the number of users, f_{ij}^p is the average influence of x factor on y , and a denotes the number of factors; here, $a = 14$ and $a \times a$ is a matrix. $\frac{1}{a^2} \sum_{x=1}^a \sum_{y=1}^a \frac{|f_{ij}^x - f_{ij}^{x-1}|}{f_{ij}^p} \times 100\% = 0.95\% < 5\%$; the significant confidence is 99.05%.

According to the total effect matrix, we assess how each of the influencing factors are related to individual factors (see Table 3). This illustrates the extent of indirect or direct effects and contrasts them with the other factors in Table 4. $PU(f_1)$ is the most significant factor for consideration; moreover, $IB(f_{14})$ has the smallest effect on the other factors. Table 4 also shows that $UB(f_{10})$ has the strongest effect on the other factors and that $TFC(f_6)$ is the most strongly affected by other factors.

4.3. Using the DANP Model for Analyzing the Influential Weights

We applied DEMATEL to determine the most influential relationships among the factors and to acquire the most accurate weightings. The objective of DANP is to explain the feedback regarding the interdependence and interrelationships among factors. Hence, we developed this quality estimation model by applying the DEMATEL method according to the concepts of ANP, so that our DANP could determine the weight of influence of each factor (see Tables 4 and 5).

We also considered whether these important factors in user behavior are compatible with legal AI bot $SA(f_8)$, $PU(f_1)$, and $TB(f_9)$. In addition, the weights of influence are integrated with the DEMATEL method to evaluate the significance of problem-solving according to the gaps recognized by using the M-*VIKOR* technique and INRM (shown as Figure 2).

Table 4. The weights, the sum of effects, and ranking per factor.

Aspects/Factors		p_x	q_x	p_i+q_i	p_x-q_x	Influential Weights (Global Weights)
ARB	A_1					0.270
PU	f_1	1.423	1.430	2.853	-0.006	0.094
PEOU	f_2	1.365	1.351	2.716	0.014	0.089
Complexity	f_3	1.331	1.338	2.669	-0.007	0.088
PBC	A_2					0.244
SE	f_4	1.095	1.108	2.203	-0.012	0.080
RFC	f_5	1.207	1.132	2.339	0.075	0.082
TFC	f_6	1.082	1.145	2.227	-0.063	0.081
TRB	A_3					0.277
DT	f_7	1.371	1.355	2.726	0.017	0.090
SA	f_8	1.395	1.434	2.829	-0.038	0.096
TB	f_9	1.388	1.367	2.755	0.022	0.091
IR	A_4					0.210
UB	f_{10}	1.809	3.603	-0.016	1.809	0.045
VB	f_{11}	1.719	3.526	0.089	1.719	0.043
RB	f_{12}	1.690	3.547	0.166	1.690	0.042
TB	f_{13}	1.593	3.099	-0.088	1.593	0.039
IB	f_{14}	1.593	3.034	-0.151	1.593	0.040

4.4. Using M-VIKOR for Assessing the Total Gaps

We used M-VIKOR to enhance legal AI bot services and to estimate the total accreditation gaps in users' behavior at the intention and adoption stages, as shown in Table 5. Administrators can classify problem-solving topics followed by the integrated index from this aspect of the factors as individual aspects.

Applying these indices to the four aspects and 14 factors, gaps in values can be evaluated by means of the priority sequence enhancement for attaining the ideal values. $TB(f_{13})$ with a larger gap (0.750) at the intention stage is the primary factor to be enhanced, followed by $IB(f_{14})$ and $UB(f_{10})$. Of all the factors, administrators of legal advisory institutions are the most focused on $TB(f_{13})$ (tradition barrier) at the intention step; $TB(f_{13})$ with a larger gap (0.625) is the primary factor to be enhanced in the adoption step, followed by $IB(f_{14})$ and $TFC(f_6)$. Supervisors pay the most attention to $TB(f_{13})$ (tradition barrier) in the adoption stage. The findings show the enhancement priority sequence required for the overall factors to achieve the ideal value, from the most to the least significant factors.

Priorities for enhancement can also be used for individual aspects. In $ARB(A_1)$, for example, the sequence of values of the priority gap is complexity (f_3), PEOU (f_2), and PU (f_1). In $PBC(A_2)$ of the intention stage, the sequence of values of the priority gap is TFC (f_6), SE (f_4), and RFC (f_5). In $TRB(A_3)$ of the intention stage, the sequence of the enhancement priorities is SA (f_8), DOT (f_7), and TB (f_9). In $IR(A_4)$ of the intention stage, the sequence of the enhancement priorities is TB (f_{13}), IB (f_{14}), UB (f_{10}), VB (f_{11}), and RB (f_{12}). In the adoption stage, the sequence of the enhancement priorities is (f_2), (f_3), and (f_1) in $ARB(A_1)$; (f_6), (f_4), and (f_5) in $PBC(A_2)$; (f_9), (f_7), and (f_8) in $TRB(A_3)$; and (f_{13}), (f_{14}), (f_{10}), (f_{11}), and (f_{12}) in $IR(A_4)$. Applying the values of gaps offered by the sample of users, these enhancement primacy schemes are comprehensive and unique, both in terms of their separate aspects and overall (see Table 5). Administrators will be able to understand users' behavior in adopting legal AI bots and to recognize the gaps in the stages (of multiple intention and adoption).

Table 5. The evaluation of legal artificial intelligence (AI) bot for multiple stages by M-VIKOR.

Aspects/Factors		Local Weight	Global Weight (DANP)	Legal AI Bot Gap (h_{kj})	
				Intention (H_1)	Adoption (H_2)
ARB	A_1	0.270		0.314	0.206
PU	f_1	0.347	0.094	0.175	0.100
PEOU	f_2	0.328	0.089	0.375	0.275
Complexity	f_3	0.325	0.088	0.400	0.250
PBC	A_2	0.244		0.508	0.410
SE	f_4	0.330	0.080	0.500	0.379
RFC	f_5	0.336	0.082	0.400	0.375
TFC	f_6	0.334	0.081	0.625	0.475
TRB	A_3	0.277		0.413	0.209
DOT	f_7	0.325	0.090	0.413	0.225
SA	f_8	0.347	0.096	0.425	0.175
TB	f_9	0.329	0.091	0.400	0.229
IR	A_4	0.210		0.608	0.419
UB	f_{10}	0.215	0.045	0.700	0.350
VB	f_{11}	0.207	0.043	0.500	0.325
RB	f_{12}	0.199	0.042	0.375	0.225
TB	f_{13}	0.188	0.039	0.750	0.625
IB	f_{14}	0.191	0.040	0.725	0.600
S_A		Total gaps		0.450	0.301

4.5. Results and Discussion

From our DEMATEL method, we have identified the interrelationships between factors or aspects by applying IRNM (as shown in Figure 2). As shown in Figure 2, $IR(A_4)$ affects other aspects like $PBC(A_2)$, $ARB(A_1)$, and $TRB(A_3)$. It can be seen that $IR(A_4)$ plays a significant role and has the strongest effect on the other aspects. Hence, administrators need to focus on enhancing this aspect, followed by $PBC(A_2)$, $ARB(A_1)$, and $TRB(A_3)$ sequentially, when evaluating the behavior of users and improving their implementation of legal AI bots.

After investigating the aspects, we next identified the factors considered in all aspects. Based on these outcomes, we show IRNM of the factors in Figure 2. When considering the relationships of influence among the factors, in the ARB aspect, it was shown that $PEOU(f_2)$ was the most influential factor and should be the first to be enhanced, followed by $PU(f_1)$ and complexity (f_3) (see Figure 2: the causal relationship A_1). In the PBC aspect, $RFC(f_5)$ was the most influential factor and is the most important to be enhanced, followed by $SE(f_4)$ and $TFC(f_6)$ (see Figure 2: causal relationships in A_2). In the TRB aspect, $TB(f_9)$ was the most influential factor and is the most important to enhance, followed by $DOT(f_7)$ and $SA(f_8)$ (see Figure 2 causal relationships in A_3); in the IR aspect, $RB(f_{12})$ was the most influential factor and is the most important to enhance, followed by $VB(f_{11})$, $UB(f_{10})$, $TB(f_{13})$, and $IB(f_{14})$ (see Figure 2: causal relationships in A_4).

The findings of the aspects and factors offer crucial information for understanding user behavior and what will affect their use of legal AI bots for sustainable development at institutions of legal advisory. Administrators need to consider all the aspects and factors presented in Figure 2. Although the estimation technique can be applied at legal advisory institutions and nonacademic real-life situations, administrators of the former will need to bear in mind that some modifications of the model will need to be applied at individual institutions. Because the value of significance for the 14 factors can change according to the particular situation and user behavior, supervisors will need to consider the typical behavior of their users before determining the ideal implementation technique.

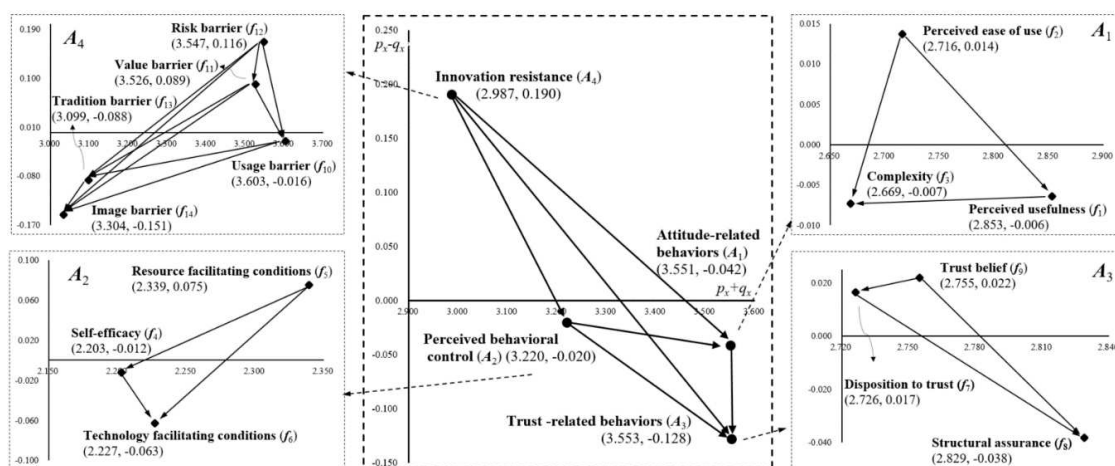


Figure 2. The influential network relationships map (INRM) per aspect and factor.

The most significant factor identified by means of DANP when estimating a legal AI bot and that affects users’ decisions was TRB(A_3) weighted at 0.277 in the aspects of SA(f_8) and PU(f_1) and weighted at 0.096 and 0.094 in the factors (see Table 5). Trust is a significant element that determines the key to any connection. Trust is formed when a user believes in the integrity and reliability of the other party. Trust is a main element in a reciprocal connection [43]. Structural assurance represents a belief in the guarantees/promises, encryptions, regulations, and other processes of a new legal AI bot, the expectations caused via the user in uncertain surroundings, and their effects on significant events. This study found that the interviewees gave a certain particular psychological response to legal AI bots and formed a one-way emotional bond: trust [44]. Hence, the trust of users in legal AI bots forms the structural assurance of robots in legal services. On the other hand, legal AI bots can help improve personnel performance, can enhance operational efficiency, and can reduce costs. These experts contended that a legal AI bot is helpful and efficient. They agreed that resolving the legal issues of users quickly, conveniently, and at a lower cost establishes the fundamental facet of perceived usefulness, which is an important influence issue [4,16]. “Structural assurance” and “perceived usefulness” are therefore the most significant factors when evaluating legal AI bot implementation.

The overall gap values in Table 5 that show room for enhancement are 0.450 in the intention stage and 0.301 in the adoption stage. From the stages, IR(A_4) featured the largest gap (0.608) in the intention stage while IR(A_4) featured the largest gap value (0.419) in the adoption stage; clearly, it needs to be a priority for enhancement if administrators wish to attain an ideal value. In terms of long-term improvement, administrators should carefully consider their intentions regarding introducing legal AI bots for the reasons given above. Assessing legal AI bots according to users’ behavior by means of a multiple-stage pattern, as offered by this approach, can be introduced to legal advisory institutions. However, managers need to be cautious about using this pattern because the significance of these 14 factors may vary according to the situation. Supervisors need to associate the legal AI bot with users’ behavior and to describe this difference before judging whether this would be the ideal service to offer.

5. Conclusions

Legal AI bots have a significant role to play at legal advisory institutions, but the strategies for their use are complex and there is not always overall clarity on how they should be implemented for sustainable development. Different situations may require different conditions for their use. Based on previous research and the opinion of experts, we established four aspects with 14 factors that align with legal AI bot implementation according to user behavior. We used an integrated MCDM model, DDANPV, which is very powerful technique, and a combination of DEMATEL, DANP, and M-*VIKOR* in a case study. The key motivations among these various methods are available for conflict resolution. When various criteria are to be considered, integrated MCDM is one of the most widespread methods.

M-VIKOR is an MCDM technique that is based on assessing established criteria and on reaching a compromise for generating the best solution. VIKOR ranks the criteria to establish the solution that is closest to the ideal for sustainable development.

In our decision-making procedure, we applied weightings to local and global alternatives to allow the leaders of legal advisory institutions to choose the features that would best assist them at implementing legal AI bots for sustainable development. We have not only chosen the best elements but also have established how to narrow gaps to attain the ideal values for legal AI bots. It is argued that the methodology used in this research is capable of handling intricate problems related to the sustainable development of legal AI bots. This study not only has deep significance for related specialists but also offers an adequate and feasible approach to the sustainable development of legal AI bots under an approach that offers management support when targeting enhancement of legal AI bot usage.

The limitations of this paper offer direction for future research. The primary data were obtained from a limited number of users in the field of legal AI bots. Though adjusted, there were some dissimilar estimations in the assembly of the data of the primary matrix of influence due to variances in specialized viewpoints. Further research will be able to expand the channels and scope to obtain more extensive primary data, which will increase the accuracy of the final results. Following further developments in the implementation of legal AI bots, it will be necessary to undertake more studies in the field. Research should investigate the core reasons for variances in order to fully understand the interrelationships among a wide range of factors.

Author Contributions: Methodology, M.-T.L.; Formal analysis, M.-T.L.; Investigation, J.-H.H.; Writing—original draft, M.-L.T. and J.-H.H.; Writing—review & editing, G.-G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was not funded.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Boynton, S. DoNotPay, 'World's First Robot Lawyer,' Coming to Vancouver to Help Fight Parking Tickets. *Global News*. 1 November 2017. Available online: <https://globalnews.ca/news/3838307/donotpay-robot-lawyer-vancouverparking-tickets> (accessed on 22 June 2020).
2. Aguilo-Regla, J. Introduction: Legal Informatics and the Conceptions of the Law. In *Law and the Semantic Web*; Benjamins, R.P., Casanovas, J., Gangemi, A., Eds.; Springer: Berlin, Germany, 2005; pp. 18–24.
3. Bench-Capon, T.; Araszkievicz, M.; Ashley, K.; Atkinson, K.; Bex, F.; Borges, F.; Wyner, A.Z. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artif. Intell. Law* **2012**, *20*, 215–319. [CrossRef]
4. Xu, N.; Wang, K.J. Adopting robot lawyer? The extending artificial intelligence robot lawyer technology acceptance model for legal industry by an exploratory study. *J. Manag. Organ.* **2019**, *13*, 1–19. [CrossRef]
5. Hilt, K. What does the future hold for the law librarian in the advent of artificial intelligence? *Can. J. Inf. Lib. Sci.* **2017**, *41*, 211–227.
6. Adamski, D. Lost on the digital platform: Europe's legal travails with the digital single market. *Common Mkt. Law Rev.* **2018**, *55*, 719–751.
7. Goodman, J. Meet the AI Robot Lawyers and Virtual Assistants. 2016. Available online: <https://www.lexisnexis-es.co.uk/assets/files/legal-innovation.pdf> (accessed on 22 June 2020).
8. Papakonstantinou, V.; De Hert, P. Structuring modern life running on software. Recognizing (some) computer programs as new "digital persons". *Comput. Law Secur. Rev.* **2018**, *34*, 732–738. [CrossRef]
9. Alarie, B.; Niblett, A.; Yoon, A.H. How artificial intelligence will affect the practice of law. *Univ. Toronto Law J.* **2018**, *68*, 106–124. [CrossRef]
10. Castell, S. The future decisions of RoboJudge HHJ Arthur Ian Blockchain: Dread, delight or derision? *Comput. Law Secur. Rev.* **2018**, *34*, 739–753. [CrossRef]
11. D'Amato, A. Can/should computers replace judges. *Georgia Law Rev.* **1976**, *11*, 1277.

12. Von der Lieth Gardner, A. *An Artificial Intelligence Approach to Legal Reasoning*; MIT Press: Cambridge, MA, USA, 1987.
13. McGinnis, J.O.; Pearce, R.G. The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services. *Fordham Law Rev.* **2014**, *82*, 3041–3066. [CrossRef]
14. Almaiah, M.A. Acceptance and usage of a mobile information system services in University of Jordan. *Educ. Inf. Technol.* **2018**, *23*, 1873–1895. [CrossRef]
15. Roca, J.C.; Chiu, C.M.; Martinez, F.J. Understanding e-learning continuance intention: An extension of the Technology Acceptance Model. *Int. J. Hum. Comput. Stud.* **2006**, *64*, 683–696. [CrossRef]
16. Sarkar, S.; Chauhan, S.; Khare, A. A meta-analysis of antecedents and consequences of trust in mobile commerce. *Int. J. Inf. Manag.* **2020**, *50*, 286–301. [CrossRef]
17. Kim, J.B. An empirical study on consumer first purchase intention in online shopping: Integrating initial trust and TAM. *Electron. Commer. Res.* **2012**, *12*, 125–150. [CrossRef]
18. Gefen, D.; Karahanna, E.; Straub, D.W. Trust and TAM in online shopping: An integrated model. *MIS Q.* **2003**, *27*, 51–90. [CrossRef]
19. Abroud, A.; Choong, Y.V.; Muthaiyah, S.; Fie, D.Y.G. Adopting e-finance: Decomposing the technology acceptance model for investors. *Serv. Bus.* **2015**, *9*, 161–182. [CrossRef]
20. Holder, C.; Khurana, V.; Harrison, F.; Jacobs, L. Robotics and law: Key legal and regulatory implications of the robotics age (Part I of II). *Comput. Law Secur. Rev.* **2016**, *32*, 383–402. [CrossRef]
21. Greenleaf, G.; Mowbray, A.; Chung, P. Building sustainable free legal advisory systems: Experiences from the history of AI & law. *Comput. Law Secur. Rev.* **2018**, *34*, 314–326.
22. Rogers, E.M. *The Diffusion of Innovations*, 4th ed.; Free Press: New York, NY, USA, 1995.
23. Lu, M.T.; Tzeng, G.H.; Cheng, H.; Hsu, C.C. Exploring mobile banking services for user behavior in intention adoption: Using new hybrid MADM model. *Serv. Bus.* **2015**, *9*, 541–565. [CrossRef]
24. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, *13*, 319–340. [CrossRef]
25. Moore, G.C.; Benbasat, I. Development of an instrument to measure the perceptions of adopting an information technology innovation. *Inf. Syst. Res.* **1991**, *2*, 192–222. [CrossRef]
26. Bandura, A. *Social Foundations of Thought and Action: A Social Cognitive Theory*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1986.
27. Taylor, S.; Todd, P. Decomposition and crossover effects in the theory of planned behavior: A study of consumer adoption intentions. *Int. J. Res. Mark.* **1995**, *12*, 137–155. [CrossRef]
28. McKnight, D.H.; Choudhury, V.; Kacmar, C. Developing and validating trust measures for e-commerce: An integrative typology. *Inf. Syst. Res.* **2002**, *13*, 344–359. [CrossRef]
29. McKnight, D.H.; Chervany, N.L. What trust means in e-commerce customer relationships: An interdisciplinary conceptual typology. *Int. J. Electron. Commun.* **2001**, *6*, 35–59. [CrossRef]
30. Ma, L.; Lee, C.S. Understanding the barriers to the use of MOOCs in a developing country: An innovation resistance perspective. *J. Educ. Comput. Res.* **2019**, *57*, 571–590. [CrossRef]
31. Fain, D.; Roberts, M.L. Technology vs. consumer behavior: The battle for the financial services customer. *J. Direct Mark.* **1997**, *11*, 44–54. [CrossRef]
32. Kuisma, T.; Laukkanen, T.; Hiltunen, M. Mapping the reasons for resistance to internet banking: A means-end approach. *Int. J. Inf. Manag.* **2007**, *27*, 75–85. [CrossRef]
33. Laukkanen, T.; Lauronen, J. Consumer value creation in mobile banking services. *Int. J. Mobile Commun.* **2005**, *3*, 325–338. [CrossRef]
34. Mohammadi, M.; Rezaeia, J. Ensemble ranking: Aggregation of rankings produced by different multi-criteria decision-making methods. *Omega* **2020**, *96*, 102254. [CrossRef]
35. Lu, M.T.; Hsu, C.C.; Liou, J.J.H.; Lo, H.W. A hybrid MCDM and sustainability-balanced scorecard model to establish sustainable performance evaluation for international airports. *J. Air Transp. Manag.* **2018**, *71*, 9–19. [CrossRef]
36. Lu, M.T.; Lin, S.W.; Tzeng, G.H. Improving RFID adoption in Taiwan’s healthcare industry based on a DEMATEL technique with a hybrid MCDM model. *Decis. Support Syst.* **2013**, *56*, 259–269. [CrossRef]
37. Feng, G.C.; Ma, R. Identification of the factors that influence service innovation in manufacturing enterprises by using the fuzzy DEMATEL method. *J. Clean. Prod.* **2020**, *253*, 120002. [CrossRef]



38. Opricovic, S.; Tzeng, G.H. Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *Eur. J. Oper. Res.* **2004**, *156*, 445–455. [CrossRef]
39. Opricovic, S.; Tzeng, G.H. Extended VIKOR method in comparison with outranking methods. *Eur. J. Oper. Res.* **2007**, *178*, 514–529. [CrossRef]
40. Acuña-Soto, C.M.; Liern, V.; Pérez-Gladish, B. A VIKOR-based approach for the ranking of mathematical instructional videos. *Manag. Decis.* **2019**, *57*, 501–522. [CrossRef]
41. Kumar, A.; Aswin, A.; Gupta, H. Evaluating green performance of the airports using hybrid BWM and VIKOR methodology. *Tour. Manag.* **2020**, *76*, 103941. [CrossRef]
42. Garg, C.P.; Sharma, A. Sustainable outsourcing partner selection and evaluation using an integrated BWM–VIKOR framework. *Environ. Dev. Sustain.* **2020**, *22*, 1529–1557. [CrossRef]
43. Nora, L. Trust, commitment, and customer knowledge: Clarifying relational commitments and linking them to repurchasing intentions. *Manag. Decis.* **2019**, *57*, 3134–3158. [CrossRef]
44. Arbib, M.A.; Fellous, J.M. Emotions: From brain to robot. *Trends Cogn. Sci.* **2004**, *8*, 554–561. [CrossRef]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

Artificial Intelligence Applications in Military Systems and Their Influence on Sense of Security of Citizens

Marta Bistron *  and Zbigniew Piotrowski 

Faculty of Electronics, Military University of Technology, 00-908 Warsaw, Poland;
zbigniew.piotrowski@wat.edu.pl

* Correspondence: marta.bistron@wat.edu.pl

Abstract: The paper presents an overview of current and expected prospects for the development of artificial intelligence algorithms, especially in military applications, and conducted research regarding applications in the area of civilian life. Attention was paid mainly to the use of AI algorithms in cybersecurity, object detection, military logistics and robotics. It discusses the problems connected with the present solutions and how artificial intelligence can help solve them. It briefly presents also mathematical structures and descriptions for ART, CNN and SVM networks as well as Expectation–Maximization and Gaussian Mixture Model algorithms that are used in solving of discussed problems. The third chapter discusses the attitude of society towards the use of neural network algorithms in military applications. The basic problems related to ethics in the application of artificial intelligence and issues of responsibility for errors made by autonomous systems are discussed.

Keywords: neural networks; artificial intelligence; AI in military; CNN; social robots

Citation: Bistron, M.; Piotrowski, Z. Artificial Intelligence Applications in Military Systems and Their Influence on Sense of Security of Citizens. *Electronics* **2021**, *10*, 871. <https://doi.org/10.3390/electronics10070871>

Academic Editors: Imre J. Rudas and Jun Liu

Received: 16 February 2021

Accepted: 2 April 2021

Published: 6 April 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

One of the main pillars determining the position of a state in the international arena is its military potential. As defined by the U.S. Department of Defense, military capability means [1] “the ability to achieve a specified wartime objective (win a war or battle, destroy a target set).” Military capability is determined by structure, modernization, readiness and sustainability. The level of modernization depends mainly on technical sophistication, weapon systems and equipment.

A typical war known from the Second World War is slowly fading into oblivion and goes into cyberspace. As research shows [2] hacker attacks on both private companies and government institutions have become a common phenomenon. According to researchers [3,4], artificial intelligence (AI) and innovative automatic systems will become an inseparable element of future armed conflicts.

Most modern AI algorithms require large amount of data [5] for example AI algorithms to natural language processing [6]. They can work better, faster and more efficient, but they cannot work well without access to large databases. Access to extensive data sources and the increasing computing power of machines have enabled the development of this field of science. Nowadays, interest in using neural networks is still growing, which can be observed by analyzing scientific publications on various topics from the last few years—development of ITS (Intelligent Transport Systems) [7,8], prediction and evaluation of atmospheric phenomena [9–11], distinguish information tweets (containing relevant facts) from non-information ones (containing rumors or non-detailed information) [12] and predicting dynamic FX markets [13] and the real estate market [14]. In military sector, AI algorithms can be used, among others, for speech recognition systems [15] or object detection and recognition [16].

Artificial intelligence has a wide range of applications, resulting in its enormous and multi-faceted impact on society. In recent years AI has taken part in creating new standards of social behavior, people-to-people contacts and even politics and functioning

of state, what the authors write about in [17–19]. New technologies give hope but also inspire fear, which is discussed in more detail later in this paper. The rest of the paper is organized as follows: Chapter 2 presents used methodology; Chapter 3 discusses examples of applications based on the use of artificial intelligence in military; Chapter 4 is dedicated to the impact of using these algorithms in military on state politics, state defense and sense of security of citizens; Chapter 5 shows conclusions relating to the overview.

2. Methodology

Firstly, selected examples of artificial intelligence algorithms used in military systems are presented. Said examples are related to areas of the army particularly important in terms of ensuring the proper functioning and the security of the state and all citizens and play a crucial role in conducting modern combat activities on the battlefield. The used taxonomy (showed in detailed way in Figure 1) presents the overview of literature coming from the type of military applications or area of applications to the different algorithms of artificial intelligence used to solve this problem. All areas and specific algorithms are shortly described in order to better understanding of the development of AI in military.

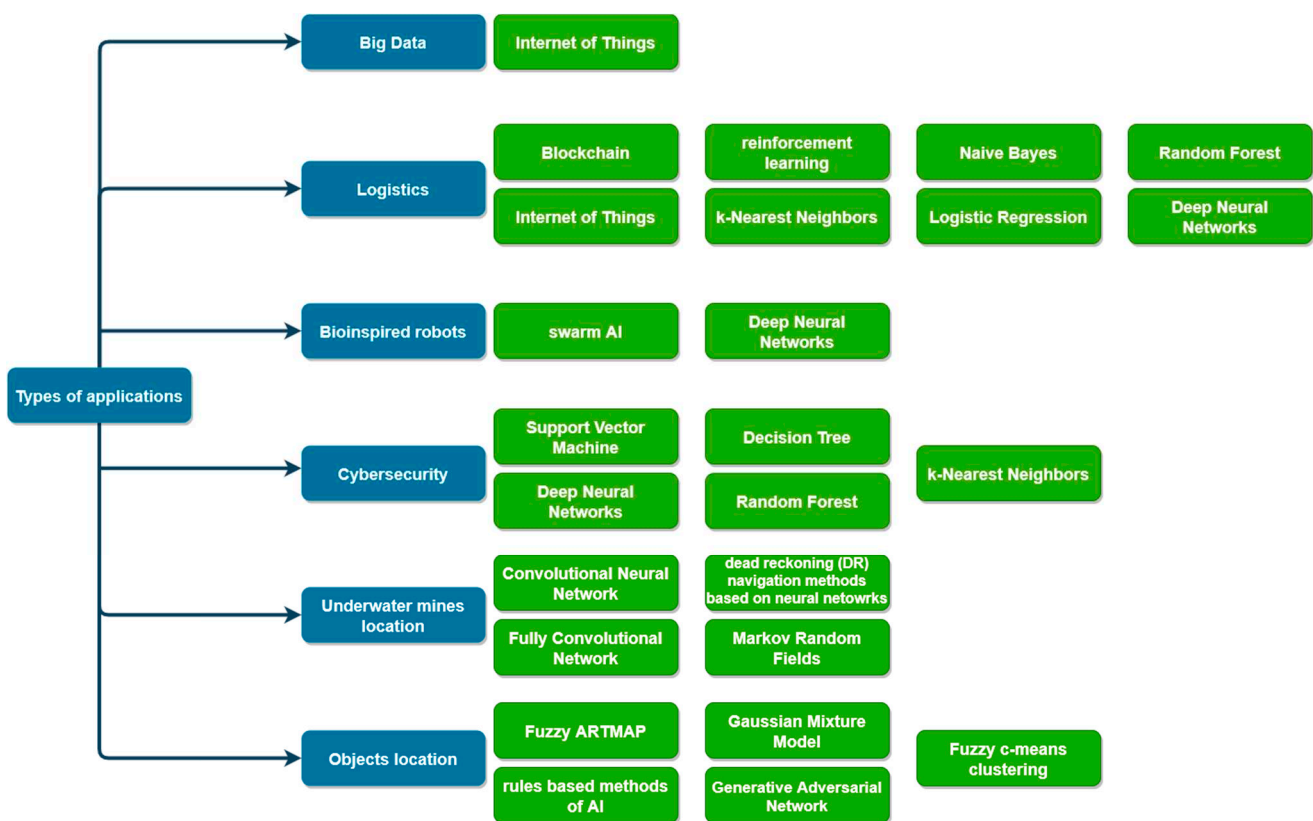


Figure 1. Taxonomy proposed in the overview of military applications.

With reference to the described military applications, an analysis of the main concerns of the society regarding the development of AI in the context of ethics and social behavior was performed. The surveys conducted in 2011 and 2019 on the attitudes of citizens from various social groups to the development of artificial intelligence algorithms in various areas and applications were compared.

3. Practical Using of Neural Networks in Military Applications

Artificial neural networks in military applications have great potential in every field; they can provide support during land, sea, air and information warfare. Artificial intelligence finds military applications in logistics, transport, armed attack analysis and in

communication, which was presented in the report [20]. The existence of a high demand for applications using AI in the defense sector is confirmed by the AIE (Artificial Intelligence Exploration) program launched by DARPA (Defense Advanced Research Projects Agency) in July 2018 [21], which is a key element of the organization's investment strategy focused on providing advantage to the United States in this area of technology. Moreover, the European Defence Agency supports the use of artificial intelligence in the field of defense, especially for tasks related to the processing of large amounts of data [22]. The following chapter provides some examples of the use of AI technology in military applications.

3.1. Application of Neural Networks in Object Location

Classic methods of location at sea include among others the use of various types of radar stations, air patrols, maritime patrols, remotely controlled drones or satellites, e.g., CleanSeaNet—the European satellite object and pollution detection service, developed and monitored by EMSA [23]. In recent years, the automatic identification system (AIS) has also become very popular. The system provides a lot of information about marine traffic; however, due to the large amount of processed data, it is not always effective. Instead, various machine learning approaches are used to monitor and inform about any anomaly—the movement that deviates from established standards.

One of the methods used in AISs is Fuzzy ARTMAP [24]. It is an architecture that combines fuzzy logic elements and Adaptive Resonance Theory (ART) neural networks. By default, ART uses unsupervised learning technique. The algorithm of the network operation consists in maintaining readiness to learn new patterns while preventing the rejection or modification of previously learned ones [25]. The system must be able to maintain stability when facing non-significant events and the ability to update on significant events. The basic ART network consists of two layers and a reset module, which are shown in Figure 2. The first layer, called the comparison field, receives normalized input data, processes it and transfers it to the second layer with appropriate weights. The second layer, otherwise the recognition field is a competitive layer according to the WTA principle—"winner takes all" [26], in which the unit with the best match (the highest product of the input vector and weight) becomes a candidate for learning a new pattern; the other units are ignored. The reset module decides whether a new unit can learn a pattern based on its similarity to the prototype vector; this is called the vigilance test.

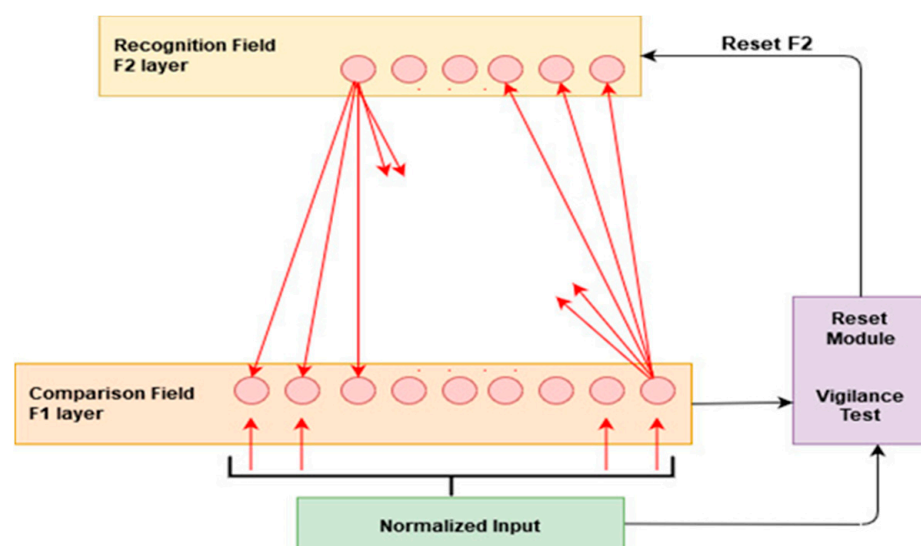


Figure 2. Structure of the Adaptive Resonance Theory (ART) network with emphasis on existing layers and mechanisms [25].

The authors modified the basic algorithm to obtain greater speed and efficiency of training, which allows it to be adapted in real time and in interactive conditions (the operator supports model training).

Other solution of AIS was presented in [27]. Authors proposed using the old-fashioned artificial intelligence method of data integrity assessment based on set of rules. The work was developed in cooperation with representatives of French military units—officers of the French Navy and cadets of the French naval academy.

In [28], authors present another method of grouping data based on similarity—Gaussian Mixture Model (GMM). Most of the data can be modelled using the Gaussian distribution. The idea of this model is to assume that the data for grouping comes from different Gaussian distributions, so the data set can be modeled as a mixture of different Gaussian distributions. The Gaussian distribution is described as follows [29]:

$$f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad (1)$$

where μ is a location parameter, equal to the distribution mean and σ is the standard deviation.

According to the method described in [30], it is assumed that there are K units (clusters) with estimated values of μ and σ parameters, for which the probability density function is defined as a linear probability density function of all distributions:

$$p(X) = \sum_{k=1}^K \mu_k G(X|\mu_k, \Sigma_k) \quad (2)$$

where

Σ —covariance matrix;

$G(X|\mu_k, \Sigma_k)$ —the probability density function of a Gaussian Distribution.

Defining an example variable $\varphi_k(X) = p(k|X)$ in accordance with the Bayes' theorem that will be used in further calculations:

$$\varphi_k(X) = p(k|X) = \frac{p(X|k)p(k)}{\sum_{k=1}^K p(k)p(X|k)} = \frac{p(X|k)\pi_k}{\sum_{k=1}^K \pi_k p(X|k)} \quad (3)$$

For the probability function to be maximum, its derivative of $p(X|\mu, \Sigma, \pi)$ with respect to π , Σ , μ should equal zero. After substitution $\varphi_k(X)$ in equations, the following are obtained:

$$\mu_k = \frac{\sum_{n=1}^N \varphi_k(x_n) x_n}{\sum_{n=1}^N \varphi_k(x_n)}, \quad (4)$$

$$\Sigma_k = \frac{\sum_{n=1}^N \varphi_k(x_n) (x_n - \mu_k)(x_n - \mu_k)^T}{\sum_{n=1}^N \varphi_k(x_n)}, \quad (5)$$

$$\pi_k = \frac{1}{N} \sum_{n=1}^N \varphi_k(x_n) x_n, \quad (6)$$

where the sum $\sum_{n=1}^N \varphi_k(x_n)$ is the total number of sampling points in the k -th set.

Parameters cannot be estimated in closed form; that is why the iterative Expectation-Maximization algorithm is used together with the GMM. This is a frequently used method [31] that helps identify the maximum probability estimates when the data is incomplete or contains hidden variables.

The crucial issue in object detection problem is the separation of moving and stationary targets. Synthetic aperture radars (SAR) are very often used for this purpose in view of their capability to removing the ambiguity stemming from inevitable moving targets in stationary scene imaging and suppressing clutter in moving target imaging. In [32] the author proposed using a novel solution—shuffle GAN (generative adversarial networks) with autoencoder separation method to separate the moving and stationary targets in SAR

imagery. One of the biggest advantages of this method is working in a totally unsupervised way, which allows to train the model without the dataset containing mixed and separated SAR images. The idea of GAN is a “combat” between two networks working recursively. The first of them—generator—generates new data, and the second one—discriminator—works in slightly similar way to the classifier. It assesses results of work of the generator. The whole training process is repeated as long as the discriminator will evaluate the results of the generator as true (it is not possible to distinguish produced images from original images) [33].

Image segmentation relies on partitioning an image into multiple segments that are related to various groups of objects, for example, neutral objects and threats. There are a lot of different methods of segmentation: thresholding, region of interest based, clustering, compression-based, Histogram-based, etc. Clustering is a method of grouping of unlabeled data that determines a feature vector for each pixel of the image and uses a similarity metric for clustering vectors that have similar features. One of the popular methods used for this purpose is fuzzy *c*-means clustering (FCM) developed by J.C. Dunn in 1973 [34]. The idea of the algorithm is very similar to *k*-means [35] and based on computing the centroid for each cluster and coefficients of being in the clusters. The procedure is repeated until the algorithm has converged, i.e., until the change of coefficients between two iterations is no more than threshold. Authors in [36] proposed using an approach based on extracting texture and geometry structure features to detect objects like planes, tanks and vehicles in natural background using FCM. Object detection and recognition is a very important area of the modern warfare. The future research in this domain should be focused on achieving better results by electronic armed forces in

- Automatization of the localization process and increase in the accuracy of the localization;
- Operation of location systems in conditions of targeted environmental disturbances;
- Achieving operational reliability of location modules based on distributed networks.

3.2. Location of Underwater Mines Using Deep Convolutional Neural Network

Underwater mines pose a very high threat to passing ships. Various types of mine countermeasures are used to localize and neutralize the threat [37,38]. The purpose of future countermeasures is to ensure unrestricted freedom of movement for naval forces and to rapidly remove mines when needed. For this goal, Unmanned Airborne Vehicle (UAV) and Unmanned Undersea Vehicle (UUV), also known as autonomous underwater vehicles (AUV), are being developed. UAVs are mainly used by the armed forces for observation and reconnaissance, which is why they are usually equipped with observation equipment in the form of optoelectronic heads. Armed drones designed to perform combat operations are often referred to as Unmanned combat air vehicles (UCAVs). UUVs, sometimes known as underwater drones, are equipped with sonars that create seabed maps based on the collected data. Such vehicles are developed, among others, by the Monterey Bay Aquarium Research Institute (MBARI) [39], which uses them for collecting information from underwater areas. AUVs create and accumulate large amounts of photos, that must then be classified to distinguish mines from other objects. Deep convolutional neural networks can be used for this purpose.

A neural network consists of an input layer, output layer and optionally also hidden layers (in Figure 3). Each neural network that has at least one hidden layer is called deep neural network DNN.

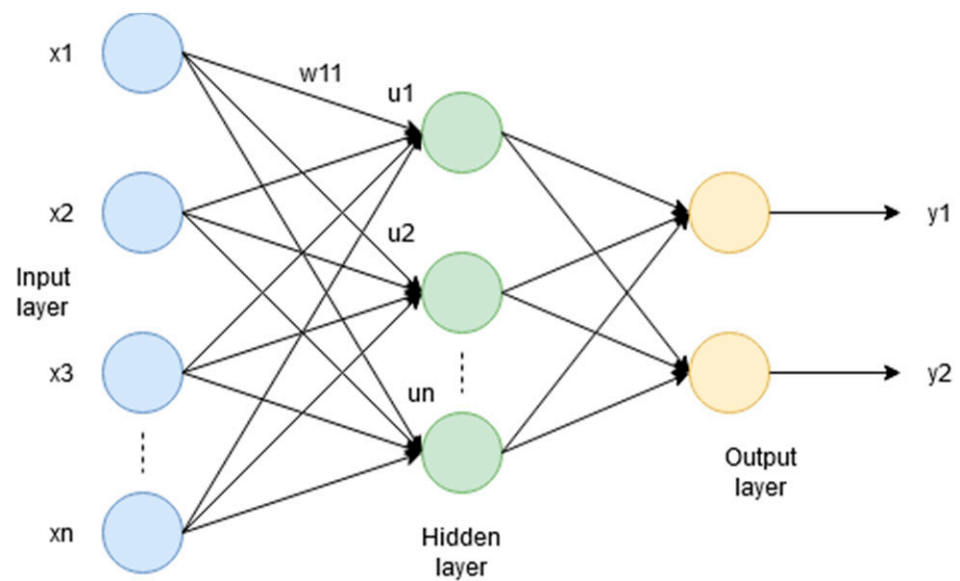


Figure 3. A three-layers deep neural network with one hidden layer.

In [40], authors proposed an autonomous underwater vehicle with side scan sonar (SSS). The sensor carries out image segmentation using convolutional neural networks CNN (CNN). It is one of the deep networks variants often used to process digital images. The main advantage of the convolutional network compared to traditional algorithms is that there is no need to perform feature extraction. The network consists of a convolution base and a classifier built of the so-called fully connected layers (FC). The most important element of the convolution base are the convolution layers, which, using sets of filters with different sizes of filter mask—firstly, bigger filters are used that filter each channel of the input image to extract features, thus creating feature maps whose dimensions are smaller and smaller, while the depth gets bigger. Mathematically, convolution operation is presented as follows [41]:

$$y(n) = x(n) \otimes h(n) = \sum_{k=-\infty}^{\infty} x(k)h(n-k) \quad (7)$$

The scheme of the convolution layer operation was shown in Figure 4.

Solution proposed by authors is a Fully Convolutional Network (FCN). In this method convolution operations are used in the fully connected layer. For better results also Markov random fields (MRF) were used. The combination of the methods allowed to obtain an overall accuracy of 90%.

Authors in [42] presented a solution of AUV with dead-reckoning (DR) navigation method based on neural networks, called NN-DR which is perfect for rapidly changing conditions. The training was carried out on a network consisting of 3 hidden layers because of limited computational ability and energy of AUVs.

Location underwater mines is a one of the crucial area of object detection in military applications affecting on the sense of security both of civilians and military. The future research in this domain should be concentrated on the following:

- Effective and fast location of underwater mines in real time;
- Increasing the reliability in detecting and distinguishing between hazardous and neutral objects;
- Effective cooperation of detection systems with systems that neutralize the threat.

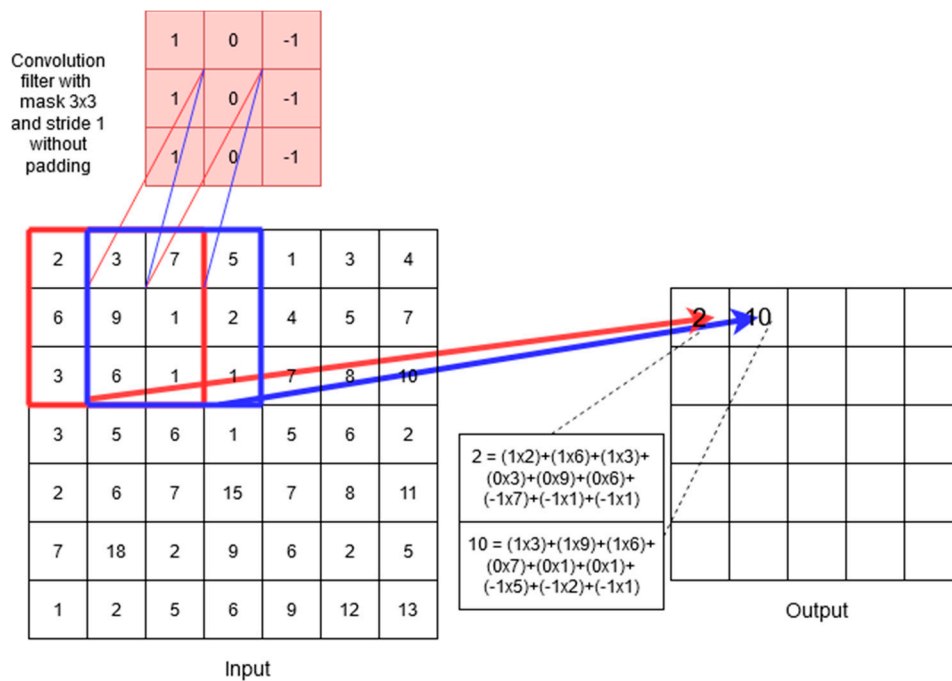


Figure 4. The idea of the convolution layer operation.

3.3. Application of Neural Networks in Cybersecurity

Hackers' attacks are becoming more and more common and dangerous with every year. As reports and studies show [43], both commercial companies as well as public, defense and government institutions of various countries are threatened by them. Incident detection is carried out via the IDS (intrusion detection system), that analyses network traffic, classifies it as intrusive or normal and in the event of danger sends a notification [44]. The normal network traffic often has a similar signature to attacks, making classification difficult. In addition, the method is often slow and expensive, which gave rise to the idea of using artificial intelligence algorithms for this purpose. One of the techniques that is being tested for IDS support is the Support Vector Machine (SVM) [45].

SVM is an algorithm that aims to find a hyperplane in the N -dimensional space that clearly classifies (separates) data points. There are many such hyperplanes, but the algorithm looks for the one with the maximum margin, i.e., the maximum distance between points of individual classes that provides better generalization capacity and greater resistance to overlearning. The hyperplanes can have different shapes as was shown in Figure 5.

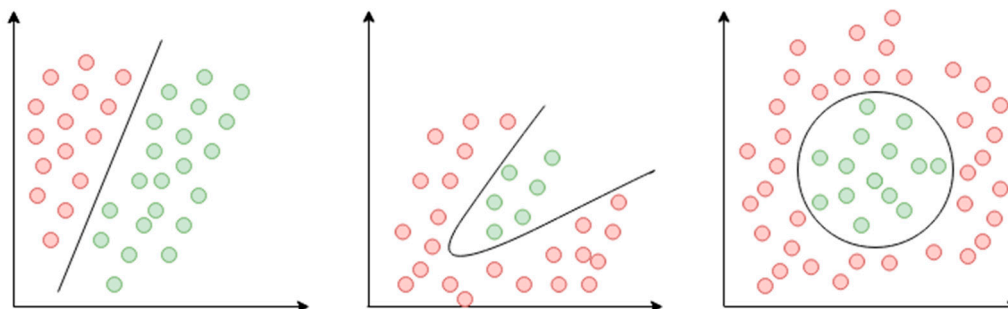


Figure 5. Different types of separating hyperplanes.

A linear hyperplane is described by a linear equation [46] that maximizes the distance between the extreme points of both classes:

$$y(x) = wx + b = 0, \quad (8)$$

x —vector $x = (x_1, x_2, \dots, x_n)$;

w —vector $w = (w_1, w_2, \dots, w_n)$;

b —bias.

The hyperplane equation as a classifier assigns points to the appropriate classes (+1 or -1) according to the equations [46]:

$$wx_i + b \geq 0, \quad h(x_i) = +1, \quad (9)$$

$$wx_i + b \leq 0, \quad h(x_i) = -1, \quad (10)$$

The purpose of the learning is to maximize the separation margin $\frac{2}{\|w\|}$, which means that the following condition is fulfilled:

$$\max\left(\frac{2}{\|w\|}\right) \rightarrow \min\left(\frac{\|w\|}{2}\right) \rightarrow \min\left(\frac{1}{2}\|w^2\|\right), \quad (11)$$

Each SVM network can distinguish only two classes. There are two techniques for solving multiclass problems: one-against-all (there are as many 2-classes classifiers as there are classes) and one-against-one there are as many classifiers as there are pairs of classes) [47].

Authors in [48] presented a comparison of different AI algorithms that can be used in intrusion detection systems. They tested machine learning classification algorithms such as Decision Tree, k-Nearest Neighbors, Random Forest and SVM and also two models of neural networks with the same architecture and different types of optimizers—Adam optimizer and stochastic gradient descent [49]. Feature selection was performed with using a Support Vector Classifier with a linear kernel as the estimator and forward feature selection with cross validation to rank the features. Proposed system based on anomaly detection does not only distinguish network traffic packets signatures but to also determines whether a network intrusion was obfuscated.

Nowadays, a large percentage of military operations has been transferred to cyber space. One of the key factors determining the success of a military mission was ensuring the security and secrecy of combat data. The issues that should be developed and improved in this domain in the nearest future are the following:

- Improving the operation of systems securing access to key data—authorization and authentication modules;
- Support for systems identifying unauthorized access to data in real time.

3.4. Bioinspired AI Robots on the Battlefield

One of the main goals of modern technology on the battlefield is to protect the health and lives of soldiers. A solution often proposed in this regard is to bring the machines onto the battlefield. The world leader in the field of mobile robots is the American company Boston Dynamics [50]. The robots can move independently, detect and avoid obstacles, follow a predetermined route, as well as recognize and respond to voice messages coming from the environment. Enrico Guizzo in [51] describes his impressions of the visit to the Boston Dynamics headquarters and presents some of their solutions—Spot and Atlas.

Spot is a nimble quadruped which can move over almost any terrain. On the front, back and sides of the robot, there are sensors with cameras that allow you to use the SLAM (simultaneous localization and mapping) navigation method. SLAM algorithms are very often employed in problems combining the need to update the map of an unknown environment while tracking a moving object, including objects localization [52], pedestrians recognition [53] or localization unmanned aerial vehicles position [54]. Spot behaves

completely autonomously or can be controlled remotely, while retaining a great deal of autonomy. The robot is presented in Figure 6.



Figure 6. Spot: (a) overall look, (b) going up the stairs [55].

Atlas is a 150-cm-tall humanoid. The control software uses mathematical models of the robot's physics and the integration of its body with the environment, so that the movements performed are as natural as possible, inspired by the behavior of athletes. The first version of the robot was developed as part of a competition DARPA Robotics Challenge in 2015. Atlas was presented in Figure 7.



Figure 7. Atlas: (a) overall look, (b) jumping over an obstacle [55].

Even higher results on the battlefield can be achieved with the cooperation of humanoid robots, e.g., with the use of Swarm AI, as shown in [56] in the example of cleaning robots.

Using bioinspired robots in military operations can become the new standard of warfare in a short time. Robots are resistant to fatigue, lack of food and water and extreme weather conditions, but their proper functioning can be easily disturbed by hacker attacks. The main goal of future research related to this area should be ensuring reliability and resistance to hostile interference in the software of robot.

3.5. AI Applications for Military Logistics

Logistics, distribution and supply chain are parts of a very sophisticated and advanced process that refers to the movement of products or services to a designated location at an

agreed time. The history of logistics is inextricably linked with military. Already, ancient Romans organized efficient logistic systems to supply legions [57].

Currently, logistics especially in military domain includes many different functions related to the processing of large amounts of data and making of decisions related to transport, delivery and communication, supporting combat units and many others. Organizing an efficient supply chain is very important both in times of peace and war. In [58], authors proposed a method of military logistics management based on the Internet of Things (IoT) that allows to shorten the logistical response time and improve the speed of actions. Authors in [59] noted an important issue of ensuring safety and reliability of supply chain which is crucial operational capability of military forces. They proposed new solution Military Supply Chain Cyber Implications Model (M-SCCIM) which combines logistic and cybersecurity. The presented model uses the newest technologies such as Internet of Things (IoT) and smart contracts. Smart contracts are “pieces of software that represent a business arrangement and execute themselves automatically under pre-determined circumstances” [60]. In implementation of smart contracts, some blockchain technologies which ensure decentralization, persistency, anonymity and auditability are used. In military supply chains, smart contracts can be responsible for checking product flows throughout the supply chain or ensuring the integrity of the chain. Using IoT devices in supply chains allows for better tracking movement of goods (also tracking speed and traffic flow of movement) that makes easier other administrative actions associated to supply chain. However, some properties of IoT structure that are beneficial in commercial environment can be big challenges to implement in military network architectures, which was described in [61]. The military needs a weapon, repair parts, fuel and a lot of other equipment; that is why one of the most important elements of military logistics is a process of management of supply chain. Authors in [62] present a technology to control a supply chain that ensures speed and safety and maximizes the military and economic benefits. Important elements of each supply chain are data analysis and decision-making process. Nowadays, a reinforcement learning is a very common technique supporting decision-making, also in the pre-war planning stage [63] which includes planning supplies to the battlefield and other issues related to military logistics. The reinforcement learning is one of the three main types of machine learning approach. In this approach, the user does not prepare a large training dataset to learn the model, but he uses the environment that allows to collect learning data automatically. The idea of training is interactions of agents with the environment in order to maximize the reward returned by the environment [64]. The idea of the reinforcement learning was presented in Figure 8.

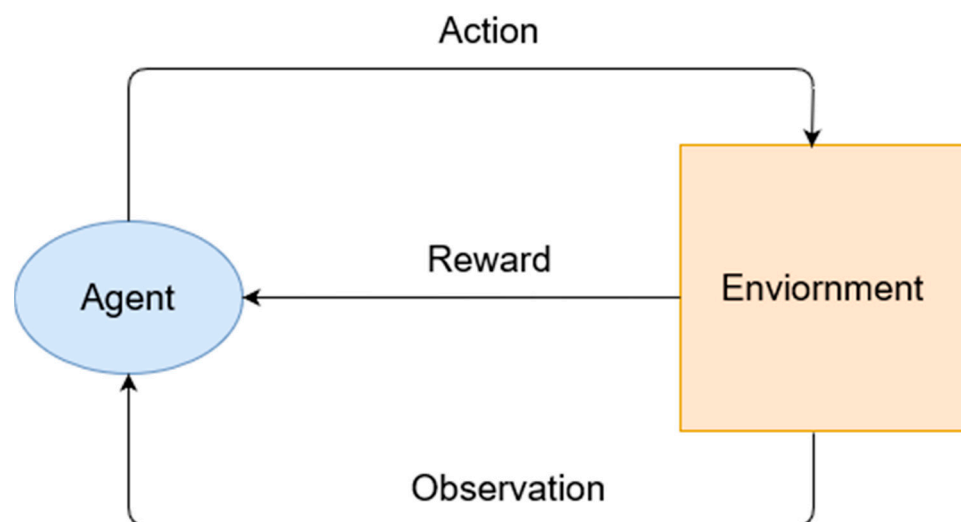


Figure 8. Reinforcement learning.

Another approach to supply chain management was presented in [65]. Authors compared artificial neural network (ANN) and machine learning algorithms like k-Nearest Neighbors, Logistic Regression, Random Forest and Naive Bayes in solving the problem of prediction of availability and possible reorder level of military logistics in an example of ensuring the availability of petroleum products.

Another important issue in the field of military logistics is efficient and quick organization of medical aid. It is planned that artificial intelligence will be a significant help in monitoring, diagnosing and segregating the wounded on the battlefield, to provide help to all those in need with limited resources. Charles River Analytics [66] is working on a semi-automatic support software for doctors, used when the evacuation of a soldier from the battlefield is not possible. The application called Automated Ruggedized Combat Casualty Care (ARC3) is implemented on behalf of the U.S. Army's Telemedicine & Advanced Technology Research Center (TATRC). This system is part of the strategy of trauma care on the battlefield, known as Tactical Combat Casualty Care (TCCC) [67].

Over the years, logistics has been one of the crucial parts of the military that influences the course of hostilities. In coming years, research in this area should be focused on the following:

- Acceleration of logistics processes, in particular in the supply chain, by applying deep learning algorithms that enable the processing of large amounts of data;
- Improving the timeliness of logistics deliveries;
- Cooperation of logistic data analysis systems with systems ensuring data security in order to ensure reliability and protection against diversionary activities.

3.6. Big Data in Military Data Processing and Modeling

All previously described AI solutions very often require big amounts of digital data. Its storage, transfer, analysis and visualization generate a lot of problems related to restricted computational power capabilities even for military hardware. Remedies for this problem can be found by using innovative modern architectures including Big Data solutions. In order to better understanding Big Data methodologies, techniques and their potential influence the development of the defense domain; in September 2016, the European Defence Agency (EDA) launched the "Big Data in Defence Modelling and Simulation" (BIDADEMS) [68]. According to recommendations from research, future modelling and simulations military applications should be designed using Cloud Computing. It also seems necessary to focus on education of analysts on new data analytics techniques and providing developers with Big Data tools when developing future models. The study results have led to a new research project (Modelling and Simulation Methodologies for Operations Research - MODSIMMET) analyzing very complex scenarios like hybrid warfare supported by Big Data and Artificial Intelligence.

Authors in [69] analyzed the application of military big data in equipment quality information management. They showed how the more effective flow of equipment quality information in the process could improve the management of information in comparison to scattered systems based on information from people. Other research presented in [70] showed possibilities of optimization of the education model in the military campus using Big Data systems to store and analyze students and teachers data. The proposed system is based on using Internet of Things and directed acyclic graph (DAG) to data processing. The aim of the described solution is the optimization of decision-making processes. The research highlights the reforming trend in education under the "military reform of Chinese characteristics" [71].

The need to process very large amounts of data requires the use of Big Data solutions and data collection in cloud systems. The most important thing that should be improved and developed in this area is ensuring the security of processed data. Another crucial issue is concentrating on introducing cloud solutions and cooperation between systems in each area of military operations.

4. Impact of Using Artificial Intelligence in Military on Society

As the examples show, the neural network applications can be very useful and effective also in the military sector. Actions that have been carried out by people in recent years are now fully automated. Algorithms decide what is good and what is bad, what is safe and what is dangerous, when we should react and when we should wait. The problem is even more important when we talk about AI applications in the military, because their decisions will affect the lives of all citizens. Can people feel safe when machines decide upon their lives? Some people are ready to fully trust the machine and nominate it to presidential election. This happened in Russia in 2017 when forty thousand Russians nominated a piece of AI software called “Alice” to stand against Vladimir Putin in the 2018 election. The virtual assistant created by Yandex could work 24/7 and used only logic to make decisions without emotions and seeking personal advantages [72]. A similar situation happened in Tokyo where a machine named “Michihito Matsuda” placed third in the election of mayor and in New Zealand where “Sam” was created—the world’s first Virtual Politician [73]. “Sam” was designed to run in the 2020 general election to analyze everyone’s opinions and to promote better policy for every citizen, but some people are still very skeptical and have a lot of fears related to AI.

4.1. Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) is a new trend that relates to the methods and techniques of applying artificial intelligence technology, so that the obtained solution results are understandable to the average person [74,75]. As the authors say in [76], this “research field holds substantial promise for improving trust and transparency of AI-based systems.”

The Centre for the Governance of AI (GovAI), part of the Future of Humanity Institute at the University of Oxford [77], is an organization that supports society in reaping the benefits and risk management of artificial intelligence. They conduct extensive research on important and neglected issues in AI management using political science, international relations, computer science, economics, law, and philosophy. Below is a brief summary of the surveys conducted by GovAI in 2011 [78] and 2019 [79] regarding the public’s attitude towards AI.

As stated in research carried out in 2011, according to respondents, artificial intelligence will reach the level of human intelligence at 50% around 2050 and 90% around 2150. Organizations from the area of industry, the military and academic centers will have the biggest contribution to development. In the case of the question about the probability of positive and negative effects of developing human-level artificial intelligence, the highest probability was indicated for the answer “extremely bad”. However, the answer “extremely good” came second which proves the presence of both extreme threats and benefits of AI. The survey results are presented in Figure 9.

According to research carried out in 2019, 41% of respondents support or strongly support the development of artificial intelligence, while 21% are somewhat or definitely against it. Much greater support (57%) is expressed by university graduates than people with lower education. There are clear differences in the level of trust in organizations working on the development and management of artificial intelligence. University researchers and the US military are the most trusted—50% and 49%. As for the impact of high-level machine intelligence on society, 22% of respondents think that the technology will be “on balance bad”, 12% think that it would be “extremely bad” (possibility of human extinction), 21% think it will be “on balance good”, and 5% think it will be “extremely good.” The results of a survey from 2019 are presented in Figure 10.

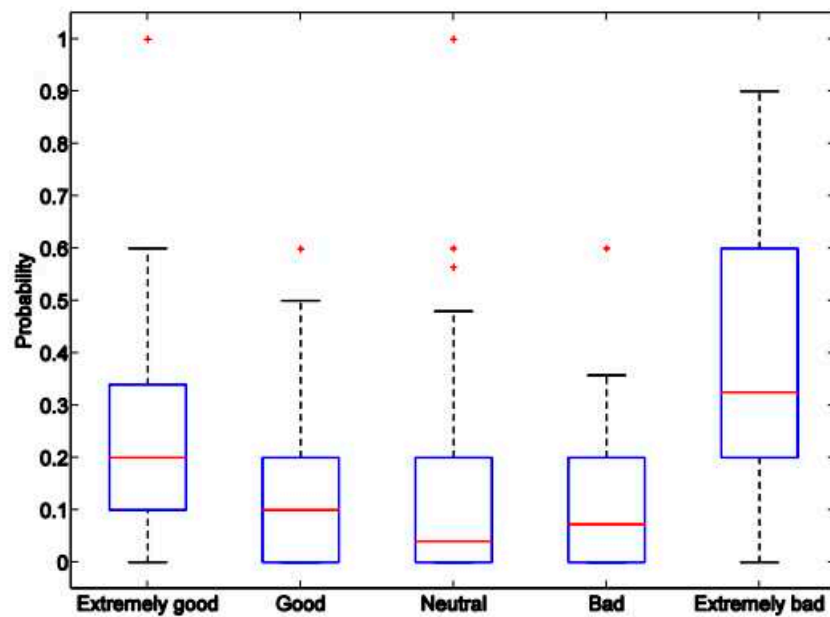


Figure 9. The results of the 2011 survey [78].

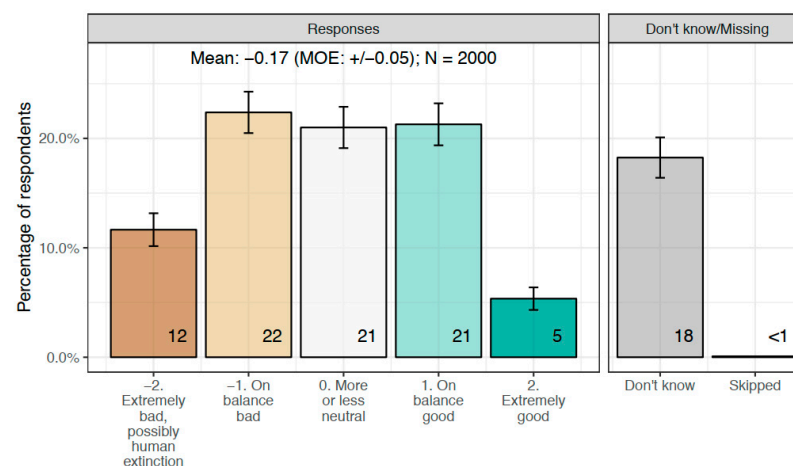


Figure 10. The results of the 2019 survey [79].

As research shows, according to society, military applications are among the main ones responsible for the development of AI, at the same time enjoying great public trust. A positive trend is also the declining percentage of society predicting “extremely bad” impact of AI on humanity.

4.2. Cooperation with Robots

When the average person thinks about artificial intelligence, he has a vision of an intelligent robot performing typical human activities. A robot is a programmable machine which, in accordance with ISO 8373, to some extent autonomously performs the assigned tasks based on the given state—without human intervention [80]. The increasingly popular social robots [81,82] are a specific type of robots. Social robots are defined as autonomous or semi-autonomous robots which, when communicating and interacting with people, behave in accordance with accepted social and behavioral norms adopted by people [83]. How do machines know what behavior people expect from them?

In [84], the authors propose an interesting way of learning ethical behavior by robots using data from social media (Twitter, Facebook, Instagram), records of court cases, and related available data sources. The training data set should include both ethical and

unethical patterns of behavior. The first phase of training also involves the presence of a human trainer who would supervise and provide feedback to teach the machine the appropriate behavior in each scenario.

However, the continuous improvement of social robots causes public concern. People worry that in the near future robots can replace them at work and significantly increase unemployment. However, as the authors write in [85], “this perception implicitly overestimates the social skills of the robots, which despite being continually upgraded, are still far from being able to dominate humans entirely”. The authors in [86] presented a proposal on how to bring man and machine closer together and increase people’s trust in the machines with which they must cooperate. Pilot studies were conducted at the United States Air Force Academy to show that building a human relationship with an AI agent earlier can be beneficial for military missions. In the cases of building relationships, the robot asked people questions about their favorite food, type of music, while in cases without building relationships, the robot was only simply introduced to people. The participants felt much more comfortable with the robot if a relationship had been established beforehand.

4.3. Ethics

In addition to trust, another important problem with artificial intelligence is machine ethics, especially in military applications. People were afraid that the thinking machines could harm them and thought about moral status of the machines themselves. Time of war often requires morally difficult decisions from commanders. By definition, neural networks applications should rely only on logic and programmed algorithms without any emotions. However, in the case of military operations, logic and efficiency cannot be the sole determinant. Ethics in the context of military operations was discussed extensively by Helen Frowe in [87], which also touches on the topic of remote warfare.

The authors in [88] raise the important issue of Lethal Autonomous Weapon Systems (LAWS) and strive to answer the question of why artificial intelligence systems should not have the right to decide about killing people as part of warfare. The main problem the authors point to is the lack of perfect, non-error-making AI systems. In the case of deciding about human life, even the accuracy at the level of 99% is too small. Another issue is interpretability. Some decisions made by the algorithm are incomprehensible to people, which resulted in the introduction of the so-called “right to explanation” meaning a right to be given an explanation for an output of the algorithm [89].

People are afraid of what enemy army can do them and what their army can do civilian from other countries. Therefore, the Pentagon announced it has adopted “ethical principles” for AI in warfare in February 2020 [90]. Decisions made by automata should also be “traceable” and “governable.” This means the possibility of deactivation systems that behavior raises concerns or threatens. Earlier the guideline included only an obligation of involving people in all AI military decisions. The actions of the Pentagon can be due to Google’s resignation from renewing contract called “Project Maven” under pressure from employees [91,92].

4.4. Consequences of Errors

As it was written in the previous subsection, people are afraid that a malfunctioning algorithm may harm them. What about the consequences of such an error? When people make mistakes, it is easy to decide who is responsible for this situation. But when the machine makes a mistake, a situation is a bit more sophisticated. Who is to blame?

The main problem is the definition of mistake. Suppose a hypothetical situation that we are planning an elegant party. Our artificial intelligent system has developed a menu that has many dishes from very fashionable fusion cuisine, but we like traditional meals and are not satisfied with the choice of machine. Has the system made an error? Is the creator guilty of not having programmed our preferences in the machine algorithm? Maybe it is our fault because we did not control the machine during the menu selection process. This situation does not have very serious consequences in contrast to military decisions,

e.g., regarding armed attacks, but it perfectly illustrates the problem of responsibility for errors made by artificial intelligence systems. The use of “ethical principles” can help, but every situation is different, and every person has a different moral system, so people are not ready to totally trust “intelligent systems” especially in state defense sector. They may be afraid that nobody will be responsible for any mistakes made.

In general, artificial intelligence is programmed to do useful tasks and help people, but malfunction can cause a very serious errors. The author in [93] tried to explain this correlation by using dynamic programming (division of the problem to be solved into sub-problems with regard to several parameters [94]). The article analyzed which elements of the AI system can be causes of errors and disastrous consequences and tried to answer the question of when AI can be dangerous.

5. Conclusions

The aim of the submission was to present main areas of use of AI algorithms in the military sector, especially in objects detection, cybersecurity, robotics and logistic and discuss their impact on people’s sense of security. The article shortly describes well-known algorithms of neural networks but in new, atypical applications. The authors wanted to point to the huge popularity of neural networks, which is increasing day by day thanks to the possibility of using big databases in the learning process. This applies both to commercial, research, educational and pure entertainment applications. The popularity of programs, such as AIE, shows how important this field of knowledge is. As research shows, people are still afraid of the possible effects of these technologies. This is understandable because even experts do not have a clear opinion on the future and development of artificial intelligence. As Prof. Stephen Hawking said: “The rise of powerful AI will be either the best, or the worst thing, ever to happen to humanity”.

Author Contributions: M.B. contributed to theoretical formulation, design methodology, dataset development, original draft preparation and revision. The other author, Z.P., contributed to project supervision, theoretical formulation and revision of the initial draft. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was supported and funded by The National Centre for Research and Development, grant no. CyberSecIdent/381319/II/NCBR/2018.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Definition of the Term ‘Military Capability’ Per Official Documentation of the United States Department of Defense. Available online: https://www.militaryfactory.com/dictionary/military-terms-defined.asp?term_id=3357 (accessed on 25 January 2021).
2. Cyber Attack Trends: 2020 Mid-Year Report. Available online: <https://research.checkpoint.com/2020/cyber-attack-trends-2020-mid-year-report/> (accessed on 25 January 2021).
3. Wong, Y.H.; Yurchak, J.; Button, R.W.; Frank, A.; Laird, B.; Osoba, O.A.; Steeb, R.; Harris, B.N.; Bae, S.J. *Deterrence in the Age of Thinking Machines*; RAND Corporation: Santa Monica, CA, USA, 2020. [CrossRef]
4. Johnson, J. Artificial Intelligence & Future Warfare: Implications for International Security. *Def. Secur. Anal.* **2019**, *35*, 147–169. [CrossRef]
5. Big Data Is Too Big Without AI—Maryville University Online. Available online: <https://online.maryville.edu/blog/big-data-is-too-big-without-ai/> (accessed on 25 January 2021).
6. Zhu, S.; Cao, R.; Yu, K. Dual Learning for Semi-Supervised Natural Language Understanding. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1936–1947. [CrossRef]
7. Deshmukh, P.S. Travel Time Prediction using Neural Networks: A Literature Review. In Proceedings of the 2018 International Conference on Information, Communication, Engineering and Technology, Pune, India, 29–31 August 2018; pp. 1–5. [CrossRef]
8. Dogru, N.; Subasi, A. Traffic accident detection using random forest classifier. In Proceedings of the 2018 15th Learning and Technology Conference (L&T), Jeddah, Saudi Arabia, 25–26 February 2018; pp. 40–45. [CrossRef]
9. Berkhahn, S.; Neuwaeiler, I.; Fuchs, L. Real-Time Water Level Prediction Based on Artificial Neural Networks. In *New Trends in Urban Drainage Modelling*; Mannina, G., Ed.; UDM 2018. Green Energy and Technology; Springer: Cham, Switzerland, 2018; pp. 603–607. [CrossRef]
10. Ghorbani, M.A.; Deo, R.C.; Karimi, V.; Yaseen, Z.M.; Terzi, O. Implementation of a hybrid MLP-FFA model for water level prediction of Lake Egirdir, Turkey. In *Stoch Environ Res Risk Assess*; Springer: Heidelberg, Germany, 2018; Volume 32, pp. 1683–1697. [CrossRef]

11. Williams, D.P. Underwater target classification in synthetic aperture sonar imagery using deep convolutional neural networks. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 2497–2502. [CrossRef]
12. Madichetty, S.; Sridevi, M. Detecting Informative Tweets during Disaster using Deep Neural Networks. In Proceedings of the 2019 11th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, India, 7–11 January 2019; pp. 709–713. [CrossRef]
13. Ranjit, S.; Shrestha, S.; Subedi, S.; Shakya, S. Foreign Rate Exchange Prediction Using Neural Network and Sentiment Analysis. In Proceedings of the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida (UP), India, 12–13 October 2018; pp. 1173–1177. [CrossRef]
14. Varma, A.; Sarma, A.; Doshi, S.; Nair, R. House Price Prediction Using Machine Learning and Neural Networks. In Proceedings of the 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 20–21 April 2018; pp. 1936–1939. [CrossRef]
15. Lotfidereshgi, R.; Gournay, P. Speech Prediction Using an Adaptive Recurrent Neural Network with Application to Packet Loss Concealment. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5394–5398. [CrossRef]
16. Pietrow, D.; Matuszewski, J. Objects detection and recognition system using artificial neural networks and drones. In Proceedings of the 2017 Signal Processing Symposium (SPSymo), Jachranka, Poland, 12–14 September 2017; pp. 1–5. [CrossRef]
17. Yanke, G. Tying the knot with a robot: Legal and philosophical foundations for human–artificial intelligence matrimony. *Ai Soc.* **2020**. [CrossRef]
18. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
19. Laufer, R. The social acceptability of AI systems: Legitimacy, epistemology and marketing. *Ai Soc.* **1992**, *6*, 197–220. [CrossRef]
20. Svenmarck, P.; Luotsinen, L.; Nilsson, M.; Schubert, J. Possibilities and Challenges for Artificial Intelligence in Military Applications. In Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists’ Meeting, Bordeaux, France, 31 May 2018.
21. DARPA—Accelerating the Exploration of Promising Artificial Intelligence Concepts. Available online: <https://www.darpa.mil/news-events/2018-07-20a> (accessed on 25 January 2021).
22. Sanchez, S.L. Artificial Intelligence (AI) Enabled Cyber Defence. Available online: [https://www.eda.europa.eu/webzine/issue14/cover-story/artificial-intelligence-\(ai\)-enabled-cyber-defence](https://www.eda.europa.eu/webzine/issue14/cover-story/artificial-intelligence-(ai)-enabled-cyber-defence) (accessed on 25 January 2021).
23. EMSA—European Maritime Safety Agency. Available online: <http://www.emsa.europa.eu/> (accessed on 25 January 2021).
24. Rhodes, B.J.; Bomberger, N.A.; Seibert, M.; Waxman, A.M. Maritime situation monitoring and awareness using learning mechanisms. In Proceedings of the MILCOM 2005-2005 IEEE Military Communications Conference, Atlantic City, NJ, USA, 17–20 October 2005; pp. 646–652. [CrossRef]
25. Al Salam, M. Adaptive Resonance Theory Neural Networks. Available online: https://www.academia.edu/38067953/Adaptive_Resonance_Theory_Neural_Networks (accessed on 25 January 2021).
26. Mao, Z.; Massaquoi, S.G. Dynamics of Winner-Take-All Competition in Recurrent Neural Networks with Lateral Inhibition. *IEEE Trans. Neural Netw.* **2007**, *18*, 55–69. [CrossRef]
27. Iphar, C.; Ray, C.; Napoli, A. Data integrity assessment for maritime anomaly detection. *Expert Syst. Appl.* **2020**, *147*. [CrossRef]
28. Laxhammar, R. Anomaly detection for sea surveillance. In Proceedings of the 2008 11th International Conference on Information Fusion, Cologne, Germany, 30 June–3 July 2008; pp. 1–8.
29. Walck, C. *Hand-Book on Statistical Distributions for Experimentalists*; Universitet Stockholms: Stockholm, Swede, 2007; p. 119.
30. GeeksforGeeks—Gaussian Mixture Model. Available online: <https://www.geeksforgeeks.org/gaussian-mixture-model/> (accessed on 25 January 2021).
31. Grefl, K.; van Steenkiste, S.; Schmidhuber, J. Neural Expectation Maximization. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Hong Kong, China, 4–9 December 2017; pp. 6694–6704. [CrossRef]
32. Pu, W. Shuffle GAN With Autoencoder: A Deep Learning Approach to Separate Moving and Stationary Targets in SAR Imagery. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–15. [CrossRef]
33. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* **2014**, *3*. [CrossRef]
34. Dunn, J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *J. Cybern.* **1973**, *3*, 32–57. [CrossRef]
35. Forgy, E.W. Cluster analysis of multivariate data: Efficiency versus interpretability of classifications. *Biometrics* **1965**, *21*, 768–769.
36. Fei, C.; Honghui, C.; Jianwei, M. Man-made Object Detection Based on Texture Clustering and Geometric Structure Feature Extracting. *Int. J. Inf. Technol. Comput. Sci. (Ijitcs)* **2011**, *3*, 9–16. [CrossRef]
37. The Future of Mine Countermeasures. Available online: <https://fas.org/man/dod-101/sys/ship/weaps/docs/mcmfuture.htm> (accessed on 25 January 2021).
38. THALES. The Future of Mine Warfare: A Quicker, Safer Approach. Available online: <https://www.thalesgroup.com/en/united-kingdom/news/future-mine-warfare-quicker-safer-approach> (accessed on 25 January 2021).

39. MBARI—Autonomous Underwater Vehicles. Available online: <https://www.mbari.org/at-sea/vehicles/autonomous-underwater-vehicles/> (accessed on 25 January 2021).
40. Song, Y.; Zhu, Y.; Li, G.; Feng, C.; He, B.; Yan, T. Side scan sonar segmentation using deep convolutional neural network. In Proceedings of the OCEANS 2017—Anchorage, Anchorage, AK, USA, 18–21 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
41. Zieliński, T.P. *Cyfrowe Przetwarzanie sygnałów. Od teorii do zastosowań*; Wydawnictwo Komunikacji i Łączności: Warszawa, Poland, 2004; pp. 17–21.
42. Song, S.; Liu, J.; Guo, J.; Wang, J.; Xie, Y.; Cui, J.H. Neural-Network-Based AUV Navigation for Fast-Changing Environments. *IEEE Internet Things J.* **2020**, *7*, 9773–9783. [CrossRef]
43. Center for Strategies & International Studies—Significant Cyber Incidents. Available online: <https://www.csis.org/programs/technology-policy-program/significant-cyber-incidents> (accessed on 25 January 2021).
44. Pratt, M.K. What Is an Intrusion Detection System? How an IDS Spots Threats. Available online: <https://www.csoonline.com/article/3255632/what-is-an-intrusion-detection-system-how-an-ids-spots-threats.html> (accessed on 25 January 2021).
45. Ghanem, K.; Aparicio-Navarro, F.J.; Kyriakopoulos, K.G.; Lambotharan, S.; Chambers, J.A. Support Vector Machine for Network Intrusion and Cyber-Attack Detection. In Proceedings of the 2017 Sensor Signal Processing for Defence Conference (SSPD), London, UK, 6–7 December 2017; pp. 1–5. [CrossRef]
46. Kowalczyk, A. *Support Vector Machines*; Synclution: Research Triangle, NC, USA, 2017; pp. 25–26.
47. Osowski, S. *Neural Networks for Information Processing*, 4th ed.; Oficyna Wydawnicza Politechniki Warszawskiej: Warsaw, Poland, 2020; pp. 193–194. ISBN 978–83-7814-923-1.
48. Shah, A.; Clachar, S.; Minimair, M.; Cook, D. Building Multiclass Classification Baselines for Anomaly-Based Network Intrusion Detection Systems. In Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Sydney, Australia, 6–9 October 2020; pp. 759–760. [CrossRef]
49. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2017**, arXiv:1609.04747v2.
50. FAQs about Boston Dynamics. Available online: <https://www.bostondynamics.com/about> (accessed on 25 January 2021).
51. Guizzo, E. By Leaps and Bounds: An exclusive look at how Boston dynamics is redefining robot agility. *IEEE Spectr.* **2019**, *56*, 34–39. [CrossRef]
52. Loesch, A.; Bourgeois, S.; Gay-Bellile, V.; Gomez, O.; Dhome, M. Localization of 3D objects using model-constrained SLAM. *Mach. Vis. Appl.* **2018**, *29*, 1041–1068. [CrossRef]
53. Ding, Z.; Huang, R.; Hu, B. Robust Indoor SLAM based on Pedestrian Recognition by Using RGB-D Camera. In Proceedings of the 53rd Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 292–297. [CrossRef]
54. Rizk, M.; Mroue, A.; Farran, M.; Charara, I. Real-Time SLAM Based on Image Stitching for Autonomous Navigation of UAVs in GNSS-Denied Regions. In Proceedings of the 2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS), Genova, Italy, 31 August–2 September 2020; pp. 301–304. [CrossRef]
55. ROBOTS—Your Guide to the World of Robotics. Available online: <https://robots.ieee.org/> (accessed on 25 January 2021).
56. Hiejima, T.; Kawashima, S.; Ke, M.; Kawahara, T. Effectiveness of Synchronization and Cooperative Behavior of Multiple Robots based on Swarm AI. 2019 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Bangkok, Thailand, 11–14 November 2019; pp. 341–344. [CrossRef]
57. Campbell, D.; Roth, J. The Logistics of the Roman Army at War (264 B.C.–A.D. 235). *J. Rom. Stud.* **2000**, *90*, 224. [CrossRef]
58. Wang, J.; Cao, L.; Shen, Y.; Zheng, G. Research on Design of Military Logistics Support System Based on IoT. In Proceedings of the 2018 Prognostics and System Health Management Conference (PHM-Chongqing), Chongqing, China, 26–28 October 2018; pp. 829–832. [CrossRef]
59. Sobb, T.M.; Turnbull, B. Assessment of Cyber Security Implications of New Technology Integrations into Military Supply Chains. In Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 21–21 May 2020; pp. 128–135. [CrossRef]
60. Kim, H.M.; Laskowski, M. Toward an ontology—driven blockchain design for supply—chain provenance. *Intell. Syst. Account. Financ. Manag.* **2018**, *25*, 18–27. [CrossRef]
61. Tortonesi, M.; Morelli, A.; Govoni, M.; Michaelis, J.; Suri, N.; Stefanelli, C.; Russell, S. Leveraging Internet of Things within the military network environment—Challenges and solutions. In Proceedings of the 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), Reston, VA, USA, 12–14 December 2016; pp. 111–116. [CrossRef]
62. Wang, W.; Guan, Y.; Jiang, D.; Yao, P. Analysis of information flow control in military supply chain management. In Proceedings of the 2010 8th International Conference on Supply Chain Management and Information, Hong Kong, China, 6–9 October 2010; pp. 1–4.
63. Yin, C.; Yang, R.; Zou, X. Research of Command Entity Intelligent Decision Model based on Deep Reinforcement Learning. In Proceedings of the 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), Nanjing, China, 23–25 November 2018; pp. 552–556. [CrossRef]
64. Lapan, M. *Deep Reinforcement Learning Hands-on: Apply Modern RL Methods to Practical Problems of Chatbots, Robotics, Discrete Optimization, Web Automation, and More*; Packt Publishing Ltd.: Birmingham, UK, 2020, ISBN 978-1-83882-004-6.

65. Ajakwe, S.O.; Nwakanma, C.I.; Lee, J.M.; Kim, D.S. Machine Learning Algorithm for Intelligent Prediction for Military Logistics and Planning. In Proceedings of the 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea (South), 21–23 October 2020; pp. 417–419. [CrossRef]
66. Charles River Analytics. Available online: <https://www.cra.com/> (accessed on 25 January 2021).
67. *Tactical Combat Casualty Care Handbook Lessons and Best Practices*; Version 5; Center for Army Lessons Learned: Fort Leavenworth, KS, USA, 2017; pp. 13–17.
68. European Defence Agency—Big Data Analytics for Defense. Available online: <https://eda.europa.eu/webzine/issue14/cover-story/big-data-analytics-for-defence> (accessed on 16 March 2021).
69. Xiang, Z.; Xiaofang, L.; Weigang, G. Analysis of the Application of Military Big Data in Equipment Quality Information Management. In Proceedings of the 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, China, 12–15 April 2019; pp. 68–71. [CrossRef]
70. Liu, M.; Ma, J.; Jin, L. Analysis of Military Academy Smart Campus Based on Big Data. In Proceedings of the 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), Hangzhou, China, 25–26 August 2018; pp. 105–108. [CrossRef]
71. Weitz, R. PLA Military Reforms: Defense Power with Chinese Characteristics. Available online: <https://www.worldpoliticsreview.com/articles/18215/pla-military-reforms-defense-power-with-chinese-characteristics> (accessed on 16 March 2021).
72. Petit, H. Could Russia's President One Day be a ROBOT? Alisa AI Software that Claims 'Enemies of the People will be Shot' wins the Backing of 40,000 to Stand against Vladimir Putin. Available online: <https://www.dailymail.co.uk/sciencetech/article-5166847/Russian-AI-Alisa-wins-backing-40-000-election-run-up.html> (accessed on 25 January 2021).
73. Meet Your Politician of the Future. Available online: <http://www.politiciansam.nz/> (accessed on 25 January 2021).
74. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Se, J.; Benneto, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
75. Luotsinen, L.; Oskarsson, D.; Svenmarck, P.; Wickenberg, B.U. Explainable Artificial Intelligence: Exploring XAI Techniques in Military Deep Learning Applications. Available online: <https://www.foi.se/report-summary?reportNo=FOI-R--4849--SE> (accessed on 25 January 2021).
76. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2020**, *6*, 52138–52160. [CrossRef]
77. Centre for the Governance of AI. Available online: <https://www.fhi.ox.ac.uk/GovAI> (accessed on 22 January 2021).
78. Sandberg, A.; Bostrom, N. Machine Intelligence Survey. In *Technical Report 2011–1*; Future of Humanity Institute, Oxford University: Oxford, UK, 2011; pp. 1–12.
79. Zhang, B.; Dafoe, A. *Artificial Intelligence: American Attitudes and Trends*; Center for the Governance of AI, Future of Humanity Institute, University of Oxford: Oxford, UK, 2019; Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3312874 (accessed on 25 January 2021).
80. International Federation of Robotics. Available online: <http://www.ifr.org/service-robots> (accessed on 22 January 2021).
81. Onyeulo, E.B.; Gandhi, V. What Makes a Social Robot Good at Interacting with Humans? *Information* **2020**, *11*, 43. [CrossRef]
82. Lazzeri, N.; Mazzei, D.; Cominelli, L.; Cisternino, A.; De Rossi, D.E. Designing the Mind of a Social Robot. *Appl. Sci.* **2018**, *8*, 302. [CrossRef]
83. Bartneck, C.; Nomura, T.; Kanda, T.; Suzuki, T.; Kennsuke, K. A cross-cultural study on attitudes towards robots. In Proceedings of the HCI International, Las Vegas, NV, USA, 22–27 July 2005. [CrossRef]
84. Buenfil, J.; Arnold, R.; Abruzzo, B.; Korpela, C. Artificial Intelligence Ethics: Governance through Social Media. In Proceedings of the 2019 IEEE International Symposium on Technologies for Homeland Security (HST), Woburn, MA, USA, 5–6 November 2019; pp. 1–6. [CrossRef]
85. Van Greunen, D. User Experience for Social Human-Robot Interactions. In Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI), Dubai, United Arab Emirates, 4–6 February 2019; pp. 32–36. [CrossRef]
86. Bellas, A.; Perrin, S.; Malone, B.; Rogers, K.; Lucas, G.; Phillips, E.; Tossel, C.; de Visser, E. Rapport Building with Social Robots as a Method for Improving Mission Debriefing in Human-Robot Teams. In Proceedings of the 2020 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 24 April 2020; pp. 160–163. [CrossRef]
87. Frowe, H. *The Ethics of War and Peace: An Introduction*, 2nd ed.; Routledge: London, UK, 2015. [CrossRef]
88. Šimák, V.; Gregor, M.; Hruboš, M.; Nemeč, D.; Hrbček, J. Why Lethal autonomous weapon systems are unacceptable. In Proceedings of the 2017 IEEE 15th International Symposium on Applied Machine Intelligence and Informatics (SAMII), Herľany, Slovakia, 26–28 January 2017; pp. 359–364. [CrossRef]
89. De La Torre, L.F. The 'Right to an Explanation' under EU Data Protection Law. Available online: <https://medium.com/golden-data/what-rights-related-to-automated-decision-making-do-individuals-have-under-eu-data-protection-law-76f70370fcd0> (accessed on 25 January 2021).
90. Mind for Minds—US military Adopts 'Ethical' AI guidelines. Available online: <https://www.dw.com/en/us-military-adopts-ethical-ai-guidelines/a-52517260> (accessed on 25 January 2021).
91. Google's Departure from Project Maven Was a 'Little Bit of a Canary in a Coal Mine'. Available online: <https://www.fedscoop.com/google-project-maven-canary-coal-mine/> (accessed on 18 March 2021).

92. Artificial Intelligence (AI) Becomes the Latest Arms Race as Adversaries Seek to Perfect Machine Learning. Available online: <https://www.militaryaerospace.com/computers/article/14189468/artificial-intelligence-ai-project-maven-arms-race> (accessed on 18 March 2021).
93. Mathew, A. The Peril of Artificial Intelligence. In Proceedings of the 2020 Fourth International Conference on Inventive Systems and Control (ICISC), Coimbatore, India, 8–10 January 2020; pp. 919–924. [CrossRef]
94. Bellman, R. On the Theory of Dynamic Programming. In Proceedings of the National Academy of Sciences (USA), Santa Monica, CA, USA, 1 August 1952.

Article

Autonomous Weapons Systems and the Contextual Nature of *Hors de Combat* Status

Steven Umbrello ^{1,*}  and Nathan Gabriel Wood ²¹ Department of Philosophy and Educational Sciences, University of Turin, 10124 Turin, Italy² Department of Philosophy and Moral Sciences, Ghent University, Blandijnberg 2, 9000 Gent, Belgium; Nathan.Wood@UGent.be

* Correspondence: steven.umbrello@unito.it; Tel.: +39-3518238010

Abstract: Autonomous weapons systems (AWS), sometimes referred to as “killer robots”, are receiving ever more attention, both in public discourse as well as by scholars and policymakers. Much of this interest is connected to emerging ethical and legal problems linked to increasing autonomy in weapons systems, but there is a general underappreciation for the ways in which existing law might impact on these new technologies. In this paper, we argue that as AWS become more sophisticated and increasingly more capable than flesh-and-blood soldiers, it will increasingly be the case that such soldiers are “in the power” of those AWS which fight against them. This implies that such soldiers ought to be considered *hors de combat*, and not targeted. In arguing for this point, we draw out a broader conclusion regarding *hors de combat* status, namely that it must be viewed contextually, with close reference to the capabilities of combatants on both sides of any discreet engagement. Given this point, and the fact that AWS may come in many shapes and sizes, and can be made for many different missions, we argue that each particular AWS will likely need its own standard for when enemy soldiers are deemed *hors de combat*. We conclude by examining how these nuanced views of *hors de combat* status might impact on meaningful human control of AWS.

Citation: Umbrello, S.; Wood, N.G.Autonomous Weapons Systems and the Contextual Nature of *Hors de**Combat* Status. *Information* **2021**, *12*,216. [https://doi.org/10.3390/](https://doi.org/10.3390/info12050216)[info12050216](https://doi.org/10.3390/info12050216)**Keywords:** autonomous weapons; meaningful human control; *hors de combat* status; killer robots; military ethics

Academic Editor: Luis Martínez López

Received: 18 March 2021

Accepted: 18 May 2021

Published: 20 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Autonomous weapons systems (AWS) are likely to become a mainstay of modern advanced militaries. These systems come in different forms, shapes, and sizes, and imbued with different levels of autonomy, and thus have different capabilities in the field. There are numerous reasons for the rise of AWS in contemporary militaries. Aside from their use reducing the number of humans on the battlefield and thus reducing possible casualties, such systems can feasibly be capable of waging war more ethically, given their potential speed, efficiency, and precision. Moreover, it has been argued that with the advancement of computer vision and other means of surveillance, the epistemic gap in selecting targets for engagement can be closed even more so than in the case where human operators manually select their targets [1,2].

However, despite these advantages, there remain objections to AWS and their potential deployment. In particular, there is heavy opposition to AWS which are capable of lethal engagement, with organizations like the Campaign to Ban Killer Robots, the International Committee for Robot Arms Control, and the International Committee of the Red Cross (among others) advocating for either partial or full bans on the development and deployment of such systems. The arguments against AWS range from principled objections concerning their deleterious effects on human dignity, to more practical concerns about command and control over systems with opaque machine learning algorithms, or targeting systems which are technically insufficient for meeting the principle of discrimination [3–8]. Proponents of a ban often argue that these points alone require prohibition. However, even

granting that there currently exist technical obstacles to properly meet these challenges, this does not preclude such a possibility in the future as technology surpasses those barriers.

In this paper we avoid these principled and practical arguments, and instead explore how existing law might impact on AWS. In particular, we examine the notion of *hors de combat* status, as it is defined in the Geneva Protocol I Additional to the Geneva Conventions (AP I), arguing that advances in AWS will necessitate a more nuanced and varied approach to determining this status for enemy combatants. Our aim is to show that the law demands a contextualized appraisal of whether or not an enemy is in fact *hors de combat*, or “out of combat”, and that, based on this appraisal, more combatants will be deemed *hors de combat* than is usually taken to be the case. These factors imply that AWS will need to be capable of making finer-grained distinctions of *hors de combat* status of enemies based on subtle contextual factors. In making this final point, we also pay heed to the fact that there are many different types of AWS, and that the determination of *hors de combat* status may vary depending on the particular AWS under examination. As a final remark of clarification, our argument is concerned with how the law impacts on AWS, and not on what the capabilities of current or near-future AWS are. Thus, we do not argue that current or near-future AWS necessarily can make such nuanced distinctions with regards to *hors de combat* status. Rather, we only defend the much more modest claim that the law requires that they be able to do so. As such, our arguments focus on the legal and moral side of the equation, and do not examine the exact science of programming morality into machines, or questions of that nature.

The paper is structured as follows. In Section 2 we discuss the legal foundations of *hors de combat* status, in particular, AP I, Article 41 and its commentary. We show that *hors de combat* status is likely broader than one might initially suppose, and more importantly, that determining *hors de combat* status requires a contextualized approach, where the relative power of opposing belligerent agents will impact upon the final judgment. Section 3 applies this finding to AWS, showing how different autonomous weapons systems will need to be able to make different *hors de combat* determinations depending on their own abilities and on their precise mission objectives. In Section 4, we discuss how these points effect the notion of “meaningful human control” over AWS, and indicate future research which might further the debate.

2. The Safeguards of *Hors de Combat* Status

Contemporary discussions on the ethics and laws of war often place heavy emphasis on the protections afforded to noncombatants. However, in addition to safeguards for noncombatants, historically, the development of many ethical and legal norms was centered on minimizing, or at least mitigating, harm to combatants [9–11]. This should not come as a surprise, given that in the 19th century—when the laws of armed combat (LOAC) were being first codified—wars were “conducted by professional armies and the civilian population was not involved to any great extent” [10]. And it was against this backdrop that the legal status of being *hors de combat*, and its associated protections, was codified.

The currently binding legal articulation of *hors de combat* status can be found in AP I, Art. 41, which maintains that

1. A person who is recognized or who, in the circumstances, should be recognized to be *hors de combat* shall not be made the object of attack.
2. A person is *hors de combat* if:
 - (a) he is in the power of an adverse Party;
 - (b) he clearly expresses an intention to surrender; or
 - (c) he has been rendered unconscious or is otherwise incapacitated by wounds or sickness; and therefore is incapable of defending himself; provided that in any of these cases he abstains from any hostile act and does not attempt to escape.

For the purposes of our argument, we will primarily be concerned with 2(a) and 2(c), but before moving onto our discussion, there are two general points worth indicating with regards to this treaty instrument. First, the protections associated with *hors de combat* status

are extended not only when a person is recognized to be *hors de combat*, but also when they *should be recognized* to hold that status. With this, the protections of *hors de combat* status are made more objective than some other protections of the LOAC, as they are in force even when an adversary fails to recognize that his opponent is truly *hors de combat* [Appendix A]. Second, the conditions (a)–(c) are disjunctive, meaning that one is *hors de combat* so long as any one of those conditions is met, “provided that in any of these cases he abstains from any hostile act and does not attempt to escape”.

The core question we will now address is who precisely, given the above treaty instrument, is to be deemed *hors de combat* during the course of hostilities. There are some obvious classes of individuals specifically picked out in Art. 41, such as those who are unconscious, those who are so severely sick or wounded as to be utterly incapable of defending themselves, and those who have surrendered. However, what of individuals who are generally combat-effective, yet for some reason cannot defend themselves for contextual reasons, or cannot defend themselves during the duration of a combat engagement (even though they might be threatening later)? Or what of individuals who are unarmed but still dangerous, or those who are armed but so laughably underequipped or undertrained as to be virtually incapable of creating a threat? In what follows, we examine these questions in full, and argue that in most cases, the deciding factor will be contextual, and that by and large, the set of individuals deemed to be *hors de combat* will likely be larger than is usually taken to be the case.

In or Out of Combat?

The core intent of Art. 41 is to provide a defense for persons who are no longer a part of combat, yet who have not yet been taken into custody and made prisoners of war [Appendix A]. It is for this reason that it picks out individuals who cannot defend themselves due to unconsciousness, wounds, or sickness—as these persons could be taken into custody at any time—or picks out individuals who have expressed an intent to surrender (but who have not yet been taken into custody). By the same line of reasoning, combatants “in the power of an adverse Party” are taken to be protected due to the fact that they cannot be seen to be in combat any longer (otherwise they would not be in the power of their enemy); a man with effective means to defend himself and thwart capture cannot be seen to be in anyone’s power. Moreover, the AP I official Commentary maintains that “[a]defenseless adversary is *hors de combat* whether or not he has laid down arms” [11], showing that the core point of Art. 41 is not whether or not enemies have surrendered (or even intend to surrender), but whether or not they can still be seen to be a part of combat. If they cannot be seen that way, that is, if they are out of combat (*hors de combat*), then they are protected under Art. 41 [Appendix A].

An important implication of this element of Art. 41 is that whether or not an enemy is to be deemed *hors de combat* is likely to depend upon very specific contextual factors in a given case. To see this, let us consider a handful of examples.

First consider a hypothetical encounter set in the First Gulf War, which we will call *Tanker*. Suppose a Coalition tank brigade is rolling through the desert, far from any other friendly soldiers or civilians, and they happen to come across an Iraqi platoon of footmen. The Iraqi soldiers have small arms, but nothing besides that, something which the Coalition tankers can see through their sights. The tankers also do not see any means of communication with which the footmen could call for support or radio in the tanks’ position, nor do the tankers have any evidence to suggest there is such equipment out of sight somewhere near. The tanks approach the Iraqis, and are presented with a number of choices: they could continue on their way toward their objective and ignore the Iraqis; they could disarm the Iraqis and then continue on their way; they could take the Iraqis prisoner and return to base; or they could engage and kill the Iraqis (something they can do with impunity, given that the Iraqis have no available means for effectively attacking tanks). (Importantly, in this case the tankers are relatively certain that the Iraqis are powerless against them. If there is any uncertainty, the following argument will not hold. However, it

should also be noted that there are many instances in modern warfare where one party can have such knowledge, and may even have such knowledge long before making contact with an enemy, due to the fact that certain nations possess weapons and armaments which have no effective counter from particular enemies) [Appendix A].

In such a case, the Iraqis are armed, and have shown no intent to surrender, but in this precise case, they are also clearly defenseless. Moreover, given that they could fire their rifles at the tanks all day to no effect, they are, from the tanker's point of view, out of combat. That is, to use the French, they are *hors de combat*. And importantly, they are *hors de combat* from the tanker's perspective not because they are wounded or unarmed (in fact they are both healthy and armed), but because they could not possibly engage the tanks in a meaningful manner. It is the utter and complete irrelevance they hold with regard to the tanks' mission that makes them effectively out of combat, or *hors de combat*.

However, one may object that soldiers are often targeted even when they have no defense, and that this is perfectly correct. More strongly, one might object that it is the goal of military men and women to do their utmost to outmatch their enemies, such that they may strike with impunity. After all, that is the point of calling in air or artillery strikes, to hit an enemy in such a way that they cannot defend themselves, and thus preventing any risk to your own people in the process [Appendix A]. However, the point in the above example is not precisely that the Iraqis are defenseless, but rather that they are out of combat, or *hors de combat*. To put it differently, in order for the words of the AP I Commentary to sensibly capture the customary legal understanding of *hors de combat* status, it should not say that a "defenseless Adversary is *hors de combat*", but rather that a "powerless Adversary is *hors de combat*" [Appendix A]. This is because whether or not a soldier can effectively defend himself is beside the point as to whether or not he is deemed "in combat". However, a soldier who is powerless, that is, one who can have no impact on his enemy, truly is out of combat, or *hors de combat*. To better see this, consider a variant of the above example, which we will call *Tanker with Trooper*.

Imagine that everything is as described above, except that the tank brigade is providing escort to a group of Coalition foot soldiers. These additional Coalition troops are militarily on par with the Iraqis encountered, with both groups being capable of inflicting casualties on the other.

In this scenario, the Iraqis are just as outmatched as before. In fact, with a smattering of foot soldiers alongside the tanks, the tank brigade is, if anything, more overpowering to the Iraqis, making the Iraqis, if anything, more defenseless. However, the Iraqis, though clearly outgunned and without hope, can still inflict casualties on this attacking force. As such, the Iraqis are not properly powerless against this enemy (despite being effectively defenseless), and so the Iraqis are clearly not out of combat. That is, they cannot be seen to be *hors de combat*, despite the fact that they are just as, if not more, defenseless than in the previous case [Appendix A].

If one agrees that the Iraqis are not *hors de combat* in the *Tanker with Trooper* case, then it follows that there is more to being *hors de combat* than merely being defenseless. Moreover, it seems that the essential component to being deemed *hors de combat* turns out to be the very intuitive notion of whether or not a soldier is in fact in or out of combat, which itself hinges on whether or not a soldier holds any power to harm his or her enemy. If I can harm my enemy, then I am obviously a part of combat, but if I may be killed and can do nothing to prevent that, and I cannot do anything to harm my enemy in other ways (say, by killing enemy soldiers other than my attacker), then I am no longer a part of combat. (It is also worth pointing out that there is little military reason to kill such an individual, as it costs time and money—in the form of spent ammunition—to kill such a person, but provides only minute military gain. Given this, there is a potential case to be made that killing such men would violate the principle of necessity, as embodied in AP I, Art. 35 [12–17].

At this point, one may object that our position is too broad, in that many soldiers far from the front lines may permissibly be targeted, even though they are not part of combat and may also be defenseless against the attacks on them. For example, one might imagine

a military supply train bringing tanks and their crews to the front, but which is still very far from the front and will take some time to get there. Even though it is far away, and even though it won't be dropping off its tanks and soldiers anytime soon, it is perfectly legal to target such a train. As such, basing *hors de combat* status on whether or not a soldier is in combat *at that moment* seems to provide an unfounded (and extreme) widening of the protected status.

To this objection, there is no direct response which can be given. However, it is worth pointing out that *hors de combat* status is not understood in a fully coherent manner. For example, a combatant rendered unconscious during fighting is protected under Art. 41, yet one who is sleeping in his barracks is not so protected. Thus, it is not unthinkable that *hors de combat* status may be extended in one situation and withheld in another which is structurally quite similar. Moreover, even if the military supply train might permissibly be targeted in the above example, it almost certainly should not be targeted in a minor variant of the case, to which we now turn.

Imagine the case as above, but further suppose that the supply train is across a deep river, and as part of the war effort, every strong bridge over that river has been destroyed. In this scenario, the train is trapped and cannot move to the front lines. Every tank aboard the train is still in essence combat effective, but there is no way for them to reach the actual fighting. Given this, the tanks are, for all intents and purposes, out of combat. In this case, we should regard such units as *hors de combat*, not because they are necessarily defenseless (though they may also be defenseless against certain forms of attack), but because they are quite simply out of combat, *hors de combat*. And if this granted, what makes them out of combat in this situation has nothing to do with the tanks or soldiers themselves, but is rather a feature of the context of this scenario. They are out of combat because the bridges are out [Appendix A].

Thus, whether or not an individual (or unit) is deemed to be *hors de combat* will depend upon contextual factors, most importantly, whether or not the individual (or unit) can actually contribute to the fighting. In some cases, combatants will be rendered *hors de combat* because there is no way they can harm their adversary, and in others it will be because they are excluded from the fighting altogether due to the actions of others (or possibly due to environmental hindrances). But in any case, one's status as *hors de combat* will be impacted upon by both that fighter's capabilities, and the capabilities of his enemy. And as a related point, whether or not one is armed need not impact on *hors de combat* status in the least. An unarmed man may still be very capable of killing his enemies, and there may also be armed men who are utterly powerless against their enemies, because their enemies are far better armed and armored. The important point is not what one carries or what one can do, but rather what both you and your enemy carry, and what both you and your enemy can do to one another. The greater the gap between the two, the more likely the weaker party will have to be viewed to be out of combat, simply because the stronger will be impervious to harm from the weaker.

At this point, one may grant all of the above arguments, but still wonder what the precise upshot of this is for commanders and soldiers on the ground. Put differently, assuming the Iraqis are *hors de combat* in the *Tanker* case, what does that mean for the tankers themselves? Must they take the Iraqis prisoner? Must they disarm the Iraqis? Even more strongly, if we have assumed the Iraqis are "in the power" of the tankers, does that mean the tankers are required to provide the protections and supplies which would be demanded if they had taken the Iraqis prisoner?

In answer, we would stress that whether or not the Iraqis (or anyone else) are deemed to be *hors de combat* provides no guidance with regards to these further questions. All that *hors de combat* status demands is that the tankers (or anyone else) refrain from making the Iraqis the object of attack, provided the Iraqis abstain from hostile acts and do not attempt to escape. However, it is up to the tankers themselves whether they choose to simply continue on with their mission, or instead to disarm the Iraqis, or instead to take the Iraqis prisoner. What the tankers may not do is simply engage the Iraqis, as they are

to be treated as *hors de combat* in that situation. Importantly, however, they are only *hors de combat* for as long as they act in accordance with the final proviso of Art. 41. Thus, if the tankers demand that the Iraqis throw down their weapons so that the tankers may run over their rifles (effectively disarming those footmen), the Iraqis must comply with this or accept that they have made themselves liable to attack; they must act the prisoner or the enemy, but cannot enjoy the protections of the former while remaining the latter. In this way, Art. 41 provides a protection for persons who are powerless, but does not give blanket permission for powerless individuals to exploit their weakness. It guards them against initial and unnecessary violence, but simultaneously demands that they either peaceably comply with their victors' requests, or accept that they have forfeited the safeguards of *hors de combat* status.

As a final point, it is worth clarifying that we do not think that the foot soldiers in *Tanker* must necessarily be viewed to be *hors de combat*. Moreover, we accept that there is, and will continue to be, reasonable disagreement about when one party is "in the power of an adverse Party", or when one party is truly defenseless in the face of its enemy's might. Our point is that what renders one defenseless or "in the power" of another cannot be determined with a one-sided assessment of a single parties' capabilities; it must be put in the context of a relational assessment based on the capabilities and limitations of both parties. Thus, being armed or unarmed, being wounded or healthy, or being conscious or not, rarely provides adequate information for determining *hors de combat* status. Rather, that will virtually always demand a view to the capabilities of combatants on both sides of a conflict, and to the differences between them. Moreover, one's status as *hors de combat* (or not) may depend on factors wholly outside one's control, like whether or not bridges have been destroyed, or whether or not the enemy you face has men who are vulnerable to your arms. The core lesson is that persons are deemed out of combat, *hors de combat*, in virtue of a myriad of factors, many of which will be derived from the context within which the assessment is being made.

3. *Hors de Combat* Status and AWS

What exactly does this understanding of *hors de combat* then mean for AWS? There will likely be many implications, but we contend that in light of the widely varied, and potentially dynamic situations in which AWS will be deployed, these systems must be capable of responding to changing and contextualized evaluations of an enemy's status as *hors de combat* (or not). AWS should also be treated individually, given that the contexts of their use—airial, naval, and ground-based—are substantively different. And even within these broader categories, different types of AWS will come with their own capabilities and limitations, something that will (potentially) change when an enemy is deemed *hors de combat*. For example, a lightly armored autonomous drone may often encounter enemies who are neither defenseless nor powerless, while a heavily armored autonomous assault platform will likely encounter individuals fitting both of those descriptions. This, in turn, requires that such systems possess a level of technical sophistication high enough to allow for calculations which are sensitive to the many contextual factors that will impact upon the relative strength and power of all belligerent groups. Although this is not technically impossible, the viability as well as necessity of this is beyond technical plausibility at present. Nonetheless, what this betrays is that *hors de combat* status is fundamentally tailored by the entities making such evaluations, given both their capabilities and limitations. A foot soldier makes different evaluations than a tank commander, who makes different evaluations than an autonomous turret sentry, all of which would likewise make different evaluations from a Reaper drone.

In order to illustrate this for AWS, we can take the examples above as inspiration. To begin, we will compare two cases to show how mission objectives might alter what AWS would determine as *hors de combat*.

First, imagine a fully autonomous Reaper drone designated to neutralize an insurgent leader, a case we will call *High-Value Target*. The commander of a forward operating base,

alongside his tacticians, legal professionals, and other experts, determines that the most efficient plan is to neutralize the target via an aerial strike, and that such a strike is lawful. The commander has a fully autonomous Reaper drone outfitted to undertake the mission. The drone is tasked with taking off, arriving at the target's location, confirming that the target is present, confirming that the target is not in the vicinity of so many noncombatants as to render the strike disproportionate, releasing its payload, and then flying back to base. However, suppose that while en route to its target, the drone passes over a company of heavily armed enemy combatants who are isolated in the hills. Despite the fact that such a group is heavily armed, their offensive and defensive capacity against a Reaper drone is functionally irrelevant. In such a case, the hostile party is rendered, as in the *Tanker* case, *hors de combat*.

However, how would an AWS fare in a case similar to *Tanker with Trooper*? Let us imagine that the base commander, instead of sending the Reaper alone, decides to deploy a team of Navy SEALs to neutralize the target, and that they are to travel using ground vehicles. In this case, *High-Value Target with SEALs*, an autonomous Reaper drone is deployed to provide close air support for the SEALs, but all other factors are the same as in *High-Value Target*, with the SEAL team encountering the same heavily armed company of enemy troops. In this case, the Reaper should arguably not view the enemy combatants as being *hors de combat*, because those enemies can inflict casualties on the SEALs, and thus are not powerless and are therefore legitimate targets for the Reaper drone. Yet in this case, the Reaper drone plus the SEALs forms an even greater asymmetric advantage over the enemy combatants. However, like the *Tanker with Trooper* case, the factors that determine *hors de combat* status are not simply whether or not one is able to defend oneself, but rather whether or not one has power to affect one's enemy. In this case, despite the advantage held by the SEALs and Reaper, the enemy troops are nonetheless able to inflict casualties, whereas in *High-Value Target* they are powerless against the Reaper, and thus are (arguably) to be deemed *hors de combat*.

Taken together, these points demonstrate that certain classes of people are not to be treated as *hors de combat a priori*. Rather, contextual factors can change the status of the same group of individuals, all other things remaining equal, simply by changing the other actors (machine or human) involved in a given scenario. For AWS, this means that it would be nonsensical, if not technically unfeasible, to create a blanket method for such systems to determine whether enemies should be classed as *hors de combat*. Moreover, this would be the case even for specific types of AWS, because any given combat scenario is marked by dynamism, something which must be reflected in the way AWS operate in order for them to accurately determine whether or not enemies are *hors de combat*. Given the current technical obstacles for such nuanced programming in AWS, we contend that commanders should hold the final say regarding the rules of engagement and adequate standards of due care in such engagements.

To see the value of this in practice, consider the differences between when military forces are taking a city held by enemy forces as compared to when they are occupying said city. (As a real world example, we might envision the differences between taking Mosul from ISIS versus holding Mosul afterward.) During the course of a large-scale operation to clear enemy combatants from an area, and where there are large numbers of noncombatants who might be unintentionally or intentionally harmed by the enemy, it is sensible to treat all enemy combatants as "in combat", regardless of whether they are at a disadvantage or not. This is because, even if they cannot strike directly at their opposing forces, they still hold power to harm noncombatants, and so, unless they fit one of the stricter categories of *hors de combat* (they are seriously wounded or sick, or are attempting to surrender), they should be seen as legitimate targets. However, once the city has been taken, arguably all persons within its environs should be viewed as "in the power" of the forces currently holding the city. As such, many more persons will need to be considered to be *hors de combat*. Moreover, the degree to which persons may be seen to be in the power of attacking forces may change on a daily basis, with districts shifting hands regularly.

As such, there may be a need to alter rules of engagement regularly, or to utilize rules which are conditional on certain contextual factors like one's control of an area. All of this points to the need for commanders and combatants on the ground to be able to quickly and effectively alter the way AWS acting in their theater of operations view the *hors de combat* status of the enemy. And this, in turn, argues against trying to determine single overarching means of determining this status for all AWS, or even for all AWS in a given environment. Each battle is unique, and commanders should have the ability and means to ensure that their soldiers and hardware, AWS included, are compliant with the laws of war.

To reiterate, what these illustrations aim to demonstrate is that there is no base standard that works for all scenarios. As such, the governing norms in place during deployments of AWS will be highly contextual. In order for AWS to be lawfully deployed, they must be able to pay due heed to the dynamism necessary for determining *hors de combat* status, and it must likewise be assured that humans, in this case deployment commanders, retain the ability to make on-the-fly changes to the targeting principles that are in force in such systems.

In sum, AWS can come in many forms, some more or less fragile or vulnerable. As such, they need to be able to evaluate whether enemies are *hors de combat* in nuanced and contextual ways, paying heed to not just their abilities and their enemies', but also to how changing mission parameters may affect whether or not any enemy is in combat. This, in turn, has implications for our understanding of meaningful human control in the domain of AWS governance and deployment, a point to which we now turn.

4. Regaining Meaningful Human Control (MHC)

Meaningful Human Control (MHC) is a complex, albeit relatively modern concept that emerged from the growing discourse on the ethical and legal issues of AWS. Although there are a number of different frameworks for what constitutes MHC, all frameworks agree that humans must remain in control, or at least have oversight over the decision-making of a system in a non-arbitrary, and thus "meaningful" way. What constitutes this meaningfulness remains debated by scholars, leading to (at least) six different approaches to MHC:

1. Preserving MHC through proper preparation and legitimate context for use, viz. through current NATO targeting procedures [18];
2. Attaining MHC by having a human agent make "near-time decision[s]" in AWS engagement [19];
3. Preserving MHC through adequately training commanders in the deployment and function of AWS to ensure proper attribution of responsibility [20];
4. Attaining MHC through apprising designers/programmers of their moral role in the architecture of AWS [21];
5. Attaining MHC through design requirements involving necessary conditions to *track* the relevant moral reasons for agent actions and *trace* the relevant lines of responsibility through design histories [22,23]; and
6. Preserving MHC by distributing responsibility for decisions through the entirety of the military-industrial complex [24,25].

This paper does not aim to propose or endorse any of these approaches to MHC. However, it does merit noting that human agents are never extricated from decision-making in any of the above approaches, nor do they abdicate full decision-making to the system, even if the system can technically be designed with 'full autonomy' (i.e., the ability to select and engage a target without contemporaneous human input). This central tenet is supported by the results of this paper's exploration of *hors de combat*; as we have demonstrated, it would be incorrect and possibly even dangerous to have a single set of targeting principles for AWS designed to determine *hors de combat* status in enemy combatants. This also applies to specific and narrow types of AWS, showing that control must necessarily remain in the hands of commanders for AWS to be lawfully deployed.

This, of course, does not mean that lawfully fielded AWS must be devoid of any programming to that end more generally. On the contrary, it is imperative that AWS possess some general abilities to discern the status of enemy combatants. However, such general patterns will rarely be sufficient alone, and responsible deployment of AWS will have to address this fact. This means that the human overseers must be able to impact on the way AWS view enemy combatants, and the way AWS determine whether enemy combatants are still in combat.

5. Conclusions

Autonomous weapons systems remain a hotly debated topic in both academia, but also in international public spheres. The debate over the ethics and legality of their design and deployment is further complicated by how *actual* military operations are currently carried out, how the letter of the law regarding military operations relates to and differs from the spirit of the law, as well as how different AWS change how they legally relate to potential combatants. In an effort to clear up some of these complications, this paper drew on the legal articulation of *hors de combat* status as found in AP I, Art. 41, showing how this legal principle might impact on the design and use of AWS. What we aimed to show is that *hors de combat* status cannot be wholly reduced to a set of clear categories that can be programmed into AWS, and that some relevant categories are contingent on dynamically changing contexts in the field. These contexts change when one can be seen to be *hors de combat* and thus systems, in order to remain lawful, must be sensitive to this changing status. As such, this paper concludes that military commanders must retain control over AWS' general targeting behavior, in order to be able to respond to the shifting legality of certain targets due to changing contexts. Meeting this imperative would also provide a greater degree of meaningful human control of AWS, an aim which has independent merit and may prove necessary for addressing other legal and ethical concerns related to autonomous weapons systems.

Author Contributions: The authors have contributed equally to the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We would like to thank Maciej Zajac for providing useful feedback on an earlier draft. All remaining errors are the authors' alone. The views expressed in this paper are not necessarily those of the authors' affiliations.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Note that though one is to be protected whenever one should be recognized to be *hors de combat*, it is only a war crime to make "a person the object of attack in the knowledge that he is *hors de combat*" (AP I, Art. 85.3(e)). Thus, it is possible to fail to meet the demands of Art. 41, while still not acting in a criminal manner, as long as that failure is based in ignorance. Once one knows an enemy is *hors de combat*, any targeting of that enemy will constitute a "grave breach" of the treaty, and thus be deemed a war crime.

ICRC 1987, paragraphs 1601–1603 (pp. 480–482) Note also that the relevant factor is whether or not a person is him- or herself in combat, but not necessarily whether or not a person is still in a military engagement. An individual may still be (partially) involved in a military engagement in virtue of being physically present where hostilities continue to unfold, yet be *hors de combat* due to wounds. As a result, it is possible that some *hors de combat* persons may suffer harm simply due to their proximity to battle, but this will not mean that those who harm them have breached Art. 41, so long as the *hors de combat* persons are not made the object of attack.

Recent debates on the killing of so-called "naked soldiers" make arguments that are similar to those presented here. However, these debates reach a much stricter conclusion

than what we are advancing, and moreover seem to be far less sensitive to the particularities and importance of context in determining the permissibility of targeting (lethal or otherwise). For these reasons, we view these positions with caution, as they seem liable to prove too much. For such arguments, see the recent position developed in [26–28]. For a response to this see the forthcoming [29].

For example, in the UNOSOM II mission to Somalia, U.S. General William Garrison had requested for the delivery of M1 Abrams tanks and an AC-130 Specter gunship. These armaments were ultimately not delivered, but had they been, they would almost certainly have been impervious to any weapon the Somali militias possessed. As such, U.S. forces could have easily assessed that in any future engagements, those units (M1 Abrams and AC-130 gunships) would face enemies who would be powerless against them. The enemies would not be powerless if the fighting was in populated zones where civilians might be caught in a cross-fire, but any engagement in the countryside between an M1 Abrams and Somali militiamen would be certain to see the tankers as facing an enemy who was utterly powerless against them.

For example, such arguments are put forward in Zając n.d. Zając's points are made in relation to the so-called "naked soldier" debates (supra, note 3 above), but with very minor adjustments they would be relevant here as well.

Alternatively, one might also simply ignore the words of the Commentary. However, it is our belief that if a possible (and plausible) interpretation of the Commentary allows one to reconcile its central findings with customary international law, then such is preferable, especially in light of Art. 38(1)(d) of the Statute of the International Court of Justice, which maintains that the court shall apply "judicial decisions and the teachings of the most highly qualified publicists of the various nations, as subsidiary means for the determination of rules of law". Given that the International Committee of the Red Cross represents one of the most respected authorities on International Humanitarian Law, we hold that its Commentary cannot be simply ignored in those instances where it appears to run against common opinion or state practice, but should instead be interpreted in the most charitable light which might allow it to be reconciled with such positions. Thanks to Maciek Zając for pushing me on this point.

This point reflects the facts of many cases in modern warfare. For example, aircraft providing close air support nearly always target individuals who are defenceless against them. However, these craft are providing support to ground troops who are in harm's way, and are under threat from precisely those individuals being targeted by the aircraft providing support. Thus, though the individuals targeted by close support aircraft are defenceless against such craft, those individuals are not powerless, as they can still harm the ground troops being supported by air cover. Perhaps more controversially, aircraft or drones carrying out kill strikes in counterinsurgency environments may be viewed as striking enemies who are defenceless and out of combat. However, we would argue that if those targeted are in a position to harm others (friendly or third party individuals), then they may be justifiably deemed to be in combat, or at least close enough to combat to be permissibly made the object of attack. At any rate, whether or not they are *hors de combat* will depend on the precise nature of the situation, and most importantly (for our purposes) on a number of contextual factors.

If the state contemplating such a strike has a legitimate war aim of disarming its enemy, then it may still be permissible to destroy these tanks. However, this would arguably only be permissible if the tanks could be struck without harming the crews of said tanks. The reason for this is because, so long as the tanks have no possibility of joining the fighting, the war may be carried out without doing anything with regards to these units, and they may still be destroyed after the war has ended (such disarmament may even be made an explicit condition of the peace settlement). It is also worth emphasizing that this example is predicated on their being some degree of certainty that the bridges will stay down. Such certainty will be possible in some cases, but less so in others, and whether or not one knows the bridges will stay down will certainly affect the permissibility of targeting decisions


against units across the river. Again, thanks to Maciek Zajac for suggesting these points (among others).

References

1. Arkin, R.C. *Governing Lethal Behavior in Autonomous Robots*; CRC Press: Boca Raton, FL, USA, 2009; p. 256.
2. Guetlein, M.A. *Lethal Autonomous Weapons: Ethical and Doctrinal Implications*; Technical Report Naval War College: Newport, RI, USA, 2005; p. 34.
3. Johnson, A.S.; Axinn, S. The morality of autonomous robots. *J. Mil. Ethics* **2013**, *12*, 129–141. [CrossRef]
4. Purves, D.; Jenkins, R.; Strawser, B.J. Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory Moral Pract.* **2015**, *18*, 851–872. [CrossRef]
5. Sparrow, R. Killer robots. *J. Appl. Philos.* **2007**, *24*, 62–67. [CrossRef]
6. Sparrow, R. Robots and respect: Assessing the case against autonomous weapon systems. *Ethics Int. Aff.* **2016**, *30*, 93–116. [CrossRef]
7. Roff, H.M. Killing in war: Responsibility, liability, and lethal autonomous robots. In *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century*; Allhoff, F., Evans, N.G., Henschke, A., Eds.; Routledge: Milton Park, UK, 2013; pp. 352–364. [CrossRef]
8. Guarini, M.; Bello, P. Robotic warfare: Some challenges in moving from noncivilian to civilian theaters. In *Robot Ethics: The Ethics and Social Implications of Robotics*; Lin, P., Abney, K., Bekey, G.A., Eds.; MIT Press: Cambridge, MA, USA, 2012; pp. 129–144.
9. Best, G. Restraints on war by land before 1945. In *Restraints on War: Studies in the Limitation of Armed Conflict*; Howard, M.E., Ed.; Oxford University Press: Oxford, UK, 1979; pp. 17–37.
10. Gardam, J.G. Proportionality and force in international law. *Am. J. Int. Law* **1993**, *87*, 391–413. [CrossRef]
11. ICRC. *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949*; Martinus Nijhoff Publishers: Geneva, Switzerland, 1987.
12. Garraway, C. To kill or not to kill? Dilemmas on the use of force. *J. Confl. Secur. Law* **2009**, *14*, 499–510. [CrossRef]
13. Goodman, R. The power to kill or capture enemy combatants. *Eur. J. Int. Law* **2013**, *24*, 819–853. [CrossRef]
14. Mayer, C. Minimizing harm to combatants: Nonlethal weapons, combatants' rights, and state responsibility. In *Routledge Handbook of Ethics and War: Just War Theory in the 21st Century*; Allhoff, F., Evans, N.G., Henschke, A., Eds.; Routledge: Milton Park, UK, 2013; pp. 301–311. [CrossRef]
15. Ohlin, J.D. The duty to capture. *Minn. Law Rev.* **2012**, *97*, 1268–1342. Available online: <https://scholarship.law.umn.edu/mlr/356>. (accessed on 18 March 2021). [CrossRef]
16. Schmitt, M.N. Wound, capture, or kill: A reply to Ryan Goodman's the power to kill or capture enemy combatants. *Eur. J. Int. Law* **2013**, *24*, 855–861. [CrossRef]
17. Wood, N.G. The problem with killer robots. *J. Mil. Ethics* **2020**, *19*, 220–240. [CrossRef]
18. Roorda, M. NATO's Targeting Process: Ensuring Human Control Over (and Lawful Use of) 'Autonomous' Weapons. In *Autonomous Systems: Issues for Defence Policymakers*; Andrew Williams, A., Scharre, P., Eds.; NATO Headquarters Supreme Allied Command Transformation: Norfolk, UK, 2015; pp. 152–168.
19. Asaro, P. On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *Int. Rev. Red Cross* **2012**, *687*–709. [CrossRef]
20. Saxon, D. Autonomous drones and individual criminal responsibility. In *Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*; Di Nucci, E., Santoni de Sio, F., Eds.; CRC Press: Boca Raton, FL, USA, 2016; pp. 17–46.
21. Leveringhaus, A. Drones, automated targeting, and moral responsibility. In *Drones and Responsibility: Legal, Philosophical, and Socio-Technical Perspectives on the Use of Remotely Controlled Weapons*; Di Nucci, E., Santoni de Sio, F., Eds.; CRC Press: Boca Raton, FL, USA, 2016; pp. 169–181.
22. Mecacci, G.; Santoni de Sio, F. Meaningful human control as reason-responsiveness: The case of dual-mode vehicles. *Ethics Inf. Technol.* **2020**, *22*, 103–115. [CrossRef]
23. Santoni de Sio, F.; van den Hoven, J. Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Front. Robot. AI* **2018**, *5*, 15. [CrossRef] [PubMed]
24. Ekelhof, M. Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation. *Glob. Policy* **2019**, *10*, 343–348. [CrossRef]
25. Umbrello, S. Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: A two-tiered approach. *Ethics Inf. Technol.* **2021**, 1–10. [CrossRef]
26. Restrepo, D. Naked soldiers, naked terrorists, and the justifiability of drone warfare. *Soc. Theory Pract.* **2019**, *45*, 103–126. [CrossRef]
27. Restrepo, D. Excuses, justifications, and the just war tradition: Are there good reasons to kill the naked soldier? *J. Glob. Ethics* **2017**, *13*, 58–69. [CrossRef]
28. Restrepo, D. In defense of mercy. *J. Mil. Ethics* **2020**, *19*, 40–55. [CrossRef]
29. Zajac, M. Spare Not a Naked Soldier: A Response to Daniel Restrepo. *J. Mil. Ethics* **2021**, forthcoming.

Article

A Framework for Using Humanoid Robots in the School Learning Environment

Deepti Mishra ^{1,*}, Karen Parish ², Ricardo Gregorio Lugo ³ and Hao Wang ¹

¹ Department of Computer Science (IDI), NTNU—Norwegian University of Science and Technology, 2815 Gjøvik, Norway; hawa@ntnu.no

² Faculty of Education, Inland Norway University of Applied Sciences, 2624 Lillehammer, Norway; Karen.Parish@inn.no

³ Department of Information Security and Communication Technology (IIK), NTNU—Norwegian University of Science and Technology, 2815 Gjøvik, Norway; ricardo.g.lugo@ntnu.no

* Correspondence: deepti.mishra@ntnu.no

Abstract: With predictions of robotics and efficient machine learning being the building blocks of the Fourth Industrial Revolution, countries need to adopt a long-term strategy to deal with potential challenges of automation and education must be at the center of this long-term strategy. Education must provide students with a grounding in certain skills, such as computational thinking and an understanding of robotics, which are likely to be required in many future roles. Targeting an acknowledged gap in existing humanoid robot research in the school learning environment, we present a multidisciplinary framework that integrates the following four perspectives: technological, pedagogical, efficacy of humanoid robots and a consideration of the ethical implications of using humanoid robots. Further, this paper presents a proposed application, evaluation and a case study of how the framework can be used.

Keywords: school learning environment; human–robot interaction; pedagogy; education; efficacy; ethics

Citation: Mishra, D.; Parish, K.; Lugo, R.G.; Wang, H. A Framework for Using Humanoid Robots in the School Learning Environment. *Electronics* **2021**, *10*, 756. <https://doi.org/10.3390/electronics10060756>

Academic Editors: Savvas A. Chatzichristofis, Zinon Zinonos, Ying Tan and Jungong Han

Received: 29 January 2021
Accepted: 18 March 2021
Published: 23 March 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

According to Oxford University researchers, many white and blue-collar jobs are at risk of the Fourth Industrial Revolution [1,2] with its increasing supply and demand of industrial robots globally [3]. According to the Economist Intelligence Unit's recently released Automation Readiness Index, not a single nation included in the study was fully prepared to address the challenge [4]. Robotics and efficient machine intelligence are the building blocks for the coming revolution [5,6]. Countries need a long-term strategy to deal with the challenges of automation and education must be at the center of it. Countries must provide students with a grounding in certain technical skills, such as computational thinking, which are likely to be required in many future roles [4]. Many such roles will also require an understanding of robotics [4].

Humanoid robots have already been used with children to examine various phenomena [7–9]. However, the use of humanoid robots in classrooms is a recent development [10]. The understanding of how children use and learn with these robots is beginning to display signs of future potential [10]. Much of the research to date has focused on the technological capabilities of robots to act as educational tools, focusing for example on language acquisition, Science, Technology, Engineering and Mathematics (STEM) and the basic principles of programming [11,12].

Educational robotics (ER) offer the possibility both of the facilitation and the evaluation of learning as “pedagogical agents” [13]. Through human–robotic interactions and targeted feedback, ER can be programmed to help with learning and develop technical skills through individual and collaborative learning [14]. In particular, ER can be used to target specific

learning outcomes of subject knowledge (i.e., math), skills (i.e., programming and critical thinking) [15]. A recent meta-analysis [16] has shown that ER has been shown to improve knowledge and skills, help with transferring skills to other domains, increase creativity and motivation, increase the inclusion of broad and diverse populations and have an added benefit of increasing teacher development. ER has also shown benefits in STEM subjects [17], but in general, there are mixed findings on the effectiveness of ER [18]. This may be due to methodological shortcomings in design and evaluation [19].

In the context of educational robotics, there have been many efforts made to improve the teaching work in STEM programs to aid both teachers and learners; however, there is a lack of clear-cut guidelines or standards [20]. While ER is a growing field, the benefits to learning outcomes and the evaluations of these interventions need standardized and validated frameworks to assess the efficacy of ER in schools.

Robots have also been used as educational agents with a focus on developing social psychological skills. For example, the iCat robot has been used to teach children to play chess [21] and the Keepon robot for robot-assisted therapy with children on the autistic spectrum [22,23]. Research with the NAO, RoboVie and Tiro humanoid robots have provided insights into the psychological dynamics characterizing social human-robot interaction (HRI) in educational settings [24]. However, multiple studies [25,26] have acknowledged a lack of understanding of the efficacy of humanoid robots in school learning environments (SLEs).

In recent reviews, it has been found that humanoid robots largely act as novices, tutors or peers in educational settings to support learning and that the majority of these applications are driven by technological feasibility and not grounded in didactical theory [12,26]. When theory has provided some didactical frame-working for working with robots in educational contexts, the following approaches have been used: project-based learning, experiential learning and constructionist learning [27].

From the technological perspective, the social element of the interaction between robot and human is difficult to automate and fully autonomous social tutoring behavior in unconstrained environments remains elusive [28]. The robots are limited by the degree to which they can accurately interpret the learner's social behavior [28]. Building artificial "social interaction requires a seamless functioning of a wide range of cognitive mechanisms and their interfaces" ([28] p. 7). This social element of the interaction is especially difficult to automate [12] and needs further research.

In Reference [27], the benefits of incorporating robotics as an educational tool in different areas of knowledge are explored. Another study [29] investigated how robots in the classroom reshape education and foster learning. A recent study has reported that students are generally motivated and have a very positive reaction to the introduction of educational robotics in the academic curriculum [30]. Although humanoid robots have the potential to bring benefits, the incorporation of such technology into SLEs brings its own set of challenges for teachers. These are due to the robot's presence in the social and physical environment and the expectations that the robot creates in the user [28]. In Reference [31], the influence of robots on children's behavior and development and their reaction to the robot's appearance and visual characteristics were examined. There is a call for research into people's interactions with and social reactions towards humanoid robots as a way to shape ethical, social and legal perspectives on these issues as a prerequisite to the successful introduction of robots into our societies [32].

There is a lack of empirical research involving the use of robots in SLEs; therefore, there is a need for more effective analysis of the potential of robotics as a teaching tool for schools [27]. A recent review of the literature [16] observed that the majority of the existing studies lacked an experimental or quasi-experimental design. Another study [33] proposed having more intervention studies with focused research design in K–12 spaces. Recently emphasis has been put on the importance of conducting these interventions with effective robotic pedagogies and underlying theoretical foundations that are required for educational modules in STEM education to make robot-based pedagogies more efficient [16].

Further to this, it has been argued that educational robotics allows for an integrated, multi-disciplinary approach and it is essential to provide a more holistic portrayal of the research on educational robots [16]. In response, this article contributes to the field by presenting a multidisciplinary framework. The multidisciplinary nature of the framework acknowledges that the use of humanoid robots in SLEs must be holistic, rather than focusing on just the technical, or the pedagogical for example. As a position paper, our intention is to present the framework with a proposed application, evaluation and case study by way of an illustration. In particular, we propose that the introduction and evaluation of technology in the classroom should be explored from the following four perspectives: pedagogical, technological/human robot interaction, psycho-social development and a consideration of the ethical implications of using humanoid robots.

Firstly, from an educational perspective and in light of the United Nations Sustainable Development Goal 4 which seeks to “Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all.” [34], can humanoid robots contribute anything to the promotion of quality education? Can a humanoid robot offer a learning experience tailored to the learner, supporting and challenging students in ways unavailable in current resource-limited classrooms? Can humanoid robots contribute to adapted or differentiated education? Can robots be used and thereby “... free up precious time for human teachers, allowing the teacher to focus on what people still do best: providing a comprehensive, empathic, and rewarding educational experience” [12]? What are the pedagogical and didactical foundations or frameworks for the use of humanoid robots in educational settings?

Secondly, how can Artificial Intelligence (AI) and robotic technology be integrated to develop humanoid robots to teach children in SLEs?

Thirdly, how do the human factors interaction with humanoid robotics influence psycho-social development in children (i.e., motivation, self-efficacy, resilience)?

Finally, as AI technology develops and the social interactions between robots and students become more complex, what are the ethical implications of using humanoid robots in educational settings and how do we address these?

This article firstly in Section 2 presents the multidisciplinary framework for using humanoid robots in SLEs. Section 3 includes concrete suggestions on how the proposed framework could be applied and evaluated by researchers and practitioners in different contexts and settings. Section 4 describes a case study related to the application of this framework in a real setting followed by a conclusion and future work.

2. A Multidisciplinary Framework for Humanoid Robots in School Learning Environments

In this section, we present the presuppositions upon which the framework is built. We then present an outline of the framework, including a brief description of each of the four aspects.

2.1. Presupposition

The framework is grounded in the values of inclusive education and the right to education for all. The foundations of inclusive education are built upon the principles of universal human rights and supported by international organizations, such as UNICEF, UNESCO, the Council of Europe, the United Nations and the European Union [35]. The Salamanca Declaration includes all groups of students in danger of marginalization highlighting the right to participate in common learning activities within the ordinary school system, regardless of special needs, gender, ethnicity, culture, social background, etc. [36]. If inclusive education is to become a reality, we must develop learning environments to embrace diversity. For example, some students understand quickly through images, others may prefer texts and readings. Some may deal well with theories, others may learn through experiments and examples and some may have specific learning difficulties [37]. Some learn through engaging in discussion with others, whilst some learn through having the opportunity to work alone. What are the potential ways in which humanoid robots

can contribute to the development of SLEs that embrace diversity and help to promote inclusive education?

With the focus on the learning of each individual, the student is placed at the center of our proposed framework as shown in Figure 1.

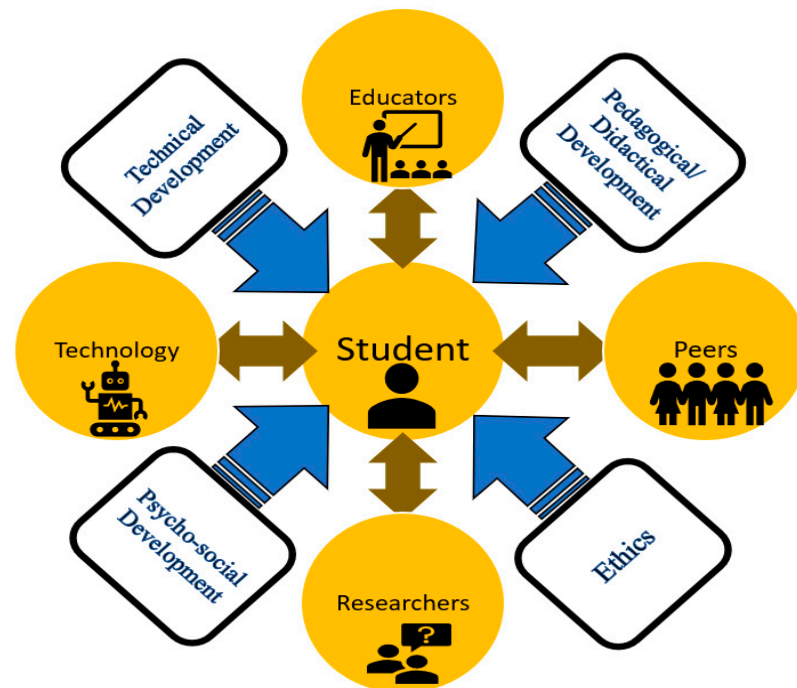


Figure 1. A framework for introducing humanoid robots in school learning environments.

In a two-way collaboration with the student, educators (teachers and assistants), technology (humanoid robots), peers and researchers contribute to the SLE. Through the development of this collaborative learning environment, we seek to explore the following areas.

2.2. Pedagogical/Didactical Development

It is proposed that the pedagogical/didactical aspect of the framework should be grounded in experiential learning theory (ELT) which defines learning as “the process whereby knowledge is created through the transformation of experience. Knowledge results from the combination of grasping and transforming experience” [38] (p. 41). With the focus on learning as a “process”, the ELT model proposes two dialectically related modes of grasping experience—Concrete Experience (CE) and Abstract Conceptualization (AC) [39]. In addition, the ELT model proposes two dialectically related modes of transforming experience—Reflective Observation (RO) and Active Experimentation (AE). The ELT model allows for a diversity of learning styles in students and acknowledges that for some, concrete experience helps them to grasp, perceive and gain new knowledge. However, for others, grasping or taking hold of new information happens through symbolic representation or abstract conceptualization. In the same way, some of us transform or process experiences by watching others and reflecting on the observation of others who are involved in the experience, whereas others actively experiment, jumping right in and doing things [39].

We propose that the ELT model be used as the theoretical foundation for the didactical approach. Further, the didactical approach must be developed as part of an iterative process in collaboration with those working in the specific SLE context.

2.3. Technological Development for Human–Robot Interaction

In order to realize a successful human–robot interaction, a key element—a spoken dialog system—needs to be implemented. A spoken dialog system consists of multiple

components: speech recognition, natural language understanding, dialog management, natural language generation and speech synthesis [40]. On the other hand, social signal processing [41], social expression generation, turn-taking [42] and physical action generation including pose, hand, arm, head movements [43] are also important elements of a spoken dialog system, especially in multiparty dialogs. In order to maintain a multi-turn dialog, the robot has to maintain and understand the conversational history and context [44].

2.4. *Psycho-Social Development*

We propose that the individual and social behaviors, capabilities, constraints and limitations should be explored as the humanoid robot is incorporated into the SLE. The development of behavioral prediction models for user-behavior and performance outcomes can then be used to develop the framework further for human cognition in socio-technical systems. We propose the modeling of user–task interaction at the individual and group level of the SLE through systematic experimentation and naturalistic testing. The research findings then have the potential to be used in the development of evidence-based training modules that cover both the needs of the students and teachers.

2.5. *Ethical Development*

We recognize the need for applied ethical engagement when it comes to the use of humanoid robots in social settings such as learning environments. In particular, we wish to see research with humanoid robots that moves beyond the question of “what can we do technically?” to “what should we do, ethically?”

This framework requires a theoretical contribution by developing a didactical approach that can be used and evaluated through working with humanoid robots in SLEs. The proposed framework allows for the expansion of the boundaries of artificial intelligence by implementing various components of spoken dialog systems for humanoid robots. Further, we propose that the key performance indicators, to assess different aspects of HRI in SLEs, are identified to determine the efficacy of existing HRI metrics and propose new HRI metrics if required. Finally, we propose the development and evaluation of the humanoid robot’s efficacy to help pupils to learn. The framework enables the promotion of students and teachers learning about how robots work, but it also uses robots to help them to learn competencies needed for a future with robots. In particular, the framework incorporates applied ethical engagement as an important aspect of the competencies needed for a future with robots.

3. Proposed Application and Evaluation of the Framework

In this section, we first present our methodological standpoint for the framework followed by an outline of how the framework can be applied, evaluated and executed.

3.1. *Methodology*

The proposed framework requires a multidisciplinary and multiple-methods approach that will include applied, qualitative and quantitative aspects. Whilst respecting the integrity of the different paradigms, we propose the utilization of different ways of knowing to expand our understanding of the potential ways in which humanoid robots can be used in SLEs to promote student learning. With such a research design, we can expand the scope of our understanding as different methods will be used to assess different aspects of the phenomenon [45]. By combining qualitative and quantitative aspects in our evaluation of humanoid robots in the SLE, we incorporate both subjective experiences and objective observations [46,47].

3.2. *Methodological Implications*

Research into understanding and learning the effects of human–robotic interactions in schools is still in the early stages. The applied nature and real-world complexity of this field mean this research is multidisciplinary. The use of a mixed-methods research design

that includes qualitative, quantitative and theory can lead to insights and discoveries in this novel domain. There are few existing theoretical frameworks in the literature encompassing these research questions and validated approaches. This requires using validated approaches from different disciplines, that is, psychology, human factors and educational research.

This framework also promotes using naturalistic settings over laboratory settings due to the nature of the domain studied. Socio-technical domains incorporate human-technology interactions while in social settings (i.e., classroom) but research frameworks need to be validated across domains. Experimental laboratory settings are applicable to identify the impact of variable manipulation on outcome variables and may give high internal validity, but it is limited in generalizations. Naturalistic design allows the observation of participants in their natural settings and observes for outcomes. While this approach may have low internal validity, it is high in ecological validity, therefore the findings can be generalized to other populations.

Both quantitative and qualitative approaches need to include their respective approaches to validity (See for qualitative approaches [48]). By using a mixed-methods design and triangulation methods, new insights on ELT approaches can be validated and form the foundations for future work that are applicable to all four domains (technological, psychological, educational and ethical). This approach will allow for the reflective observation and active experimentation of the ELT framework.

3.3. Preparation

We propose that the framework must be situated within the specific context and take into account the needs of the teachers and SLE. In particular, the needs of the SLE must be established regarding the identification and definition of scenarios related to existing educational contents suitable for the use of humanoid robotics, for example, grade/age, types of school, state/private, types of learning formats, group/individual/whole class. In order to complete this task, the researcher will need to engage in a period of consultation and information gathering with school teachers. This activity may take multiple sessions as the researchers learn about existing educational content to be able to develop a set of scenarios involving humanoid robots depending upon the learner profile(s) to deliver context-appropriate and tailored educational content.

This preparation stage also involves organizing information sessions for teachers and parents along with obtaining necessary permissions from relevant ethical boards and parents since these activities involve children.

In addition, in this preparation stage, the researchers must identify and design appropriate data collection tools that measure learning outcomes, performance, user interface experience and psychosocial skill development.

3.4. In-Context Development of Various Aspects and Evaluation Instruments

3.4.1. Pedagogical

As stated in Section 2.2, we propose the development of a didactical approach to working with humanoid robots in SLEs based on ELT [38]. The didactical approach should, however, be developed in collaboration with the teachers and based on the needs of the specific SLE context. We propose that this should be an iterative process to allow for the investigation of both how the development of a didactical approach can contribute to more effective working with humanoid robots in specific SLEs, and in what ways educational activities with humanoid robots can promote learning.

We propose that to evaluate the effectiveness of the pedagogical aspect of the framework the main approach should be qualitative and exploratory. Since programming robots for social interaction and for teaching is highly creative, it requires co-design and development with stakeholders, and an iterative development methodology will be highly beneficial. Semi-structured interviews could be used to evaluate humanoid robots in

SLEs with respect to HRI, robot behavior, natural language understanding and social signal processing.

In addition, a qualitative approach could be used to focus on both student and teacher experiences of introducing and working with humanoid robots in the classroom. The advantage of adopting a qualitative approach is that it allows us to explore how the students and teachers interact with the humanoid robots, including feelings, strengths/challenges and ethical considerations of working with humanoid robots.

3.4.2. Technical

The main approach for technical development can be iterative, requiring continuous qualitative and quantitative evaluation. We propose to implement a spoken dialog system consisting of various components (as shown in Figure 2) to create engaging educational activities with humanoid robots.

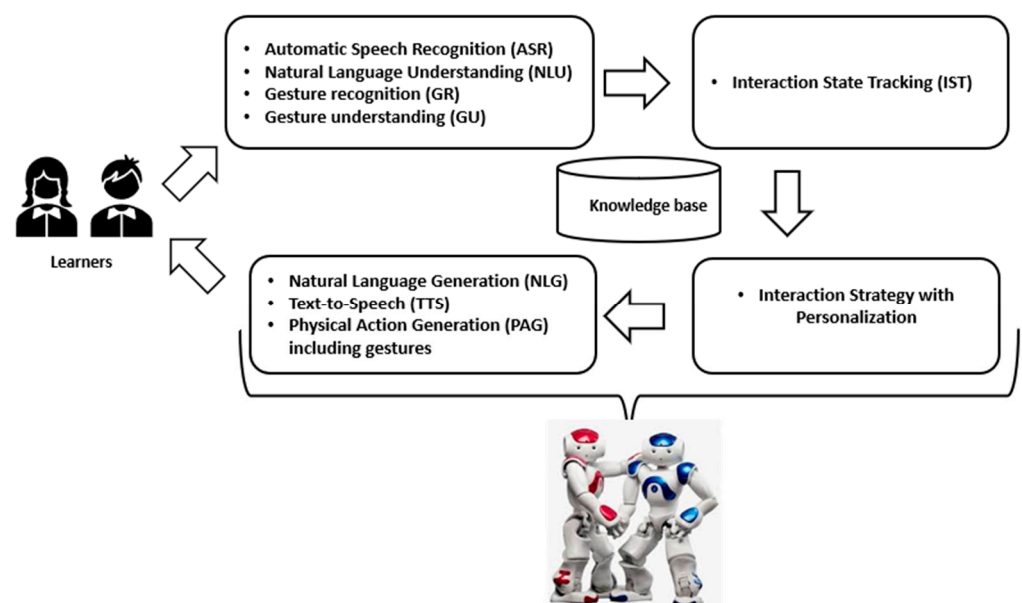


Figure 2. Proposed implementation of spoken dialog system (SDS) in SLEs.

- Automatic speech recognition, natural language understanding, gesture recognition and understanding so that the robot can perceive the learning environment and human participants;
- Interaction state tracking so that the robot can determine the current state comprising of the dialog act and/or gesture by maintaining a “memory” to store interaction history and contextual information;
- The robot will then form an interaction strategy plan consisting of various actions with personalization;
- Natural language generation, text to speech and physical action generation including gestures with personalization for adaptive learning customized according to the level and learning speed of the user.

The above-mentioned activities can be designed for two settings, individual educational activities and multi-party educational activities with group interactions and teamwork between peers.

Existing tools and libraries provided with commercially available humanoid robots can be explored for components such as automatic speech recognition and generation, natural language understanding and generation, text to speech synthesis and the main focus can be on components such as creating a knowledge base for efficient dialog management to be used with the humanoid robot in SLEs. Other available techniques and methods such as for natural language understanding, deep learning methods involving Convolutional Neural

Networks [49] or Recurrent Neural Networks [50] and for leveraging external knowledge for natural language understanding [51] and natural language generation [52], knowledge graphs can also be explored.

Various metrics (e.g., cognitive interaction, degree of monotonicity, human awareness—human recognition, characterization and adaptation, robots' self-awareness, safety) have been discussed to evaluate and assure functionality of humanoid robots [53]. However, a key factor that limits the success of human–robot teams is the lack of consistent test methods and metrics for assessing the effectiveness of HRI [54] since existing metrics are not sufficient to capture all aspects of HRI [53] in every setting [55]. Therefore, HRI metrics in conjunction with observations, quantitative (e.g., questionnaire) and qualitative methods (e.g., semi-structured interviews) can be used to evaluate humanoid robots in SLEs.

3.4.3. Psycho-Social

We propose the development of behavioral prediction models for user-behavior and performance outcomes that are situated in the specific context of the SLE. This can be achieved through the modeling of user-task interaction at the individual and group level of the SLE through systematic experimentation and naturalistic testing.

We propose that by using validated approaches from human factors and cognitive engineering, we can evaluate the efficacy of humanoid robots on the psycho-social development of learners (i.e., motivation, self-efficacy, resilience). This can be achieved by developing and validating applied interventions based on human factors and cognitive engineering aspects where the interaction of individual aspects of human behavior (microcognition; i.e., self-efficacy, resilience, metacognition) and naturalistic environments (macrocognition; i.e., shared situational awareness, communication) are considered in both human–robot interaction and human–human interactions. These measures will be analyzed using social science paradigms (i.e., statistical analysis, cognitive task analysis, qualitative interviews).

3.4.4. Ethics

Careful consideration must be given to ethics and it is proposed that these considerations are situated in the specific context in which the research is taking place. Some considerations to be taken are, first, what are the implications for the students and teachers/assistants in introducing humanoid robots into the SLE? As researchers, we have an ethical responsibility to “do no harm” to those who participate in such studies. Secondly, as the technological advancement of artificial intelligence continues and humanoid robots become more autonomous, what ethical applications apply to the robots? Thirdly, and related to the above two, how do we prepare students and teachers/assistants for a future with robots which are founded upon ethical considerations?

4. Case Study

This section presents an example of how the framework can be implemented.

Aim: To explore how humanoid robots can assist teachers to promote Mathematics and programming skills.

Sample: Grade 6 students ($n = 20$) and teachers ($n = 2$)

Preparation: Researchers have two meetings with the grade 6 teachers to prepare the content of the three-day workshop, including discussion surrounding the learning needs of the students. Ethical consent is gained from the relevant body to conduct the research. An information meeting is held for teachers and parents/guardians of participants under the age of 16. Informed consent is gained from participants and the parents/guardians of participants under the age of 16. The discussion related to the selection of evaluation methods (e.g., observations, quantitative and qualitative) and instruments is also initiated at this stage.

Didactical approach: Execution of a three-day workshop which involves the following activities for the participants:

Activity 1—Introduction to robots—including a presentation and class discussion led by the researchers. Informed consent is explained to the participants.

Activity 2—Participants complete a pre-test structured questionnaire of their metacognitive judgment on how they expect to do working with the robot, math and programming.

Activity 3—Participants are separated into groups of four or five by the regular class teachers. Each group participates in a one-hour practical session led by the researchers. The session includes basic programming and math tasks using the robot.

Activity 4—Participants complete a post-test structured questionnaire of their metacognitive judgment about how well they think they did working with the robot, math and programming.

Activity 5—The researchers conduct semi-structured group interviews with each of the four groups of grade 6 students to gather in-depth data about the experiences of working with the robot.

Activity 6—Plenary—including a presentation and class discussion surrounding the experiences of working with robots, what a future with robots looks like and the ethical considerations to working with robots, led by the researchers.

Activity 7—The researchers conduct semi-structured group interviews with the grade 6 teachers to gather in-depth data about the experiences of working with the robot.

Technical development: The robots are programmed for activities related to mathematics and programming tasks. This is done in multiple iterations so that other researchers and teachers can provide feedback in order to improve these activities before the workshop with the participants. Questionnaire and semi-structured interviews are used to evaluate human–robot interaction along with participants’ views on the current technical capabilities, limitations and potential improvements in robot activities for future workshops.

Psycho-social development: This is explored during the three-day workshop and in particular through the collecting of pre- and post-test data that explores the participants’ self-efficacy and meta-cognition.

Ethical development: This occurs primarily through the discussions during Activity 6 and in the semi-structured interviews. This is also covered through following ethical guidelines such as informed consent.

Evaluation: Both qualitative and quantitative analysis of the interviews and pre- and post-test data can be analyzed using validated methodologies. Inferential statistics can be used for quantitative data, while qualitative approaches such as Interpretive Phenomenological Analysis or Thematic Analysis can be used to analyze interview data. These approaches have been validated across social and technical domains to measure experiences, interactions and outcomes.

5. Conclusions and Future Work

This position paper has proposed a framework that addresses an under-researched and not well-understood aspect of humanoid robots in SLEs. Rapid technological progress in SLEs needs to be balanced with a holistic approach to research that attempts to support human adaptation in rapidly changing socio-technical system dynamics. With such a multidisciplinary framework, we offer the possibility to move beyond extending the technical possibilities to evaluating how technological advancements can be used in an ethical way to benefit individuals and society through education. In particular, the multidisciplinary framework presented here integrates the technological, pedagogical, psycho-social and ethical aspects of HRI. Further, this paper has presented a possible way to apply and evaluate the framework, methodologically, along with an example of a case study. It is hoped that readers will be inspired to adopt this interdisciplinary framework as their starting point for research into how humanoid robots can be used effectively in SLEs and contribute to the development of the research base within this field.

Although this study includes concrete suggestions regarding the application and evaluation of the proposed interdisciplinary framework along with a case study describing its application in a real setting with a focus on learning mathematical and programming

concepts, it is beyond the scope of this paper to include empirical data. Further research is needed to empirically evaluate the framework in order to derive more grounded conclusions. Therefore, future work will report on the comparative analysis, both by longitudinal research and by comparison with the results of experiments designed within different courses and also at other schools.

If humanoid robots can contribute positively towards the SLE and increased learning opportunities (motivation, self-efficacy, resilience) then this will benefit both students in the short and long-term, and in turn society. This framework has the potential to impact the teaching and training of future generations of students that can be reached and benefit from the implementation of the proposed framework. The addition of humanoid robotics in the classroom may facilitate the learning process in students who struggle and may decrease apprehensive behaviors in students, allowing for cognitive processes to open up for more efficient learning and the promotion of inclusive education for all.

Author Contributions: Conceptualization, D.M., K.P. and R.G.L.; Investigation, D.M., K.P., R.G.L. and H.W.; Methodology, D.M., K.P., R.G.L. and H.W.; Visualization, D.M. and K.P.; Writing—original draft, D.M. and K.P.; Writing—review and editing, D.M., K.P., R.G.L. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. The APC cost is covered by NTNU—Norwegian University of Science and Technology.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the NSD—Norwegian Centre for Research Data (reference number: 238991 and date of approval: 28 September 2020).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. BBC. Written Evidence to UK Parliament Artificial Intelligence Select Committee's Publications. Available online: <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10001.htm> (accessed on 23 March 2021).
2. Frey, C.B.; Osborne, M.A. The future of employment: How susceptible are jobs to computerisation? *Technol. Forecast. Soc. Chang.* **2017**, *114*, 254–280. [CrossRef]
3. Robots double worldwide by 2020. In Proceedings of the International Federation of Robotics Press Conference, Tokyo, Japan, 18 October 2018.
4. Economist Intelligence Unit. *The Automation Readiness Index: Who Is Ready for the Coming Wave of Automation?* Economist Intelligence Unit: London, UK, 2018.
5. Accenture UK Limited. Written Evidence to UK Parliament Artificial Intelligence Select Committee's Publications. Available online: <https://www.gov.uk/government/publications/government-response-to-the-house-of-lords-select-committee-on-artificial-intelligence> (accessed on 23 March 2021).
6. Kim, J.-H.; Myung, H.; Lee, S.-M. Robot. Intelligence technology and applications. In Proceedings of the 6th International RiTA Conference 2018, Kuala Lumpur, Malaysia, 16–18 December 2018; Springer: Berlin/Heidelberg, Germany, 2019; Volume 1015.
7. Tanaka, F.; Cicourel, A.; Movellan, J.R. Socialization between toddlers and robots at an early childhood education center. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 17954–17958. [CrossRef]
8. Mazzoni, E.; Benvenuti, M. A robot-partner for preschool children learning English using socio-cognitive conflict. *J. Educ. Technol. Soc.* **2015**, *18*, 474–485.
9. Ioannou, A.; Andreou, E.; Christofi, M. Pre-schoolers' interest and caring behaviour around a humanoid robot. *TechTrends* **2015**, *59*, 23–26. [CrossRef]
10. Crompton, H.; Gregory, K.; Burke, D. Humanoid robots supporting children's learning in an early childhood setting. *Br. J. Educ. Technol.* **2018**, *49*, 911–927. [CrossRef]
11. Balogh, R. Educational robotic platform based on arduino. In Proceedings of the 1st International Conference on Robotics in Education RiE 2010, Bratislava, Slovakia, 16–17 September 2010; pp. 119–122.
12. Powers, K.; Gross, P.; Cooper, S.; McNally, M.; Goldman, K.J.; Proulx, V.; Carlisle, M. Tools for teaching introductory programming: What works? In Proceedings of the 37th SIGCSE Technical Symposium on Computer Science Education, New York, NY, USA, 3–5 March 2006; pp. 560–561.

13. Tang, A.L.; Tung, V.W.S.; Cheng, T.O. Dual roles of educational robotics in management education: Pedagogical means and learning outcomes. *Educ. Inf. Technol.* **2020**, *25*, 1271–1283. [CrossRef]
14. Scaradozzi, D.; Screpanti, L.; Cesaretti, L. Towards a definition of educational robotics: A classification of tools, experiences and assessments. In *Smart Learning with Educational Robotics*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 63–92.
15. Ronsivalle, G.B.; Boldi, A.; Gusella, V.; Inama, C.; Carta, S. How to implement educational robotics' programs in Italian schools: A brief guideline according to an instructional design point of view. *Technol. Knowl. Learn.* **2019**, *24*, 227–245. [CrossRef]
16. Anwar, S.; Bascou, N.A.; Menekse, M.; Kardgar, A. A systematic review of studies on educational robotics. *J. Pre-Coll. Eng. Educ. Res.* **2019**, *9*, 2. [CrossRef]
17. Aris, N.; Orcos, L. Educational robotics in the stage of secondary education: Empirical study on motivation and STEM skills. *Educ. Sci.* **2019**, *9*, 73. [CrossRef]
18. Zhong, B.; Xia, L. A systematic review on exploring the potential of educational robotics in mathematics education. *Int. J. Sci. Math. Educ.* **2020**, *18*, 79–101. [CrossRef]
19. Hoorn, J.F.; Huang, I.S.; Konijn, E.A.; van Buuren, L. Robot tutoring of multiplication: Over one-third learning gain for most, learning loss for some. *Robotics* **2021**, *10*, 16. [CrossRef]
20. Phan, M.-H.; Ngo, H.Q.T. A multidisciplinary mechatronics program: From project-based learning to a community-based approach on an open platform. *Electronics* **2020**, *9*, 954. [CrossRef]
21. Leite, I.; Castellano, G.; Pereira, A.; Martinho, C.; Paiva, A. Modelling empathic behaviour in a robotic game companion for children: An ethnographic study in real-world settings. In Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction, Boston, MA, USA, 5–8 March 2012; pp. 367–374.
22. Feil-Seifer, D.; Mataric, M. Robot-assisted therapy for children with autism spectrum disorders. In Proceedings of the 7th International Conference on Interaction Design and Children, Chicago, IL, USA, 11–13 June 2008; pp. 49–52.
23. Kozima, H.; Michalowski, M.P.; Nakagawa, C. Keepon. *Int. J. Soc. Robot.* **2009**, *1*, 3–18. [CrossRef]
24. Lehmann, H.; Rossi, P.G. Social robots in educational contexts: Developing an application in enactive didactics. *J. eLearn. Knowl. Soc.* **2019**, *15*, 27–41.
25. Kazakoff, E.R.; Sullivan, A.; Bers, M.U. The effect of a classroom-based intensive robotics and programming workshop on sequencing ability in early childhood. *Early Child. Educ. J.* **2013**, *41*, 245–255. [CrossRef]
26. Ros, R.; Baroni, I.; Demiris, Y. Adaptive human-robot interaction in sensorimotor task instruction: From human to robot dance tutors. *Robot. Auton. Syst.* **2014**, *62*, 707–720. [CrossRef]
27. Benitti, F.B.V. Exploring the educational potential of robotics in schools: A systematic review. *Comput. Educ.* **2012**, *58*, 978–988. [CrossRef]
28. Belpaeme, T.; Kennedy, J.; Ramachandran, A.; Scassellati, B.; Tanaka, F. Social robots for education: A review. *Sci. Robot.* **2018**, *3*, eaat5954. [CrossRef] [PubMed]
29. Karim, M.E.; Lemaignan, S.; Mondada, F. A review: Can robots reshape K-12 STEM education? In Proceedings of the 2015 IEEE International Workshop on Advanced Robotics and Its Social Impacts (ARSO), Lyon, France, 1–3 July 2015; pp. 1–8.
30. Román-Graván, P.; Hervás-Gómez, C.; Martín-Padilla, A.H.; Fernández-Márquez, E. Perceptions about the use of educational robotics in the initial training of future teachers: A study on steam sustainability among female teachers. *Sustainability* **2020**, *12*, 4154. [CrossRef]
31. Toh, L.P.E.; Causo, A.; Tzuo, P.-W.; Chen, I.-M.; Yeo, S.H. A review on the use of robots in education and young children. *J. Educ. Technol. Soc.* **2016**, *19*, 148–163.
32. De Graaf, M.M. An ethical evaluation of human-robot relationships. *Int. J. Soc. Robot.* **2016**, *8*, 589–598. [CrossRef]
33. Xia, L.; Zhong, B. A systematic review on teaching and learning robotics content knowledge in K-12. *Comput. Educ.* **2018**, *127*, 267–282. [CrossRef]
34. United Nations. *The Sustainable Development Goals Report 2019*; United Nations: New York, NY, USA, 2019.
35. Haug, P. Understanding inclusive education: Ideals and reality. *Scand. J. Disabil. Res.* **2017**, *19*, 206–217. [CrossRef]
36. Unesco. The Salamanca Statement and Framework for action on special needs education. In Proceedings of the World Conference on Special Needs Education—Access and Quality, Salamanca, Spain, 7–10 June 1994; Unesco: Salamanca, Spain, 1994.
37. Truong, H.M. Integrating learning styles and adaptive e-learning system: Current developments, problems and opportunities. *Comput. Hum. Behav.* **2016**, *55*, 1185–1193. [CrossRef]
38. Kolb, D.A. *Experiential Learning: Experience as the Source of Learning and Development*; Prentice-Hall International: Upper Saddle River, NJ, USA, 1984.
39. Kolb, D.A.; Boyatzis, R.E.; Mainemelis, C. Experiential learning theory: Previous research and new directions. *Perspect. Think. Learn. Cogn. Styles* **2001**, *1*, 227–247.
40. Lison, P.; Meena, R. Spoken dialogue systems: The new frontier in human-computer interaction. *XRDS Crossroads ACM Mag. Stud.* **2014**, *21*, 46–51. [CrossRef]
41. Funakoshi, K. A multimodal multiparty human-robot dialogue corpus for real world interaction. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 7–12 May 2018; pp. 35–39.
42. Baxter, P.; Kennedy, J.; Belpaeme, T.; Wood, R.; Baroni, I.; Nalin, M. Emergence of turn-taking in unstructured child-robot social interactions. In Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Tokyo, Japan, 4–6 March 2013; pp. 77–78.

43. Jokinen, K.; Wilcock, G. Multimodal open-domain conversations with robotic platforms. In *Multimodal Behavior Analysis in the Wild*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 9–26.
44. Yang, L.; Qiu, M.; Qu, C.; Chen, C.; Guo, J.; Zhang, Y.; Croft, W.B.; Chen, H. IART: Intent-aware response ranking with transformers in information-seeking conversation systems. In Proceedings of the Web Conference 2020, Online, 20–24 April 2020; pp. 2592–2598.
45. Greene, J.C. *Mixed Methods in Social Inquiry*; John Wiley & Sons: Hoboken, NJ, USA, 2007; Volume 9.
46. Almalki, S. Integrating quantitative and qualitative data in mixed methods research—Challenges and benefits. *J. Educ. Learn.* **2016**, *5*, 288–296. [CrossRef]
47. Golafshani, N. Understanding reliability and validity in qualitative research. *Qual. Rep.* **2003**, *8*, 597–607.
48. Flick, U. *An Introduction to Qualitative Research*; SAGE Publications: Thousand Oaks, CA, USA, 2018.
49. Kim, S.; Banchs, R.E.; Li, H. Exploring convolutional and recurrent neural networks in sequential labelling for dialogue topic tracking. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 1, pp. 963–973.
50. Yao, K.; Peng, B.; Zhang, Y.; Yu, D.; Zweig, G.; Shi, Y. Spoken language understanding using long short-term memory neural networks. In Proceedings of the 2014 IEEE Spoken Language Technology Workshop (SLT), South Lake Tahoe, NV, USA, 7–10 December 2014; pp. 189–194.
51. Heck, L.; Hakkani-Tür, D.; Tur, G. Leveraging knowledge graphs for web-scale unsupervised semantic parsing. In Proceedings of the 14th Annual Conference of the International Speech Communication Association, Lyon, France, 25–29 August 2013; pp. 1594–1598.
52. Li, W.; Peng, R.; Wang, Y.; Yan, Z. Knowledge graph based natural language generation with adapted pointer-generator networks. *Neurocomputing* **2020**, *382*, 174–187. [CrossRef]
53. Murphy, R.R.; Schreckenghost, D. Survey of metrics for human-robot interaction. In Proceedings of the 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Tokyo, Japan, 4–6 March 2013; pp. 197–198.
54. Marvel, J.A.; Bagchi, S.; Zimmerman, M.; Aksu, M.; Antonishek, B.; Wang, Y.; Mead, R.; Fong, T.; Amor, H.B. Test methods and metrics for effective HRI in collaborative human-robot teams. In Proceedings of the 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Daegu, Korea, 11–14 March 2019; pp. 696–697.
55. Begum, M.; Serna, R.W.; Kontak, D.; Allspaw, J.; Kuczynski, J.; Yanco, H.A.; Suarez, J. Measuring the efficacy of robots in autism therapy: How informative are standard hri metrics. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, Portland, OR, USA, 1–4 March 2015; pp. 335–342.

Commentary

The Social Robot in Rehabilitation and Assistance: What Is the Future?

Daniele Giansanti

Centre Tisp, Istituto Superiore di Sanità, 00131 Rome, Italy; daniele.giansanti@iss.it; Tel.: +39-06-4990-2701

Abstract: This commentary aims to address the field of social robots both in terms of the global situation and research perspectives. It has four polarities. First, it revisits the evolutions in robotics, which, starting from collaborative robotics, has led to the diffusion of social robots. Second, it illustrates the main fields in the employment of social robots in rehabilitation and assistance in the elderly and handicapped and in further emerging sectors. Third, it takes a look at the future directions of the research development both in terms of clinical and technological aspects. Fourth, it discusses the opportunities and limits, starting from the development and clinical use of social robots during the COVID-19 pandemic to the increase of ethical discussion on their use.

Keywords: e-health; medical devices; m-health; rehabilitation; robotics; organization models; artificial intelligence; electronic surveys; social robots; collaborative robots

1. Introduction

We can certainly place among the most marvelous and shocking technological developments of recent years those of collaborative robotics and, among them, those related to social robotics.

The social robot represents an important technological issue to deeply explore both from a technological and clinical point of view. It has been highlighted in an editorial in the Special Issue of the journal *Healthcare* entitled “Rehabilitation and Robotics: Are They Working Well Together?” [1]. Among the most important directions in the development of social robotics connected to assistance and rehabilitation we find, in a wider approach to the process of rehabilitation and assistance, the following:

- To invest in social robots specifically designed as support during rehabilitation phases (such as, for example, in the care of the elderly).
- To invest in social robots specifically designed as cultural mediators to support during communication/therapy activity (such as in the care of autism).
- To address the problem of empathy in robotics, especially in relation to interaction with social robots.

In fact, starting from the experiences of collaborative robotics, social robots have spread and are opening new opportunities in the field of the rehabilitation and assistance of fragile subjects with different types of problems, ranging from neuromotor disabilities to those of a communicative and psychological type. A particular acceleration in this area has also certainly been due to the COVID-19 pandemic. The need to maintain social distancing, combined with that of (a) ensuring the continuity of care and (b) giving a communicative type of support, has prompted us to look in the direction of social robots as a possible solution at hand: a real lifebuoy. We have, therefore, increasingly begun to look at social robots both, in a more futuristic way, as a potential substitute for human health care and rehabilitation and, in a more realistic and ethically acceptable way, as a reliable possible mediator/facilitator between humans in the field of rehabilitation and assistance. To tell the truth, even before the pandemic, some of the “social” potential of robots had begun to scare us. Recent challenges in some games (which involve a high degree of social

Citation: Giansanti, D. The Social Robot in Rehabilitation and Assistance: What Is the Future?. *Healthcare* **2021**, *9*, 244. <https://doi.org/10.3390/healthcare9030244>

Academic Editor: Tin-Chih Toly Chen

Received: 21 January 2021

Accepted: 14 February 2021

Published: 25 February 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

interactions based on tactics) between robots and humans have in fact shown us how the computational abilities of robots have definitively knocked out what we previously believed to be the primacy of human intelligence. In 2016, years after Deep Blue's [2–4] famous defeat of Kasparov at chess [5,6], a computer called AlphaGo [7] beat the world champion of Go [8,9], a game much more complex than chess; in fact, in this game, the possible options for the first move are 361 (20 in chess) and the second are 130,000 (400 in chess!). According to the scholars of this game, to win, it is necessary to be familiar with the models of social interaction that go far beyond simple computation! The following questions immediately emerge:

- With AlphaGo, are we crossing the threshold between the two forms of artificial and human intelligence, and what does this entail for future developments?
- What is the boundary between a social robot and a powerful computer?
- Does a social robot have at least a mechatronic body (AlphaGo does not have one)?
- Is an interactive video connected to a computer attached to a mobile body/column sufficient to characterize a social robot?
- What degree of autonomy must a social robot have in any case?
- Is all of this ethically acceptable?

As scholars in the field of assistance and rehabilitation, we also question ourselves on these points, which touch on important aspects of (a) scientific research in mechatronics, neuroscience, artificial intelligence and bioengineering; (b) bioethics; and (c) economics and politics, ranging from regulatory to organizational aspects. In light of this, taking into account the focus of this Special Issue, the goal of our study is mainly to produce a commentary that is useful in the field of research without, however, where possible, neglecting the other aspects. In particular, we wish to highlight in this study a map point and a conceptual contextualization of these technologies starting from the roots, which are based on corobotics, and understand what direction these devices are taking and what we can expect in the future.

2. The Social Robot as an Evolution of the Collaborative Robot

2.1. Collaborative Robots

The term *corobot* or *cobot* derives from the merging of the term collaborative with the term robot [10]. It appeared in the Wall Street Journal in its millennium edition on 1 January 2000 [11] and refers to technologies used since 1996 thanks to the ingenuity of two professors from Northwestern University, J. Edward Colgate and Michael Peshkin. Cobots are robots designed to interact with humans from a certain work environment and in an interaction workspace. Currently, among the robotics sectors, this sector represents one of the greatest developments.

The International Federation of Robotics [12], a professional, nonprofit organization, recognizes two types of robots: industrial robots used in automation and collaborative robots that can be of service for professional and home use. In the field of collaborative robots, there are four groupings:

1. Reactive collaboration: the robot responds to the movement of the worker in real time;
2. Cooperation: the human and robot are both in motion and work simultaneously;
3. Sequential collaboration: the human and robot share part or all of a workspace but do not work simultaneously;
4. Coexistence: there is no shared workspace, but the human and robot work together.

2.2. Social Robots

The ability to interact and work with humans is a characteristic of collaborative robots. However, if this interaction and work activity is more characterized by social interaction until it becomes the key role, then we are dealing with a social robot, also called a socially interactive robot [13].

In other words, social robots are collaborative robots evolved/specialized in social interaction, and their work is social interaction.

We must take into account that robots are and will be increasingly part of our lives. Interaction with artificial intelligence in workplaces, shops, healthcare facilities and numerous other meeting places will be increasingly frequent.

Social robots (SRs) in their collaborative interaction are capable [13] of:

- Establishing and maintaining social relationships;
- Learning social skills development and role models;
- Using “natural” signals, such as gestures and gaze;
- Expressing emotions and are able to perceive them;
- Communicating with high-level dialog;
- Expressing one’s own personality and distinctive character.

SRs can be used for a variety of purposes; for example, as educational tools and therapeutic aids. There are several examples of SRs designed for use by elderly people [14–17], in nursing homes or in hospitals, for example, to:

- (a) Support certain motor activities;
- (b) Support the elderly during feeding;
- (c) Support them in drug therapy; for example, by reminding them to take a drug;
- (d) Support them from a cognitive point of view; for example, by stimulating them with games and supporting them from the point of view of communicative interaction, even as simple company;
- (e) Or, more generally, provide support as a hospital assistant.

For this reason, SRs are being considered among the key gerontechnologies [17] for the future.

In the COVID-19 era, there has been an increase in the use of SRs in the above-listed desirable activities due to the necessary supervening obligation of social distancing to combat the pandemic [18]. One nonexhaustive example of this is the use of Pepper [7,19] in the UK in this field during the COVID-19 pandemic [20]. Social robotics can also be useful as:

- (f) Support in the rehabilitation therapy of communication disabilities such as autism or others, where the robot can represent a useful tool full of stimuli for children [18,21–28].

However, the robots can also be used in the home environment while integrated with home automation technologies by supporting the activities listed above in the elderly. Wakamaru [29], for example, can be integrated into domotics with a wide range of support possibilities. Additionally, so-called home-telepresence robots are headed in this direction. They act as home management mediators/facilitators, allowing communication with other people by means of proper devices (cameras, speakers, microphones, etc.) and improving the subject’s safety. Kuri [30] and JIBO [31] are a family of robots that includes telepresence.

3. Research Directions in Social Robots

3.1. A Possible Categorization as a Reference

In an interesting review, Sheridan [32] recently categorized the research direction in the field of SRs as follows: (1) Affect, Personality and Adaptation; (2) Sensing and Control for Action; (3) Assistance to the Elderly and Handicapped; (4) Toys and Markets. We summarize this briefly, referring to the review for an in-depth view.

3.1.1. Affect, Personality and Adaptation

The research in this direction [32–38] concerns using information about the user in order to adapt the SRs to the user’s particular needs and performance intentions, thereby improving acceptance; therefore several studies focus, for example, on how movements of the robot’s body parts imitate human emotions to express different emotions such as anger, disgust, fear, happiness, sadness and surprise.

3.1.2. Sensing and Control for Action

This section considers research that focuses more on the physical interaction between humans and SRs, with consideration to bioengineering solutions [32,39–65]. While safety is essential to human–robot collaboration for industrial manipulation and carefully avoiding collisions, in SRs, the guard is different, and great attention is given to the social tasks, such as applying makeup to the human face. More attention has been given to the problem of motion planning, not only for collision avoidance (obviously, safety remains a basic aspect to consider) but also for human likeness. The touch of a robot, in many cases, for example, induces a positive response in a human, so this aspect must be carefully considered.

3.1.3. Assistance to the Elderly and Handicapped

This is one social robot application that has received much attention [32,66–76]. For example, families coping with a relative with autism often struggle with social and emotional communication. In the case of the elderly, the research directions confirm what has been discussed above in Section 1. In the case of the research on the use of robots for children with autism, some gaps have been identified and reported by Sheridan [32], such as diversity in focus, bias in the research toward specific behavior impairments, the effectiveness of the human–robot interaction after impairment and the use of robot-based motor rehabilitation in autism.

3.1.4. Toys and the Market for Social Robots in General

Here, Sheridan [23] makes the important consideration that for user acceptance, government regulator acceptance and sales appeal, engineering/research related to social psychological and human factors should be applied to social robots. This is especially true for children’s toys because children are the most vulnerable of the various user categories. It should be considered that most of the sales of social robots today are for children’s toys as it is possible to see over the web.

3.2. Further Personal Considerations

I agree with the categorization identified by Sheridan [32], and I believe that it can be used as a reference for evaluating the future developments of social robots, with particular reference to the assistance and rehabilitation sectors. Without introducing new categorizations and focusing on the rehabilitation sector, I believe that two recent, further considerations are worthy of note. The first is the introduction of a sort of robot-based pet therapy through robots with the appearance of animals. The second is the impact of the research and clinical applications on SRs, as partly anticipated in Section 2 due to the COVID-19 pandemic. Both topics are translational with respect to the four categories described above.

3.2.1. Social-Animal-Like Robot for Pet Therapy

The pet therapy is identified as a complementary intervention that strengthens traditional treatments and can be used on patients suffering from various pathologies, with the aim to improve their state of health, thanks to the human–animal interaction. It has been proved that the presence of an animal (e.g., dog, cat, rabbit) improves both the emotional relationship and the work with the patient, favoring the interaction, attention and in general the communication channel and stimulating the active participation of the subject. Pet therapy is often used in dedicated interventions.

Pet therapy is now finding fertile ground in SRs. Two examples of this are the two social-animal-like-robots Paro and Robear. Paro was designed by Takanori Shibata in early 1993 [77]. It was designed on the basis of a puppy seal. Paro features a complex mechatronic, with tactile sensors covering its fur, touch-sensitive whiskers and actuators that quietly move its limbs and body.

Thanks to this design, it responds to cuddles by moving its tail and opening and closing its eyes, memorizes faces, follows the guard and learns actions, generating pos-

itive reactions. Among the principal applications [15,16], it is possible to find the same applications of pet therapy in (a) reducing psychological disorders such as anxiety and depression and (b) improving communication skills and (c) the levels of attention and participation. Therefore, the social robot Paro also acts as a rehabilitation therapist. It has been used in rehabilitation therapies on the elderly (for example, with dementia) and on children with autism. Paro is a social companion for those who interact with him, encouraging effects such as increased participation, increased levels of attention and new social performances, such as cooperative attention and interaction [15,16,78–85]. Robear [86] is a white, bear-shaped robot that lifts and helps patients in wheelchairs to move to bed or go to the bathroom. It is a special robot nurse made by the Riken Brain Science Institute [87] that is conquering hospitals in Japan for its efficiency and “sweetness.” Robear is driven by software and three different types of sensors, including “tactile” structures made of rubber. Weighing approximately 140 kg, Robear is strong and agile enough to (a) gently lift the patient from the bed to the wheelchair, (b) help them stand up and (c) move quickly. While the first example, represented by Paro, is a clear example of a pure robot-based pet therapy, the second, Robear, is an example of the application of both robot-based pet therapy and robot-based caregiving, which could also contribute to avoiding caregiver burning during the complex activities of assistance, especially during the COVID-19 pandemic. It should also be considered that many fragile subjects prefer more to be manipulated by a social robot (Robear in this case) than a human caregiver.

3.2.2. Social Robots and COVID-19

The COVID-19 pandemic has dramatically brought to the fore the problem of the frailty of the elderly. Often the elderly were subjected to forced isolation to avoid contagion. This has resulted in both difficulties in health care (including psychological) and the appearance of disturbing factors such as fear, anxiety and other psychological disorders. Their functional capabilities also generally declined during this period.

To try to minimize the problem, some nursing homes have started using robots to take care of the elderly to try to alleviate their loneliness while supporting them from a mental health point of view. An example of this, as briefly anticipated in Section 2, is the use of Pepper [17] in the UK. SRs, including the previously reported Robear [86], have provided an impetus in research and clinical application during the COVID-19 pandemic. At the end of the pandemic, it will be possible to completely assess this and make a map point.

4. Conclusions

The last evolution of collaborative robots (historically proposed for collaboration with human subjects) [10] is the capability to play the role of an interactive social communicator and, therefore, to be a social robot [13]. This new role is showing high potential in both the direction of rehabilitation and assistance of subjects with disabilities, especially the fragile and handicapped. SRs have particularly demonstrated potential both in the care of the elderly and children with communication disabilities, such as autism [9–22]. Recently, we have also witnessed boosted activity both in the research and clinical applications of SRs caused by the COVID-19 pandemic. In fact, SRs present a chance to allow the continuity of care and communication and psychological support in situations where there are rules/initiatives to maintain social distancing to avoid infection; in other terms, a kind of lifebuoy [17,18]. The research direction in the field of SRs has been clearly detected. In an interesting review, Sheridan [32] recently categorized the research direction in this field of SRs as follows: (1) Affect, Personality and Adaptation; (2) Sensing and Control for Action; (3) Assistance to the Elderly and Handicapped; (4) Toys and Markets. As transversal fields of this research direction, I have detected the clear introduction of robot-based pet therapy [15,16,78–86] and the impact of the COVID-19 pandemic on the research activity [17,18]. The latter opened much discussion around the use of SRs in rehabilitation and assistance, complimenting the economic and ethical sphere. Ethical issues have arisen around the key question that SRs cannot provide true selflessness, compassion and warmth,

which should be at the heart of an assistance system. Scholars of epistemology are worried that SRs, with increased use, could even increase long-term loneliness, reducing the actual contact people have with humans and increasing a sense of disconnection. This, obviously, is not applicable when SRs are used either as facilitators or mediators among humans, as in most cases in domotics or in some applications in the care of autism, such as the robot Kaspar [88–91].

It is precisely this role that makes us reflect on the further opportunities of SRs in telerehabilitation applications that can occur in three important sectors:

- As facilitators/mediators to put fragile and/or needy subjects in contact with the health system and/or family members for more complete support of rehabilitation monitoring.
- As support in a more tailored patient-centered therapy by adapting SRs to the patient's telerehabilitation needs.
- In the domiciliation of care also integrated on the basis of the previous point, with the emerging robotic rehabilitation technologies of the upper and lower limbs integrated into the telerehabilitative pathways and processes.

When we reflect on SRs, and if we are worried about the above-listed problems (increasing loneliness, reducing contacts, etc.), we must also see the flip side of the coin; that is to say that, in this pandemic season, a robot of this type could provide answers to many problems that are encountered in nursing homes and hospitals, such as lack of personnel. In times of lockdown, many elderly and disabled people are left completely alone in their homes and sometimes without adequate health care. Furthermore, even leaving out the COVID-19 pandemic, there was already a problem of assistance (worldwide and in every period) for the elderly, the frail, the disabled, the sick, the lonely and the non self-sufficient. My opinion is that, in general, robotic caregivers should not only be viewed with suspicion but also as a possible opportunity for support. There is no doubt that robotics will be an important part of the health and care of the future. The robots will assist in surgical interventions (in presence or remotely), rehabilitation, in home automation, they will take care of hospital hygiene, dispense lunch and medicines and support of various kinds in general. It is certainly true that robots are not currently able to express the emotions of a human being, however they can do a job in a precise and effective way and could be of great help in dealing with the problems of disability and many problems in health care.

From an economic point of view, it is very interesting for insurance companies under various aspects, ranging from the possibility of developing new insurance formulas that revolve around the use of care-robots, as well as the introduction of new policies that cover the risks of using robots. As for other applications of artificial intelligence, a key point for the diffusion of SRs will clearly be the opinion and the acceptance, the so-called last yard, of all the involved actors, ranging from physicians, nurses and caregivers to patients with their familiars. Therefore, it will be necessary to set up dedicated studies based on dedicated large surveys [92,93] to face the last yard, in which artificial intelligence cannot fail to play a key role [94], given that artificial intelligence will be, for example, fundamental for specifying the level and characteristics of the empathy of social robots in the near future. All this is of basic importance because, according to studies focused on bibliometric indicators, we are witnessing significant growth in this sector. In the study reported in [95], for example, it is documented that the field started growing since the mid-1990s, and after 2006 [95], we can observe a larger amount of publications. The authors [95] obtained academic article data from the robotics and the social robotics fields, highlighting the important increasing number of publications on SRs (a) by number of articles and (b) proportion in relation to all-robotics research. Furthermore, now, official studies show that the social robots market is (https://www.mordorintelligence.com/industry-reports/social_robots_market) [96] estimated to grow at a compound annual growth rate of about 14% over the forecast period 2021 to 2026 thanks to the rise of research in the field of artificial intelligence (AI), natural

language processing (NLP) and the development of platforms such as the robotic operating system, which enabled the rise of social robotics.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

References

- Giansanti, D. The Rehabilitation and the Robotics: Are They Going Together Well? *Health* **2020**, *9*, 26. [CrossRef] [PubMed]
- Heßler, M. Der Erfolg der “Dummheit”. *NTM Z. Gesch. Wiss. Tech. Med.* **2017**, *25*, 1–33. [CrossRef] [PubMed]
- Sverzellati, N.; Brillet, P.-Y. When Deep Blue first defeated Kasparov: Is a machine stronger than a radiologist at predicting prognosis in idiopathic pulmonary fibrosis? *Eur. Respir. J.* **2017**, *49*, 1602144. [CrossRef]
- Kasparov, G. Strategic intensity: A conversation with world chess champion Garry Kasparov. *Harv. Bus. Rev.* **2005**, *83*, 49–53. [PubMed]
- Deep Blue. Available online: <https://www.sciencedirect.com/science/article/pii/S0004370201001291> (accessed on 22 February 2021).
- Deep Blue Defeats Garry Kasparov in Chess Match. Available online: <https://www.history.com/this-day-in-history/deep-blue-defeats-garry-kasparov-in-chess-match> (accessed on 22 February 2021).
- Kim, H. AI, Big Data, and Robots for the Evolution of Biotechnology. *Genom. Inform.* **2019**, *17*, e44. [CrossRef] [PubMed]
- AlphaGo. Available online: <https://deepmind.com/research/case-studies/alphago-the-story-so-far> (accessed on 22 February 2021).
- The Evolution of Computing: AlphaGo. Available online: <https://ieeexplore.ieee.org/document/7499782> (accessed on 22 February 2021).
- Matthews, P.; Greenspan, S. *Automation and Collaborative Robotics: A Guide to the Future of Work*; Apress: New York, NY, USA, 2020.
- 20 Years Later: Cobots Co-Opt Assembly Lines. Available online: <https://www.mccormick.northwestern.edu/news/articles/2016/08/twenty-years-later-cobots-co-opt-assembly-lines.html> (accessed on 22 February 2021).
- International Federation of Robotics. Available online: <https://ifr.org/> (accessed on 22 February 2021).
- Korn, O. *Social Robots: Technological, Societal and Ethical Aspects of Human-Robot Interaction*; Springer: Berlin/Heidelberg, Germany, 2019.
- Ziaaetabar, F.; Pomp, J.; Pfeiffer, S.; El-Sourani, N.; Schubotz, R.I.; Tamosiunaite, M.; Wörgötter, F. Using enriched semantic event chains to model human action prediction based on (minimal) spatial information. *PLoS ONE* **2020**, *15*, e0243829. [CrossRef]
- Hirt, J.; Ballhausen, N.; Hering, A.; Kliegel, M.; Beer, T.; Meyer, G. Social Robot Interventions for People with Dementia: A Systematic Review on Effects and Quality of Reporting. *J. Alzheimer's Dis.* **2021**, *79*, 773–792. [CrossRef]
- Pu, L.; Moyle, W.; Jones, C.; Todorovic, M. The effect of a social robot intervention on sleep and motor activity of people living with dementia and chronic pain: A pilot randomized controlled trial. *Maturitas* **2021**, *144*, 16–22. [CrossRef]
- Chen, K. Use of Gerontechnology to Assist Older Adults to Cope with the COVID-19 Pandemic. *J. Am. Med. Dir. Assoc.* **2020**, *21*, 983–984. [CrossRef] [PubMed]
- Ghiță, A.Ș.; Gavril, A.F.; Nan, M.; Hoteit, B.; Awada, I.A.; Sorici, A.; Mocanu, I.G.; Florea, A.M. The AMIRO Social Robotics Framework: Deployment and Evaluation on the Pepper Robot. *Sensors* **2020**, *20*, 7271. [CrossRef]
- Pepper. Available online: <https://robots.ieee.org/robots/pepper/> (accessed on 22 February 2021).
- Robots to be Introduced in UK Care Homes to Allay Loneliness—That’s Inhuman. Available online: <https://theconversation.com/robots-to-be-introduced-in-uk-care-homes-to-allay-loneliness-thats-inhuman-145879> (accessed on 22 February 2021).
- Naffa, H.; Fain, M. Performance measurement of ESG-themed megatrend investments in global equity markets using pure factor portfolios methodology. *PLoS ONE* **2020**, *15*, e0244225. [CrossRef] [PubMed]
- Lewis, T.T.; Kim, H.; Darcy-Mahoney, A.; Waldron, M.; Lee, W.H.; Park, C.H. Robotic uses in pediatric care: A comprehensive review. *J. Pediatr. Nurs.* **2021**, *58*, 65–75. [CrossRef] [PubMed]
- Soares, E.E.; Bausback, K.; Beard, C.L.; Higinbotham, M.; Bunge, E.L.; Gengoux, G.W. Social Skills Training for Autism Spectrum Disorder: A Meta-analysis of In-person and Technological Interventions. *J. Technol. Behav. Sci.* **2020**, 1–15. [CrossRef]
- Egido-García, V.; Estévez, D.; Corrales-Paredes, A.; Terrón-López, M.-J.; Velasco-Quintana, P.-J. Integration of a Social Robot in a Pedagogical and Logopedic Intervention with Children: A Case Study. *Sensors* **2020**, *20*, 6483. [CrossRef] [PubMed]
- So, W.-C.; Cheng, C.-H.; Law, W.-W.; Wong, T.; Lee, C.; Kwok, F.-Y.; Lee, S.-H.; Lam, K.-Y. Robot dramas may improve joint attention of Chinese-speaking low-functioning children with autism: Stepped wedge trials. *Disabil. Rehabil. Assist. Technol.* **2020**, 1–10. [CrossRef]
- Sandgreen, H.; Frederiksen, L.H.; Bilenberg, N. Digital Interventions for Autism Spectrum Disorder: A Meta-analysis. *J. Autism Dev. Disord.* **2020**, 1–15. [CrossRef]
- Pontikas, C.-M.; Tsoukalas, E.; Serdari, A. A map of assistive technology educative instruments in neurodevelopmental disorders. *Disabil. Rehabil. Assist. Technol.* **2020**, *30*, 1–9. [CrossRef]
- Jain, S.; Thiagarajan, B.; Shi, Z.; Clabaugh, C.; Matarić, M.J. Modeling engagement in long-term, in-home socially assistive robot interventions for children with autism spectrum disorders. *Sci. Robot.* **2020**, *5*, eaaz3791. [CrossRef]

29. Wakamaru. Available online: <https://robots.ieee.org/robots/wakamaru/> (accessed on 22 February 2021).
30. Kuri. Available online: <https://robots.ieee.org/robots/kuri/> (accessed on 22 February 2021).
31. Jibo. Available online: <https://robots.ieee.org/robots/jibo/> (accessed on 22 February 2021).
32. Sheridan, T.B. A review of recent research in social robotics. *Curr. Opin. Psychol.* **2020**, *36*, 7–12. [CrossRef] [PubMed]
33. Cerasa, A.; Ruta, L.; Marino, F.; Biamonti, G.; Pioggia, G. Brief Report: Neuroimaging Endophenotypes of Social Robotic Applications in Autism Spectrum Disorder. *J. Autism Dev. Disord.* **2020**, 1–5. [CrossRef]
34. Martins, G.S.; Santos, L.; Dias, J. User-Adaptive Interaction in Social Robots: A Survey Focusing on Non-physical Interaction. *Int. J. Soc. Robot.* **2018**, *11*, 185–205. [CrossRef]
35. Johnson, D.O.; Cuijpers, R.H. Investigating the Effect of a Humanoid Robot’s Head Position on Imitating Human Emotions. *Int. J. Soc. Robot.* **2018**, *11*, 65–74. [CrossRef]
36. Willemse, C.J.A.M.; van Erp, J.B.F. Social touch in human–robot interaction: Robot-initiated touches can induce positive responses without extensive prior bonding. *Int. J. Soc. Robot.* **2019**, *11*, 285–304. [CrossRef]
37. Block, A.E.; Kuchenbecker, K.J. Softness, Warmth, and Responsiveness Improve Robot Hugs. *Int. J. Soc. Robot.* **2018**, *11*, 49–64. [CrossRef]
38. Palanica, A.; Thommandram, A.; Fossat, Y. Adult Verbal Comprehension Performance is Better from Human Speakers than Social Robots, but only for Easy Questions. *Int. J. Soc. Robot.* **2019**, *11*, 359–369. [CrossRef]
39. Ruijten, P.A.M.; Haans, A.; Ham, J.; Midden, C.J.H. Perceived Human-Likeness of Social Robots: Testing the Rasch Model as a Method for Measuring Anthropomorphism. *Int. J. Soc. Robot.* **2019**, *11*, 477–494. [CrossRef]
40. Lupowski, P.; Rybka, M.; Dziedic, D.; Wlodarczyk, W. The background context condition for the uncanny valley hypothesis. *Int. J. Soc. Robot.* **2019**, *11*, 25–33. [CrossRef]
41. Hoorn, J.F.; Konijn, E.A.; Pontier, M.A. Dating a synthetic character is like dating a man. *Int. J. Soc. Robot.* **2019**, *11*, 235–253. [CrossRef]
42. Bruno, B.; Recchiuto, C.T.; Papadopoulous, I.; Saffiotti, A.; Koulouglioti, C.; Menicatti, R.; Mastrogiovanni, F.; Zaccaria, R.; Sgorbissa, A. Knowledge Representation for Culturally Competent Personal Robots: Requirements, Design Principles, Implementation, and Assessment. *Int. J. Soc. Robot.* **2019**, *11*, 515–538. [CrossRef]
43. Carlson, Z.; Lemmon, L.; Higgins, M.; Frank, D.; Shahrezaie, R.S.; Feil-Seifer, D. Perceived Mistreatment and Emotional Capability Following Aggressive Treatment of Robots and Computers. *Int. J. Soc. Robot.* **2019**, *11*, 727–739. [CrossRef]
44. Stroessner, S.J.; Benitez, J. The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. *Int. J. Soc. Robot.* **2019**, *11*, 305–315. [CrossRef]
45. Wang, B.; Rau, P.-L.P. Influence of Embodiment and Substrate of Social Robots on Users’ Decision-Making and Attitude. *Int. J. Soc. Robot.* **2018**, *11*, 411–421. [CrossRef]
46. Shariati, A.; Shahab, M.; Meghdari, A.; Nobaveh, A.A.; Rafatnejad, R.; Mozafari, B. Virtual Reality Social Robot Platform: A Case Study on Arash Social Robot. In Proceedings of the International Conference on Social Robotics 2018, Qingdao, China, 28–30 November 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 551–560.
47. De Graaf, M.M.A.; Allouch, S.B. Exploring influencing for the acceptance of social robots. *Robot. Auton. Syst.* **2013**, *61*, 1476–1486. [CrossRef]
48. Homma, Y.; Suzuki, K. A Robotic Brush with Surface Tracing Motion Applied to the Face. *Lect. Notes Comput. Sci.* **2018**, *2018*, 513–522. [CrossRef]
49. Turnwald, A.; Wollherr, D. Human-Like Motion Planning Based on Game Theoretic Decision Making. *Int. J. Soc. Robot.* **2019**, *11*, 151–170. [CrossRef]
50. Erlich, S.K.; Cheng, G. A feasibility study for validating robot actions using EEG-based error-related potentials. *Int. J. Soc. Robot.* **2019**, *11*, 271–283. [CrossRef]
51. Heimerdinger, M.; Lavers, A. Modeling the Interactions of Context and Style on Affect in Motion Perception: Stylized Gaits Across Multiple Environmental Contexts. *Int. J. Soc. Robot.* **2019**, *11*, 495–513. [CrossRef]
52. Kaushik, R.; Lavers, A. Imitating Human Movement Using a Measure of Verticality to Animate Low Degree-of-Freedom Non-humanoid Virtual Characters. In Proceedings of the International Conference on Social Robotics 2018, Qingdao, China, 28–30 November 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 588–598.
53. Hamandi, M.; Hatay, E.; Fazli, P. Predicting the Target in Human-Robot Manipulation Tasks. In Proceedings of the International Conference on Social Robotics 2018, Qingdao, China, 28–30 November 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 580–587.
54. Papenmeier, F.; Uhrig, M.; Kirsch, A. Human Understanding of Robot Motion: The Role of Velocity and Orientation. *Int. J. Soc. Robot.* **2018**, *11*, 75–88. [CrossRef]
55. Liu, T.; Wang, J.; Hutchinson, S.; Meng, M.Q.-H. Skeleton-Based Human Action Recognition by Pose Specificity and Weighted Voting. *Int. J. Soc. Robot.* **2018**, *11*, 219–234. [CrossRef]
56. Kostavelis, I.; Vasileiadis, E.; Skartados, E.; Kargakos, A.; Giakoumis, D.; Bouganis, C.S.; Tzovaris, D. Understanding of human behavior with a robotic agent through daily activity analysis. *Int. J. Soc. Robot.* **2019**, *11*, 437–462. [CrossRef]
57. Kaushik, R.; Lavers, A. Imitation of Human Motion by Low Degree-of-Freedom Simulated Robots and Human Preference for Mappings Driven by Spinal, Arm, and Leg Activity. *Int. J. Soc. Robot.* **2019**, *11*, 765–782. [CrossRef]

58. Sprute, D.; Tönnies, K.; König, M. A Study on Different User Interfaces for Teaching Virtual Borders to Mobile Robots. *Int. J. Soc. Robot.* **2018**, *11*, 373–388. [CrossRef]
59. Radmard, S.; Moon, A.; Croft, E.A. Impacts of Visual Occlusion and Its Resolution in Robot-Mediated Social Collaborations. *Int. J. Soc. Robot.* **2018**, *11*, 105–121. [CrossRef]
60. Yoon, H.S.; Jang, J.; Kim, J. Multi-pose face recognition method for social robot. In Proceedings of the International Conference on Social Robotics 2018, Qingdao, China, 28–30 November 2018; pp. 609–619.
61. Karatas, N.; Tamura, S.; Fushiki, M.; Okada, M. The Effects of Driving Agent Gaze Following Behaviors on Human-Autonomous Car Interaction. In Proceedings of the International Conference on Social Robotics 2018, Qingdao, China, 28–30 November 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 541–550.
62. Li, H.; Yihun, Y.; He, H. MagicHand: In-Hand Perception of Object Characteristics for Dexterous Manipulation. In Proceedings of the International Conference on Social Robotics 2018, Qingdao, China, 28–30 November 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 523–532.
63. Yamashita, Y.; Ishihara, H.; Ikeda, T.; Asada, M. Investigation of Causal Relationship Between Touch Sensations of Robots and Personality Impressions by Path Analysis. *Int. J. Soc. Robot.* **2018**, *11*, 141–150. [CrossRef]
64. Spatola, N.; Belletier, C.; Chausse, P.; Augustinova, M.; Normand, A.; Barra, V.; Ferrand, L.; Huguet, P. Improved Cognitive Control in Presence of Anthropomorphized Robots. *Int. J. Soc. Robot.* **2019**, *11*, 463–476. [CrossRef]
65. Komatsubara, T.; Shiomi, M.; Kaczmarek, T.; Kanda, T.; Ishiguro, H. Estimating Children’s Social Status Through Their Interaction Activities in Classrooms with a Social Robot. *Int. J. Soc. Robot.* **2019**, *11*, 35–48. [CrossRef]
66. Ismail, L.L.; Verhoeven, T.; Dambre, J.; Wyffels, F. Leveraging robotics research for children with autism: A review. *Int. J. Soc. Robot.* **2019**, *11*, 389–410. [CrossRef]
67. Jonaiti, M.; Henaff, P. Robot-based motor rehabilitation in autism: A systematic review. In Proceedings of the International Conference on Social Robotics 2018, Qingdao, China, 28–30 November 2018; pp. 1–12.
68. Alhaddad, A.Y.; Cabibihan, J.-J.; Bonarini, A. Head Impact Severity Measures for Small Social Robots Thrown During Meltdown in Autism. *Int. J. Soc. Robot.* **2018**, *11*, 255–270. [CrossRef]
69. Yoshikawa, Y.; Kumazake, H.; Matsumoto, Y.; Miyao, M.; Ishiguro, H.; Shimaya, J. Communication support via a tele-operated robot for easier talking: Case/laboratory study of individuals with/ without autism spectrum disorder. *Int. J. Soc. Robot.* **2019**, *11*, 171–184.
70. Parviainen, J.; Turja, T.; Van Aerscht, L. Robots and Human Touch in Care: Desirable and Non-desirable Robot Assistance. In Proceedings of the International Conference on Social Robotics 2018, Qingdao, China, 28–30 November 2018; Springer International Publishing: Berlin/Heidelberg, Germany, 2018; pp. 533–540.
71. Karunaratne, D.; Morales, Y.; Nomura, T.; Kanda, T.; Ishiguro, H. Will Older Adults Accept a Humanoid Robot as a Walking Partner? *Int. J. Soc. Robot.* **2018**, *11*, 343–358. [CrossRef]
72. Moro, C.; Lin, S.; Nejat, G. Mihailidis: Social robots and seniors: A comparative study on the influence of dynamic social features on human-robot interaction. *Int. J. Soc. Robot.* **2019**, *11*, 5–24. [CrossRef]
73. Portugal, D.; Alvito, P.; Christodoulou, E.; Samaras, G.; Dias, J. A Study on the Deployment of a Service Robot in an Elderly Care Center. *Int. J. Soc. Robot.* **2018**, *11*, 317–341. [CrossRef]
74. Fattal, C.; Leynaert, V.; Laffont, I.; Baillet, A.; Enjalbert, M.; Leroux, C. SAM, an Assistive Robotic Device Dedicated to Helping Persons with Quadriplegia: Usability Study. *Int. J. Soc. Robot.* **2018**, *11*, 89–103. [CrossRef]
75. Zhang, J.; Zhang, H.; Dong, C.; Huang, F.; Liu, Q.; Song, A. Architecture and Design of a Wearable Robotic System for Body Posture Monitoring, Correction, and Rehabilitation Assist. *Int. J. Soc. Robot.* **2019**, *11*, 423–436. [CrossRef]
76. Wang, L.; Du, Z.; Dong, W.; Shen, Y.; Zhao, G. Hierarchical Human Machine Interaction Learning for a Lower Extremity Augmentation Device. *Int. J. Soc. Robot.* **2018**, *11*, 123–139. [CrossRef]
77. PARO Therapeutic Robot. Available online: <http://www.parorobots.com/> (accessed on 22 February 2021).
78. Kelly, P.A.; Cox, L.A.; Petersen, S.F.; Gilder, R.E.; Blann, A.E.; Autrey, A.; MacDonell, K. The effect of PARO robotic seals for hospitalized patients with dementia: A feasibility study. *Geriatr. Nurs.* **2021**, *42*, 37–45. [CrossRef]
79. Tavaszi, I.; Nagy, A.S.; Szabo, G.; Fazekas, G. Neglect syndrome in post-stroke conditions. *Int. J. Rehabil. Res.* **2020**. [CrossRef]
80. Jøranson, N.; Olsen, C.; Calogiuri, G.; Ihlebæk, C.; Pedersen, I. Effects on sleep from group activity with a robotic seal for nursing home residents with dementia: A cluster randomized controlled trial. *Int. Psychogeriatrics* **2020**, 1–12. [CrossRef] [PubMed]
81. Kolstad, M.; Yamaguchi, N.; Babic, A.; Nishihara, Y. Integrating Socially Assistive Robots into Japanese Nursing Care. *Stud. Health Technol. Inform.* **2020**, *272*, 183–186. [CrossRef] [PubMed]
82. Jones, C.; Liu, F.; Murfield, J.; Moyle, W. Effects of non-facilitated meaningful activities for people with dementia in long-term care facilities: A systematic review. *Geriatr. Nurs.* **2020**, *41*, 863–871. [CrossRef] [PubMed]
83. Kolstad, M.; Yamaguchi, N.; Babic, A.; Nishihara, Y. Integrating Socially Assistive Robots into Japanese Nursing Care. *Stud. Health Technol. Inform.* **2020**, *270*, 1323–1324. [CrossRef]
84. Geva, N.; Uzefovsky, F.; Levy-Tzedek, S. Touching the social robot PARO reduces pain perception and salivary oxytocin levels. *Sci. Rep.* **2020**, *10*, 1–15. [CrossRef]
85. Pu, L.; Todorovic, M.; Moyle, W.; Jones, C. Using Salivary Cortisol as an Objective Measure of Physiological Stress in People With Dementia and Chronic Pain: A Pilot Feasibility Study. *Biol. Res. Nurs.* **2020**, *22*, 520–526. [CrossRef] [PubMed]

86. Khan, Z.H.; Siddique, A.; Lee, C.W. Robotics Utilization for Healthcare Digitization in Global COVID-19 Management. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3819. [CrossRef] [PubMed]
87. The Strong Robot with the Gentle Touch. Available online: https://www.riken.jp/en/news_pubs/research_news/pr/2015/20150223_2/ (accessed on 22 February 2021).
88. Huijnen, C.A.G.J.; Lexis, M.A.S.; Jansens, R.; De Witte, L.P. Roles, Strengths and Challenges of Using Robots in Interventions for Children with Autism Spectrum Disorder (ASD). *J. Autism Dev. Disord.* **2018**, *49*, 11–21. [CrossRef] [PubMed]
89. Huijnen, C.A.G.J.; Lexis, M.A.S.; Jansens, R.; De Witte, L.P. How to Implement Robots in Interventions for Children with Autism? A Co-creation Study Involving People with Autism, Parents and Professionals. *J. Autism Dev. Disord.* **2017**, *47*, 3079–3096. [CrossRef]
90. Mengoni, S.E.; Irvine, K.; Thakur, D.; Barton, G.; Dautenhahn, K.; Guldberg, K.; Robins, B.; Wellsted, D.; Sharma, S. Feasibility study of a randomised controlled trial to investigate the effectiveness of using a humanoid robot to improve the social skills of children with autism spectrum disorder (Kaspar RCT): A study protocol. *BMJ Open* **2017**, *7*, e017376. [CrossRef] [PubMed]
91. Wood, L.J.; Dautenhahn, K.; Rainer, A.; Robins, B.; Lehmann, H.; Syrdal, D.S. Robot-Mediated Interviews—How Effective Is a Humanoid Robot as a Tool for Interviewing Young Children? *PLoS ONE* **2013**, *8*, e59448. [CrossRef]
92. Giansanti, D. Towards the evolution of the mHealth in mental health with youth: The cyber-space used in psychological rehabilitation is becoming wearable into a pocket. *mHealth* **2020**, *6*, 18. [CrossRef]
93. Giansanti, D.; Monoscalco, L. The cyber-risk in cardiology: Towards an investigation on the self perception among the cardiologists. *mHealth* **2020**. [CrossRef]
94. Giansanti, D.; Monoscalco, L. A smartphone-based survey in mHealth to investigate the introduction of the artificial intelligence into cardiology. *mHealth* **2021**, *7*, 8. [CrossRef]
95. Mejia, C.; Kajikawa, Y. Bibliometric Analysis of Social Robotics Research: Identifying Research Trends and Knowledgebase. *Appl. Sci.* **2017**, *7*, 1316. [CrossRef]
96. Social Robots Market—Growth, Trends, COVID-19 Impact, and Forecasts (2021–2026). Available online: <https://www.mordorintelligence.com/industry-reports/social-robots-market> (accessed on 22 February 2021).



Review

Trust, but Verify: Informed Consent, AI Technologies, and Public Health Emergencies

Brian Pickering

IT Innovation, Electronics and Computing, University of Southampton, University Road, Southampton SO17 1BJ, UK; j.b.pickering@soton.ac.uk

Abstract: To use technology or engage with research or medical treatment typically requires user consent: agreeing to terms of use with technology or services, or providing informed consent for research participation, for clinical trials and medical intervention, or as one legal basis for processing personal data. Introducing AI technologies, where explainability and trustworthiness are focus items for both government guidelines and responsible technologists, imposes additional challenges. Understanding enough of the technology to be able to make an informed decision, or consent, is essential but involves an acceptance of uncertain outcomes. Further, the contribution of AI-enabled technologies not least during the COVID-19 pandemic raises ethical concerns about the governance associated with their development and deployment. Using three typical scenarios—contact tracing, big data analytics and research during public emergencies—this paper explores a trust-based alternative to consent. Unlike existing consent-based mechanisms, this approach sees consent as a typical behavioural response to perceived contextual characteristics. Decisions to engage derive from the assumption that all relevant stakeholders including research participants will negotiate on an ongoing basis. Accepting dynamic negotiation between the main stakeholders as proposed here introduces a specifically socio-psychological perspective into the debate about human responses to artificial intelligence. This trust-based consent process leads to a set of recommendations for the ethical use of advanced technologies as well as for the ethical review of applied research projects.

Citation: Pickering, B. Trust, but Verify: Informed Consent, AI Technologies, and Public Health Emergencies. *Future Internet* **2021**, *13*, 132. <https://doi.org/10.3390/fi13050132>

Keywords: informed consent; terms of use; AI-technologies; technology acceptance; trust; public health emergency; COVID-19; big data; contact tracing; research ethics

Academic Editors: Stefano Modafferi and Francesco Lelli

Received: 20 April 2021
Accepted: 12 May 2021
Published: 18 May 2021

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Although confusion over informed consent is not specific to a public health emergency, the COVID-19 pandemic has brought into focus issues with consent across multiple areas often affecting different stakeholders. Consent, or Terms of Use for technology artefacts including online services, is intended to record the voluntary willingness to engage. Further, it is assumed to be informed: that individuals understand what is being asked of them or that they have read and understood the Terms of Use. It is often unclear, however, what this entails. For the user, how voluntary is such consent, and for providers, how much of their technology can they represent to their users? As an example from health and social care, contact tracing—a method to track transmission and help combat COVID-19—illustrates some of the confusion. Regardless of the socio-political implications of non-use, signing up for the app would imply a contract between the user and the service provider based on appropriate use of the app and limiting the liability of the provider. However, since it would typically involve the processing of personal data, there may also be a request for the user (now a data subject) to agree to that processing. In the latter case, consent is one possible legal basis under data protection law for the collection and exploitation of personal data. In addition, though, the service provider may collaborate with researchers and wish to share app usage and user data with them. This too is referred to as (research) consent, that is the willingness to take part in research. Finally, in response to an indication

that the app user has been close to someone carrying the virus, they may be invited for a test; they would need to provide (clinical) consent for the clinical intervention, namely undergoing the test. It is unclear whether individuals are aware of these different, though common, meanings of consent, or of the implications of each. Added to that, there may be a societal imperative for processing data about individual citizens, which implies that there is a balance to be struck between individual and community rights.

Such examples emerge in other domains as well. Big Tech companies, for instance, may request user consent to process their personal data under data protection law. They may intend to share that data with third parties, however, to target advertising which involves some degree of profiling, which is only permitted under European data protection regulation in specific circumstances. Although legally responsible for the appropriate treatment of their users' data, the service provider may not understand enough of the technology to meet their obligations. Irrespective of technology, the user too may struggle to identify which purpose or purposes they are providing consent for. With social media platforms, the platform provider must similarly request data protection consent to store and process their users' personal data. They may also offer researchers access to the content generated on their platform or to digital behaviour traces for research purposes. This would come under research consent rather than specifically data protection consent. In these two cases, first the user must identify different purposes under the same consent that their (service) data may be used for, but secondly they may need to review different types of consent regarding their data as used for providing the service versus content they generate or activities they engage in used for research.

In this paper, I will explore the confusions around consent in terms of common social scientific models. This provides a specifically behavioural conception of the dialogue associated with consent contextualised within an ecologically valid presentation of the underlying mechanisms. As such, it complements and extends the discussion on explainable artificial intelligence (AI). Instead of focusing on specific AI technology, though, this discussion centres on the interaction of users with technologies from a perspective of engagement and trust rather than specifically focusing on explainability.

Overview of the Discussion

The following discussion is organised as follows. Section 2 provides an overview of responsible and understandable AI as perceived by specific users, and in related government attempts to guide the development of advanced technologies. In Section 3, I introduce behavioural models describing general user decision forming and action. Section 4 covers informed consent, including how it applies in research ethics in Section 4.1 (For the purpose of this article, *ethics* is used as an individual, subjective notion of right and wrong; *moral*, by contrast, would refer to more widely held beliefs of what is acceptable versus what is not [1]). Specific issues with consent in other areas are described in Section 4.2, including technology acceptance (Section 4.3). Section 4 finishes with an introduction to trust in Section 4.4 which I develop into an alternative to existing *Informed Consent* mechanisms.

Having presented different contexts for consent, Section 5 considers a trust-based approach applied to three different scenarios: *Contact Tracing*, *Big Data Analytics* and *Public Health Emergencies*. These scenarios are explored to demonstrate trust as an explanatory mechanism to account for the decision to engage with a service, share personal data or participate in research. As such, unlike existing consent-based mechanisms, a trust-based approach introduces an ecologically sound alternative to *Informed Consent* free from any associated confusion, and one derived from an ongoing negotiated agreement between parties.

2. Responsible and Explainable AI

Technological advances have seen AI components introduced across multiple domains such as transportation, healthcare, finance, the military and legal assessment [2]. At the same time, user rights to interrogate and control their personal data as processed

by these technologies (Art 18, 21 and 22 [3]) call for a move away from a black-box implementation [2,4] to greater transparency and explainability (i.a., [5]). Explainable AI as a concept has been around at least since the beginning of the millennium [2] and has been formalised in programs such as DARPA in the USA [6]. In this section, I will consider some of the research and regulatory aspects as they relate to the current discussion.

The DARPA program defines explainable AI in terms of:

"... systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future" [6]

Anthropomorphising technology in this way has implications for technology acceptance (see [7]; and Section 4.3 below). Surveys by Adadi and Berrada [2] and Arrieta and his colleagues [5] focus primarily on mapping out the domain from recent research. Although both conclude there is a lack of consistency, common factors include explainability, transparency, fairness and accountability. Whilst they recognise those affected by the outcomes, those using technology for decision support, regulators and managers all as important stakeholders, Arrieta et al. focus ultimately on developers and technologists with a call to produce "Responsible AI" [5]. Much of this was found previously in our own 2018 Delphi consultation with domain experts. In confirming accountability in technology development and use, however, experts also called for a new type of ethics and encouraged ethical debate [8]. Adadi and Berrada meanwhile emphasise instead the different motivations and types of explainability: for control, to justify outcomes, to enable improvement, and finally to provide insights into human behaviours (*control to discover*) [2]. Došilović and his colleagues highlight a need for formalised measurement of *subjective* responses to explainability and interpretability [9], whereas Samek et al. propose a formalised, *objective* method to evaluate at least some aspects of algorithm performance [4]. Meanwhile, Khrais sought to investigate the research understanding of explainability, discovering not only terms like *explanation, model* and *use* which might be expected, but also more human-centric concepts like *emotion, interpret* and *control* [10].

Looking not at the interpretability of AI technologies, other studies seek to explore the implications of explainability on stakeholders, and especially on those dependent on its output (for instance, patients and clinicians using an AI-enabled medical decision-support system). The DARPA program seeks to support "*explanation-informed acceptance*" via an understanding of the socio-cognitive context of explanation [6]. Picking up on such a human-mediated approach, Weitz and her colleagues demonstrate how even simple methods, in their case the use of an avatar-like component, encourage and enhance perceptions of understanding the technology [7]. Taking this further and echoing [9] on trust, Israelsen and Ahmed, meanwhile, focus on trust-enhancing "*algorithmic assurances*" which echo traditional constructs like trustworthiness indicators in the trust literature (see Section 4.4) [11]. All of this comes together as positioning AI explainability as a co-construction of understanding between explainer (the advanced AI-enabled technology) and explainee (the user) [12]. This ongoing negotiation around *explainability* echoes my own trust-based alternative to the dialogue around informed consent below (Section 4.4).

Much of the research above makes explicit a link between the motivation towards explainable or responsible AI with regulation and data subject rights [2,4,5,9,11,12]. With specific regard to big data, the Toronto Declaration puts the onus on data scientists and to some degree governance structures to protect individual rights [13]. However, human rights conventions often balance individual rights with what is right for the community. For example, although the first paragraph of Art. 8 on privacy upholds individual rights and expectations, the second provides for exceptions where required by the community [14]. Individual versus community rights are significant for contact tracing and similar initiatives associated with the pandemic. While calling upon technologists for transparency and fairness in their use of data, the UK Government Digital Services guidance also tries to balance community needs with individual human rights [15]. The UK Department of Health and Social Care introduces the idea that both clinicians and patients, that is multiple stakeholders, need to be involved and to understand the technology [16]. Similarly, the

EU stresses that developers, organisations deploying a given technology, and end-users should all share some responsibility in specifying and managing AI-enabled technologies, without considering how such technologies might disrupt existing relationships [17].

The focus on transparency and explainability within the (explainable) AI literature is relevant to the idea that consent should be informed. Although the focus is often on technologists [5,8], this assumes that all stakeholders—those affected by the output of the AI component, those using it for decision support, and those developing it (cf. [5])—share responsibility for the consent process. Even where studies have focused on stakeholder interactions and the co-construction of explainability [12], there is an evident need to consider the practicalities of the negotiation between parties to that process. For contact tracing, for example, who is responsible to the app user for the use and perhaps sharing of their data? Initially, the service provider would assume this role and make available appropriate terms of use, a privacy notice and privacy policy. However, surely the data scientist providing algorithms or models for such a system at least needs to explain the technology? A Service Level Agreement (SLA) for a machine-learning component would not typically involve detail about how a model was generated or its longer term performance. If it is not clear what stakeholders are responsible for, it becomes problematic to identify who should be informing the app user or participant of what to expect. Further, with advanced, AI-enabled technologies, not all stakeholders may be able to explain the technology. A clinician, for instance, is focused on care for their patients; they would not necessarily know how a machine-learning model had been generated or what the implications would be. There would have to be a paradigm shift perhaps before they consider trying to understand AI-technologies.

Leading on from studies which situate explainable AI within a behavioural context ([6,11,12]), I take the more general discussion about the use and effects of advanced technologies into the context of planned behaviour (in Section 3) and extend discussions of trust [9,11] into the practical consideration of informed consent in a number of different domains. Starting with contact tracing and similar applications of AI technologies (Section 5), this discussion seeks to explore the confusion around consent in a practical context, evaluate the feasibility of transparency, and review responsible stakeholders for different scenarios. Consent to engage with advanced technologies highlights, therefore, the impact of AI rather than specifically on how explainable the technology might be. Focusing on a new kind of ethics, this leads to the proposal for a trust-based alternative to consent.

3. Behaviour and Causal Models

The Theory of Planned Behavior (TPB) assumes that a *decision to act* precedes the *action* or the actual activity itself. The separation between a decision to act and the action itself is important: we may decide to do something, but not actually do it. The *decision to act* is the result of a response to a given *context*. This is summarised in Figure 1. The *context* construct may include characteristics of the individual, of the situation itself, of the activity they are evaluating or of any other background factors. For instance, Figure 2 provides interpretations of *Terms of Use* ((a) the upper half of the figure) and *Research Consent* ((b) in the lower half) as behavioural responses.

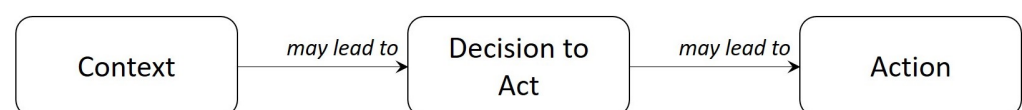


Figure 1. Schematic representation of the Theory of Planned Behavior [18].

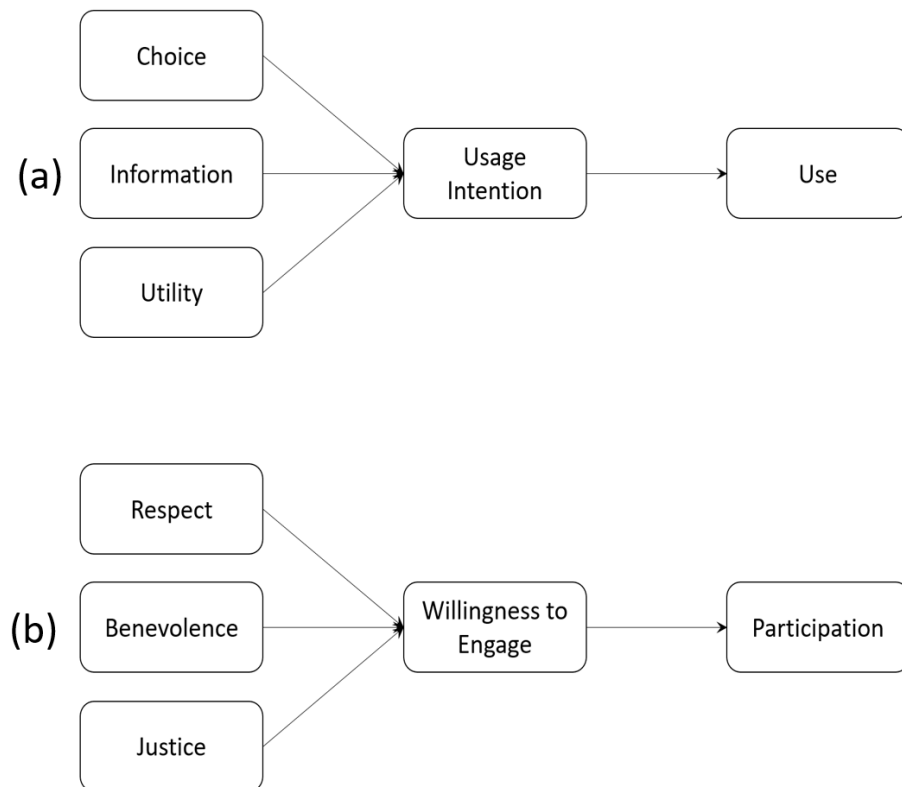


Figure 2. TPB Interpretation of Behaviours associated with (a) *Terms of Use* and (b) *Research Consent*.

Someone wishing to sign up to an online service, for instance, would be presented with the *Choice* to use the service or not, which may depend on the *Information* they are provided about the service provider and the perceived *Utility* they might derive from using the service. The *context* for *Terms of Use* therefore comprises *Choice*, *Information* and *Utility*. By contrast, a potential research participant would decide whether or not to take part (develop a *Willingness to Engage*) based on *Respect* shown to them by the researcher, whether the researcher is well disposed towards them (*Benevolence*), and that research outcomes will be shared equitably across the community (*Justice*).

4. Informed Consent

Although the concept can be traced back historically [19], the definition of informed consent was formalised more recently after World War II in the Nuremberg Code [20] and the Helsinki Declaration [21]. The emphasis is on:

“Voluntary agreement to or acquiescence in what another proposes or desires; compliance, concurrence, permission.” (<https://www.oed.com/oed2/00047775>) (accessed on 12 May 2021).

For technology, terms of use focuses on an agreement between user and provider, defining usage and limiting liability:

“[The] circumstances under which buyers of software or visitors to a public Web site can make use of that software or site” [22] as part of a *“binding contractual agreement”* [23].

Indeed, Luger and her colleagues [24] and subsequently Richards and Hartzog make explicit the link between terms of use and:

“Consent [which] permeates both our law and our lives—particularly in the digital context” (The term *consent* as used here will therefore include *terms of use*) [23].

In a medical or clinical context, and to counter the paternalism of earlier medical practice, the definition makes explicit who the main parties to the agreement are:

“[the] process of informed consent occurs when communication between a patient and physician results in the patient’s authorization or agreement to undergo a specific medical intervention”. (<https://www.ama-assn.org/delivering-care/ethics/informed-consent>) (accessed on 12 May 2021).

There are other concerns, though. For clinical treatment, the patient is reliant on the clinician to improve their well-being, and there is no guarantee that they will understand the implications of the treatment proposed. Further, concerned about their health or general prognosis, they may not be emotionally fit to think objectively about the information that they have been given. Additionally, during a public health emergency, there may be a legal or moral obligation to disclose data, such as infection status. This implies a balance to be struck between individual rights and the common good.

In research terms, consent is seen as integral to respect for the autonomy of the research participant:

“Respect for persons requires that subjects, to the degree that they are capable, be given the opportunity to choose what shall or shall not happen to them. This opportunity is provided when adequate standards for informed consent are satisfied.” (Part C (1), [25]).

Latterly and with the emphasis on the right to privacy (Art. 8 [14]), consent is defined in data protection legislation as:

“any freely given, specific, informed and unambiguous indication of the data subject’s wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her” (Art. 4(11), [3]).

Irrespective of context, all definitions assume the pre-requisite characteristics of voluntariness (freely given) and informed (understanding what another proposes or desires). It is not always clear, however, if these basic requirements are met or even possible. In a clinical context, apart from the assumption that patients are emotionally objective rather than directly reliant on the expertise of the clinician, the stakeholder relationship is unequal. Notwithstanding the right to religious tolerance [19] assuming it be based on true autonomy [26], clinical judgement is subordinate to uninformed or emotionally charged patient preference [27]. Putting clinician and patient on a more equal footing may be preferable and preserve the interests of all parties [28].

The legal context, that is the requirements governing data protection and clinical practice, limits what can be done in practical terms in regard to resolving confounding issues associated with informed consent. Any such legislation will tend to be jurisdiction-specific, and like the COPI Regulations [29] requiring the sharing of medical data, may be time- and domain-specific. It is also worth remembering in the context of data protection, that there are different requirements depending on the nature of the data themselves (Art. 6, 9 and 10 [3]). These imply different legal bases, or justifications, to allow personal data to be processed. Consent (Art. 6(1)(a)–(f), Art. 9(2)(a)–(j), [3]) is only one such legal basis, but has implications such as the data subject’s right to object to processing or to withdraw the data (Chapter 3 [3]). In a research study involving the processing of personal data, consent to participate may be different from the lawful basis covering the collection of the personal data (for instance, Art. 89, i.a., [3]). Consequently, conflating data protection and research ethics consent may confuse the data subject/participant as well as restrict what a researcher can do with the data they collect, or worse still, undermine the researcher/participant relationship. Further, an institutional Research Ethics Committee (REC; also known as Institutional Review Board, IRB) or indeed a data protection review would need to consider whether consent refers to research participation only (clinical trial or other academic research studies), or a legal basis for data protection purposes, or both. By contrast, consent for technology use (*terms of use*) may be weighted against the user and in favour of the supplier. This may result from a failure to read or be able to read the conditions [24], to differentiate technology-use contexts [30], or even from a *de facto*

assumption that app usage implies consent [22]. Notwithstanding these specific issues of the type of consent, the following sections consider first the governance of research ethics, before moving on to the various challenges regarding voluntary and informed consent from the research participant's perspective, and subsequently to implications for technology acceptance and adoption. The final subsection will give a definition of trust and describe how it applies to various different contexts where consent would normally be expected.

4.1. Applied Research Ethics

Research ethics relates to more general applied ethical principles, which helps contextualise issues pertaining to informed consent. Research ethics review is often based on recommendations from the Belmont Report [25], which echo similar values in medical ethics [31], and assumes that activity is research [32] rather than service evaluation, for instance. The primary focus includes:

- *Participant Autonomy or Respect for the Individual*: a guarantee to protect the dignity of the participant as well as respecting their willingness or otherwise to take part. This assumes that the researcher (including those involved with clinical research) have some expectation of outcomes and can articulate them to the potential participant. Deception is possible under appropriate circumstances [33], including the use of placebos. Big data requires careful thought, since it shifts the emphasis away from individual perspectives [34]. In so doing, a more societally focused view may be more appropriate [35]. Originally, autonomy related directly to the informed consent process. However, the potential for big data and AI-enabled approaches suggests that this may need rethinking.
- *Beneficence and non-malevolence*: ensuring that the research will treat the participant well and avoid harm. This principle, most obvious for medical ethics, puts the onus on the researcher (or data scientist) to understand and control outcomes (see also [15,17]). Although there is no suggestion that the researcher would deliberately wish to cause harm, unsupervised learning may impact expected outcomes. Calls for transparency and the human-in-the-loop to intervene if necessary [8,17] imply a recognition that predictions may not be fixed in advance. Once again, the informed nature of consent might be difficult to satisfy.
- *Justice*: to ensure the fair distribution of benefits. The final principle highlights a number of issues. The CARE Principles [36], originally conceived in regard to indigenous populations, stress the importance of ensuring that all stakeholders within research at least are treated equitably. For all its limitations, the trolley car dilemma [37] calls into question the assessment of justice. During a public health emergency, and inherent in contact tracing, the issue is whether justice is better served by protecting the rights of the individual especially privacy over a societal imperative.

In general ethics terms, autonomy, beneficence/non-malevolence, and justice reflect a Kantian or deontological stance: they represent the rules and obligations which govern good research practice (for a useful summary on applied ethical theories, see [38]). Utilitarianism—justifying the means on the basis of potential benefits—is subordinate to such obligations. However, Rawls' justice theory and different outcomes to the trolley-car dilemma motivate a re-evaluation of a simple deontological versus utilitarian perspective. Further, the socially isolating effect of the COVID-19 pandemic raises the question as to whether a move away from focusing on the individual and considering instead the individual as defined by the collective community is more appropriate (see [39]). Indeed, the challenge comes when applying ethical principles in specific research environments [40,41], or medical situations [27]. Ethics review must therefore balance the potentially competing interests and expectations of research participants, researchers and the potential benefit to the community at large in determining what is ethically acceptable. At the same time, it is essential to consider whether the researcher or any other stakeholder can truly assess what the potential outcomes might be. In either case—individual versus community benefit, and

the overall knowledge and transparency of the research involved—there needs to be an ongoing negotiation amongst stakeholders to agree on an acceptable approach.

4.2. *Issues with Consent*

Leading on from the earlier discussion about confusions arising from consent, in this section I return to some specific issues with consent identified in the literature. As part of the consent process, a potential participant is provided detailed information covering what the research is about, what they as participant will be expected to do, and any potential risks or benefits to them. They are typically given an opportunity to discuss with the researcher or indeed anyone else for clarification. Ultimately, it is assumed that this will address the need to provide all the detail the research participant needs to make an informed decision about participation. Although there is some evidence to suggest both researchers and their participants are satisfied with the informed consent process [42], this begs the question as to whether participants are able to make balanced decision based on that information—namely, their competence—but also whether they do in fact feel that they are free to make whatever decision they choose.

Researchers must consider the ability of their potential participants to assimilate and understand the information they can provide when requesting participation, therefore. This may be due to the capacity of the participant themselves, but also their understanding of the implications of the research on them and on others like them [43]. There are also indications that the amount of information provided [44], and of issues such as potential risks and benefits may not be satisfactory in some contexts [45,46]. At the same time, researchers tend to be concerned about regulatory compliance as opposed to balancing what the research goals are and how to manage risks and challenges [47]. Ultimately, though, participants may simply fail to understand the implications of what they are being asked to agree to [48] or be unable to engage with the information provided [49]. They do however appreciate that researchers must balance multiple aspects of the proposed research and so would be prepared to negotiate on those terms with the researchers rather than be part of a regulatory compliance exercise [47]. Finally, whether researchers themselves are fully aware of the implications of what they are asking for patient or participant willingness to engage is not always clear [50].

Just as Biros et al. [43] highlight concerns about the broader, community implications of research, there are other social dimensions which should be considered. Nijhawan et al. [51] and Kumar [52], for instance, maintain that consent is really a Western construct. In their studies in India, they also stress that an independently made decision to participate could be influenced by institutional regard: patients may be influenced by an implicit trust in healthcare services, for example [51]. The cultural aspects here echo traditional differences between individualist and collectivist societies [53,54]. Only the former would be used to putting their own wishes and needs above those of the community at large. Indeed, for some cultures, the concept of self exists only as it is part of and dependent on a collective group [39]. This may not be as simple as individualism versus collectivism though: European data protection legislation, for example, puts the rights of the individual above those of the collective, whereas the opposite is true in the USA [40].

Nijhawan and his colleagues [51] highlight a more general concern about the informed consent process: if there is implicit trust in an institution influencing the consent decision, then there are emotional issues which need consideration (see also [55] on privacy; and [56] on trust). Ethics review in the behavioural sciences was introduced to provide additional oversight for what may be seen as unnecessarily stressful for participants [57,58]. However, this misses the point that potential participants may comply, and give their consent, because of a perceived power dynamic [59,60]: the participant may feel obliged to do as they are told to please the researcher. Alternatively, they may simply trust that the researcher is competent and means well which in turn encourages them to trust the researcher [47,48]. Indeed, there is anecdotal evidence that participants would prefer just to get on with the

research [48]; and in areas like clinical treatment, consent is almost irrelevant [61]. Whether or not the consent process really reflects true autonomy is not always clear [26].

Like others, Nijhawan et al. [51] make a distinction between consent (the formal agreement to engage based on participant competence) and assent (a less formal indication of such agreement). So, while a parent or guardian must provide legally based consent for their child to take part, the child themselves should provide assent. The latter is not legally required, but without it there is no monitoring of continuing motivation and willingness to carry on with participation. Assent is an informal agreement then which should be in place to preserve the quality of the research itself though not required by law. Irrespective of whether consent or assent, there is an argument which says it should be re-negotiated throughout a given research study [48,62]. Otherwise, and learning from considerations around biobanks and the ongoing exploitation of samples, informed consent may become too broad to be effective [63] and need continued review [64].

Finally, perhaps most fundamentally, there is the question of whether consent based on full disclosure of information is practical (see Section 2 above) or even desirable. There are contexts within which deliberate deception can be justified [33], and where traditional informed consent is actually undermining research progress [65]. Similarly, as well as competence and the emotional implications of illness or duress discussed above, there are always cases where full disclosure may not be possible. One such example, which will be explored below, concerns public health emergencies and vaccination programs, as well as inadvertent third-party disclosures [66]. Even within a clinical context, O'Neill suggests informed consent be replaced with an informed request: extending that to research, the participant would effectively respect the researcher's competence and agree on that basis to proceed.

Given the nature and extent of issues raised in the literature, it is important to reconsider the purpose of the relationship between research participant and researcher in light of generalised deontological obligations such as autonomy (respect for the individual), benevolence and non-malevolence, and justice. In legal terms alone, problems with informed consent have been summarised thus [23]: consent may be *unwitting* (or unintentional) as in cases where acceptance is assumed by default [22], non-voluntary or *coerced*, and *incapacitated* in that a full understanding is not possible [23]. With that in mind, empirical research is ultimately a negotiation between the researcher and participant, or indirectly between researcher and the data they can access. So, it is unclear whether the informed consent process is adequate to capture what is needed for regulatory compliance or even the reality of empirical research. A research participant is effectively prepared to expose themselves to the vulnerability that the research protocol may not provide the expected outcomes, but believe that the researcher respects them and their input; researchers will do what they can to support participants throughout the research lifecycle and consider the implications of eventual publication. The willingness to be vulnerable in this way has been discussed repeatedly in the trust literature within the social sciences for many years. After a review of implications for technology acceptance in the next section, discussion subsequently turns to the potential benefit of a trust-based approach to research participation.

4.3. Technology Acceptance

How users decide to engage with technology, especially for services like contact tracing, needs to consider issues of consent. This is not always the case (see [23]). Traditionally, causal models predicting technology adoption have focused particularly on features of the technology itself, such as perceived ease of use and perceived usefulness [67]. The *Context* (see Section 3) derives solely from the technology. Other variables associated with technology uptake such as the demographics of potential adopters, any facilitating context and social influence have been identified as moderators, however [68]. McKnight, Thatcher and their colleagues combine these technology-based approaches with models of trust, however [69–71]. In so doing, they extend the influence of social norms identified by

Venkatesh [68] to include a trust relationship with relevant human agents in the overall context within which a particular technology is used.

For the COVID-19 Public Health Emergency (PHE), contact tracing was seen as an effective tool in managing the pandemic. Early on in the pandemic, it was identified as one method among many which would be of use, given the balance between individual rights and societal benefit [72]. Concerns around privacy simply need careful management [73]. Later empirical studies have demonstrated a willingness to engage with the technology: perceived usefulness in managing personal risk outweighs both effort and privacy concerns [74,75]. Perceived ease of use as predicted by standard causal models for technology acceptance [67], therefore, was not seen as an issue.

By contrast the overall context for the introduction of contact tracing needs to be considered. The perceived failure in France, for instance, was due in part to a lack of trust in the government and how it was encouraging uptake [76]. Indeed, Rowe and his colleagues attribute this failure to a lack of cross-disciplinary planning and knowledge sharing, with attempts to force public acceptance of contact tracing perceived as coercive and therefore likely to lead to public distrust [76]. Introducing trust here is entirely consistent with McKnight and Thatcher's work (see above, refs. [69–71]). Without trust in the main stakeholders, as Rowe et al. found, adoption will be compromised.

Examining trust in the context of contact tracing usage brings the discussion back to consent. Given the shortcomings of consent across multiple contexts including terms of use, Richards and Hartzog [23] argue for an approach which does not ignore basic ethical principles such as autonomy but rather empowers stakeholders to engage appropriately. They conclude that existing consent approaches should be replaced with a legally-based trust process. This would allow, they claim, obligations to protect and be discrete with participants and their data and to avoid manipulative practices. The present discussion takes this idea one stage further. In the next section, I review a common social psychological definition of trust as a continuous socially constructed agreement between parties. After that, this is applied to specific technology scenarios relevant to contact tracing, AI-enabled technologies and PHE.

4.4. Trust

Many of the issues outlined above imply that there is a negotiation between different actors in research, or other activities like clinical treatment. While O'Neill discourages replacing the informed consent process with a "ritual of trust" [66], and notwithstanding the importance of general trust in healthcare in making consent decisions [51], it is unclear how consent and trust relate to one another. Roache positions consent as an important part of encouraging good practice in order to introduce a debate on trust and consent [77]. Eyal, however, rejects an unqualified assumption that informed consent promotes trust in medical care in general [78]. He questions [79] the utilitarian approach proposed by Tännsjö [80] and the social good arguments of Bok [81]. However, and regardless of the relative strength of the arguments in this exchange, both Bok and Tännsjö contextualise informed consent within a social negotiation between actors: in their case, patient and clinician. In a research study, researcher and participant may similarly not be on equal footing in terms of competence and understanding. Yet they continue to engage [47,48].

Defining trust can be problematic [82,83]. However, and with specific reference to the inferred vulnerability of the research participant, for the present discussion, one helpful definition was offered by Mayer and his colleagues:

"... the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." Ref. [84]

Trust is therefore the response by a trustor to perceived trustworthiness indicators of benevolence, competence (or ability) and integrity in the trustee.

This is summarised in Figure 3. Like the TPB-based visualisations for *Terms of Use* and *Research Consent* in Figure 2 in Section 3 above, the assumption is that a *Willingness to*

Trust in Mayer et al.'s conception [84] is a response to trustworthiness indicators as *context*. The separation between *Willingness to Trust* and *Trust* itself is important. Once the trustor actually trusts the trustee, they may still continue to reassess the trustworthiness indicators. In so doing, their willingness may be undermined. They lose trust, and the trustee must now act in order to rebuild the lost trust. Trust becomes a constant negotiation, therefore.

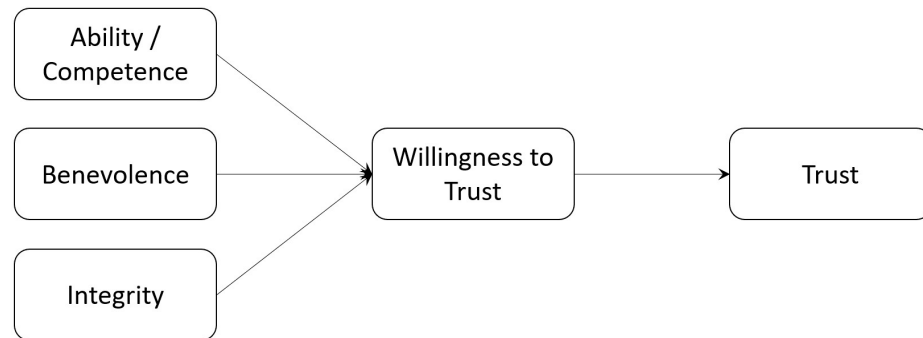


Figure 3. A Schematic Representation of Trust Behaviours.

To expand on this, trust is socially constructed [85] as an ongoing dialogue between exposure to risk and an evaluation of the behaviours of others [86]. Undermining any one of the individual trustworthiness indicators leads to a loss of trust even distrust and a need to repair the original trust response [87] if the relationship is to continue. Trust repair depends on a willingness to identify issues, take responsibility to address them and contextualise behaviours within a narrative that makes sense to the trustor [88,89]. As such, I maintain, this behavioural perspective on trust reflects well Bok’s call for one party to accept the potential limitations of another whilst continuing to evaluate and re-evaluate their behaviour [81].

In a research context, basing the agreement to participate on the assumed trust between researcher and participant can account for the observed pragmatic approach to informed consent [48]. Over time, a trust relationship in a research study will depend on the maintenance by the researcher of their reputation for integrity, competence and benevolence towards the research participant [88–92]. However, there is empirical evidence for a willingness to compromise: technology adoption would not be possible without it [70,71,93,94]. Further, trust may generalise across related areas, which would be beneficial for a research study itself and its context [95–97]. Leaving aside the specific issue of whether this is simply a different definition or approach to informed consent, a trust-based approach may well describe what happens in decisions to engage with research more closely than anticipated by regulatory control or governance frameworks [47]. Adopting a trust-based perspective derived from Mayer et al.’s definition [84] would need to consider how to demonstrate the trustworthiness indicators of integrity, benevolence and competence in pertinent research activities. This will be termed trust-based consent in the discussion below.

5. Scenarios

Notwithstanding any legal obligations under clinical practice and data protection regulations, to evaluate the concept of trust-based research consent, this section considers three different scenarios with specific relevance to the COVID-19 pandemic and beyond. Throughout the discussion above, I have used contract tracing as a starting point. Identifying the transmission paths of contagious diseases with such technology has been around as a concept for some time [98]. However, this has potential implications for privacy including the inadvertent and unconsented disclosure of third parties from a consent perspective. It has been noted that human rights instruments make provision for when community imperatives supersede individual rights (Art. 8 §2, [14]); and trust in government has had implications for the acceptance and success of tracing applications, for instance [76,99]. For

representative research behaviours, therefore, it is important to consider the implications of current informed consent procedures in research ethics as well as from the perspective of trust-based consent introduced above.

- *Contact tracing*: During the COVID-19 pandemic, there has been some discussion about the technical implementation [100] and how tracing fits within a larger socio-technical context [101]. Introduction of such applications is not without controversy in socio-political terms [76,102]. At the same time, there is a balance to be struck between individual rights and the public good [103]; in the case of the COVID-19 pandemic, the social implications of the disease are almost as important as its impact on public and individual health [104]. Major challenges include:
 - Public Opinion;
 - Inadvertent disclosure of third party data;
 - Public/Individual responses to alerts.
- *Big Data Analytics*: this includes exploiting the vast amounts of data available typically via the Internet to attempt to understand behavioural and other patterns [105,106]. Such approaches have already shown much promise in healthcare [107], and with varying degrees of success for tracing the COVID-19 pandemic [108]. There are, however, some concerns about the impact of big data on individuals and society [109,110]. Major challenges include:
 - Identification of key actors;
 - Mutual understanding between those actors;
 - Influence of those actors on processing (and results).
- *Public Health Emergency Research*: multidisciplinary efforts to understand, inform and ultimately control the transmission and proliferation of disease (see for instance [111]) as well as social impacts [99,104], and to consider the long-term implications of the COVID-19 pandemic and other PHEs [112]. Major challenges include:
 - Changes in research focus;
 - Changes introduced as research outcomes become available;
 - Respect for all potential groups;
 - Balancing individual and community rights;
 - Unpredicted benefits of research data and outcomes (e.g., in future).

Table A1 in Appendix A summarises perspectives relating to *informed consent* and *trust-based consent* relating to the three related activities: contact tracing, big data and issues pertinent to research during a PHE as described above. Each of these scenarios needs to be contextualised within different perspectives: the broader socio-political context, the wider delivery ecosystem, and historical and community-benefit aspects, respectively. Traditional informed consent for research would be problematic for different reasons in each case as summarised. If run in connection with or as part of data protection informed consent, any risk of research participants stopping their participation may result in withdrawal of research data unless a different legal basis for processing can be found.

In all three cases, it is apparent that a simple exchange between researcher and research participant is not possible. There are other contextual factors which must be taken into account and which may well introduce additional stakeholders. There are also external factors—contemporary context, a relevant underlying ecosystem setting expectations, and a dynamic and historical perspective which may introduce both types of factors from the other two scenarios—which would indicate at the very least that each contextualised agreement must be re-validated, and that the consent cannot be assumed to remain stable as external factors influence the underlying perceptions of the actors involved. Trust would allow for such contextualisation and implies a continuous negotiation.

6. Discussion and Recommendations

The existing informed consent process clearly poses several problems, not least the potential to confuse research participants about what they are agreeing to: use of an app,

the processing of their personal data, undergoing treatment, or taking part in a research study. This situation would be exacerbated where several such activities co-occur. Indeed, it is not unusual for research studies to include collection of personal data as part of the research protocol. However, there are more challenging issues. Where the researcher is unable to describe exactly what should happen, what the outcomes might be, and how data or participant engagement will be used, then it is impossible to provide sufficient information for any consent to be fully informed. The literature in this area provides some evidence too that research participants may well wish to engage without being overwhelmed with detail they do not want or may not understand. There is an additional complication where multiple stakeholders, not just the researcher, may be involved in handling and interpreting research outcomes. Any such stakeholders should be involved in or at least represented as part of the discussion with the research participant. All of this suggests that there needs to be some willingness to accept risk: participants must trust researchers and their intentions.

6.1. Recommendations for Research Ethics Review

Such a trust-based approach would, however, affect how RECs/IRBs review research submissions. Most importantly, reviewers need to consider the main actors involved in any research and their expectations. This suggests a number of main considerations during review:

1. The research proposal should first describe in some detail the trustworthiness basis for the research engagement. I have used characteristics from the literature—integrity, benevolence, and competence—though others may be more appropriate such as reputation and evidence of participant reactions in related work.
2. The context of the proposed research should be disclosed, including the identification of the types of contextual effects which might be expected. These may include the general socio-political environment, existing relationships that the research participant might be expected to be aware of (such as clinician–patient), and any dynamic effects, such as implications for other cohorts, including future cohorts. Any such contextual factors should be explained, justified and appropriately managed by the researcher.
3. The proposed dialogue between researcher and research participant should be described, how it will be conducted, what it will cover, and how frequently the dialogue will be repeated. This may depend, for example, on when results start to become available. The frequency and delivery channel of this dialogue should be simple for the potential research participant. This must be justified, and the timescales realistic. This part of the trust-based consent process might also include how the researcher will manage research participant withdrawal.

The intention with such an approach would be to move away from the burdensome governance described in the literature (see [47,48], for instance), instead focusing on what is of practical importance to enter into a trust relationship and what might encourage a more natural and familiar communicative exchange with participants. Traditional information such as the assumed benefits of the research outcomes should be confined to the research ethics approval submission; it may not be clear to a potential research participant how relevant that may be for them to make a decision to engage. Review ultimately must consider the *Context* (see Section 3 above) within which a participant develops a *Willingness to Engage*.

The ethics review process thereby becomes an evaluation not only a consideration of the typical cost–benefit to the research participant, but rather of how researcher and research participant are likely to engage with one another to collaborate effectively on an equal footing and sharing the risks of failure. The participant then becomes a genuine actor within the research protocol rather than simply a subject of observation.

6.2. Recommendations for the Ethical Use of Advanced Technologies

Official guidance tends to focus on data governance [13,15] or on the obligations of technologists to provide robust, reliable and transparent operation [16,17]. However, I have emphasised in the previous discussion that it is essential to consider the entire ecosystem where advanced, AI-enabled technologies are deployed. These technologies are an integral part of a broader socio-technical system.

The data scientist providing the technology to a service provider and the service provider themselves must take into account a number of factors:

1. Understand who the main actors are. Each domain (healthcare, eCommerce, social media, and so forth) will often be regulated with specific obligations. More importantly though, I maintain, would be the interaction between end user and provider, and the reliance of the provider on the data scientist or technologist. These actors would all influence the trust context. So how they contribute needs to be understood.
2. Understand what their expectations are. Once the main actors have been identified, their individual expectations will influence how they view their own responsibilities and how they believe the other actors will behave. This will contextualise what each expects from the service or interaction, and from one another.
3. Reinforce *competence, integrity* and *benevolence* (from [84]). As the defining characteristics of a trust relationship outlined above, each of the actors has a responsibility to support that relationship, and to avoid actions which would affect trust. Inadvertent or unavoidable problems can be dealt with ([88,89]). Further, occasional (though infrequent [23]) re-affirmation of the relationship is advantageous. So, ongoing communication between the main actors is important in maintaining trust (see also [12]).

Just as a trust-based approach is proposed as an alternative to the regulatory constraint of existing deontological consent processes, I suggest that the main actors share a responsibility to invest in a relationship. In ethical terms, this is more consistent with Floridi's concept of *entropy* [113]: each actor engages with the high-level interaction (e.g., contact tracing) in support of common beliefs. Rather than trying to balance individual rights and the common good, this assumes engagement by the main actors willing to expose themselves to vulnerability (because outcomes are not necessarily predictable at the outset) and therefore invest jointly towards the success of the engagement.

7. Future Research Directions

Based on existing research across multiple domains, I have presented here a trust-based approach to consent. This assumes an ongoing dialogue between trustor (data subject, service user, research participant, patient) and trustee (data controller, service provider, researcher, clinician). To a large extent, this echoes what Rohlffing and her colleagues describe as a co-constructed negotiation around explainability in AI between explainer and explainee [12]. However, my trust-based approach derives from social psychological terms and therefore accepts vulnerability. None of the stakeholders are assumed to be infallible. Any risk to the engagement is shared across them all. This would now benefit from empirical validation.

Firstly, and following some of the initial work by Wiles and her colleagues [47], trustors of different and representative categories could provide at least two different types of responses: their attitudes and perceptions of current consent processes, backed up with ethnographic observation of how they engage with those processes currently. Secondly, expanding on proposals by Richards and Hartzog [23] as applied not only in the US but also in other jurisdictions, engaging with service providers, researchers and clinicians asked to provide their perspective on how they currently use the consent process and what a trust-based negotiation would mean to them in offering the services or engaging with trustors as described here. Third, it is important to compare the co-construction of explainability for AI technologies (which assumes understanding is enough for acceptability) and the negotiation of shared risk implied by a trust-based approach to consent. If understanding

the technology alone proves insufficient, then informed consent to formalise the voluntary agreement to engage is not enough either.

Synthesising these findings would provide concrete proposals for policy makers, as well as a basis to critically evaluate existing guidance on data sharing and the development and deployment of advanced technologies.

8. Conclusions

In this paper, I have suggested a different approach to negotiating ongoing consent (including terms of use) from the traditional process of informed consent or unwitting acceptance of terms of use, based on the definition of trust from the social psychology literature pertaining to person-to-person interactions. This was motivated by four sets of observations: firstly, that informed consent has different implications in different situations such as data protection, clinical trials or interventions, or research, and known issues with terms of use for online services. Secondly, the research literature highlights multiple cases where the assumptions relating to informed consent do not hold, and terms of use are typically imposed rather than informed and freely given. Thirdly, there may be contexts which are more complex than a simple exchange between two actors: researcher and research participant, or service user and service provider. Finally, even explainability for AI technologies may rely on a co-constructed understanding of outputs between the main stakeholders. Reviewing common activities during the COVID-19 pandemic, but also relevant to any Public Health Emergency, I have stressed that the broader socio-political context, the socio-technical environment within which big data analytics are implemented, and the historical relevance of PHE research complicates a straight-forward informed consent process. Further, researchers may simply not be in a position to predict or guarantee expected research outcomes making fully informed consent problematic. I suggest that this might better be served by a trust-based approach. Trust, in traditional definitions in the behavioural sciences, is based on an acceptance of vulnerability to unknown outcomes, a shared responsibility for those outcomes. In consequence, a more dynamic trust-based negotiation in response to situational changes over time is called for. This, I suggest, could be handled with a much more communication-focused approach, with implications for research ethics review, as well as AI-enhanced services. Moving forward, there needs to be discussion with relevant stakeholders, especially potential research participants and researchers themselves, to understand their expectations and thereby validate the arguments presented here exploring how a trust-based consent process might meet their requirements. Finally, although I have contextualised the discussion here against the background of the coronavirus pandemic, other test scenarios need to be explored to evaluate whether the same factors apply.

Funding: This work was funded in part by the European Union's Horizon 2020 research and innovation programme under grant agreement No 780495 (project BigMedilytics). *Disclaimer:* Any dissemination of results here presented reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains. It was also supported, in part, by the Bill & Melinda Gates Foundation [INV-001309]. Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission.

Data Availability Statement: Not Applicable, the study does not report any data.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Summary of Issues across Domains.

Domain	Challenges	Informed Consent	Trust-Based Consent
Contact Tracing	The socio-political context within which the app is used or research is carried out. Media reporting, including fake news, can influence public confidence	One-off consent on research engagement or upon app download may not be sufficient as context changes. Retention may be challenging depending on trustworthiness perceptions of public authorities and responses to media reports leading to app/research study abandonment (i.e., the impact and relevance of context which may have nothing to do with the actual app/research)	Researchers (app developers) may need to demonstrate <i>integrity</i> and <i>benevolence</i> on an ongoing basis, and specifically when needed in response to any public concerns around data protection, and to any misuse or unforeseen additional use of data. Researchers must therefore communicate their own trustworthiness and position themselves appropriately within a wider socio-political context for which they may feel they have no responsibility. It is their responsibility, however, to maintain the relationship with relevant stakeholders, i.e., to develop and maintain trust.
Big Data Analytics	The potential disruption to an existing ecosystem —e.g., the actors who are important for delivery of service, such as patient and clinician for healthcare, or research participant and researcher for Internet-based research. Technology may therefore be disruptive to any such existing relationship. Further, unless the main actors are identified, it would be difficult to engage with traditional approaches to consent.	Researcher (data scientist) may not be able to disclose all information necessary to make a fully informed decision, not least because they may only be able to describe expected outcomes (and how data will be used) in general terms. The implications of supervised and unsupervised learning may not be understood. Not all beneficiaries can engage with an informed consent process (e.g., clinicians would not be asked to consent formally to data analytics carried out on their behalf; for Internet-based research, it may be impractical or ill-advised for researchers to contact potential research participants).	Data scientists need to engage in the first instance with domain experts in other fields who will use their results (e.g., clinicians in healthcare; web scientists etc. for Internet-based modelling; etc.) to understand each other’s expectations and any limitations. For a clinician or other researcher dependent on the data scientist, this will affect the perception of their own <i>competence</i> . This will also form part of trust-based engagement with a potential research participant. Ongoing communication between participants, data scientists and the other relevant domain experts should continue to maintain perceptions of <i>benevolence</i> and <i>integrity</i> .
Public Health Emergency	The difficulty in identifying the scope of research (in terms of what is required and who will benefit now, and especially in the future) and therefore identify the main stakeholders, not just participants providing (clinical) data directly	The COVID-19 pandemic has demonstrated that research understanding changed significantly over time: the research community, including clinicians, had to adapt. Policy decisions struggled to keep pace with the results. Informed consent would need constant review and may be undermined if research outcomes/policy decisions are not consistent. In the latter case, this may result in withdrawal of research participants. Further, research from previous pandemics was not available to inform current research activities	A PHE highlights the need to balance individual rights and the imperatives for the community (the common good). As well as the effects of fake news, changes in policy based on research outcomes may lead to concern about <i>competence</i> : do the researchers know what they are doing? However, there needs to be an understanding of how the research is being conducted and why things do change. So, there will also be a need for ongoing communication around <i>integrity</i> and <i>benevolence</i> . This may advantageously extend existing public engagement practices, but would also need to consider future generations and who might represent their interests. There is a clear need for an ongoing dialogue including participants where possible, but also other groups with a vested interest in the research data and any associated outcomes, including those who may have nothing to do with the original data collection or circumstances.

References

1. Walker, P.; Lovat, T. You Say Morals, I Say Ethics—What’s the Difference? In *The Conversation*; IMDb: Seattle, WA, USA, 2014.
2. Adadi, A.; Berrada, M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* **2018**, *6*, 52138–52160. [CrossRef]
3. European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, 2016*; European Commission: Brussels, Belgium, 2016.
4. Samek, W.; Wiegand, T.; Müller, K.R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv* **2017**, arXiv:1708.08296.
5. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
6. Gunning, D.; Aha, D.W. DAPRA’s Explainable Artificial Intelligence Program. *AI Mag.* **2019**, *40*, 44–58.
7. Weitz, K.; Schiller, D.; Schlagowski, R.; Huber, T.; André, E. “Do you trust me?”: Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In Proceedings of the IVA ’19: 19th ACM International Conference on Intelligent Virtual Agents, Paris, France, 2–5 July 2019; ACM: New York, NY, USA, 2019; pp. 7–9. [CrossRef]
8. Taylor, S.; Pickering, B.; Boniface, M.; Anderson, M.; Danks, D.; Følstad, A.; Leese, M.; Müller, V.; Sorell, T.; Winfield, A.; et al. *Responsible AI—Key Themes, Concerns & Recommendations For European Research and Innovation*; HUB4NGI Consortium: Zürich, Switzerland, 2018. [CrossRef]
9. Došilović, F.K.; Brčić, M.; Hlupić, N. Explainable Artificial Intelligence: A Survey. In Proceedings of the 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 21–25 May 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 0210–0215. [CrossRef]
10. Khrais, L.T. Role of Artificial Intelligence in Shaping Consumer Demand in E-Commerce. *Future Internet* **2020**, *12*, 226. [CrossRef]
11. Israelsen, B.W.; Ahmed, N.R. “Dave...I can assure you ...that it’s going to be all right ...” A Definition, Case for, and Survey of Algorithmic Assurances in Human-Autonomy Trust Relationships. *ACM Comput. Surv.* **2019**, *51*, 113. [CrossRef]
12. Rohlfing, K.J.; Cimiano, P.; Scharlau, I.; Matzner, T.; Buhl, H.M.; Buschmeier, H.; Eposito, E.; Grimminger, A.; Hammer, B.; Häb-Umbach, R.; et al. Explanation as a social practice: Toward a conceptual framework for the social design of AI systems. *IEEE Trans. Cogn. Dev. Syst.* **2020**. [CrossRef]
13. Amnesty International and AccessNow. The Toronto Declaration: Protecting the Right to Equality and Non-Discrimination in Machine Learning Systems. 2018. Available online: <https://www.accessnow.org/the-toronto-declaration-protecting-the-rights-to-equality-and-non-discrimination-in-machine-learning-systems/> (accessed on 14 May 2021).
14. Council of Europe. *European Convention for the Protection of Human Rights and Fundamental Freedoms, as Amended by Protocols Nos. 11 and 14*; Council of Europe: Strasbourg, France, 2010.
15. UK Government Digital Services. Data Ethics Framework. 2020. Available online: <https://www.gov.uk/government/publications/data-ethics-framework> (accessed on 14 May 2021).
16. Department of Health and Social Care. *Digital and Data-Driven Health and Care Technology*; Department of Health and Social Care: London, UK, 2021.
17. European Commission. *Ethics Guidelines for Trustworthy AI*; European Commission: Brussels, Belgium, 2019.
18. Ajzen, I. The theory of planned behavior. *Organ. Behav. Hum. Decis. Process.* **1991**, *50*, 179–211. [CrossRef]
19. Murray, P.M. The History of Informed Consent. *Iowa Orthop. J.* **1990**, *10*, 104–109.
20. USA v Brandt Court. The Nuremberg Code (1947). *Br. Med. J.* **1996**, *313*, 1448. [CrossRef]
21. World Medical Association. *WMA Declaration of Helsinki—Ethical Principles for Medical Research Involving Human Subjects*; World Medical Association: Ferney-Voltaire, France, 2018.
22. Lemley, M.A. Terms of Use. *Minn. Law Rev.* **2006**, *91*, 459–483.
23. Richards, N.M.; Hartzog, W. The Pathologies of Digital Consent. *Wash. Univ. Law Rev.* **2019**, *96*, 1461–1504.
24. Luger, E.; Moran, S.; Rodden, T. Consent for all: Revealing the hidden complexity of terms and conditions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Paris, France, 27 April–2 May 2013; pp. 2687–2696.
25. Belmont. *The Belmont Report: Ethical Principles and Guidelines for The Protection of Human Subjects of Research*; American College of Dentists: Gaithersburg, MD, USA, 1979.
26. Beauchamp, T.L. History and Theory in “Applied Ethics”. *Kennedy Inst. Ethics J.* **2007**, *17*, 55–64. [CrossRef] [PubMed]
27. Muirhead, W. When four principles are too many: Bloodgate, integrity and an action-guiding model of ethical decision making in clinical practice. *Clin. Ethics* **2011**, *38*, 195–196. [CrossRef]
28. Rubin, M.A. The Collaborative Autonomy Model of Medical Decision-Making. *Neuro. Care* **2014**, *20*, 311–318. [CrossRef]
29. The Health Service (Control of Patient Information) Regulations 2002. 2002. Available online: <https://www.legislation.gov.uk/uksi/2002/1438/contents/made> (accessed on 14 May 2021).
30. Hartzog, W. The New Price to Play: Are Passive Online Media Users Bound By Terms of Use? *Commun. Law Policy* **2010**, *15*, 405–433. [CrossRef]
31. Beauchamp, T.L.; Childress, J.F. *Principles of Biomedical Ethics*, 8th ed.; Oxford University Press: Oxford, UK, 2019.

32. OECD. *Frascati Manual 2015*; OECD: Paris, France, 2015. [CrossRef]
33. BPS. *Code of Human Research Ethics*; BPS: Leicester, UK, 2014.
34. Herschel, R.; Miori, V.M. Ethics & Big Data. *Technol. Soc.* **2017**, *49*, 31–36. [CrossRef]
35. Floridi, L.; Taddeo, M. What is data ethics? *Philos. Trans. R. Soc.* **2016**. [CrossRef]
36. Carroll, S.R.; Garba, I.; Figueroa-Rodríguez, O.L.; Holbrook, J.; Lovett, R.; Materechera, S.; Parsons, M.; Raseroka, K.; Rodriguez-Lonebear, D.; Rowe, R.; et al. The CARE Principles for Indigenous Data Governance. *Data Sci. J.* **2020**, *19*, 1–12. [CrossRef]
37. Thomson, J.J. The Trolley Problem. *Yale Law J.* **1985**, *94*, 1395–1415. [CrossRef]
38. Parsons, T.D. *Ethical Challenges in Digital Psychology and Cyberpsychology*; Cambridge University Press: Cambridge, UK, 2019.
39. Murove, M.F. Ubuntu. *Diogenes* **2014**, *59*, 36–47. [CrossRef]
40. Ess, C. *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee*; IGI Global: Hershey, PA, USA, 2002.
41. Markham, A.; Buchanan, E. *Ethical Decision-Making and Internet Research: Recommendations from the Aoir Ethics Working Committee (Version 2.0)*. 2002. Available online: <https://aoir.org/reports/ethics2.pdf> (accessed on 14 May 2021).
42. Sugarman, J.; Lavori, P.W.; Boeger, M.; Cain, C.; Edson, R.; Morrison, V.; Yeh, S.S. Evaluating the quality of informed consent. *Clin. Trials* **2005**, *2*, 34–41. [CrossRef]
43. Biros, M. Capacity, Vulnerability, and Informed Consent for Research. *J. Law Med. Ethics* **2018**, *46*, 72–78. [CrossRef]
44. Tam, N.T.; Huy, N.T.; Thoa, L.T.B.; Long, N.P.; Trang, N.T.H.; Hirayama, K.; Karbwang, J. Participants' understanding of informed consent in clinical trials over three decades: Systematic review and meta-analysis. *Bull. World Health Organ.* **2015**, *93*, 186H–198H. [CrossRef] [PubMed]
45. Falagas, M.E.; Korbila, I.P.; Giannopoulou, K.P.; Kondilis, B.K.; Peppas, G. Informed consent: How much and what do patients understand? *Am. J. Surg.* **2009**, *198*, 420–435. [CrossRef] [PubMed]
46. Nusbaum, L.; Douglas, B.; Damus, K.; Paasche-Orlow, M.; Estrella-Luna, N. Communicating Risks and Benefits in Informed Consent for Research: A Qualitative Study. *Glob. Qual. Nurs. Res.* **2017**, *4*. [CrossRef]
47. Wiles, R.; Crow, G.; Charles, V.; Heath, S. Informed Consent and the Research Process: Following Rules or Striking Balances? *Sociol. Res. Online* **2007**, *12*. [CrossRef]
48. Wiles, R.; Charles, V.; Crow, G.; Heath, S. Researching researchers: Lessons for research ethics. *Qual. Res.* **2006**, *6*, 283–299. [CrossRef]
49. Naarden, A.L.; Cissik, J. Informed Consent. *Am. J. Med.* **2006**, *119*, 194–197. [CrossRef] [PubMed]
50. Al Mahmoud, T.; Hashim, M.J.; Almahmoud, R.; Branicki, F.; Elzubeir, M. Informed consent learning: Needs and preferences in medical clerkship environments. *PLoS ONE* **2018**, *13*, e0202466. [CrossRef]
51. Nijhawan, L.P.; Janodia, M.D.; Muddukrishna, B.S.; Bhat, K.M.; Bairy, K.L.; Udupa, N.; Musmade, P.B. Informed consent: Issues and challenges. *J. Adv. Phram. Technol. Res.* **2013**, *4*, 134–140. [CrossRef]
52. Kumar, N.K. Informed consent: Past and present. *Perspect. Clin. Res.* **2013**, *4*, 21–25. [CrossRef]
53. Hofstede, G. *Cultural Dimensions*. 2003. Available online: www.geert Hofstede.com (accessed on 12 May 2021).
54. Hofstede, G.; Hofstede, J.G.; Minkov, M. *Cultures and Organizations: Software of the Mind*, 3rd ed.; McGraw-Hill: New York, NY, USA, 2010.
55. Acquisti, A.; Brandimarte, L.; Loewenstein, G. Privacy and human behavior in the age of information. *Science* **2015**, *347*, 509–514. [CrossRef] [PubMed]
56. McEvily, B.; Perrone, V.; Zaheer, A. Trust as an Organizing Principle. *Organ. Sci.* **2003**, *14*, 91–103. [CrossRef]
57. Milgram, S. Behavioral study of obedience. *J. Abnorm. Soc. Psychol.* **1963**, *67*, 371–378. [CrossRef] [PubMed]
58. Haney, C.; Banks, C.; Zimbardo, P. *Interpersonal Dynamics in a Simulated Prison*; Wiley: New York, NY, USA, 1972.
59. Reicher, S.; Haslam, S.A. Rethinking the psychology of tyranny: The BBC prison study. *Br. J. Soc. Psychol.* **2006**, *45*, 1–40. [CrossRef] [PubMed]
60. Reicher, S.; Haslam, S.A. After shock? Towards a social identity explanation of the Milgram 'obedience' studies. *Br. J. Soc. Psychol.* **2011**, *50*, 163–169. [CrossRef]
61. Beauchamp, T.L. Informed Consent: Its History, Meaning, and Present Challenges. *Camb. Q. Healthc. Ethics* **2011**, *20*, 515–523. [CrossRef]
62. Ferreira, C.M.; Serpa, S. Informed Consent in Social Sciences Research: Ethical Challenges. *Int. J. Soc. Sci. Stud.* **2018**, *6*, 13–23. [CrossRef]
63. Hofmann, B. Broadening consent - and diluting ethics? *J. Med Ethics* **2009**, *35*, 125–129. [CrossRef]
64. Steinsbekk, K.S.; Myskja, B.K.; Solberg, B. Broad consent versus dynamic consent in biobank research: Is passive participation an ethical problem? *Eur. J. Hum. Genet.* **2013**, *21*, 897–902. [CrossRef] [PubMed]
65. Sreenivasan, G. Does informed consent to research require comprehension? *Lancet* **2003**, *362*, 2016–2018. [CrossRef]
66. O'Neill, O. Some limits of informed consent. *J. Med. Ethics* **2003**, *29*, 4–7. [CrossRef] [PubMed]
67. Davis, F.D. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* **1989**, 319–340. [CrossRef]
68. Venkatesh, V.; Morris, M.G.; Davis, G.B.; Davis, F.D. User acceptance of information technology: Toward a unified view. *MIS Q.* **2003**, *27*, 425–478. [CrossRef]

69. McKnight, H.; Carter, M.; Clay, P. Trust in technology: Development of a set of constructs and measures. In Proceedings of the Digit, Phoenix, AZ, USA, 15–18 December 2009.
70. McKnight, H.; Carter, M.; Thatcher, J.B.; Clay, P.F. Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manag. Inf. Syst. (TMIS)* **2011**, *2*, 12. [CrossRef]
71. Thatcher, J.B.; McKnight, D.H.; Baker, E.W.; Arsal, R.E.; Roberts, N.H. The Role of Trust in Postadoption IT Exploration: An Empirical Examination of Knowledge Management Systems. *IEEE Trans. Eng. Manag.* **2011**, *58*, 56–70. [CrossRef]
72. Hinch, R.; Probert, W.; Nurtay, A.; Kendall, M.; Wymant, C.; Hall, M.; Lythgoe, K.; Cruz, A.B.; Zhao, L.; Stewart, A. Effective Configurations of a Digital Contact Tracing App: A Report to NHSX. 2020. Available online: https://cdn.theconversation.com/static_files/files/1009/Report_-_Effective_App_Configurations.pdf (accessed on 14 May 2021).
73. Parker, M.J.; Fraser, C.; Abeler-Dörner, L.; Bonsall, D. Ethics of instantaneous contact tracing using mobile phone apps in the control of the COVID-19 pandemic. *J. Med. Ethics* **2020**, *46*, 427–431. [CrossRef]
74. Walrave, M.; Waeterloos, C.; Ponnet, K. Ready or Not for Contact Tracing? Investigating the Adoption Intention of COVID-19 Contact-Tracing Technology Using an Extended Unified Theory of Acceptance and Use of Technology Model. *Cyberpsychol. Behav. Soc. Netw.* **2020**. [CrossRef]
75. Velicia-Martin, F.; Cabrera-Sanchez, J.-P.; Gil-Cordero, E.; Palos-Sanchez, P.R. Researching COVID-19 tracing app acceptance: Incorporating theory from the technological acceptance model. *PeerJ Comput. Sci.* **2021**, *7*, e316. [CrossRef]
76. Rowe, F.; Ngwenyama, O.; Richet, J.-L. Contact-tracing apps and alienation in the age of COVID-19. *Eur. J. Inf. Syst.* **2020**, *29*, 545–562. [CrossRef]
77. Roache, R. Why is informed consent important? *J. Med. Ethics* **2014**, *40*, 435–436. [CrossRef] [PubMed]
78. Eyal, N. Using informed consent to save trust. *J. Med. Ethics* **2014**, *40*, 437–444. [CrossRef]
79. Eyal, N. Informed consent, the value of trust, and hedons. *J. Med. Ethics* **2014**, *40*, 447. [CrossRef] [PubMed]
80. Tännsjö, T. Utilitarianism and informed consent. *J. Med. Ethics* **2013**, *40*, 445. [CrossRef] [PubMed]
81. Bok, S. Trust but verify. *J. Med. Ethics* **2014**, *40*, 446. [CrossRef]
82. Rousseau, D.M.; Sitkin, S.B.; Burt, R.S.; Camerer, C. Not so different after all: A cross-discipline view of trust. *Acad. Manag. Rev.* **1998**, *23*, 393–404. [CrossRef]
83. Robbins, B.G. What is Trust? A Multidisciplinary Review, Critique, and Synthesis. *Sociol. Compass* **2016**, *10*, 972–986. [CrossRef]
84. Mayer, R.C.; Davis, J.H.; Schoorman, F.D. An Integrative Model of Organizational Trust. *Acad. Manag. Rev.* **1995**, *20*, 709–734. [CrossRef]
85. Weber, L.R.; Carter, A.I. *The Social Construction of Trust*; Clinical Sociology: Research and Practice; Springer Science+Business Media: Berlin/Heidelberg, Germany, 2003.
86. Ferrin, D.L.; Bligh, M.C.; Kohles, J.C. Can I Trust You to Trust Me? A Theory of Trust, Monitoring, and Cooperation in Interpersonal and Intergroup Relationships. *Group Organ. Manag.* **2007**, *32*, 465–499. [CrossRef]
87. Schoorman, F.D.; Mayer, R.C.; Davis, J.H. An integrative model of organizational trust: Past, present, and future. *Acad. Manag. Rev.* **2007**, *32*, 344–354. [CrossRef]
88. Fuoli, M.; Paradis, C. A model of trust-repair discourse. *J. Pragmat.* **2014**, *74*, 52–69. [CrossRef]
89. Lewicki, R.J.; Wiethoff, C. Trust, Trust Development, and Trust Repair. *Handb. Confl. Resolut. Theory Pract.* **2000**, *1*, 86–107.
90. Bachmann, R.; Gillespie, N.; Priem, R. Repairing Trust in Organizations and Institutions: Toward a Conceptual Framework. *Organ. Stud.* **2015**, *36*, 1123–1142. [CrossRef]
91. Bansal, G.; Zahedi, F.M. Trust violation and repair: The information privacy perspective. *Decis. Support Syst.* **2015**, *71*, 62–77. [CrossRef]
92. Memery, J.; Robson, J.; Birch-Chapman, S. Conceptualising a Multi-level Integrative Model for Trust Repair. In Proceedings of the EMAC, Hamburg, Germany, 28–31 May 2019.
93. Lee, J.D.; See, K.A. Trust in automation: Designing for appropriate reliance. *Hum. Factors J. Hum. Factors Ergon. Soc.* **2004**, *46*, 50–80. [CrossRef]
94. Lee, J.-H.; Song, C.-H. Effects of trust and perceived risk on user acceptance of a new technology service. *Soc. Behav. Personal. Int. J.* **2013**, *41*, 587–598. [CrossRef]
95. Cheshire, C. Online Trust, Trustworthiness, or Assurance? *Daedalus* **2011**, *140*, 49–58. [CrossRef] [PubMed]
96. Pettit, P. Trust, Reliance, and the Internet. *Inf. Technol. Moral Philos.* **2008**, *26*, 161.
97. Stewart, K.J. Trust Transfer on the World Wide Web. *Organ. Sci.* **2003**, *14*, 5–17. [CrossRef]
98. Eames, K.T.D.; Keeling, M.J. Contact tracing and disease control. *Proc. R. Soc. Lond.* **2003**, *270*, 2565–2571. [CrossRef]
99. Jetten, J.; Reicher, S.D.; Haslam, S.A.; Cruwys, T. *Together Apart: The Psychology of COVID-19*; SAGE Publications Ltd.: Thousand Oaks, CA, USA, 2020.
100. Ahmed, N.; Michelin, R.A.; Xue, W.; Ruj, S.; Malaney, R.; Kanhere, S.S.; Seneviratne, A.; Hu, W.; Janicke, H.; Jha, S.K. A Survey of COVID-19 Contact Tracing Apps. *IEEE Access* **2020**, *8*, 134577–134601. [CrossRef]
101. Kretzschmar, M.E.; Roszhnova, G.; Bootsma, M.C.; van Boven, M.J.; van de Wijgert, J.H.; Bonten, M.J. Impact of delays on effectiveness of contact tracing strategies for COVID-19: A modelling study. *Lancet Public Health* **2020**, *5*, e452–e459. [CrossRef]
102. Bengio, Y.; Janda, R.; Yu, Y.W.; Ippolito, D.; Jarvie, M.; Pilat, D.; Struck, B.; Krastev, S.; Sharma, A. The need for privacy with public digital contact tracing during the COVID-19 pandemic. *Lancet Digit. Health* **2020**, *2*, e342–e344. [CrossRef]

103. Abeler, J.; Bäcker, M.; Buermeyer, U.; Zillesen, H. COVID-19 Contact Tracing and Data Protection Can Go Together. *JMIR Mhealth Uhealth* **2020**, *8*, e19359. [CrossRef] [PubMed]
104. Van Bavel, J.J.; Baicker, K.; Boggio, P.S.; Caprano, V.; Cichocka, A.; Cikara, M.; Crockett, M.J.; Crum, A.J.; Douglas, K.M.; Druckman, J.N.; et al. Using social and behavioural science to support COVID-19 pandemic response. *Nat. Hum. Behav.* **2020**, *4*, 460–471. [CrossRef] [PubMed]
105. Ackland, R. *Web Social Science: Concepts, Data and Tools for Social Scientists in the Digital Age*; SAGE Publications Ltd.: Thousand Oaks, CA, USA, 2013.
106. Papacharissi, Z. *A Networked Self and Platforms, Stories, Connections*; Routledge: London, UK, 2018.
107. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*, 1–10. [CrossRef]
108. Agbehadji, I.E.; Awuzie, B.O.; Ngowi, A.B.; Millham, R.C. Review of Big Data Analytics, Artificial Intelligence and Nature-Inspired Computing Models towards Accurate Detection of COVID-19 Pandemic Cases and Contact Tracing. *Int. J. Environ. Res. Public Health* **2020**, *17*, 5330. [CrossRef]
109. Cheney-Lippold, J. *We Are Data: Algorithms and the Making of Our Digital Selves*; New York University Press: New York, NY, USA, 2017.
110. O’Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*; Crown: New York, NY, USA, 2016.
111. Austin, C. RDA COVID-19 Zotero Library—March 2021. Available online: <https://www.rd-alliance.org/group/rda-covid19-rda-covid19-omics-rda-covid-19-epidemiology-rda-covid19-clinical-rda-covid19> (accessed on 14 May 2021).
112. Norton, A.; Sigfrid, L.; Aderoba, A.; Nasir, N.; Bannister, P.G.; Collinson, S.; Lee, J.; Boily-Larouche, G.; Golding, J.P.; Depoortere, E.; et al. Preparing for a pandemic: Highlighting themes for research funding and practice—Perspectives from the Global Research Collaboration for Infectious Disease Preparedness (GloPID-R). *BMC Med.* **2020**, *18*, 273. [CrossRef] [PubMed]
113. Floridi, L. On the intrinsic value of information objects and the infosphere. *Ethics Inf. Technol.* **2002**, *4*, 287–304. [CrossRef]

Article

Reconciling Remote Sensing Technologies with Personal Data and Privacy Protection in the European Union: Recent Developments in Greek Legislation and Application Perspectives in Environmental Law

Maria Maniadaki ^{1,*}, Athanasios Papathanasopoulos ¹, Lilian Mitrou ² and Efpraxia-Aithra Maria ¹

¹ School of Environmental Engineering, Technical University of Crete, 73100 Chania, Greece; apapathanasopoul@isc.tuc.gr (A.P.); emaria@isc.tuc.gr (E.-A.M.)

² Department of Information and Communication Systems Engineering, University of the Aegean-Greece, 81100 Mitilini, Greece; L.Mitrou@aegean.gr

* Correspondence: mmaniadaki@isc.tuc.gr

Citation: Maniadaki, Maria, Athanasios Papathanasopoulos, Lilian Mitrou, and Efpraxia-Aithra Maria. 2021. Reconciling Remote Sensing Technologies with Personal Data and Privacy Protection in the European Union: Recent Developments in Greek Legislation and Application Perspectives in Environmental Law. *Laws* 10: 33. <https://doi.org/10.3390/laws10020033>

Received: 7 March 2021

Accepted: 7 May 2021

Published: 11 May 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Using remote sensing technologies to ensure environmental protection responds to the need of protection of a right and a public good and interest. However, the increasing introduction of these technologies has raised new challenges, such as their interference with the rights of privacy and personal data, which are also protected fundamental rights. In this paper the importance of remote sensing technologies as tools for environmental monitoring and environmental law enforcement is analyzed, while legal issues regarding privacy and data protection from their use for environmental purposes are presented. Existing legislation for reconciling emerging conflicts is also examined and major European Court of Human Rights (ECtHR) and Court of Justice of the European Union (CJEU) case law on the issue is approached. Finally, recent developments in Greek legislation and their application perspectives in environmental law are presented as a timely “case study”.

Keywords: Remote Sensing; personal data; privacy; drones; UAV; satellites; environmental monitoring; environmental law

1. Introduction

The development of remote sensing technologies, has led to numerous applications in several sectors. Remote sensing “provides tools for gathering data and solving real world problems¹”. Especially in the field of environmental monitoring, the development of remote sensing technologies has been proven more than crucial, as it enables the collection of a wealth of data for Earth’s current and future state, affecting directly the decision making process as well as the environmental law enforcement sector (Mertikas et al. 2021). However, the transformation of collected data into useful information in the scope of environmental law, raises new challenges, such as their interference with the rights of privacy and personal data (Coffer 2020; Santos and Rapp 2019; Finn and Wright 2016; Sandbrook 2015; Doldirina 2014; Purdy 2011). Although it has become common knowledge that environmental problems have a global impact, calling thus for global action, nations still have their own role in legislation and regulation. In this sense, embracing new technologies such as remote sensing technologies in the case of Greece responds not only to Article 37 of EU Charter of Fundamental Rights² but also to the need of protection of a—in Greece constitutionally anchored—right and a public good and interest for environmental protection (Article 24 of the Greek Constitution). At the same time, key questions arise: is

¹ Available online: http://gsp.humboldt.edu/OLM/Courses/GSP_216_Online/lesson8-2/future.html (accessed on 5 April 2021).

² Article 37 of EU Charter of Fundamental Rights: “A high level of environmental protection and the improvement of the quality of the environment must be integrated into the policies of the Union and ensured in accordance with the principle of sustainable development”.

the protection of privacy and personal data a normative restriction thereof and vice versa? How could a fair and balanced reconciliation of all rights be achieved? Does the law provide the instruments for striking this balance? What is the role of the existing ECtHR and CJEU case law for such an interpretation? Further, what is more: does national legislation play a role for a successful regulation? The paper is structured in four parts, as follows: in the first part, the importance of remote sensing technologies as tools for environmental monitoring and environmental law enforcement is analyzed. In the second part, legal issues regarding privacy and data protection from the use of remote sensing technologies for environmental purposes are presented. In the third part, existing legislation for reconciling emerging conflicts from the application of remote sensing technologies between the right for a high level of environmental protection and the rights for privacy and personal data protection is examined. In addition, major ECtHR and CJEU case law on the issue is approached focusing on the application of the principle of proportionality. In the fourth part, recent developments in Greek legislation and their application perspectives in environmental law are presented as a timely “case study”. Greece, one of the oldest members of EU, with 80% of its surface belonging to mountainous areas and with thousands of islands, faces difficulties in the collection of data for its territory. As a result, the use of remote sensing technologies in Greece seems inevitable and therefore this country may become an excellent example for studying emerging challenges from the application of remote sensing technologies in the environmental sector.

2. Remote Sensing Technologies as Tools for Environmental Monitoring and Environmental Law Enforcement

2.1. Definitions-Brief Description of Current and Future Capacities

“Remote sensing may be broadly defined as the collection of information about an object without being in physical contact with the object. Aircraft and satellites are the common platforms from which remote sensing observations are made. The term remote sensing is restricted to methods that employ electromagnetic energy as the means of detecting and measuring target characteristics” (Sabins 1978). Remote sensing systems are based on signals and images acquired by sensors installed on artificial satellites or aircraft and are used for vast geographical phenomena (di Vimercati et al. 2013). The advancement of satellite technologies and unmanned aerial vehicles has been remarkable last decades. The technological development of satellite technologies on one hand has led to on-demand satellite constellations, which deliver high resolution data (0.75 m) with a daily revisit interval anywhere around the globe. In addition to the high resolution, they can acquire a sequence of images with a small time interval (video persistent mode) due to their unique rapid sensor depointing agility (Almar et al. 2019). Furthermore, as more countries gain their own Earth observation capability, commercialization is a common theme (Harris and Baumann 2021). On the other hand, unmanned aerial vehicles or “drones”, although initially used almost exclusively for military applications, it is now to mention their rapid development for civil applications, and it has even been said that “we are entering the drone age” (Anderson 2012). The surveillance capabilities of drones are rapidly advancing and cheap storage is now available³. The capabilities of drones depend on what they are able to carry. Due to the growing commercialization of drones, commercial UAV manufacturers will increasingly improve their products following the needs of their clients. Additionally, a service sector will evolve to offer UAV services such as leased systems, on-demand flights, or consultation for choosing appropriate platforms or analyzing UAV-generated data (Watts et al. 2012).

To sum up, the future of remote sensing technologies can be described into three words: development, privatization, commercialization.

³ Drones and Environmental Monitoring. 2017. Environmental Law Institute, Washington, DC, USA.

2.2. Applications of Remote Sensing Technologies in Environmental Monitoring and Environmental Law Enforcement

Remote sensing is used in numerous fields for environmental purposes. Remote sensing has provided the means for detecting and quantifying the rates of pollution, as well as for mapping and monitoring sources of pollution and the degree of remediation for their management. It has the means to respond and facilitate environmental management, and makes sound and evidence-based decisions in relation to Earth's resources at a global scale and across different continents, nations, and domains (Mertikas et al. 2021). Such a collection of environmental monitoring data through remote sensing technologies is undoubtedly essential for the effective decision making of environmental authorities.

Simultaneously, the most important applications of remote sensing technologies in environmental law enforcement consist of their use from public authorities for their work (duty) known as "environmental compliance assurance". Environmental compliance assurance describes all the ways in which public authorities promote, monitor and enforce compliance with environmental law. Through the Copernicus program and the relevant EU action plan, the EU Commission promotes the use of satellite images and other geospatial data resources to detect illegal disposal of waste, illegal land use and other breaches⁴. Earth observation technology may also contribute to implementing and ensuring compliance with multilateral environmental agreements (Kuriyama 2005) and they have been actually used to monitor the implementation of environmental agreements such as the World Heritage Convention, the Convention of Biological Diversity, the Ramsar Convention, the UN Convention to Combat Desertification, and the UN Framework Convention on Climate Change. In some countries, such as the Netherlands, earth observation technology is also used in the preparation of 'environmental impact reports' to obtain permits for new water projects, in order to verify their compliance with the legal framework⁵. Another significant application of remote sensing technologies in environmental law enforcement refers to collecting reliable information that can provide solid evidence to combat environmental crime (Patias et al. 2020). However, remote sensing technologies as means of proof are subject to certain limitations and are therefore preferably used as complementary means of proof. In particular, data collected by remote sensing technologies are of digital nature which means that they are subject to alterations and thus need to be verified⁶. In addition, strict control of the whole process of data collection and interpretation is essential, from the moment the data is obtained, in order to avoid wrong evidence (Laituri 2018).

3. Privacy and Data Protection: Legal Issues from the Use of Remote Sensing Technologies for Environmental Monitoring and Environmental Law Enforcement

Technology has always been a threat to the right to privacy, in other words, to "the right to be le(f)t alone" (Warren and Brandeis 1890). In spite of several attempts that have been made to define privacy, no universal definition of privacy could be created. Although the claim for privacy is universal, its concrete form differs according to the prevailing societal characteristics, the economic and cultural environment (Lucács 2016). There are—among others—the following forms of privacy: information privacy and location privacy. Informational privacy indicates much more as informational seclusion, a refuge for the individual. Informational privacy rests on the premise that information about ourselves is something over which individuals may exercise autonomy. Location privacy refers to the right of individuals to move in their "home" and other public or semi-public places without being identified, tracked or monitored (Mitrou 2009). In this sense, the use of remote sensing technologies in the current era may interfere with the rights to informational and location privacy. Observation of private spaces with remote sensing technologies or the location of a person (even without collection of data) or even the correlation of collected data with other

⁴ Available online: https://ec.europa.eu/environment/legal/compliance_en.htm (accessed on 5 April 2021).

⁵ ESA Workshop Evidence from Space, Document ESA-ISPL/EO 47, 5 October 2010, Available on line: <https://www.space-institute.org/wp-content/uploads/2010/10/Workshop-Information-Package-Final.pdf> (accessed on 5 April 2021).

⁶ Ibid.

data may reveal information about individuals' (private) life. Especially when using drones also the so called "bodily privacy" could be affected. As "bodily privacy" we understand also the right to keep bodily functions and body characteristics private (Mitrou 2009). Indicatively, regarding the use of remote sensing technologies for monitoring compliance with environmental legislation on vegetation clearance, in a survey of UK and Australian farmers about their attitudes to being monitored using satellite imagery, most farmers were happy to be monitored this way in principle, however, 58% of Australian respondents and 75% of UK respondents agreed that satellite monitoring was "an invasion of their privacy" (Purdy 2011). Similarly, even if people are aware that certain drones are used for conservation purposes, for example for combatting illegal hunting in South Africa, they may nonetheless feel aggrieved (Sandbrook 2015). The use of remote sensing technologies may interfere also with the right to data protection. Privacy and data protection are closely linked but they are not identical. Data protection serves the protection of private life but the relevant rules apply also to personally identified information, which does not fall under the scope of "private life" even in its broad interpretation. Data protection rules are applicable, whenever personal data are processed (Mitrou 2009). The right to data protection will only protect individuals when remote sensing technologies process personal data (which includes collection of personal data). The collection of images, videos, sounds, and the geo-localization data related to an identified or identifiable natural person (according to the definition of Article 4 (1) of General Data Protection Regulation—GDPR) that has been collected by remote sensing technologies and may also be processed by using suitable methods is subject to data protection legislation. According to CJEU case law, personal data are those that "allow very precise conclusions to be drawn concerning the private lives of the persons whose data has been retained, such as the habits of everyday life, permanent or temporary places of residence, daily or other movements, the activities carried out, the social relationships of those persons and the social environments frequented by them"⁷.

In this sense, very high resolution (VHR) satellite imagery creates considerable challenges for personal data protection, since contextualizing satellite imagery in reference to geographical locations, such as neighborhoods or even houses, can transform an individual in an image from arbitrary to distinguishable (Coffer 2020). Additionally, interactive maps that integrate various types of data, including satellite Earth observation data, into GIS, as well as zooming function available when browsing GIS, may make available personal information linked to a specific geographic location or even an individual (Doldirina 2014). In addition, the application of facial recognition technology or big data analytical software in data collected by remote sensing technologies puts in danger the protection of personal data when it constitutes process of personal data. With regard to drones the threats are more direct, since they can easily observe persons and private spaces and collect personal data, such as persons' locations, relationships etc. Further, what is more: if data subjects are not informed about the use of remote sensing technologies for monitoring purposes their right to informational self-determination and to autonomous and informed decision making is affected. Furthermore, if they are not adequately informed about the data processing equipment, about the purposes of data collection and the identity of who is collecting data as well as the agency's or company's location, that would result in an increased feeling of being under surveillance and a subsequent possible decrease in the legitimate exercise of civil liberties and rights, best known as "chilling effect"⁸.

For this reason, personal data protection law is applicable, so that personal data procession may be only under strict requirements allowed (see below under Section 4.2). Before applying personal data protection law, it must be first checked whether personal data concerns are raised by the use of remote sensing technologies in each particular case.

⁷ C-293/12 and C-594/12 *Digital Rights Ireland* para 27, C-203/15 and C 698/15 *Tele 2* para 99 and C-207/16 *Ministerio Fiscal* para 60.

⁸ On the chilling and panopticon effect syndrome arising from a large-scale use of drones, see Rachel L. Finn, David Wright and Anna Donovan (Trilateral Research & Consulting, LLP), Laura Jacques and Paul De Hert (Vrije Universiteit Brussel), 2014, Privacy, data protection and ethical risks in civil RPAS operations, 7 November 2014, Available online: [http://ec.europa.eu\T1\textgreater{}translations\T1\textgreater{}renditions\T1\textgreater{}pdf](http://ec.europa.eu/T1\textgreater{}translations\T1\textgreater{}renditions\T1\textgreater{}pdf) (accessed on 5 April 2021).

For example, regarding the use of remote sensing technologies for the detection of planning breaches, it is remarkable that the Belgium Privacy Commission in its Opinion no. 26/2006 stated that: “The Privacy Commission considered that the satellite images, insofar as they concerned property of natural persons, constituted information about identified or identifiable natural persons which qualified as personal data for the purposes of privacy law, and that the processing of that information by the planning authorities had to be treated as processing of personal data within the meaning of privacy law” (Billiet 2012).

4. Setting the Limits between Conflicting Rights

It is clear so far, that the importance of remote sensing technologies as tools for environmental monitoring and environmental law enforcement is undoubtable, however, the same time their use may cause considerable threats to the rights for privacy and personal data protection. In the following section, it is examined how a fair and balanced reconciliation of all rights could be achieved before technology significantly outpaces legislation⁹.

4.1. Specific Legislation on Remote Sensing Technologies

Satellite remote sensing is subject to international space law. The Outer Space Treaty and the four follow-on treaties consist the most important documents for international space law. They have not been recently modified. There is to observe a lack of relevant and precise guidance in the Outer Space Treaty on issues of privacy related to VHR satellite data. Further, in the four follow-on treaties on space no specific provision is included, as no consideration has been given to privacy aspects and the respective protection. This is due to the fact that at the time these major space treaties were drafted no consideration was given to privacy protection (Dunk 2013). Only the Convention on International Liability for Damage Caused by Space Objects rules in Article II that “A launching State shall be absolutely liable to pay compensation for damage caused by its space object on the surface of the earth or to aircraft in flight¹⁰”. Taking into account that the term “damage” in Article I (a) is defined as the “loss of life, personal injury or other impairment of health”, it can be claimed that a violation of an individual’s privacy right can be potentially construed as an impairment of health under this Convention. Such an interpretation is based on the World Health Organization’s definition of health¹¹, according to which health is “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” (Santos and Rapp 2019). In this sense, targeted surveillance or even the fear of constant surveillance by satellite remote sensing may disturb people’s mental and social well-being and cause “damage” under the Convention on International Liability for Damage Caused by Space Object. Finally, neither the Resolution 41/65 on the Principles of Remote Sensing of the Earth from Outer Space focuses at all on privacy matters.

International law regarding unmanned aircraft systems clearly states a need for harmonization comparable to that of manned operations, even though drones are subject to national civil aviation law of the member States¹². However, in such international contexts there is again no clear reference to privacy and personal data matters.

Nevertheless, especially for drones, there is to mention a recent trend for detailed regulation in European level. Regulation (EU) 2018/1139 clearly recognizes the threats for privacy and personal data protection by the use of drones: “The rules regarding unmanned aircraft should contribute to achieving compliance with relevant rights guaranteed under

⁹ According to the Collingridge dilemma ‘Regulators having to regulate emerging technologies face a double- bind problem: the effects of new technology cannot be easily predicted until the technology is extensively deployed. Yet once deployed they become entrenched and are then difficult to change’ (Collingridge 1980).

¹⁰ Convention on International Liability for Damage Caused by Space Objects (1972), Available online: https://www.unoosa.org/pdf/gares/ARES_26_2777E.pdf (accessed on 5 May 2021).

¹¹ Preamble to the Constitution of the World Health Organization, reprinted in Final Acts of the International Health Conference, U.N. Doc. E/155, at 11 (1946).

¹² See: ICAO Cir 328, Unmanned Aircraft Systems (UAS), Available online: https://www.icao.int/meetings/uas/documents/circular%20328_en.pdf (accessed on 5 April 2021).

Union Law, and in particular the right to respect for private and family life, set out in Article 7 of the Charter of Fundamental Rights of the European Union, and with the right to protection of personal data, set out in Article 8 of that Charter and in Article 16 TFEU, and regulated by Regulation (EU) 2016/679 of the European Parliament and of the Council¹³. Generally, the Regulation (EU) 2018/1139 serves for the protection of privacy in such use by setting what should be achieved. Recent Commission Delegated Regulation (EU) 2019/945¹⁴ which applies since 1 July 2020 has divided drones into classes in terms of their technical characteristics (open, specific and certified category) and lays down the requirements for the remote identification of drones, which is very important in helping to determine the operator of the drone, serving thus for more effective privacy protection in the use of drones (Puraite and Silinske 2020). However, for classes C0 and C4, which are technically simpler and therefore more accessible to the majority of people, no requirement of a direct remote identification equipment is included. In addition, Commission Implementing Regulation (EU) 2019/947 of 24 May 2019¹⁵ on the rules and procedures for the operation of unmanned aircraft, being in effect and applying since 1 July 2020, includes requirements for the implementation of three foundations of the U-space system, namely registration, geo-awareness and remote identification, which will need to be further completed. According to the Preamble of this Regulation par. 14 and 16: “Operators of unmanned aircraft should be registered where they operate an unmanned aircraft which, in case of impact, can transfer, to a human, a kinetic energy above 80 Joules or the operation of which presents risks to privacy, protection of personal data, security or the environment” . . . “Considering the risks to privacy and protection of personal data, operators of unmanned aircraft should be registered if they operate an unmanned aircraft which is equipped with a sensor able to capture personal data”. This is a clear safeguard clause but it is still questionable how alone the registration of operators would be effective for privacy issues if for classes C0 and C4, there is no requirement of a direct remote identification equipment. In addition, Article 11 of the Regulation 2019/947 states the rules for conducting an operational risk assessment while Article 18 (h) and (i) of the Regulation imposes the development of a risk based oversight system and an audit planning for certain drone operators, but it seems difficult to perceive how Article 35 GDPR¹⁶ vis a vis Article 11 and 18 of the Regulation 2019/947 could complement each other (Pagallo and Bassi 2020). To sum up, the new legislation at EU level, namely Regulations 2019/945 and 2019/947, establish registration and remote identification requirements in the use of drones, making thus a huge contribution to the effectiveness of privacy and personal data protection, but with exceptions that could possibly undermine this goal, while there are still some unclear points of the risk assessment mechanism set.

4.2. Parallel Application of International and European Union Law on the Protection of Privacy and Personal Data

Apart from the above mentioned specific legislation on remote sensing technologies, it is important to assess the parallel application of International and European Union Law on the protection of privacy and personal data when using remote sensing technologies.

Protection of privacy on international level is ruled by Article 8 of the European Convention on Human Rights (ECHR): “Everyone has the right to respect for his private and family life, his home and his correspondence”. According to Paragraph 2 of the Article 8 ECHR “There shall be no interference by a public authority with the exercise of this right

¹³ Regulation (EU) 2018/1139 of the European Parliament and of the Council of 4 July 2018 on common rules in the field of civil aviation and establishing a European Union Aviation Safety Agency, and amending Regulations (EC) No. 2111/2005, (EC) No. 1008/2008, (EU) No. 996/2010, (EU) No. 376/2014 and Directives 2014/30/EU and 2014/53/EU of the European Parliament and of the Council, and repealing Regulations (EC) No. 552/2004 and (EC) No. 216/2008 of the European Parliament and of the Council and Council Regulation (EEC) No. 3922/91 Preamble para 28.

¹⁴ Commission Delegated Regulation (EU) 2019/945 of 12 March 2019 on unmanned aircraft systems and on third-country operators of unmanned aircraft systems.

¹⁵ Commission Implementing Regulation (EU) 2019/947 of 24 May 2019 on the rules and procedures for the operation of unmanned aircraft.

¹⁶ In Article 35 GDPR data protection impact assessment is ruled in 11 paragraphs. In particular, it is ruled when and how a data protection impact assessment is conducted in Member States.

except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic wellbeing of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others". Therefore, the right to private life is not guaranteed in ECHR as an absolute right but it must be balanced against and reconciled with other legitimate interests, either private or public, while any interference with the right to privacy has to comply with the so-called "democracy test" (Mitrou 2009).

On European Union level, Article 16 of the Treaty on the Functioning of the European Union (TFEU) in accordance with Article 8 of the Charter of Fundamental Rights of the European Union, they rule together the protection of personal data. Article 7 of the Charter of Fundamental Rights of the European Union declares respect for private and family life. Furthermore, according to Article 52 (1) of the Charter of Fundamental Rights of the European Union, the principle of proportionality is introduced as a tool for balancing fundamental rights. According to the last Article, limitations on the exercise of the rights and freedoms recognized by the Charter must be necessary and appropriate.

In this sense, a limitation may be necessary if there is a need to adopt measures for the public interest objective pursued. If a limitation is proven to be strictly necessary, there must be also be assessed whether it is proportionate. Proportionality means that the advantages resulting from the limitation should outweigh the disadvantages the latter causes on the exercise of the fundamental rights at stake. To reduce disadvantages and risks to the enjoyment of the rights to privacy and data protection, it is important that limitations contain appropriate safeguards¹⁷.

Furthermore, Union Law contains since very early specialized legislation on the protection of personal data. The current basic legislative acts for the protection of personal data in the EU is GDPR¹⁸ on one hand, and Police and Criminal Justice Authorities Directive¹⁹ on the other hand.

GDPR's territorial scope according to Article 3 par. 2 b covers the processing of data (which includes collection) both from satellites and drones, as long as they collect or process data of EU residents, even if they collect or process such data from satellites under the jurisdiction and control of a non-EU country provided that processing activities are related to the monitoring of the behavior of EU residents as far as their behavior takes place within the Union. Police and Criminal Justice Authorities Directive applies to the processing of personal data by competent authorities of member states for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security. It also covers collected data both from satellites and drones, as long they are processed by competent authorities of member states.

Following the above mentioned legislation, and especially Article 52 (1) of the Charter and Article 8 (2) ECHR any limitation to the exercise of rights and freedoms recognized by the Charter must be provided for by law ("in accordance with the law"), made only if it is necessary and genuinely meets objective of general interest recognized by the Union or the need to protect the rights and freedoms of others ("in pursuit of one of the legitimate aims set out in Article 8 (2) of the ECHR and necessary in a democratic society")²⁰.

As a result, the police and other environmental authorities when using remote sensing technologies should first assure themselves that they have a valid legal basis for processing personal data. This also stems directly from Article 8 of Police and Criminal Justice

¹⁷ Handbook on European data protection law. 2018. Available online: <https://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law-2018-edition> (accessed on 5 April 2021).

¹⁸ Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

¹⁹ Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA.

²⁰ See also: Opinion 01/2015 on Privacy and Data Protection Issues relating to the Utilization of Drones. Available on line: https://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=640602 (accessed on 5 April 2021).

Authorities Directive as well as from Article 6 of GDPR. In this point, it is important to underline that Police and Criminal Justice Authorities Directive and GDPR supplement each other as they operate in different sectors but cooperate in the areas where they overlap (Pajunoja 2017). CJEU case law also identifies this relation between Police and Criminal Justice Authorities Directive and GDPR²¹. Therefore, police and the Criminal Justice Authorities Directive are applied when limitations to rights are imposed by the State for personal data collected directly by competent authorities only in order to serve their work (duty) for the prevention, investigation, detection or prosecution of environmental criminal offences. In cases when data are collected by third parties (private entities etc.) for other reasons but it happens them to be necessary also for the purposes of the prevention, investigation, detection or prosecution of environmental criminal offences, Article 23d of GDPR is applicable. Finally, Article 6e of GDPR is applicable, when administrative official authorities, such as forest services, environmental departments, environmental inspectors etc., that are authorized to protect the environment and impose administrative sanctions for law infringements, may according to a certain legal basis process personal data, for example inspect protected areas with drones.

Police and other environmental authorities when using remote sensing technologies should afterwards follow all principles stemming from Article 4 of Police and Criminal Justice Authorities Directive either from Article 5 of GDPR, namely their actions should comply with the principles of lawfulness, fairness and transparency, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality (security), and accountability. This means that data subjects must be aware of the collection and processing of their personal data and therefore data controllers have the obligation to inform them according to the relevant Articles of Police and Criminal Justice Authorities Directive or of GDPR. Especially for drones, signposts or information sheets for an event could be easily used for drone operations in fixed locations, also social media, public display areas, flashing lights, buzzers and bright colors could be envisaged. Drone operators could also publish information on their website or on dedicated platforms in order to inform constantly about the different operations that take place²². In addition, remote sensing technologies shall be used from police and other environmental authorities when the necessity and appropriateness for the specific purposes is justified. A strict assessment of the necessity and proportionality of the processed data should take place.

Furthermore, data controllers and processors, where applicable, must implement the appropriate technical and organizational measures to protect personal data from accidental or unlawful destruction according to the security principle (Article 29 of Police and Criminal Justice Authorities Directive or Article 32 of GDPR). Finally, it seems that a data protection impact assessment of Article 35 of GDPR is necessary (only when GDPR is applicable because such an impact assessment is not included in Police and Criminal Justice Authorities Directive), since remote sensing technologies, especially the use of drones, in the environmental law enforcement sector are likely to result in a high risk to the rights and freedoms of natural persons as stated above. Simultaneously, decisions that produce legal effects concerning the natural person, such as imposition of environmental administrative fines, can be based on processed remote sensing data, making a data protection impact assessment in these cases absolutely essential.

4.3. Relevant ECtHR and CJEU Case Law on Lawful Limitations of Privacy and Personal Data Protection

Under this rather complicated legislative background, finding relevant case law, seems to be more than vital for a successful interpretation of lawful limitations of privacy

²¹ C- 623/17 *Privacy International*, para 47–48.

²² WP29 apart from these also acknowledges the need for the creation of a national or cross-national information resource to enable individuals to identify the missions and operators associated with individual drones (Working Group on Data Protection in Telecommunication, Working Paper on Privacy and Aerial Surveillance, 54th meeting, Berlin, September 2013. Available online: <https://www.datenschutz-berlin.de/infotek-und-service/veroeffentlichungen/working-paper/> (accessed on 5 April 2021).

and personal data protection when using remote sensing technologies for environmental purposes. In this sense, relevant ECtHR and CJEU case law is of high priority.

A first observation is that the structure and wording of ECHR is different than that of the Charter. The Charter as already mentioned above does not use the notion of interferences with guaranteed rights, but contains a provision on limitation(s) on the exercise of the rights and freedoms recognized by the Charter. However, despite different wording, in their case law, the CJEU and the ECtHR often refer to each other's judgments, as part of the constant dialogue between the two courts to seek a harmonious interpretation of data protection rules²³.

According to the jurisprudence of ECtHR, interference is in accordance with the law if it is based on a provision of domestic law, which must be "accessible to the persons concerned and foreseeable as to its effects". Since very early the ECtHR had judged that the "notion of necessity implies that the interference corresponds to a pressing social need and, in particular, that it is proportionate to the legitimate aim pursued"²⁴. In its following jurisprudence the ECtHR considers further an interference "necessary in a democratic society" for a legitimate aim if it answers a "pressing social need" and, in particular, if it is proportionate to the legitimate aim pursued and if the reasons adduced by the national authorities to justify it are "relevant and sufficient"²⁵. More recently, the ECtHR interpreted the requirement of "necessity in a democratic society", as "including whether it is proportionate to the legitimate aims pursued, by verifying, for example, whether it is possible to achieve the aims by less restrictive means" while there is settled an obligation for domestic law for providing "adequate and effective safeguards and guarantees against abuse"²⁶.

The jurisprudence of the CJEU also recognizes the same necessity for adequate and effective safeguards and guarantees or in other words the "existence of clear and precise rules" and "minimum safeguards" to protect personal data against the risk of abuse and against any unlawful access and use of that data²⁷. The CJEU also considers that only the objective of fighting serious crime is capable of justifying restrictions in personal data protection such as data retention measures or access to data protected by Articles 7 and 8 of the Charter²⁸. However, the definition of what may be considered to be 'serious crime' is left to the discretion of the member states, since depending on the national legal system, the same offence may be penalized more or less severely. Therefore, it is finally the correlation between the seriousness of the interference and the objective pursued under certain criteria, such as the categories of data concerned and the duration of the period in respect of which access is sought, that is decisive for justifying a potential restriction²⁹.

In this sense, the CJEU often³⁰ refers directly to the principle of proportionality as the appropriate tool for properly balancing the objective of general interest against the rights at issue and underlines that exceptions that allow limitations on the protection of personal data must remain exceptions and not be transformed to the rule. Of special importance is C-73/16, *Peter Puškár* case, where the CJEU judged³¹ that the processing of personal data by the authorities of a member state for the purpose of collecting tax and combating tax fraud without the consent of the data subjects is legitimate, provided that, those authorities were invested by the national legislation with tasks carried out in the public interest and

²³ Handbook on European data protection law. 2018. Available online: <https://fra.europa.eu/en/publication/2018/handbook-european-data-protection-law-2018-edition> (accessed on 5 April 2021).

²⁴ ECHR *Leander v Sweden* No. 9248/81, 26 March 1987, para 50 and 58.

²⁵ *S. and Marper v the UK* (GC), 30562/04 & 30566/04, 4 December 2008, para 101.

²⁶ *Roman Zakharov v. Russia* (GC), 47143/06, 4 December 2015, Para 260, 236, *Szabo and Vissy v. Hungary*, 37138/14, 12 January 2016, para 57, *P.N v. Germany*, 74440/17, 11 June 2020, para 74.

²⁷ C-293/12 and C-594/12 *Digital Rights Ireland* para 54, C-203/15 and C 698/15 *Tele 2* para 109.

²⁸ C-203/15 and C 698/15 *Tele 2* para 102, C-207/16 *Ministerio Fiscal* para 56 and 57.

²⁹ C-746/18, *H. K. v. Prokuratuur* para 87–97.

³⁰ C- 623/17 *Privacy International*, para 64, 67, Joined cases C-511/18 *La Quadrature du Net and Others*, C- 512/2018 *French Data Network and Others* and C- 520/2018 *Ordre des barreaux francophones et germanophone and Others*.

³¹ C-73/16, *Peter Puškár* para 112–117.

the principle of proportionality is respected. According to the decision such processing is proportionate only if there are sufficient grounds to suspect the person concerned for the alleged crimes. The court stated in this decision that the protection of the fundamental right to respect for private life at the European Union level requires that derogations from the protection of personal data and its limitations should be carried out within the limits of what is strictly necessary. In order to prove that such limitations are carried out within the limits of what is strictly necessary the CJEU requires from the national court to ascertain that there is no other less restrictive means in order to achieve the authority's objectives.

To sum up, it stems from all previous mentioned decisions of ECtHR and CJEU that limitations of privacy and personal data protection are lawful as long as they are proportionate to the legitimate aims pursued and they are imposed with sufficient safeguards against abuse or in other words as long as they are proportionate in so far as they apply only as it is strictly necessary under clear and precise rules with sufficient guarantees of the effective protection of privacy and personal data against the risk of misuse. Finally, it is obvious that although the objective of fighting serious crimes clearly justifies restrictions of privacy or personal data in areas of prevention, investigation, detection and prosecution of criminal offences, the condition of proportionality and strong safeguards to guarantee the rights are to be the same time fulfilled.

In regards with remote sensing technologies, although no ad hoc case law concerning the balance between the right for a high level of Environmental Protection and the rights for privacy and personal data exists, the use of the previously mentioned ECtHR and CJEU case law by analogy seems more than appropriate. Consequently, remote sensing technologies can be used for environmental purposes, especially for combatting serious environmental crime, however with sufficient guarantees for the effective protection of privacy and personal data, provided that no other less restrictive means exist.

In the following section, recent developments and first "concrete" steps in Greek legislation regarding the reconciliation of remote sensing technologies with personal data and privacy protection are presented, as well as their application perspectives in environmental law, in an attempt of a primary approach. However, it must be underlined even from this early point, that the new Greek regulatory framework is limited to certain crimes, covering thus only a small part of environmental crime, that is below analyzed. Police and Criminal Justice Authorities Directive (and its harmonization national law) as well as GDPR still regulate the majority of emerging legal issues from the use of remote sensing technologies for environmental monitoring and environmental law enforcement in Greece. Nonetheless, despite the limited scope of the new legislation, its value remains of great importance since it opens the path and the dialogue for a consistent regulatory framework of remote sensing technologies in national level.

5. The Case of Greece

5.1. *The Special Features of Greece*

Greece can be considered as a most interesting case for applying remote sensing technologies for environmental purposes. This is not only due to the natural features of Greece but also due to rules of constitutional protection of the environment, of privacy and personal data constitutional protection as well as due to the recent introduction of a specific regulatory framework for the use of remote sensing technologies in public places.

5.1.1. Natural Features and Remote Sensing Technologies

When it comes to the use of remote sensing technologies, Greece seems to be an "ideal" case study. This country is characterized by its unique relief, its alpine character, the great length of its coastline, its large number of islands, and its remarkable biodiversity, with habitats and species subject to a special protection status. Therefore, remote sensing technologies have great potential when it comes to covering the needs that arise from the purpose of environmental protection by replacing human physical presence, whenever such presence is inadequate or impossible.

The use of modern technological tools for the purpose of environmental protection is different from the former know-how employed by the Greek administration, which involved the “static” use of older technologies to address special technical issues (e.g., for purposes of public works³² or for forest mapping³³), and from the more recent one concerning the attainment of objectives of a wider range (National Cadastre³⁴, forest maps-Forest Register³⁵) through modern technologies, which, however, are in these cases again used in a technocratic and mechanistic manner.

The usability of the most modern technologies, such as satellite imagery and UAVs, is nowadays examined in a ‘dynamic’ manner, i.e., for the purpose of systematically recording and using data where and when required, depending on the needs of an overall environmental protection strategy. Such a use, based on a real-time monitoring strategy, exceeds the existing administrative experience, on the one hand, and raises crucial questions about human rights and especially privacy and personal data protection, on the other hand.

5.1.2. Constitutional Protection of Conflicting Rights and the Principle of Proportionality as Counterbalance

Greek legal order has the particularity that provides a constitutional protection to the environment, and, especially to the forest environment, which is subject to a special status of enhanced constitutional protection (Article 24 par. 1 and Article 117 par. 3 of the Constitution) (Maria et al. 2020). At the same time, the rights of personal data, privacy, and personality protection are also constitutionally anchored (Articles 9, 9A, 5 of the Constitution).

Finally, any conflict between protected human rights in the Hellenic Constitution system is resolved through the implementation of the principle of proportionality (Article 25 par. 1 of the Constitution³⁶), which is the essential counterbalance³⁷. In the Greek legal order, the principle of proportionality was initially acknowledged by the Hellenic Council of State as a constitutional principle derived from the concept of State of justice³⁸, and after the constitutional revision of the year 2001, it was explicitly incorporated in Article 25 par. 1 of the Constitution.

5.2. Privacy and Data Protection in Greece

The inviolable nature of private and family life is explicitly guaranteed by Article 9 of the Constitution as well as by civil and criminal legislation, which protect these rights against infringements either by state authorities or by other citizens (Dagtoglou 1991). Moreover, the protection of privacy is further guaranteed by the Constitution through Article 19 (Confidentiality of letters, free correspondence and communication) and Article 21 (protection of family, marriage, motherhood and childhood, rights of persons with disabilities), while especially the confidentiality of letters and free correspondence and communication are supervised by the independent Communications Privacy Authority.

³² Legislative Decree 3879/1958, PD 696/1974.

³³ Law 248/1976.

³⁴ Law 4512/2018.

³⁵ Law 3889/2010.

³⁶ Article 25 par. 1 “1. The rights of the human being as an individual and as a member of the society and the principle of the welfare state rule of law are guaranteed by the State. All agents of the State shall be obliged to ensure the unhindered and effective exercise thereof. These rights also apply to the relations between individuals to which they are appropriate. Restrictions of any kind which, according to the Constitution, may be imposed upon these rights, should be provided either directly by the Constitution or by statute, should a reservation exist in the latter’s favor, and should respect the principle of proportionality”.

³⁷ About the principle of proportionality and its adoption and evolution by the different national legal orders, the European Law, the CJEU case law and the ECHR case law: see Scaccia G. Proportionality and the Balancing of Rights in the Case-law of European Courts. 2019. federalismi.it, 4/2019, Available on line: <https://www.sipotra.it/wp-content/uploads/2019/03/Proportionality-and-the-Balancing-of-Rights-in-the-Case-law-of-European-Courts.pdf> (accessed on 5 April 2021).

³⁸ Hellenic Council of State 1341/1982, 2112/1984, 2261/1984, 3682/1986.

Personal data protection, which is inextricably connected to remote sensing technologies³⁹, is established in Article 9A of the Constitution⁴⁰ and currently regulated by Law 4624/2019, through which national law has been harmonized with Directive (EU) 2016/680. Privacy and personal data are also protected through criminal law, in Chapter 22 of the Penal Code regarding “infringements of personal confidentiality and communication” (Manoledakis 2008) and through civil law in Article 57 of the Civil Code regarding the protection of personality (Alexandropoulou-Egiptiadou 2007). Personal data protection in Greece is simultaneously directly subject to GDPR regulation.

Proper implementation of the personal data protection framework is under the supervision of the independent Data Protection Authority (hereinafter DPA). In the event of conflict between the necessity of safeguarding the environment and the protection of personal data, the necessary balance shall be pursued through the implementation of the principle of proportionality. In this sense, DPA in its Opinion 2/2010 considers that restrictions in personal data protection for the purpose of protecting the environment (as a whole, not only with regard to environmental crime), which is an explicit constitutional provision, are legitimate, as long as requirements set by the principle of proportionality (necessity, appropriateness, *stricto sensu* proportionality) are met.

5.3. *The National Implementation of the Principle of Proportionality*

5.3.1. The National Legal Framework on the Principle of Proportionality

Although Article 25 par. 1 of the Constitution establishes the principle of proportionality horizontally, namely in all cases of individual rights’ restrictions, without any further distinctions or clarifications, the implementation of the principle itself is related to the particular characteristics of each restricted right and its specific legal frame. As foresaid, the protection of personal data is ensured by specific legislation, at international, EU and national level and the proper implementation of this legislation is supervised by DPA. Any derogation to the protection of personal data is subject to special strict rules, because personal data are connected to elements of human personality and in particular the private sphere of the individual. Therefore, collection and procession of such data is permitted only exceptionally, when and to the extent necessary to serve another legitimate interest, in accordance with the principle of proportionality⁴¹.

Particularly in the monitoring technologies context, DPA issued the Directive No. 1/2011 regarding the use of video surveillance systems. Article 5 of this Directive, entitled “the principle of proportionality”, provides that the lawfulness of personal data procession is examined with regard to the legitimate aim pursued as well as in accordance with the principle of proportionality. Video surveillance systems must be thus appropriate and necessary in relation to the aim pursued. This aim should the same time not be possible to be achieved by means equally effective but less restrictive for individual rights.

With regards to environmental protection, the principle of proportionality intervenes with an ecological role, allowing the restriction of other rights for the sake of environmental protection, and preventing any disproportionate infringement of the environment in the course of pursuing other lawful purposes⁴². Furthermore, it ensures the protection of other public or private interests against an intensive implementation of the precautionary

³⁹ The reason for the creation of a special legal framework for personal data protection lies on the special nature of the information produced by modern technologies, which may relate to certain individuals as well as important aspects of their identity (Wagner De Cew 2004; Solove 2003; Akrivopoulou 2011).

⁴⁰ Article 9A: All persons have the right to be protected from the collection, processing and use, especially by electronic means, of their personal data, as specified by law. The protection of personal data is ensured by an independent authority, which is constituted and operates as specified by law.

⁴¹ DPA, Opinion 4/2020, Decision 31/2019.

⁴² Thus, when examining compliance of distortion of forest vegetation with the Constitution, while pursuing a lawful purpose, the protection of forest vegetation must be weighed against the objective pursued, and it must be examined whether the specific goal can be achieved by other means (Hellenic Council of State 293/2009, Perivallon and Dikeo (In Greek) 2009, p. 494, Hellenic Council of State 2763/2006, Perivallon and Dikeo (In Greek) 2007, p. 70), since even if the change of the forest form is deemed to be permitted, it must be implemented with the “least possible loss of forest wealth” (Hellenic Council of State 3816/2010, Perivallon and Dikeo (In Greek) 2011, p. 123), and only to the “absolutely necessary extent” (Hellenic Council of State 2972/2010).

principle, which would systematically exclude the protection of other rights in the name of environmental protection, as well as the avoidance of excessive sanctions in case of violation of environmental protection measures⁴³ (Veinla 2004; Thomas 2000; McNelis 2000; Siouti 2018; Nikolopoulos 2000).

In particular, with regard to environmental crimes, the Greek environmental criminal laws, and especially, both Law 4042/2012⁴⁴ transposing Directive 2008/98/EC into the Greek legislation, and special statutes⁴⁵ respect the principle of proportionality, aiming at the implementation of preventive, effective, and proportionate sanctions, which will safeguard environmental protection more effectively.

5.3.2. The National Case Law on the Principle of Proportionality

Greek case law on the principle of proportionality is quite rich. According to national case-law, no right is absolute, not even the constitutional ones, therefore even a constitutional right, such as the right to personal data protection, may be restricted for reasons of public interest, such as the protection of other constitutional rights, in accordance with the criteria imposed by the principle of proportionality⁴⁶.

Particularly, in the monitoring technologies context, the Council of State considers in line with DPA's guidelines, that personal data may only be lawfully taken and processed when a legal interest is to be satisfied, provided that this legal interest obviously outweighs the rights and interests of the personal data subject and only if the legal order does not provide any other way for satisfying the specific legal interest⁴⁷.

Individual rights' restrictions for environmental protection is a special case of implementation of the principle of proportionality particularly important for national case law. Due to the paramount importance of environmental protection, due to environmental degradation throughout the planet and natural phenomena described as "climate change" as well as due to the need for decisive measures to ensure the effective protection of the environment, measures restricting other rights that are considered proportionate to this purpose may be very intensive, reaching even "the core" of restricted rights. In this sense, the substantial deprivation of the use of a property by its owner for environmental purposes, may be considered lawful, but the same time may lead to lawful compensation claim by the owner in proportion to the imposed deprivation⁴⁸. Similarly, an absolute prohibition of hunting in an area of the Natura 2000 network, as long as there is a need for such a strict prohibition as an appropriate measure to protect wildlife in that area, is in line with the principle of proportionality⁴⁹. Moreover, the Hellenic Supreme Court applies the principle of proportionality in order to resolve the question of procedural use, before civil and criminal courts, of evidence obtained through illegal means, despite Article 19 par. 3 of the Constitution which explicitly prohibits the use of illegal evidence. According to national case law, securing the exercise of the right to judicial protection of a party (Article 20 par. 1 of the Constitution) consists a legal reason for the use of evidence obtained through illegal means in accordance with the principle of proportionality, i.e., if the data collected are absolutely necessary and appropriate for the recognition, exercise or defense of a right before the court, to the extent absolutely necessary and insofar as this purpose cannot be achieved by other less restrictive means⁵⁰.

⁴³ Hellenic Council of State 1393/2016, which ruled that in determining the environmental fine, while determining the unified fine, the principle of proportionality is applied, through the co-assessment of the elements determining and restricting the amount of the fine, which are provided for in the substantive provisions of the environmental laws.

⁴⁴ Government Gazette, Series I, No. 24/ 2012.

⁴⁵ e.g., in accordance with article 94 §§ 1 and 8a' of law 4495/2017 for administrative and criminal sanctions in case of illegal constructions, it is considered that during the measurement of the imposed penalty, the value of the illegal construction and the degree of environmental degradation are to be taken into account.

⁴⁶ Hellenic Supreme Court (Plen. Sess.) 1/2017, Hellenic Council of State 1616/2012, 2254/2005.

⁴⁷ Hellenic Council of State 265/2017, 2254/2005.

⁴⁸ Hellenic Council of State 488/2018, 2428/2016, 2133/2016, 2601/2005.

⁴⁹ Hellenic Council of State 875, 876/2019.

⁵⁰ Hellenic Supreme Court (Plen. Sess.) 1/2017, Hellenic Supreme Court 901/2019, 653/2013.

5.4. *The Establishment of a Modern Legal Framework*

In view of the aforementioned parameters, and in the light of the CJEU case law, the current EU laws (GDPR, Directive 2016/680) and the opinions and guidelines of the national Independent Data Protection Authority) and pursuant to Law 3917/2011 (regarding the use of surveillance systems with sound and picture recording in public places), innovative legislation on the use of monitoring technologies in public places has been recently established in Greece, via the Presidential Decree 75/2020⁵¹ (hereinafter PD). The PD 75/2020 does not provide for a general monitoring policy or a specific policy for environmental purposes, it only provides rules for the use of such technologies for crime prevention and repression and for traffic management. However, these provisions despite not aiming at the special regulation of the use of monitoring technologies for environmental purposes, contain, inter alia, rules applying on environmental crime prevention and repression. Therefore, even though the scope of the new legislation may be limited, it is important that these provisions, reflect all current European and national trends and needs regarding the exploitation of remote sensing technologies. Therefore, the analysis of these new specific rules can be the axis for the establishment of an integrated monitoring national legal framework for environmental purposes.

In this point, it must be noted that PD 75/2020 is a very recent law and therefore no related national case law has been produced yet, so its present analysis is only theoretical and cannot be based to any case law interpretation.

5.4.1. Overview of the Provisions of the Presidential Decree 75/2020

PD 75/2020 governs all the surveillance systems installed and operating at public spaces, provided that they process personal data, regardless of their technical specifications (Articles 1 and 2).

The restrictively designated public authorities that are competent for the prevention, investigation, detection, or prosecution of crimes, or the enforcement of criminal sanctions, namely the Hellenic Police, the Hellenic Fire Service, and the Hellenic Coast Guard, are considered as data controllers (Article 4).

The installation and operation of surveillance systems in public spaces is permitted only to the extent necessary, and when the objectives pursued cannot be achieved equally effectively using less restrictive means, in a specific place and for a specific period of time, following a reasoned decision of the competent authority. This decision has a validity term of no longer than three years, is subject to periodical evaluation and is issued following the conduct of an impact assessment study. Finally, it is promptly sent to the competent public prosecutor for district court judges. In particular, with regard to crime prevention or repression, it is required that there is adequate evidence that the offences subject to the PD were committed (Articles 5 and 12).

The collection and processing of sound data is only exceptionally allowed, following a specifically reasoned decision of the data controller, which is approved by the competent public prosecutor, for the purpose of detecting and recognizing the persons involved in specific punishable acts, including forest arson by negligence (Article 7).

Strict rules have also been established concerning the retention period, the complete and automatic destruction of the data without the right to retrieve them, and the anonymization of the data kept exceptionally for educational purposes (Article 8), the data recipients, and the safe and unimpeachable transfer of data (Article 9), and the rights of the data subjects, especially the right of information (Article 10).

Furthermore, organisational and technical safety measures are imposed with regard to the technical specifications and the operation of the surveillance systems, for the purpose of minimizing the impact on the right to personal data protection, in accordance with the accepted international standards, as well as the minimum safety measures (users' training,

⁵¹ Government Gazette, Series I, No. 173/ 10 September 2020.

creation of separate accounts, and user authentication, data encryption, etc.) are explicitly provided for (Article 11).

Harmonisation of the Presidential Decree 75/2020 with the GDPR and the Police Directive

PD 75/2020 makes explicit reference to the general application of Regulation (EU) 2016/679 (GDPR) and Directive (EU) 2016/680 (Police and Criminal Justice Authorities Directive), but it further specifies special rules, which are harmonised with the principles derived from Article 5 of GDPR and Article 4 of the Directive, as analysed above.

Firstly, as far as the principles of lawfulness, fairness, and purpose limitation are concerned, the PD limits the collection and processing of personal data exclusively to the purposes restrictively specified by the authorising legal provision of Article 14 of Law 3917/2011 (Articles 1 and 3). Such a procession is subject to a decision provided by the competent public authority (Article 12) when the above objectives cannot be achieved equally effectively using less restrictive means, and, in particular, with regard to crime prevention or repression, provided that there is adequate evidence that the crime was committed, and, in any case, provided that the collection and processing is necessary (Articles 5 and 6).

Secondly, referring to the implementation of the principle of transparency, according to the PD, data collection and processing is contingent upon the prior notification to the public prosecutor, the gathering organiser, the data subjects, and the public, as appropriate, with any expedient means, and primarily with the means explicitly specified in its provisions (Articles 6 and 10). The foregoing obligation to notify the public prosecutor and the public also includes the notification of the decision of the competent public authority on the operation of a surveillance system (Article 12). Data subjects always have the right to request and receive information about the data concerning them and any recipients of the processing (Article 10 par. 3).

Thirdly, data minimisation principle is clearly reflected in the PD, which limits the installation and operation of surveillance systems to the specific necessary space, and prohibits expansion thereof to a broader area and collection of data from non-public spaces or homes, image focus is allowed only for the detection of crimes (Article 5), while sound data collection and processing is in principle prohibited (Article 7).

Furthermore, specific provisions have been set in order to ensure storage limitation. According to the PD, the maximum data retention period is, in principle, 15 days, with certain exceptions that serve the needs of the criminal court procedure, while specifically in the case of public gatherings, the maximum data retention period is 48 hours. In addition, integrity and confidentiality (security) are pursued through specific provisions in the PD. The automatic destruction of personal data is provided in a manner that precludes retrieval thereof, and in the case of their exceptional retention for educational purposes. The PD includes also provisions for data anonymization and compliance with the confidentiality obligation (Articles 6 and 8), and for ensuring, using suitable technical means, not only secure transfer of data, but also that the transferred data cannot possibly be distorted in an unperceivable manner (Article 9). Moreover, the data controller is subject to all the necessary organisational and technical security measures (Article 11), which are aligned with Article 25 of the Regulation, or Article 20 of the Directive.

Finally, the designation of the public authorities acting as data controllers, the establishment of the legislative framework of their liability (Article 4), and the establishment of special requirements for the issuance of a decision on the installation and operation of surveillance systems (Article 12) integrate the principle of accountability in the PD.

Critical Assessment of the Provisions of Presidential Decree 75/2020

The draft PD 75/2020 was submitted to the DPA, in accordance with the law, which issued its Opinion No. 3/2020, where, presenting an analysis of the Greek and European legal framework on personal data protection, and having particularly focused to ECtHR and CJEU case law, it stressed that certain provisions needed to be amended in order

to be compatible with the International and European Union Law. Compliant with the recommendations of the DPA, the final text of the PD constitutes a strict set of rules that integrate the principles of modern protection of personal data at an international and EU level.

Although the principle of proportionality is not explicitly mentioned at any point in the text of PD 75/2020, Article 5, which sets the conditions and criteria for the installation and operation of surveillance systems, introduces the special condition of implementation of the principle of necessity and the principle of appropriateness, as manifestations of the principle of proportionality. In addition, Article 8, with respect to the retention period and the destruction of data, also follows the recommendations of the DPA regarding the respect of the principle of proportionality⁵². Besides, the authorizing legal provision of PD 75/2020 explicitly stipulates that this PD should aim at setting the criteria for complying with the principle of proportionality⁵³.

It is also to underline that Articles 11 (Organizational and Technical Security Measures) and 12 (Decision on the Installation and Operation of Surveillance Systems) provide not only for the conduct of an impact assessment study at the stage of personal data processing, but also for the conduct of an impact assessment study concerning the installation, commissioning, and procurement of the surveillance systems, the software, and the additional equipment in general. Therefore, impact assessment accompanies the surveillance system and any accompanying item or equipment already from the stage of procurement thereof until installation, operation, and processing of the collected material. Such a provision is of great importance, since impact assessment at the time of the determination of the means for processing is essential for data protection by design and by default. In this sense, legal framework set by the PD not only follows in a timeliest manner current European trends on personal data protection but also forms the necessary legal background for any other future laws regarding the use of remote sensing technologies, including possible specialized legislation for environmental protection.

However, there are some points in which PD 75/2020 did not fully comply with the recommendations of the DPA. Thus, contrary to DPA's recommendations, Article 5 (installation and operation of surveillance systems) did not encompass any provision specifying clearly the criteria based on which surveillance in a specific space is evaluated as necessary, or the precise procedural requirements and the necessitated guarantees of supervision and control of the relevant measure. Similarly, Article 9 (data recipients) did not incorporate DPA's recommendation for a procedure of control and supervision by an independent authority in the case of transfer of data (except for the cases of transfer to administrative authorities acting as third parties where the transfer is approved by the public prosecutor). Finally, in Article 10 (Rights of data subjects), DPA's recommendations for special provisions for each surveillance system, and for persons who have lost their eyesight, so that the obligation of informing data subject could be most successfully achieved, were not taken into account.

Moreover, even at the points where the PD conforms to the DPA's recommendations, it is not certain that the final wording of the provisions is always correct. Thus, despite adding to Article 8 (Data retention period and destruction) the criteria on which the justified suspicions for preparing or committing in the future offences are assessed, pursuant to the Authority's recommendations, as a reason for exceptional extension of the data retention period, the criteria encompass the wording "any kind of relevant information"⁵⁴, which is rather ambiguous, and possibly leaves room for unauthorized extension of the data retention period. These shortcomings are indicative of the necessary adjustments for the lawful use of remote sensing technologies for all purposes and especially for environmental purposes.

⁵² DPA, Opinion 3/2020, Available online: https://www.dpa.gr/sites/default/files/2020-07/gnomodotisi%203_2020.pdf (accessed on 5 April 2021).

⁵³ Law 3917/2011, Article 14 (4).

⁵⁴ Article 8 of the PD: "... justified suspicions for preparing or committing in the future the above criminal acts may stem from witnesses' testimonies or from any kind of relevant information".

5.4.2. Application of PD 75/2020 in Environmental Crimes

As already mentioned, PD 75/2020 does not specifically regulate the use of surveillance systems for the prevention and repression of environmental crime, however, its purpose, as described in Article 3, includes a large number of environmental offences, referring to the relevant provisions of the Criminal Code.

In particular, the scope of PD 75/2020 encompasses:

- organized environmental crime, in particular, felonies and misdemeanors committed for the purpose of pursuing financial gain (Article 187 of the Criminal Code);
- assault by a large crowd against environmental goods (Article 189 of the Criminal Code);
- arson in forests, forest and reforestable areas (Article 265 of the Criminal Code);
- flooding (Article 265 of the Criminal Code);
- destruction or damage to works or installations intended for protection from natural disasters (Article 273 of the Criminal Code);
- poisoning of sources, wells, and water tanks (Article 279 of the Criminal Code);
- destruction or damage to public environmental goods (Article 378 of the Criminal Code).

Therefore, PD 75/2020 offers, to a large extent, the possibility of using modern remote sensing technologies for environmental protection, since its scope primarily involves the protection of public environmental goods, including public forests, coastal and riparian zones, rivers, large lakes, sea, as well as the protection of all forest and reforestable ecosystems from arson. Furthermore, such technologies can be used both for preventive and for repressive protection of the above areas and elements (Article 3a).

5.5. Concluding Remarks for Greek Legislation and Future Perspectives in Environmental Law

Although the regulatory framework of PD 75/2020 includes many and significant offences of environmental relevance in its scope, it is found to be inadequate for facing emerging legal issues from the use of remote sensing technologies for environmental monitoring and environmental law enforcement. This is because it not only addresses certain environmental offences but also addresses them in a fragmentary manner. From this point, it even fails to regulate effectively issues related exclusively to environmental crime. It is a telling sign that Article 4 does not designate the competent environmental protection authorities as data controllers. Similarly, the provisions of Article 10 on information to the data subjects fail to take into account and to respond to the particularity of supervision of broad and freely accessible areas such as forest and coastal zones. In addition to this, the scope of PD 75/2020 is limited to the use of remote sensing technologies in public spaces, leaving private environmental goods (e.g., private forests, lakes, private coastal areas) unprotected.

Thus, it is recommended that a special legislative and regulatory framework is established, which will adjust the technical features offered by modern remote sensing technologies not only to the preventive and repressive treatment of environmental crime in its whole but also to their use in environmental monitoring and all aspects of environmental law enforcement. Lessons learned from the regulatory framework of PD 75/2020 for the protection of the infringed human rights, in accordance with the principle of proportionality, which calls for a special weighting based on the particular features of each environmental good, the special enhanced constitutional protection of forest ecosystems, and human rights' risks emerging from the use of technical means for environmental surveillance, should be taken into account, when forming such a special framework.

6. Conclusions

Remote sensing technologies provide tools for gathering data, which are extremely useful for ensuring a high level of environmental protection and the improvement of the quality of the environment. However, the same time they raise new difficult challenges,

such as their interference with the rights of privacy and personal data, which are also protected fundamental rights.

It stems from existing legislation and case law interpretation that remote sensing technologies in the European Union can be used for environmental purposes, especially for combatting serious environmental crime, however with sufficient guarantees for the effective protection of privacy and personal data, provided that no other less restrictive means exist.

The case study of Greece clearly shows that despite recent developments in the field of surveillance systems' legislation, there is still a gap in special legislative and regulatory framework which will envisage the lawful use of remote sensing technologies in the environmental sector.

However, the path has been opened and the great demand for a wider use of remote sensing technologies for supporting environmental law enforcement, for combatting environmental crime and for collecting environmental monitoring data will inevitably lead to a consistent regulatory framework in European and national level.

Author Contributions: Conceptualization, M.M. and E.-A.M.; Funding acquisition, M.M., A.P., E.-A.M. and L.M.; Investigation, M.M. and A.P.; Project administration, E.-A.M.; Supervision, E.-A.M. and L.M.; Writing—original draft, M.M. and A.P.; Writing—review & editing, M.M., A.P., E.-A.M. and L.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme “Human Resources Development, Education and Lifelong Learning 2014-2020” in the context of the project “Legal issues derived from the use of monitoring and earth observation technologies to ensure environmental compliance in the Hellenic legal order-HELLASNOMOSAT” (grant number MIS 5047355).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Akrivopoulou, Hristina. 2011. The right to personal data protection through the lens of the right to privacy. *Theoria kai Praxi Diiketikou Dikeou* 7: 2. (In Greek).
- Alexandropoulou-Egiptiadou, Evgenia. 2007. *Personal Data*. Athens-Komotini: Ant. N. Sakkoulas, p. 115. (In Greek)
- Almar, Rafael, Erwin W. J. Bergsma, Philippe Maisongrande, and Luis Pedro Melo de Almeida. 2019. Wave-derived coastal bathymetry from satellite video imagery: A showcase with Pleiades persistent mode. *Remote Sensing of Environment* 231: 111263. [CrossRef]
- Anderson, Chris. 2012. Here Come the Drones! August Issue: *Wired Magazine*. Available online: <https://www.wired.co.uk/article/here-come-the-drones> (accessed on 5 April 2021).
- Billiet, Carole. 2012. Satellite Images as Evidence for Environmental Crime in Europe: A Judge's Perspective. In *Evidence from Earth Observation Satellites Emerging Legal Issues*. Edited by Leung Denise and Purdy Ray. Leiden: Brill, pp. 321–55.
- Coffer, M. Megan. 2020. Balancing Privacy Rights and the Production of High Quality Satellite Imagery. *Environmental Science and Technology* 54: 6453–55. [CrossRef] [PubMed]
- Collingridge, David. 1980. *The Social Control of Technology*. Birmingham: The University of Aston, Technology Policy Unit, New York: St. Martin's Press.
- Dagtoglou, Prodromos. 1991. *Individual Rights*. Athens-Komotini: Sakkoulas, vol. 1, p. 323. (In Greek)
- di Vimercati, Sabrina De Capitani, Angelo Genovese, Giovanni Livraga, Vincenzo Piuri, and Fabio Scotti. 2013. Privacy and Security in Environmental Monitoring Systems: Issues and Solutions. In *Computer and Information Security Handbook*. Edited by John R. Vacca. Burlington: Morgan Kaufmann, pp. 835–53.
- Doldirina, Catherine. 2014. Privacy, earth observations and legal ways to reconcile the two. Paper presented at the 65th International Astronautical Congress, Toronto, ON, Canada, September 29–October 3.
- Dunk, Frans G. 2013. Outer Space Law Principles and Privacy. In *Evidence from Earth Observation Satellites: Emerging Legal Issues*. Edited by Leung Denise and Purdy Ray. Leiden: Brill, pp. 243–58.
- Finn, L. Rachel, and David Wright. 2016. Privacy, data protection and ethics for civil drone practice: A survey of industry, regulators and civil society organisations. *Computer Law & Security Review* 32: 577–86.
- Harris, Ray, and Ingo Baumann. 2021. Satellite Earth Observation and National Data Regulation. *Space Policy* 56: 101422. [CrossRef]

- Kuriyama, Ikuko. 2005. Supporting multilateral environmental agreement with satellite Earth observation. *Space Policy* 21: 151–60. [CrossRef]
- Laituri, Melinda. 2018. Satellite Imagery Is Revolutionizing the World. But Should We Always Trust What We See? Available online: <https://theconversation.com/satellite-imagery-is-revolutionizing-the-world-but-should-we-always-trust-what-we-see-95201> (accessed on 5 April 2021).
- Lucács, Adrienn. 2016. What Is Privacy? The History and Definition of Privacy. Available online: <https://www.semanticscholar.org/paper/What-is-Privacy-The-History-and-Definition-ofAdrienn/430bfacbab89c0033b6dcccddc18ba9bbc02c5f> (accessed on 5 May 2021).
- Manoledakis, Ioannis. 2008. Penal protection of personality. *Piniki Dikeosini*, 334. (In Greek)
- Maria, Efpraxia-Aithra, Athanasios Papathanasopoulos, and Maria Maniadaki. 2020. Natura 2000 Forest areas in Greece: A national implementation review. *Zeitschrift für Europäisches Umwelt-und Planungsrecht (EurUP)* 18: 68–85.
- McNelis, Natalie. 2000. EU Communication on the Precautionary Principle. *Journal of International Economic Law* 3: 545. [CrossRef]
- Mertikas, P. Stelios, Panagiotis Partisinelos, Constantine Mavrocordatos, and Nikolai A. Maximenko. 2021. Environmental applications of remote sensing. In *Pollution Assessment for Sustainable Practices in Applied Sciences and Engineering*. Edited by Abdel-Mohsen O. Mohamed, Evan K. Paleologos and Fares Howari. Oxford: Butterworth-Heinemann, pp. 107–163. [CrossRef]
- Mitrou, Lilian. 2009. The Commodification of the Individual in the Internet Era: Informational Self-determination or “Self-alienation”? Paper presented at the 8th International Conference Computer Ethics: Philosophical Enquiry, Corfu, Greece, June 26–28.
- Nikolopoulos, Takis. 2000. The Principles Of Community Environmental Law. Available online: <https://nomosphysics.org.gr/7034/oiarxes-tou-koinotikou-dikaiou-periballontos-noembrios-2000/> (accessed on 5 May 2021). (In Greek).
- Pagallo, Ugo, and Eleonora Bassi. 2020. The Governance of Unmanned Aircraft Systems (UAS): Aviation Law, Human rights, and the Free Movement of Data in the EU. *Minds and Machines* 30: 439–55. [CrossRef] [PubMed]
- Pajunoja, J. Lauri. 2017. The Data Protection Directive on Police Matters 2016/680 Protects Privacy-The Evolution of EU’s Data Protection Law and Its Compatibility with the Right to Privacy. Master’s thesis, University of Helsinki, Helsinki, Finland. Available online: <https://core.ac.uk/download/pdf/84363684.pdf> (accessed on 5 April 2021).
- Patias, Petros, Georgios Mallinis, Vassilios Tsioukas, Charalampos Georgiadis, Dimitrios Kaimaris, Maria Tassopoulou, Natalia Verde, Mario Dohr, and Michael Riffler. 2020. Earth observations as a tool for detecting and monitoring potential environmental violations and policy implementation. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 43: 1491–96.
- Puraite, Aurelija, and Neringa Silinske. 2020. Privacy Protection in the New EU Regulations on the use of unmanned aerial systems. *Public Security and Public Order* 24: 173–83. [CrossRef]
- Purdy, Ray. 2011. Attitudes of UK and Australian farmers towards monitoring activity with satellite technologies: Lessons to be learnt. *Space Policy* 27: 202–12. [CrossRef]
- Sabins, F. Floyd. 1978. *Remote Sensing: Principles and Interpretation*. San Francisco: W. H. Freeman.
- Sandbrook, Chris. 2015. The social implications of using drones for biodiversity conservation. *Ambio* 44: S636–47.
- Santos, Cristiana, and Lucien Rapp. 2019. Satellite Imagery, Very High-Resolution and Processing-Intensive Image Analysis: Potential Risks under the GDPR. *Air and Space Law* 44: 275–96.
- Siouti, Glikeria. 2018. *Manual of Environmental Law*. Thessaloniki: Sakkoulas, p. 58. (In Greek)
- Solove, J. Daniel. 2003. *Information Privacy Law*. New York: Aspen Publishers, pp. 47–51.
- Thomas, Robert. 2000. *Legitimate Expectations and Proportionality in Administrative Law*. Oxford: Hart Publishing, p. 78.
- Veinla, Hannes. 2004. Determination of the level of Environmental Protection and the Proportionality of environmental measures in Community Law. *Juridica International* 9: 89. Available online: https://www.juridicainternational.eu/public/pdf/ji_2004_1_89.pdf (accessed on 5 April 2021).
- Wagner De Cew, Judith. 2004. Privacy and Policy for Genetic Research. *Ethics and Information Technology* 6: 5–14. [CrossRef]
- Warren, D. Samuel, and Luis D. Brandeis. 1890. The right to privacy. *Harvard Law Review* 4: 193–220. [CrossRef]
- Watts, C. Adam, Vincent G. Ambrosia, and Everett A. Hinkley. 2012. Unmanned aircraft systems in remote sensing and scientific research: Classification and considerations of use. *Remote Sensing* 4: 1671–92. [CrossRef]

Article

Artificial Intelligence Systems-Aided News and Copyright: Assessing Legal Implications for Journalism Practices

Javier Díaz-Noci 

Communication Department, Pompeu Fabra University, 08018 Barcelona, Spain; javier.diaz@upf.edu;
Tel.: +34-93-542-1220

Received: 29 March 2020; Accepted: 6 May 2020; Published: 8 May 2020

Abstract: Automated news, or artificial intelligence systems (AIS)-aided production of news items, has been developed from 2010 onwards. It comprises a variety of practices in which the use of data, software and human intervention is involved in diverse degrees. This can affect the application of intellectual property and copyright law in many ways. Using comparative legal methods, we examine the implications of them for some legal categories, such as authorship (and hence required originality) and types of works, namely collaborative, derivative and, most especially, collective works. *Sui generis* and neighboring rights are also considered for examination as being applicable to AIS-aided news outputs. Our main conclusion is that the economics intellectual property rights are guaranteed in any case through collective works. We propose a shorter term of duration before entering public domain. Still, there is a place for more authorial, personal rights. It shows, however, more difficulty when coming to moral rights, especially in Common Law countries.

Keywords: automated news; intellectual property; copyright law

1. Introduction

In recent times, the automated production of news items has joined the traditional human-only creation of information. It has adopted many forms, at least from 2014 and 2015. This has been called algorithm journalism [1,2] or robot journalism, but, in general terms, it is mentioned as automated journalism [3,4], and it is the type of news item created with the aid of autonomous intelligent systems (AIS) [5–7]. The technology “shall eventually lead to autonomous technology that can perceive, learn, decide and create without any human intervention” [8]. The question of a gradual (even if partial, at least for now) substitution of human by machines is in the background. The generation of new professional skills and profiles is a trend, as well, that will continue in the coming years, as stated, for instance, by [9].

This is clearly beyond the traditional use of some tools to produce copyrightable works, for instance, photography [8] (p. 321) or word processors. To the extent journalists and the media use software to help human to produce news, this is to be considered a protectable work. When machines are able to produce news for themselves, with no human help—except for the design of the software itself—then we are dealing with a different question. This kind of tool, which most likely will be improved in the near future, poses some questions on intellectual property, which is the topic of this paper.

Intellectual creation is a human activity in which we must include the varied forms of journalistic works, from simple news to more elaborated features and articles. These fall within the works protected by copyright and, in general terms, intellectual property law. We have to make an important clarification of the terms to be used relating to intellectual property. While in the Common Law legal tradition, it means copyright (which is called by legal doctrine in the other great legal tradition of

the world, the so-called Civil Law tradition, authors' rights) and designs, patents and trademarks, in the Civil Law tradition, however, intellectual property is almost synonymous to copyright-authors' right law, and the rest is considered as being included within the industrial property denomination. This may add some difficulty when comparing both legal traditions from a transnational perspective, but it may, on the other hand, help in distinguishing the many implications of the arrival of the different outputs produced using artificial intelligence on news reporting.

We have mentioned this aspect in some previous works [10] altogether with some other issues of copyright law applied to the media and news reporting. It is our intention to explore the specific aspects involved in automated news, or news produced with the aid of artificial intelligence. As a starting point: to the extent that some human, substantial intervention is needed before the news item is delivered to the public, copyright law may be applicable to the protection of such works. When human intervention is minimal, unnecessary or accessorial, then we are dealing with a different legal nature and protection. The degree of originality—understood as the application of intellectual human skills in order to obtain a work—is a central requirement in copyright law, and even more so in the Civil Law legal tradition, in which the author is at the very core of its conception. However, and even if some national legal systems, such as the Spanish one (Art. 5.1 of the Spanish Copyright Act, TRLPI 1/1996, which states that only natural persons can be considered creators of literary, artistic or scientific works), insist in considering that the only possible author needs to be human, there are some other layers of rights applicable to other agents of intellectual creation. This is the case of the collective works, singularly important in journalism, since media outputs are considered precisely like that: a collection of works commissioned to (usually hired) journalists offered to the public as a bunch produced under the investment and coordination of a corporate entity, instead of a natural person. Those corporate entities have several rights under copyright law, and ultimately they have fought, and gained to a great extent, a legal battle in the European Union to get an exclusive exploitation right to confront the great news aggregators, like Google News [11].

2. What is Automated Journalism? A Typology of AIS-Aided Created News

The already short development of AIS-aided, or automated news, appears, generally speaking, as a practice of presenting or producing news items out of previously gathered and structured data, normally using templates and applying some more or less complex algorithms. We follow in this respect the definition provided by A. Graefe in 2016: "It is the process of using software or algorithms to automatically generate news stories without human intervention—after the initial programming of the algorithm, of course. Thus, once the algorithm is developed, it allows for automating each step of the news production process, from the collection and analysis of data, to the actual creation and publication of news" [7] (p. 9).

The history of AIS-aided news production is scarcely ten years old. The (initially British, now global) newspaper *The Guardian* started with software in 2010 to produce some news on sport statistics and graphics, and in 2014 experimented in a similar way with Guarbot, a tool to produce news on financial information. The real effective experimentation can be traced back to 2014, when a journalist then hired by *The Los Angeles Times*, Ken Schwencke, designed an algorithm to produce some news on a low-intensity earthquake that happened that year, based on data from the United States Geological Survey service. One year later, the main French daily newspaper, *Le Monde*, used another algorithm designed by Data2Content and Syllabs companies to produce some news on election results, using numeric data as well. One year later, in 2015, a Chinese tool, Dreamwriter, was created by a videogames company, Tencent, to produce news on consumer prices that was 916 words long in just one minute, with apparently no mistakes. From then onwards, many other ones have appeared: Heliograf (used by *The Washington Post* since 2016), Quill, Soccerbot, Wordsmith by Automate Insights, used since 2014 by the Associated Press agency, Recount, StatsMonkey, Media Brain, Kognetics and RADAR are some of them [2].

RADAR is a rather interesting case. It was created in 2017 by a news agency, the Press Association, which in three months produced, using this software, more than 50,000 items. The software was developed by Urbs Media and financed with EUR 150 M by the Google Digital News Initiative Innovation Fund. It uses open access datasets on topics such as transport, education, health, crime and education, and it is able to produce several versions of every item adapted to the necessities of their clients. Behind the RADAR working flow, there was a team of six journalists who identified interesting topics and conducted the production of automated news.

Some topics and sections seem more appropriate to use AIS-aided newswriting. Finances, election results and especially sports coverage have appeared as the most widespread topics in which algorithms are used to produce news. Media have sometimes used chatbots to communicate with users, and these tools are able to write their own sentences, based on patterns and on topics and terms recognition. In 2017, the Innovation Lab of the Spanish native-digital journal *El Confidencial* created a software named AnaFut which creates football chronicles of the lower categories. Sport coverage also combines documentation and bots [12], in the case of BeSoccer. Most of those systems use as a primary source data provided by official institutions: the Spanish public broadcast service, Radio Televisión Española, decided in 2020 to experiment with data extracted from the Spanish Football Federation to offer short news on results of the lower leagues, “interpreting them and presenting a text in natural language, related to the selected event with no personal intervention”, using HTML format and as a mere news, and not penalizing the SEO positioning of the source itself. Similar systems were used by *The Washington Post* to cover the Olympic Games in 2016.

Actually, according to Beckett [13], artificial intelligence systems can help journalists and the media in three phases: gathering, production and distribution. This can lead, however, to a wide variety of results and, which is our point from a legal perspective, intervention by human journalists. These results can be:

1) Mash-up news items can be produced, aggregating several previously published works. This results, using legal terminology, in derivative works which are obliged to mention the works and authors on which the new items are based. As an example, this is the case of Adrian Holovaty’s mashed up infographics produced for the website *Chicagocrime.org*, [14,15]. Producing intelligent infographics is also the method developed by Intelygenz and Prodigioso Volcán in Spain from 2018 onwards (see <http://losdelvolcan.com/grafia/web/>): while the journalist creates his or her item, a machine-learning comprehension software scans the words and, relating them all, it creates some graphics with no human intervention—so it can be defined, in legal terms, as a derivative work—to complement them, so as to produce, once again using legal terms, a collaborative work.

2) Automatization of processes can help journalists to provide more context, data and even links (internal or external) to their items. It is common practice to search in the documentation service of the newsroom to find related news to be used and linked. Contextualization of news seems to be exclusive of human journalists, although interfaces and search engines can help in extracting that which is needed from massive databases [16] (p. 179).

3) Another relevant use of AIS in news production is the verification of information, automatically tracing sources and facts. One example is Truthmeter, “a tool that automatically scores the journalistic credibility of social media contributors in order to inform overall credibility assessments. The Truthmeter computes credibility scores based on data made available through the Twitter API” [17].

4) Content curation is another purpose in which AIS can help journalists in their search for scoops on relevant topics. This is one of the ways used by RADAR (‘Reporters And Data And Robots’), a software system used by the British Press Association, which combines humans and machines to create localized stories at scale on topics such as “crime figures, hospital waiting times and pupil absences from school” [13] (p. 25).

5) An interesting application of AIS is the adaptation or customization of messages to different users, producing several versions—thus, and from a legal point of view, derivative works; any one of them is protectable under copyright law. The Swedish newspaper *Svenska Dagbladet*, for instance,

designed a tool to create different homepages for users, measuring the number of clicks, time or permanence or preferences introduced by the user [18]. Out of our focus, but to be considered for its commercial interest, AIS can help with subscriptions, which is in 2020—until the coronavirus crisis, at least—a common movement in media industry, following models like that of the *New York Times'* paywalls. This is called a “dynamic paywall”, in which companies like Deep Bi are working.

6) Most usually, artificial intelligence systems-aided or automated news production is based on database exploitation and automatization of raw data using patterns, which results in what some authors have named database journalism [19] (p. 5). The results of those practices can be considered protectable by copyright laws as sui generis rights. Patterns are provided by humans; only when the system is able to learn, improve and create new patterns may a non-human authorship appear.

By now, then, artificial intelligence systems-aided journalism is reduced to factual content coverage, while in principle only humans are able to elaborate more complex and contextualized pieces. Factual content coverage using AIS is, however, appealing for media companies, because it provides “a cost-effective way to create high-quality factual content that does well in SEO terms” [20].

The key of all those systems, and the ones that could be improved in the next future, is whether they need post processing by humans or not. Some of them, like Monok or RADAR, do not seem to need it to produce simple news items, with no great context. Artificial intelligence systems are not able to generate text of complex or non-predictable nature, a hypothesis placed by Ufarte and Manfredi [21]; or, as Belz says, “with a lot of unpredictability in the output” [20].

3. Methods

The main method used in this paper is a legal, comparative analysis. Since, as mentioned, very few cases have been examined in court to this point, we will base our analysis more on doctrine than on jurisprudence or on an examination of specific legal provisions. Copyright acts do not deal, to our understanding, concretely with these issues, and most of these acts do not mention the automated production of intellectual works at all. In other words, if new legal problems appear because of the use of new tools and techniques, to this point no legal reform has been practiced to include some new provisions to cover specifically those situations, and so intellectual property principles as we know them are to be applied. Even though some legal reforms have mentioned artificial intelligence beyond the automated analysis of data, which is important, they seem not to cover the automated production of works, or at least they can only be applied to some steps of it. This is the case of the *Directive of the European Union 2019/790, on copyright and related right in the digital single market*, for instance Article 3.2., on text and data mining. As some relevant scholars have underlined, “the fact that artificial intelligence and robotics are much more than science fiction becomes apparent” on the working documents of the European Commission, but at the same times it appears that it is considered just “the next step in the development of a sustainable information society” [8] (p. 3). Alternatively, automated systems are a concern whenever they are used by platforms like Facebook or Youtube for users' identification and filtering [22] (p. 267), which has a reflection on the literal interpretation of Article 17 of the aforementioned European directive, to oblige Internet services to detect unauthorized (and usually derivative) works uploaded by users without copyright permission.

We will focus on two main legal categories related to copyright and intellectual property law. First, the question of authorship and, closely related to it, originality as a sine qua non requirement for the law to consider a work copyrightable. Second, the type of work. We have advanced some of them: the individual work, a single piece created usually by a single (human) author. A collaborative work, alternatively, is when two or more authors can create a combined piece. In this case, we can consider many cases in which AIS aids a human journalist to complete his or her work. A derivative work, in which a new one is created based on one or more previously existing works, is an increasingly widespread type of work. The derivative work can be created, in turn, by both humans or machines based on both human or machine-made pre-existing works. Finally, and this is probably the central part of our analysis, the collective work, is composed by many works gathered and structured under

the coordination of another (natural or corporate) person. This is the case of newspaper, magazines, broadcasting services and webpages.

Since automated news are created (or transformed) with the help of both data, structured normally as a database, and software, some other categories of intellectual property rights are to be considered as well: the so-called, at least in some jurisdictions, *sui generis* rights, normally applicable to databases as a structure, equally created under the requirements and necessities of a corporate entity in order to produce intellectual works, and not necessarily to data themselves; and related, ancillary rights. This whole panorama means a complex superposition of layers of rights, some of which are accumulative and not mutually exclusive, to be considered in the many cases we will examine in the following sections. Even though if very few cases have been decided in court, it is expected that media organizations and practitioners of news reporting—journalists, photographers, infographics designers or even cartoonists, to mention some of them—will be facing some of these scenarios soon.

4. Results

The cross-examination of the aforementioned cases, which covers the most common practices of artificial intelligence systems-aided journalism to this point, their classification considering the legal axis of authorship (and originality) and typology of work, and the phase of journalistic work (gathering, production and dissemination) could help us to determine to what extent copyright law can cover these new products.

First of all, it is to be noted that all of them are produced due to the initiative and investment of a company, a corporate entity which is typically considered the one under whose coordination a collective work is made. This is to say that media companies as corporate entities are the producers of a collective work, not the authors of it. This is a characteristic more relevant in Civil Law countries than in the Common Law legal tradition, in which an entrepreneurial point of view is more explicit than in the most authorial, based on the creativity of individuals, approach of Civil Law countries. This, which was in the origin of copyright and authors' rights legal systems, has been modulated over time, and the importance of producers is evident in the case of, e.g., the audiovisual work. There are some movements to extend this consideration to the producers of multimedia works. One of those movements is the aforementioned lobbying activity of the major newspaper companies in Europe or the European Union to enact an ancillary exploitation right for press publishers, materialized in Article 15 of the *Directive on Copyright and Related Rights in the Digital Single Market*, 2019, to be implemented by state members (as of mid-2020, the only one to do so was France). Article 15 is intended to protect press publications "concerning online uses", for two years after publication. The duration of rights is notably shorter than the protection given to personal creations (all the authors' life term plus 70 years after his or her death), but is perfectly suitable for automated creations. There is an advantage in such artificial intelligence systems-aided works, since, when an author's name is not mentioned, no one has to receive "an appropriate share of the revenues". The *Resolution of the 2019 AIPPI World Congress on Copyright in artificially generated work*, one of the most developed documents on this area, agrees with this conception, and considers that "the term should be shorter than for the other copyrightable works" [23] (p. 19). This is an important thing to be remarked, since non-authored works may enter the public domain much earlier than authored ones. Madeleine de Cock Buning made an interesting reflection on this: "Without any form of intellectual property protection, these works can be used, reproduced, changed and distributed to the benefit of all. One can argue in favor of this option where Artificial Intelligence Systems creation is a positive consequence of Artificial Intelligence to the benefit of society as a whole" [8].

This also avoids also the application of moral rights, especially important in Civil Law, authors' right countries, but not so much in Common Law countries: in the United Kingdom, for instance, journalists are an exception of moral rights and the companies have no obligation to mention the name of their hired workers—although they usually recognize them as authors, all the way. This legal provision, and the similar one planned in Australia in the *Final Report on the Digital Platforms Inquiry* by

the Australian Competition and Consumer Commission (ACCC), published on 26 July 2019, states the importance of enacting such a right for press publishers, to help them monetize content.

The question of authorship, which is an unwaivable moral right in many countries, most especially in Civil Law ones, is of crucial importance when examining the changes that artificial intelligence systems-aided news production can cause to journalists and companies. In many Civil Law countries, authorship is only applicable to human creators, not to corporate entities or to software. Also, in the United States, even if such provision does not appear in the Copyright Act, we can consider that there is a similar principle, since the Copyright Office has repeatedly declared that it will “register an original work of authorship, provided that the work was created by a human being”. In the gathering phase of the journalistic work, artificial intelligence systems act as a simple tool—no matter how complex is their design, they are manipulated by human people—and produce no final work to be published. Facts and data, it should be remembered, are not protectable by themselves. We agree with Lin Weeks: “At the highest level of abstraction, automated journalism stories consist of an algorithm, or input (known in the industry as clean data), and of prose output” [19] (p. 85). Copyright law can only protect the second.

Copyright law only covers the final output placed into public knowledge using intellectual skills. Moreover, human authorship is to be recognized when artificial intelligence systems are used for data gathering, text mining, searches or verification. Since the initial work to be improved is made by human people, the final result is also due to them, and not to machines. When AIS is used for content curation as well, as a starting point for news items creation, the final authorship is of human journalists. A similar case happens when a journalist or editor revises the mistakes made by artificial intelligence systems. The final responsibility before the final publication of work is due to an individual or to the corporate entity, in any case.

It is obvious that the development of software can be authored by someone and commissioned by a company, which is the usual case. Following Lin Weeks, the protection of the algorithm itself, considered, we should add, as a form of software, is uncontroversial; more problematic is how to protect the output itself [19]. Whenever software is an instrument for creation, the final responsibility of the output is due to human authors. It is unusual, but not impossible, that just one individual is the inventor of the AIS software and the creator of the work. At least, there is one early example of it, the aforementioned Ken Schwencke, a journalist who both programmed an algorithm and exploited the results of it in 2014. Since he controlled the whole process, he signed the news. There is some way to attribute authorship to the programmer in countries such as Hong Kong, the United Kingdom, Ireland, India or New Zealand, all of the Common Law countries. For instance, section 9(3) of the *Copyright, Designs and Patents Act (CDPA)* states that “in the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken”, but it adds in section 178 that to be considered as that, it must be “generated by computer in circumstances such that there is no human author of the work” There is a parallel comparison, for instance, with generative music. A musician can use software (one example is *Wotja*, developed after the suggestion of musicians such as Brian Eno) to create music adjusting some parameters and patterns, and once done that, the artificial intelligence system starts creating music, which in turn can be modified whilst playing. The authorship of such musical pieces is of those human beings who decide which parameters must be adjusted, when and how. We agree with Andrés Guadamuz when he said that “the idea behind such a provision [referring to the UK Copyright Act] is to create an exception to all human authorship requirements by recognizing the work that goes into creating a program capable of generating works, even if the creative spark is undertaken by the machine” [24].

When the creation of a work is only the result of artificial intelligence systems, with no human intervention at all, which is thought to be only possible in randomly created outputs, this may be possible in music but hardly in news reporting, since it may result in a lack of sense. Anyway, there are some cases, for instance the Australian *Acohs Pty Ltd v Ucorp Pty Ltd*, which declare that a work

not produced by a human cannot be protected by copyright (Gadamez, 2017). Some other cases, for instance in the European jurisdiction, insist in the decisive intervention of human beings in the final result for a work to be considered copyrightable: the Court of Justice of the European Union dealt with that question in C-145/10, *Eva-Maria Painer/Standard Verlags* [2011], C-604/10 *Football Dataco/Yahoo!* [2012] [9] (p. 321).

The main point is, then, originality, and how to define it. In the mentioned European cases, it is required that the work is “the author’s own intellectual creation”, so in some way it must be (concurrently with a Civil Law, authors’ right legal tradition) an *oeuvre de l’esprit*: some personal touch must be found in the work. Even if artificial intelligence systems can show some creativity, meaning that they can generate works using data, patterns and algorithms, it is far more difficult to find some originality in them. Once again, the CJUE has insisted on this point, e.g., in C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, declaring that it is essential to find some elements of personality in a work to be protected under copyright law. Whenever a human intervention is a *sine qua non* condition to produce the work, such a personality characteristic can be detected. Human intervention is always needed: software cannot create software, so, according to the World Intellectual Property Organization (WIPO), there are only two ways to face this problem: to deny copyright protection to works created exclusively by computers or to attribute it to the creator of the program. There is a third way in the case of news: to attribute it to the corporate entity responsible for the collective work in which this artificial intelligence systems-aided work is inserted. In this case, and the *Resolution of the 2019 AIPPI World Congress on Copyright in artificially generated works* insists on it, a related, neighboring, ancillary or *sui generis* right is needed for publishers.

To our knowledge, this has not happened yet, but it is not impossible to think that artificial intelligence systems-aided news, which may use third parties’ data, could infringe copyright whenever the origin and eventually author of the original works from which the derivative one is developed are not properly mentioned, and the corporate person who publishes it could be sued for it. Such a case will help to clarify positions, and it has been mentioned, but not developed, in the Resolution adopted in September 2019 by the AIPPI World Congress. It is not impossible, we should add, that in some cases such practices could be considered under quotation exception—or fair use in the Common Law countries—but anyway it must be examined case by case, with no need to create new exceptions [23] (p. 11), especially in legal areas such as the European Union, in which a closed list of exceptions has to be applied. Obviously, in countries where fair use or fair dealing is applied, a case-by-case approach will be needed.

Another interesting point of view is how the media manage artificial intelligence systems to automatically display some kind of information, a question that has been examined by Jop Esneijer: “Note that automated scanning of tweets and blogs for relevant content and copy or even publishing them [...] would in principle also require the authorization of the original author as these are acts of copying or making available to the public, unless they are excepted, for example because they fall under” [25] (p. 43). This is because we are dealing with derivative works.

It is different when news items are mainly created with the aid of artificial intelligence systems, in which case the attribution of paternity is shown as anonymous or attributed to the corporate person. This is consistent, as we have already mentioned, with that old distinction of the *Berne Convention on Copyright* of 1886–1887, which in Article 2 stated that the consideration of “literary and artistic works “shall not apply to news of the day or to miscellaneous facts having the character of mere items of press information”. The Berlin Convention of 1908 defined to a greater extent which works amongst the ones published in a newspaper were copyrightable or not, the ones that could be reproduced—always mentioning author and origin—or not. The Berlin Convention protected any work published in a newspaper, which was a great advance compared with the previous conventions, which distinguished between literary works and *nouvelles du jour*. In fact, the distinction was maintained in the following conventions, those of Rome (1928), Brussels (1948) and Paris (1971; amended in 1979), now in Article 2.8. This old distinction can now have a new fashion regarding the production of news

items produced exclusively or mainly by artificial intelligence systems or produced under the final responsibility of human authors, but the rights on the economic exploitation of all of them are, anyway, to be recognized to copyright holders of the collective work. Some scholars have examined these cases, and concluded that the common practice is for the corporate entities to sign those news items using the company's name, and scarcely mentioning the fact that they have been generated with artificial intelligence systems' aid [21] (p. 13).

5. Discussion

Innovation in journalism, specially from the advent of the World Wide Web in the mid-1990s, is a central point for companies and researchers. The media industry is facing a major crisis, most especially from 2008 onwards, in which companies are trying to redefine a successful business model, searching for a viability for an activity to that point sustained mainly by advertisement. Optimizing all economic resources is thus crucial for this industry, as some many scholars have insisted, and in this respect automated tools may be "the key to the viability of news media in the digital age" [26]. It is in this context that we must situate the discussion on the role of intellectual property law when applied to the outputs of automated journalism. Companies need to monetize content, and developing artificial intelligence systems to help journalistic work for gathering data, elaborating news and disseminating them—even to commercialize them more efficiently—can help in this purpose. In most cases, as we have examined, artificial intelligence systems need human intervention at some point and this leaves some personality clue which leads to considering the output a characteristic of originality, needed for copyright law to be applied. Investment should, on the other hand, be enhanced. The most developed proposal to this point, the *Resolution of the 2019 AIPPI World Congress con Copyright in artificially generated works*, after consultation with many national groups all over the world, concludes that the majority of them "consider that the investor (natural or legal person) should be the original owner of the artificially-generated works [23] (p. 16)."

New profiles are appearing in news reporting: journalists incorporate new skills to traditional ones, and one of them, regarding to automated news, is to be a designer, programmer, supervisor or editor of news items created with the help of software [12] (p. 284), so adaptation of skills and training seems more necessary than ever [27].

The legal recognition of the journalist as an author, laboriously developed through history, is jeopardized once again. An individual approach to intellectual property (ultimately, an authors' rights approach) is more difficult to defend and the central role in copyright law is now that of the collective, and even of the derivative work. In the more optimistic views, this is good news for journalists, since artificial intelligence systems-aided production may free human journalists from heavy tasks and reserve them for an extra level of coverage [21] (p. 5,6) (reports and features, basically) with a more added value and, following the old Berne Convention literal, a more "literary" approach. Anyway, there are some ways to alternatively attribute authorship, or related rights, to a natural person or a corporate entity, and in every case AIS-aided news should be attribute to someone. Preserving the notion of authorship is extremely important in this regard. It is probably more difficult to preserve moral rights when the weight of the tasks to produce a news item is shared between software and a journalist, and in some way that weight should be balanced, but regarding economic rights we agree with the conclusions of Osha et al, 2019, that they "should not differ between artificially-generated works and regular works" [23] (p. 10). It is quite difficult, anyway, to attribute moral rights to the inventors of designers of artificial intelligence systems, and it is even difficult to attribute them to journalists who help produce artificial intelligence systems-aided news in countries like the United Kingdom, for the aforementioned reasons: Article 79 of the *Copyright, Designs and Patents Act, 1988*, states that the moral right "does not apply to a computer program [. . .] any computer generated work" or "in relation to the publication in a newspaper, magazine or similar periodical". Even if it may seem a quite restrictive provision, it provides a clue about how things can be held regarding to the specific topic of this paper.

To this extent, it seems that there is a general agreement that, following the definition of the *Resolution of the 2019 AIPPI World Congress on Copyright in artificially generated works*, “AI generated works should only be eligible for protection by Copyright if there is human intervention in the creation of the work and provided that the other conditions for protection are met. AI generated works should not be protected by Copyright without human intervention”. The extreme case is when artificial intelligence systems (AIS) are able to learn for themselves and create news autonomously, in which case the so-called “creative agents” are machines [9], or, using the title of a symposium held in Alicante (Spain) on the topic in 2019 [28], whether it might happen that robots can invent and create. As we have examined before, this is not the most usual situation in media, and when it happens the output is usually mere news, as stated in the Berne Convention on copyright, not attributable to any author, but of economic interest for corporate entities as part of a collective work. This appears to be the main category in these times, in which, trying to combat a structural crisis, the media industry is aiming to defend its interests by enforcing this legal category. Another example is how the major newspaper industry in Europe has managed to include a new ancillary, exclusive exploitation right in the European Union’s Directive on Copyright of 2019, the so-called press publishers’ rights. Even though in Article 17 of the Directive automated news is not mentioned, this literal could be eventually used to defend the media’s economic interests, with no need to compensate any human author.

The general trend should be, in our opinion, to concede that there is some originality whenever some human intervention is required at some step of the journalist routine, and some guidance, pattern providing, instruction, deep revision of news items is provided before publication. Personal authorship should not be conceived as a romantic conception of the sole creation of a work due to an individual inspiration, but to any intellectual skill required to place in the market a work to be properly and coherently understood by human people. Even in some more unclear cases, the responsibility of the corporate entity in the production and insertion of such a product in a collective work should be a sufficient condition to secure a neighboring, ancillary right or even a sui generis right generated by the responsibility in providing instructions to structure databases (not such other things are digital media in these days) and design interfaces to exploit them [23] (p. 7). A balance between the rights of the investors, the inventors and the workers is needed, as they are the rights of the audience and public knowledge. In this sense, a revision of the duration of rights is needed, and much shorter rights are foreseen to help enter artificial intelligence systems-aided news into the public domain.

Funding: This research was funded by the Ministry of Science and Innovation of Spain, grant number RTI2018-095775-B-C43 (Project: News, Networks and users in the hybrid media system. Transformation of the media industry and the news in the post-industrial era).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Diakopoulos, N. Algorithmic accountability. Algorithmic accountability. Journalistic investigation of computational power structures. *Digit. Journal.* **2015**, *3*, 398–415. [CrossRef]
2. Dörr, K. Mapping the field of algorithmic journalism. *Digit. Journal.* **2016**, *4*, 700–722. [CrossRef]
3. Diakopoulos, N. *Automating the News. How Algorithms are Rewriting the Media*; Harvard University Press: Cambridge, MA, USA, 2019.
4. Marconi, F. *Newsmakers: Artificial Intelligence and the Future of Journalism*; Columbia University Press: New York, NY, USA, 2020.
5. Hansen, M.; Roca Sales, M.; Keegan, J.M.; King, G. *Artificial Untelligence: Practice and Implications for Journalism*; Tow Center for Digit. Journal, Columbia University Libraries: New York, NY, USA, 2017.
6. Ford, M. Could artificial intelligence create an unemployment crisis? *Commun. ACM* **2013**, *56*, 37–39. [CrossRef]
7. Graefe, A. *Guide to Automated Journalism*; Columbia Journalism School, Tow Center for Journalism: New York, NY, USA, 2016. Available online: <https://pdfs.semanticscholar.org/c56d/609b3cb2ff85a3e657d2614a6de45ad2d583.pdf> (accessed on 7 May 2020).

8. De Cock Buning, M. Autonomous Intelligent Systems as Creative Agents under the EU framework for Intellectual Property. *EJRR 2. Spec. Issue Man Mach.* **2016**, *7*, 310–322. [CrossRef]
9. Newman, R. *Journalism, Media, and Technology Trends and Predictions*; Reuters Institute for the Study of Journalism: Oxford, UK, 2018.
10. Díaz-Noci, J. Authors' rights and the media. In *Interaction in Digital New Media*; Pérez-Montoro, M., Ed.; Palgrave: Gram, Switzerland, 2018; pp. 147–173.
11. Díaz-Noci, J. Copyright and User-Generated Contents for Mobile Devices: News, Entertainment, and Multimedia. In *Between the Public and Private in Mobile Communication*; Serrano, A., Ed.; Routledge: London, UK, 2017; pp. 199–217.
12. Segarra-Saavedra, J.; Cristòfol, F.J.; Martínez-Sala, A.M. Inteligencia artificial (IA) aplicada a la documentación informativa y redacción periodística deportiva. El caso de BeSoccer. *Doxa Comunicación* **2019**, *29*, 275–286. [CrossRef]
13. Becket, C. *New Powers, New Responsibilities. A Global Survey of Journalism and Artificial Intelligence*; London School of Economics: London, UK, 2019.
14. Leone, R. The Mashup Man. An Online Innovator Uses an Ingenious Fusion of Imagery and Databases to Present Information in Exciting New Ways. *Am. Journal. Rev.* **2007**, *28*, 10–14. Available online: <http://ajrarchive.org/Article.asp?id=4258> (accessed on 25 March 2020).
15. Kelley, M. *Web 2.0 Mashups and Niche Aggregators*; O'Reilly: Sebastopol, CA, USA, 2018.
16. Boyles, J.L.; Meisinger, J. Automation and Adaptation: Reshaping journalistic labor in the newsroom library. *Converg. Int. J. Res. New Media Technol.* **2020**, *26*, 178–192. [CrossRef]
17. Fletcher, R.; Schifferes, S.; Thurman, N. Building the 'Truthmeter': Training algorithms to help journalists assess the credibility of social media sources. *Converg. Int. J. Res. New Media Technol.* **2020**, *26*, 19–34. [CrossRef]
18. Stern, R. FL#195: A home page designed by algorithm. *Reynolds Journal. Inst.* **2017**, *24*. Available online: <https://www.rjionline.org/stories/fl195-a-homepage-designed-by-algorithm> (accessed on 20 March 2020).
19. Weeks, L. Media Law and Copyright Implications of Automated Journalism. *J. Intellect. Prop. Entertain. Law* **2014**, *4*, 67–94.
20. Belz, A. Fully Automatic Journalism: We Need to Talk About Nonfake News Generation. In Proceedings of the Conference for truth and trust online—BMA House, London, UK, 4–5 October 2019.
21. Ufarte-Ruiz, M.J.; Manfredi, J.L. Algorithms and bots applied to journalism. The case of Narrativa Inteligencia Artificial: Structure, production and informative quality. *Doxa Comunicación* **2019**, *29*, 213–233. [CrossRef]
22. Hugenholtz, B. *Copyright Reconstruct: Rethinking Copyright's Economic Rights in a Time of Highly Dynamic Technological and Economic Change*; Wolters Kluwer: Amsterdam, The Netherlands, 2018.
23. Osha, J.P.; Verschuur, A.M.; Laakkonen, A.; Guillaume, G.; Nack, R.; Shen, L. *AIPPI. Copyright in Artificially Generated Works: Resolution*; AIPPI World Congress: London, UK, 2019.
24. Guadamuz, A.; Artificial Intelligence and Copyright. *WIPO Magazine*. 2017. Available online: https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html (accessed on 25 March 2020).
25. Esneijer, J.; Nieuwenhuis, O.; Mijs, C.; Versloot, C.; Helberger, N.; Van der Sloot, B.; McGonagle, T. *Making User Created News Work*; TNO 2012 R11277; IViR: Amsterdam, The Netherlands, 2012.
26. Pavlik, J. Innovation and the future of journalism. *Digit. Journal.* **2013**, *1*, 181–193. [CrossRef]
27. Small, J. *NewsLab '20 Gathers Human Brainpower to Ponder Roles for AI in Journalism, Media Industry*; Local Media Association: Lake City, MI, USA, 2020. Available online: <https://www.localmedia.org/news-lab-20-gathers-human-brainpower-to-ponder-roles-for-ai-in-journalism-media-industry/> (accessed on 25 March 2020).
28. Fernández-Lasquetty, J.; López-Tarruella, A. *Summary of the Congress: Can Robots Invent and Create? A Dialogue between Artificial Intelligence and Intellectual Property*; University of Alicante: Alicante, Spain, 2019.



© 2020 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Article

On the Regulatory Framework for Last-Mile Delivery Robots

Thomas Hoffmann ^{1,*}  and Gunnar Prause ² 

¹ Tallinn Law School, Tallinn University of Technology, Ehitajate tee 5, 12616 Tallinn, Estonia

² Department of Business Administration, Tallinn University of Technology, Ehitajate tee 5, 12616 Tallinn, Estonia; gunnar.prause@ttu.ee

* Correspondence: thomas.hoffmann@ttu.ee

Received: 23 May 2018; Accepted: 26 July 2018; Published: 1 August 2018

Abstract: Autonomously driving delivery robots are developed all around the world, and the first prototypes are tested already in last-mile deliveries of packages. Estonia plays a leading role in this field with its, start-up Starship Technologies, which operates not only in Estonia but also in foreign countries like Germany, Great Britain, and the United States of America (USA), where it seems to provide a promising solution of the last-mile problem. But the more and more frequent appearance of delivery robots in public traffic reveals shortcomings in the regulatory framework of the usage of these autonomous vehicles—despite the maturity of the underlying technology. The related regulatory questions are reaching from data protection over liability for torts performance to such mundane fields as traffic law, which a logistic service provider has to take into account. This paper analyses and further develops the regulatory framework of autonomous delivery robots for packages by highlighting legal implications. Since delivery robots can be understood as cyber-physical systems in the context of Industry 4.0, the research contributes to the related regulatory framework of Industry 4.0 in international terms. Finally, the paper discusses future perspectives and proposes specific modes of compliance.

Keywords: delivery robots; autonomous transport; last-mile distribution; regulatory framework; Industry 4.0

1. Introduction

Within the last years, many initiatives towards smart manufacturing have been initiated in order to re-establish and regain a significant industrial share in the economy [1,2]. A promising concept is the fusion of the virtual and the real worlds of manufacturing to realize concepts for smart manufacturing and logistics by using cyber-physical systems (CPS) and dynamic production networks in order to achieve flexible and open value chains in the manufacturing of complex mass customization products. The German concept of smart production and logistics, Industry 4.0, goes even beyond these objectives, as Industry 4.0 aims to comprise also energy and resource efficiency, increased productivity, shortening of innovation, and time-to-market cycles [3]. Internet-based linked machine-to-machine (M2M) communication and interaction pave the way to cross-company production and logistics processes enabling the design and control of the entire supply chain of a product during its full life time, i.e., from product design to logistics, distribution, and post-production services [4,5]. Consequently, Industry 4.0 leads to new supply chain paradigms on complex and intertwined manufacturing networks with a high degree of fragmentation and low entry barriers for small and medium enterprises (SME) as well as new R&D strategies, cross-national value chains, and new business models [2,6–8].

Despite the importance of logistics and distribution for the Industry 4.0 concept, a concise literature review reveals that the last-mile problem has not been discussed by scholars in the context of

Industry 4.0 so far. Nevertheless, the last-mile problem has been subject to a large number of scholars in the context of retailing and e-commerce [9–12]. Lee and Whang point out innovative e-fulfilment strategies for orders to contribute to solve the last-mile problem [13]. Song et al. addressed the last-mile problem from the transport point of view (here transport impacts of collection and delivery points), whereas Boyer et al. conducted their research by utilizing a simulation on the base of vehicle routing software, investigating the relationship between customer density, delivery window length, and delivery efficiency [14,15].

Existing M2M concepts are mainly researched by technical scholars so far. Cha et al. discuss different use cases for M2M communication together with common security requirements to clarify the security requirements on M2M systems [16]. Their business cases include logistics-related M2M applications and come to the conclusion that information security and trustworthiness of the operations involved grows from the predictability and observability of the behaviour of the devices. Wu et al. investigated M2M systems in the context of embedded internet and identified low cost/high performance devices, scalable connectivity, and cloud-based device management and services as vision for the Internet of Things [17]. By considering M2M cases for mobility support, they investigated frame conditions for standards and M2M networks. Zhang et al. highlighted—besides security issues—self-organization and quality of service support as important factors for M2M-communication [18]. Self-organization was stressed due to low human intervention as a major requirement for M2M systems, which for that aim ought to comprise self-configuration, self-management, and self-healing. The quality of service requirements was emphasized against the background of possible applications, which could become life critical (e.g., in medical contexts).

This paper will focus on the niche of autonomous self-driving package delivery robots that are used for intra-supply chain transport in Industry 4.0 networks as well as for the delivery to the client on the last mile. The research concentrates on the regulatory framework for those autonomous delivery robots, as existing research rather addresses self-driving passenger vehicles [19]. Since the research issue of this paper is placed in a global business sector of high dynamics and ongoing innovation activities, it is impossible to give a comprehensive insight in all facets and to discuss all topics in detail. Therefore, the authors study the case of one important player in the field of delivery robots, the Tallinn-based start-up Starship Technologies, which produces and develops delivery robots that can technically be considered as self-driving package box vehicles bridging the last mile.

The paper highlights the current status of autonomous delivery robots for distribution and investigates the related regulatory framework for the use of these self-driving vehicles. As the authors consider delivery robots as parts of M2M supply chains in the context of Industry 4.0, the starting point of the research is an overview on the existing regulatory framework for Industry 4.0. The research uses an empirical analysis based on semi-structured expert interviews, research group meetings, secondary data, and results from case studies that are gathered from Estonian start-up companies. Starship Technologies is placed in Tehnopol in Tallinn and maintains a close research cooperation with Tallinn University of Technology, which allows for an empirical validation of the research results.

2. Theoretical Background

The still growing e-commerce market volumes raise the question of efficient product delivery to the client. The last-mile delivery includes three stakeholders, namely the seller, an intermediary and the client. Punakivi et al. still discussed the last mile-issue in the traditional context of B2C and e-commerce; they proposed an unattended reception of goods which could reduce home delivery costs by up to 60% [10]. The unattended delivery approach is based on two main concepts, being the reception box concept and the delivery box concept: The reception box is installed at the customer's garage or home yard, whereas the delivery box is an insulated secured box that is equipped with a docking mechanism. Based on simulation results, the authors came to the result that home delivery solutions enabling secure unattended reception are operationally the most cost-efficient model for last-mile distribution. They also confirmed that a secured delivery box solution potentially enables

a faster growth rate and higher flexibility of the investments because of a smaller investment being required per customer.

2.1. Industry 4.0

Within the last years, many innovative manufacturing initiatives have been started all over the world, driving for re-establishing and regaining a significant industrial share in the economy. Many of them embrace the fusion of the virtual and the real world based on cyber-physical systems (CPS), leading to smart manufacturing and logistics networks towards flexible and open value chains to be able to meet the demands of mass customization products in series up to lot size 1 [3,4]. In Germany, the most important industrial EU country, this approach has been called “Industry 4.0”. A deeper analysis of the objectives of Industry 4.0 reveals that Industry 4.0 targets beyond the use of cyber-physical systems and dynamic supply chain networks also to energy and resource efficiency, shortening of innovation, and time-to-market cycles, as well as a rise in productivity through internet-linked machine-to-machine (M2M) communication and interaction [3–5]. In this sense, Industry 4.0 represents nothing less than the fourth industrial revolution, comprising three-dimensional (3D) printing, big data, Internet of Things, and Internet of Services, i.e., all of the ingredients that are needed to facilitate smart manufacturing and logistics processes [3,4].

Meanwhile, new technological innovations implemented into new business models opened up new solutions to bridge the last-mile to the client by using drones and delivery robots, and food and grocery services gaining first experiences in the use of autonomous devices [20]. By transferring the traditional delivery box concept of Punakivi et al. into an Industry 4.0 context, a corresponding approach should take account of the options of internet-linked manufacturing and logistics [10]. Mainly technical scholar studied M2M systems and the realization of autonomous logistics agents in the context of Industry 4.0. Cha et al. studied business cases, including logistics-related M2M applications, and Wu et al. investigated M2M systems in the context of embedded internet and identified low cost/high performance devices, scalable connectivity, and cloud-based device management and services as vision for the Internet of Things [16,17]. By considering M2M cases for mobility support, they investigated frame conditions for standards of M2M networks and Zhang et al. highlighted self-organization and self-management as important factors for success M2M systems due to low human intervention as a major requirement [18]. Several scholars came to the same conclusion by stressing self-organisation and self-optimisation as success factors to cope with requirements of Industry 4.0 [2,8,21]. Based on these principles—together with wireless internet technologies, artificial intelligence concepts and M2M technologies—some entrepreneurs founded start-ups to develop autonomous transport devices on the base of Industry 4.0 related concepts in order to serve the last-mile delivery more or less autonomously.

Parallel to technical and economic issues also the discussion of a regulatory framework for Industry 4.0 enjoyed high importance. Already Kagermann et al. dedicated a full chapter to the regulatory framework in their recommendations for implementing the strategic initiative Industry 4.0. They highlighted the requirement to reconcile regulation and technology, i.e., they postulated the formulation of criteria to ensure that the new technologies comply with the law and development of the regulatory framework in a way that facilitates innovation. Special emphasis was laid on the protecting of personal and corporate data, liability issues, and trade restrictions ([3], pp. 58–61). For the implementation of such a regulatory framework, they proposed a mix of instruments comprising regulatory, technical and policy elements, and they pointed out the special importance of the inclusion of SME sector.

2.2. Delivery Robots

After an initial hype about delivery with flying drones, in recent times land-based delivery robots are in the focus for the last-mile [20]. Since these robots have to share their space with other transport devices or moving people, their preferred operation areas are suburbs and areas where the traffic is

comparatively low. In these areas, autonomous delivery robots have a competitive advantage when compared to other delivery modes, and the underlying business model emphasizes the cost advantage for the last-mile delivery, which is estimated to be less than 1€ per unit/delivery, which is—depending on the salary level of the respective location—up to 15 times less than current costs [22]. For the customer, additional convenience is gained by the aspect that robot delivery provides a 15-to-20 min delivery window as standard, which is a much more precise specification than for traditional delivery, which so far is only able to provide the a general date (calendar day) beforehand.

Today, the key players of last-mile delivery consists of established delivery companies, including traditional logistics service providers as DHL, UPS, and others, but also a range of new startups focusing on the development of delivery robots that grow all around the globe. The most important business areas of delivery robots are currently perishable goods as food and flowers, but also applications in retailing and warehousing sector are possible in the context of automated warehouses. A closer view into the main startup funding landscape reveals that about 50% of all investment sums are dedicated to enterprise robots, which comprise industrial automatization for manufacturing, heavy industry, as well as delivery robots [23]. According to a study of International Data Corporation, the industry and the manufacturing sector will continue to be the largest purchaser of robots and related services, and the worldwide spending on robotics that reached the \$100 billion level in 2017 is forecasted to be more than doubled until 2021 [24]. Nevertheless, the robot sector today realized that deliveries to costumer sector are predicted to represent the fourth largest growth till 2021, with a compound annual growth rate of about 60%. A deeper insight into the scene of land-based delivery robots shows that start-ups as Marble, Teleretail, Dispatch, or Starship Technologies were able to attract funding in the range of several million Euro [24].

Concerning the regulatory framework for delivery robots, the discussion is still open. On one hand, it is possible to build on the steps towards a regulatory framework for Industry 4.0; on the other hand, it is also possible to follow the discussions that are taking place in the context of autonomous mobility. Scheurs and Stewer worked on a regulatory framework of autonomous driving and analyzed the political, legal, social, and sustainability dimensions of mobility. Their investigations highlighted competitiveness, innovation, safety, harmonization, and coordination ([19], pp. 151–173). Their research based on empiric results from development in several countries as well as on the United Nations (UN) convention on road traffic. Basu et al. have recently researched the legal framework for small autonomous agricultural robots [25], but as “agribots” roam usually only on private land, the unresolved traffic law dimension has not been covered by their paper. This paper continues the regulatory framework path of Industry 4.0 by perceiving delivery robots as part of Industry 4.0 environment. Consequently, the research concentrates on liability issues, data protection, privacy, and legal developments around delivery robots.

3. Methodology

This paper highlights the current status of autonomous self-driving package delivery robots that are used for intra-supply chain transport in Industry 4.0 networks, as well as for the delivery to the client on the last mile. The research is based on semi-structured expert interviews, desktop and secondary data analysis, and a case study of Tallinn based start-up Starship Technologies Ltd. representing an important player in the branch of self-driving package box vehicles bridging the mast mile. The empirical activities were executed between September 2017 and May 2018.

It is not the aim of this paper to give a comprehensive overview of the sector of autonomous delivery robots, which is impossible due to the large number of developments in this sector. Nevertheless, the paper highlights technical, legal, and regulatory issues that are evolving with the development of autonomous delivery robots. Autonomous delivery robots are placed in the context of Industry 4.0 and M2M systems so that exiting concepts are firstly applied to case of delivery robots. In the sequel, actual issues that are related to liability and data protection are discussed.

Finally, social-technical aspects and possible legal solutions are discussed and an outlook for tentative developments is discussed. Therefore, the research questions are:

RQ 1: How do autonomous delivery robots work, and how are they defined in the context of liability?

RQ 2: Which regulatory frameworks apply on delivery robots?

RQ 3: Where does the current use of delivery robots conflict with these frameworks, and what shall users be advised to prevent violations?

Literature review reveals a research gap in the listed research questions. In addition, a case study of one of the most important start-ups for delivery robots is given to discuss and to empirically verify the research. For this purpose, the empirical evidence in this paper is based on the qualitative research style [26]. Here, the complexity of the research question requires personal interviews and a qualitative approach. The willingness to answer questions in a greater depth and in an open discussion can only be achieved by personal and individual conversations with selected interview partners. Furthermore, the field of delivery robots addresses a quickly developing innovative sector, so that a large part of the information is confidential; the research has to balance between novelty of science and the business secrets of the investigated companies.

For this, surveys, interviews, and workshops that have been conducted by the authors during the European projects, together with experts from business, ICT, and law, as well as from the start-up sector.

4. Case Study: Starship Technologies Ltd

Starship Technologies Ltd. was founded in 2014 by Skype co-founders Janus Friis and Ahti Heinla in Tallinn with the aim to tackle the last-mile problem by developing autonomous delivery robots. Today, Starship Technologies is a European technology startup with subsidiaries in Estonia, the United Kingdom (UK), and USA, which has built the first commercially available autonomous delivery robots in order to “revolutionize the local delivery industry” [22]. Starship claims to be environment-friendly as well, as Starship robots do not emit CO₂ (while—of course—the electric power plants do). It also claims that their robots contribute to reduce on-road traffic and thus congestions, and that Starship provides a solution for retailers and logistics firms to increase supply chain efficiencies and reduce costs.

Starship’s small self-driving vehicles with a weight of less than 20 kg are electric-powered and are designed for driving on sidewalks with a speed of maximal 6 km/h, being capable to locally deliver their goods within 15–30 min and within a radius of up to 5 km for a price of under 1 Euro per delivery. The robots are able to deliver freight of up to 10 kg for a shipment price which is up to 15 times lower than the normal price for last-mile deliveries in high-salary level economies, which makes the delivery robots interesting for e-commerce applications as well as for food deliveries or postal services. In practice, Starship delivery robots have been tested already by online food ordering service providers in Tallinn (Volt), as well as by Domino’s pizza delivery services to use them as “personal delivery devices”.

To safeguard safe circulation, the robots are equipped with a couple of sensors and tracking systems comprising nine cameras, GPS, and an inertial measurement unit (IMU) for special orientation. They are also equipped with microphones and speakers enabling them to communicate with humans. Even if the robots are called autonomous vehicles, they, at present, are only self-driving around 90% of the time; the remainder—mainly complex road crossings and the final meters to the receiver—the robot will be remote-controlled from a command centre, which is linked via Wi-Fi and telecommunication networks. While their entire journey, the robots are continuously supervised by a responsible, natural person, i.e., the contact with the command centre is not only established if the robot’s autonomous operation fails. This remote-control means that the operation of a delivery robot implies a permanent exchange of data, including life-video transfer, between the robot and the control centre via public telecommunication networks.

The underlying cost engineering strategy at Starship Technologies focusses on the use of traditional hardware engineering in order to make sure that the robots are cheap to produce and

that they require only basic maintenance. In terms of operational cost management, the company tries to generate cost advantages by targeting a hybrid autonomous robot to be operable in near future, which is able to drive entirely autonomously most of the time. In this fully-developed version, the remote-control supervisor in the command centre has only to be involved in teleoperations via live video link in a small percentage of time, which minimizes the operational costs of the robot.

In order to create a smart solution for bridging longer distances of delivery, the company started collaboration with Daimler in order to develop the “RoboVan”, which forms a mobile robot hub on the base of a MB Sprinter mini truck and would considerably extend the range of the robots. This approach for delivery realizes a “hub and spoke” concept, which is a well-known standard model in logistics [27]. A RoboVan-Mercedes-Benz Sprinter is to that aim equipped with a storage system for 54 delivery boxes and eight Starship robots. The Sprinter performs the long distance elements of transport as a mobile hub and it brings the robots together with the delivery boxes right into an area where a multitude of individual deliveries has to be performed. From this spot, the robots disembark from the RoboVan autonomously and cover the last-mile to the client in order to individually deliver the goods to the clients and return to the Sprinter afterwards. The approach realizes a “hub and spoke” concept with robot delivery for the last short distance.

Starship Technologies considers its delivery robots as a supplemental form of shipment, not as a replacement, i.e., the logistical models that can be used with robots are different than those models of traditional delivery methods. Ahti Heinla, the co-founder of Starship Technologies, illustrated in an interview the different areas of complementing delivery with bicycle couriers operating in very dense urban environments, since they are able to overcome gridlocks and traffic jams, whereas autonomous vehicle are predestinated for the delivery in suburbs with low traffic [28]. Access to the cargo in the robots is arranged by a smartphone app, which enables the client to unlock the robot cover lid and retrieve the goods. If someone tries to steal the robot, the cameras will take a photograph of the thief, and alarm will sound. Additionally, multiple tracking devices can track the robot’s location via GPS, and the remote operator is able to speak through two-way speakers with the thief; and, obviously, the robot will stop working and will not open the cargo unit unless re-programmed by Starship.

In January 2017, Starship Technologies announced \$17.2 million in seed funding for building autonomous robots that are designed to deliver goods locally. The funding round was led by Daimler AG and included a couple of other venture capital funds, among which were Shasta Ventures, Matrix Partners, ZX Ventures, Morpheus Ventures, Grishin Robotics, Playfair Capital, and others [22]. This amount of seed funding makes Starship Technologies rank among the worldwide leading companies of delivery robots for the last-mile.

5. Legal Challenges

Despite the fact that delivery robots are called autonomous, they are—for the time being—only partly self-driving, i.e., they are remote-controlled from a control centre. This remote-control is maintained via a permanent exchange of data between the robot and the control centre, resulting in serious issues in terms of data protection—issues this paper intends to discuss. But initially, the fact that the delivery makes use of public traffic area designated to pedestrians shall be analyzed from a legal perspective – especially in terms of tort liability for eventual accidents.

5.1. Liability for Torts Inflicted by Traffic Accidents

General tort law in most legal systems provides a general claim for damages caused by any tortious action, i.e., a civil wrong resulting in loss or damage to another person, and based on these principles implemented into positive law in all national legal systems individually, the legal or natural person steering the delivery robot and being in charge also of its supervision (in our case study Starship Technologies) would be held liable for any tortious action the legal/natural person committed via its tools—here the delivery robot—itself.

In general, tortious liability is in many legal systems fault-based (see e.g., sec. 823 I Bürgerliches Gesetzbuch, i.e., the German Civil Code, hereafter BGB) or subject to exculpation if the tort has not been committed directly by the tort-feasor, but a third person for whom the tort-feasor is responsible and who has been picked and supervised with due care (see sec. 831 BGB). In our case study, this could be an employee of Starship working in the command centre.

In contrast to that, two constellations are generally marked by strict (i.e., non-fault-based) liability for damages—product liability and liability under traffic law.

5.1.1. Product Liability

For the context of delivery robots, it is important that also the manufacturer of a product that caused damage/personal injury to the user can be held liable for the tort of negligence in most Western legal systems. In the European Union (EU) legal space, it is the Directive 85/374/EEC (Product Liability Directive), which regulates liability for defective products, and which has been implemented, respectively, in all EU member states national legal systems. The directive defines, “product” as all movables—even if incorporated into another movable or an immovable (see art 2 of amendment to directive)—which are considered by design as a completed product and imposes strict liability for any damage that is caused by the defective product on the producer, “defective” being any product that “does not provide the safety which a person is entitled to expect, considering, all of the circumstances, including, the presentation of the product, such as adequacy of the warning, the use to which it could reasonably be expected that the product would be put, and the time when the product was put into circulation are factors” (art 6), making the standard thus objective.

As Product liability can arise from constructional defect, fabrication defects, user instruction defects and product supervision defects—i.e., all spheres under the complete control of the producer—a sound production and product specification, user instruction, and supervision by the producer can limit the risks of strict liability as producer.

5.1.2. Tortious Liability under Traffic Law

This is considerably less the case for traffic law, which in most legal systems extends this liability according to the special circumstances of public traffic. In that respect, traffic law does not only extend the circle of debtors—i.e., not only the owner of a vehicle can be held liable, but also the driver separately, but also imposes generally strict liability onto the vehicle owner, i.e., the owner will be held liable for any damages caused by his vehicle in public traffic even if he did not act with intent or negligence.

Any victims of accidents in which delivery robots were involved will thus try to apply traffic law liability than product liability (or standard tort liability, which is usually fault-based) in order to maximize liability, if they can. The question is thus whether delivery robots can be qualified as vehicles participating in public traffic in standard traffic laws.

Delivery robots are starting from existing definitions for motorized vehicles, which are (only) permitted to operate in pedestrian areas (pavements) due to their low speed and weight, a comparable vehicle would be motorized wheelchairs (part a). The difference between e.g., these motorized wheelchairs and delivery robots are identical to those between human-steered cars and automatic cars. As the second difference has already been subject to regulation in various legal regimes, it can serve as a model for a respective definition of transport robots as well (part b).

Existing Definitions for Motorized Vehicles Operating on Pavements

If the usage of motorized wheelchairs is regulated—which is not always the case—most legal systems provide respective definitions in their street traffic and/or driving license acts. The German StVO (Straßenverkehrsordnung/street traffic act), for instance, already provides for a very detailed definition of motorized wheelchairs, which states in § 24 par 2 (special means of transportation),

that “motorized wheelchairs or with wheelchairs other than those referred to in paragraph 1 may be used wherever pedestrian traffic is permissible, but only at walking speed” [29].

The motorized wheelchair itself, however is defined in the Fahrerlaubnisverordnung (driving licence act) in § 4 II 1 e, being a “one-seated electrically driven vehicle, which is designed for use by physically disabled persons, has a maximum mass of not more than 300 kg including batteries but without driver, a maximum permissible mass (including driver) not exceeding 500 kg, a maximum design speed of not more than 15 km/H and a total width of 110 cm” [30].

While many of these criteria can be applied in delivery robots just as well as on motorized wheelchairs, there are three criteria of the definition that would have, respectively, to be adapted, being

- “autonomously or partially autonomously electrically driven motor vehicle (criteria 1), which is
- designed for the transport of goods (criteria 2), and
- has a maximum mass of not more than (e.g., 10) kg including batteries but without freight, a maximum permissible mass (including freight) not exceeding (e.g., 20) kg, a maximum design speed of not more than (e.g., 6 km/h) and a total height/width/length/ of xyz. (criteria 3 = technical specifications).”

While the term “motor vehicle” is already internationally defined in Art 1 p of the Road Traffic Convention of 1958” (further: The 1958 Convention) [31], being a “power-driven vehicle which is normally used for carrying persons or goods by road or for drawing, on the road, vehicles used for the carriage of persons or goods,” the definition of autonomous or partially autonomous steering devices requires clarification.

Adapting Regulations for the Needs of Delivery Robots

An essential criteria permitting motorized wheelchairs to operate in public traffic (which includes pedestrian areas) is their conformity with the general principle “Every moving vehicle or combination of vehicles shall have a driver”, as stated in art. 8 par. 1 of the 1958 Convention; they do also comply with art. 8 par 5 “Every driver shall at all times be able to control his vehicle or to guide his animals”, and art 13 par 1 “Every driver of a vehicle shall in all circumstances have his vehicle under control so as to be able to exercise due and proper care and to be at all times in a position to perform all manoeuvres required of him”.

As all autonomously driven vehicles—i.e., vehicles which are not constantly monitored by the driver—are thus inadmissible according to the provisions of the 1958 Convention, the Working Party on Road Traffic Safety (WP.1), which is responsible for the regulation of these issues for the United Nations Economic Commission for Europe, has decided [32] in their 68 meeting (24 to 26 March 2014) to propose to adapt the 1958 Convention to the needs of automated traffic by supplementing art 8 of the 1958 convention with an additional paragraph 5b is, which states that

“Vehicle systems which influence the way vehicles are driven shall be deemed to be in conformity with paragraph 5 of this Article and with paragraph 1 of Article 13, when they are in conformity with the conditions of construction, fitting and utilization according to international legal instruments concerning wheeled vehicles, equipment and parts which can be fitted and/or be used on wheeled vehicles. Vehicle systems which influence the way vehicles are driven and are not in conformity with the aforementioned conditions of construction, fitting and utilization, shall be deemed to be in conformity with paragraph 5 of this Article and with paragraph 1 of Article 13, when such systems can be overridden or switched off by the driver.”

If transport robots are intended to operate in public traffic, then they would have to comply with these criteria as well. A definition of criterion 1 would thus have to either refer to 5 bis of the 1958 Convention or implement these definitions directly.

Against this background, delivery robots could be defined as follows:

“A transport robot is an autonomously or partially autonomously electrically driven motor vehicle, which is designed for the transport of goods, and has a maximum mass of not more than (e.g., 10) kg including batteries but without freight, a maximum permissible mass (including freight) not exceeding (e.g., 20) kg, a maximum design speed of not more than (e.g., 6 km/h) and a total height/width/length/ of xyz. A motor vehicle shall be seen as autonomously or partially autonomously operated, when its steering systems are in conformity with the conditions of construction, fitting and utilization according to international legal instruments concerning wheeled vehicles, equipment and parts which can be fitted and/or be used on wheeled vehicles. Vehicle systems which influence the way vehicles are driven and are not in conformity with the aforementioned conditions of construction, fitting and utilization, shall be deemed to be in conformity with this Article, when such systems can be overridden or switched off by the driver.”

At present, a respective adaption of national traffic laws has not taken place yet, but various States will implement the UN’s Working Party on Road Traffic Safety’s in near future, and they will define delivery robots in very similar (if not identical) terms, as proposed above, making delivery robots objects to public traffic laws as well. But even as by definition until then delivery robots will not be included in public traffic law, judges do have to the discretion—provided that their respectively applicable national traffic law provides for a sufficiently broad definition of vehicles—to include delivery robots onto the scope of liability of present-day public traffic law.

Transport companies or sellers directly delivering their goods themselves should thus be aware of an eventual strict liability under public traffic law applying on delivery robots already today and take measures by addressing local traffic authorities and asking them to clarify the “liability status” of delivery robots in the receptive jurisdiction. In the case of coverage of delivery robots by the respective traffic law, they should be aware of the risk of strict liability, and, if they do wish to take that risk, take preparative measures as e.g., insuring themselves for this liability.

5.2. Delivery Robots and EU Data Protection

The information that is collected by design by most delivery robots (Starship robots, for instance, are equipped with six cameras) for various purposes—eventual accident documentation, building up maps of efficient delivery trajectories and the like—is of considerable commercial value, not only to the user of delivery robots, but also to state authorities, competitors, or the producer of delivery robots seeking to improve their product development; data protection is thus one of the central legal issues for delivery robots.

In 2016, the European Commission, the European Parliament, and the Council of the European Union approved the General Data Protection Regulation [33], which entered into force on 25 May 2018 and replaces the Data Protection Directive of 1995 [34]. The General Data Protection Regulation (GDPR) aims to strengthen and unify data protection for all individuals within the European Union and addresses especially the export of personal data to countries outside the EU. One important highlight of the GDPR is its endeavour to “return control” to citizens and residents over their personal data and to harmonize the regulatory framework for international business by unifying the regulation within the EU. As an EU regulation, the GDPR applies directly in all EU member States, i.e., it does not require national governments to pass any enabling legislation.

The key term of the GDPR is personal data that are considered to be “sensitive” under the condition that they revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership, genetic data, biometric data processed solely to identify a human being, health-related data, and data concerning a person’s sex life or sexual orientation ([33], p. 679, Article 4(13)–(15); Article 9; Recitals (51)–(56)). The GDPR defines ‘personal data’ as any information relating to an identified or identifiable natural person (‘data subject’). An identifiable natural person is any person who can be identified, directly or indirectly, in particular, by reference to an identifier, such as a name, an identification number, location data, an online identifier, or to one or more factors that

are specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person. In addition to that, a catalogue of examples for “personal data” provides examples of information relating to an individual, whether it relates to his or her private, professional or public life, e.g., name, home address, photographs, e-mail address, bank details, posts on social networking websites, medical information, or a computer’s IP address [35].

This personal data must be processed fair, lawful and transparent, whereas consent of the data subject is the main (but not only) criteria for lawfulness and also the core principle of data processing in general: “The controller shall be able to demonstrate that the data subject has consented to the processing of his or her personal data and the data subject shall have the right to withdraw his or her consent at any time” (Article 7), and the content has to have been provided explicitly, i.e., not inferred by mere implied behaviour.

Non-compliance with the strict data protection rules can cause severe penalties of up to 4% of the global turnover of a company or €20 Million ([33], Article 83). Under GDPR, organizations in breach of GDPR can be fined up to 4% of annual global turnover or €20 Million (whichever is greater). This is the maximum fine that can be imposed for the most serious infringements e.g., not having sufficient customer consent to process data or violating the core of Privacy by Design concepts. There is a tiered approach to fines e.g., a company can be fined 2% for not having their records in order ([33], Article 28), not notifying the supervising authority and data subject about a breach or not conducting impact assessment. Besides, also individuals may bring civil actions additional to measures taken by state authorities against violators.

The GDPR differentiate between the “data subject”, the “controller”, and the “processor”. The EU resident who represents the client of the delivery takes the role of a “data subject”. In order to clarify the data controller and data processor in the case of the delivery robot it is necessary to refer to Article 4 of the GDPR that defines a ‘controller’ as the “natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law”; whereas the ‘processor’ means a “natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller” ([33], Article 4).

By applying Article 4 to the case of the delivery robot that distributes e.g., pizzas for Mario’s Pizzeria, the client who ordered and receives the pizza represents the data subject, while Mario’s Pizzeria that uses the delivery robot for distributing the pizza to the client represents the data processor. If it is now assumed that Mario’s Pizzeria subcontracted for the delivery of their pizzas via delivery robots Starship Technologies, i.e., Starship Technologies owns and controls the delivery services, then Starship Technologies is the data controller in the sense of GDPR. This distinction is important for compliance considerations, as GDPR treats the data controller as the principal party for responsibilities, such as collecting consent, managing consent-revoking, enabling right to access, and other things. Thus, a data subject who wishes to revoke consent for his or her personal data will therefore contact the data controller to initiate the request, even if such data is stored on the servers of the data processor. In the case of such a request, the data controller has then to forward the request to the data processor in order to remove the revoked data from its server. In doing so, GDPR applies to all processes, irrespective of whether the organization is located inside or outside EU, and it introduces direct obligations for data processors as well as the situation to be subject to penalties and civil claims. This represents an important difference to the old Directive that only holds data controllers liable for data protection noncompliance. Thus, by recalling again Article 28(1), data controllers, i.e., customers of data processors, should only choose processors that comply with the GDPR in order to avoid penalties themselves.

Applying the GDPR on autonomous delivery robots, a first controversial issue arises in terms of the personal data collected and transmitted during the last-mile-delivery of such robots. As in any other delivery process as well, personal data of the client are necessary to fulfil the 6R of logistics, i.e., to bring

the right product, at the right time, in the right quantity and quality, to the right destination with the right costs [27]. The corresponding personal data include the address, financial, and biographical data plus personal consumption data that result from the business relationship with the client. Anyway, the sensitive data concerning the GDPR are less than those data that are needed and collected to steer the autonomous delivery robot from the starting point of the delivery to the final destination; simple address specifications are a precondition to contract performance, and its collection and storage does thus not violate the GDPR, as it matches the purpose limitation. More problematic are pictures, sound recordings and films taken by delivery robots in order to provide evidence in case of eventual accidents in which the robots were inflicted—material that inevitably also contains visual and audio information on human individuals moving in the direct vicinity of the robots. These data are collected in public spaces, and these photos, sound recordings and video sequences of natural persons are considered as “personal data” by the GDPR. These data are exchanged via internet and telecommunication networks, before they are partly considered and analysed by control personal and their IT-systems. Later, the data is stored in databases of the delivery control centres of companies.

These robots could also violate Article 25 of the Regulation, which calls for the implementation of privacy by design or privacy by default (PbD).

Privacy by default means that data controllers have to implement appropriate and technical measures in order to ensure that, by default, only personal data necessary (and at the necessary amount, period of storage, and accessibility) for the respective specific purpose of processing are processed. Article 23 supplements this principle by the duty to ensure that, by default, this personal data is not accessible without individual intervention to an indefinite number of natural persons. Appropriate measures are mentioned in Article 28(1) to provide “sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of the regulation and ensure the protection of the rights of the data subject”. Article 32 continues demanding the “Security of processing” by “taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons. These objectives shall be implemented by appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:

- (a) the pseudonymisation and encryption of personal data;
- (b) the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;
- (c) the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident; and,
- (d) a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing.

Unfortunately, Article 32 of the GDPR is not very clear by defining suitable technical and organizational measures that a company should adopt to comply with the regulation. But, in order to supervise the compliance within organizations a Data Protection Officer has to be appointed who shall be involved in all issues relating to the protection of personal data and who shall work independently, monitor the compliance with the GDPR, report to the highest management level, is reachable by data subjects, and cooperate with the supervisory authority ([33], Article 37–39).

Once data falls into the scope of application of the GDPR, the regulation provides strict instructions on how these data may be used. As the autonomous delivery robot itself as a device collects, processes and transfers user data article 25 of the GDPR concerning data protection by design and by default, i.e., the autonomous robot system has to take appropriate technical and organisational measures for “ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall

ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons" ([33], Article 25, Recitals 78). Consequently, the producer of the delivery robot has to safeguard that data protection measures have been taken, e.g., pseudonymization of personal data by the controller in an early stage of data collection, and as the communication between the robot and the remote control centre is executed via wireless links, the personal data (including photos and video sequences) have to be encrypted. Secondly, the data collection of the robot must be limited to what is necessary and transparency has to be safeguarded "with regard to the functions and processing of personal data in order to enable the data subject to monitor the data processing and the controller to create and improve security features" ([33], Recital 78). This requires that all obtained user data must be accessible and portable in order to enable any EU resident assuming that his personal data were collected by the robot (i.e., photos and videos) is provided with the possibility to request these data in a widely-compatible format, enabling him to verify which data exactly have been obtained. Thirdly, "the principles of data protection by design and by default should also be taken into consideration when developing, designing, selecting, and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations" ([33], Recital 78).

But, it is not only the producer, processor, and the controller of the delivery robot who may be held liable for GDPR violations; also telecommunication service providers the processor and/or the controller may make use of in order to transmit data from the delivery robot via the telecommunication service provider's network has to comply with the GDPR, i.e., take respective technical protection measures and store this data only within the limits of Art 25 GDPR.

Starship's Regional Business Manager for Central Europe, Hendrik Albers, recently addressed the GDPR explicitly in the context of innovative disruptions [22]. Albers warns to not over-regulate the European data protection regime and proposes to create a feasible balance between innovation and privacy for consumers. In the case of Starship, the company has according to Hendrik Albers developed very precise routines to ensure that this balance is kept. In Albers' opinion, most companies that collect customer data do so either way rather in order to benefit the customer than to market client data in order to generate profits. He thus emphasizes a privacy approach that leaves sufficiently large freedom to companies. In the case of delivery services he points out that there is an essential need to know where the customer is located in order to deliver the goods as close as possible to the customer. In addition to that, the delivery service also requires to be informed about several personal details in order to provide for an efficient organization of the delivery of items, as the customer would not be able to receive the freight if the regulation excessively prohibited the collection of one of these essential parameters.

While it may not be surprising that Albers takes a rather liberal approach on (not) subsuming Starship robots' activities under the GDPR, it has to be admitted that, indeed, in the case of delivery services, especially in the case of autonomous delivery robots that are partly remote-controlled via telecommunication networks, the main focus in area of privacy is still on the customer data, which happens to be the least controversial aspect. The collected personal data that are collected by the sensors, microphones, and cameras, and which are transferred via telecom links are until now not on the top of the agenda—in spite of obvious violations of the GDPR by many default technical data collection settings.

The organization environment of the delivery robot control must be able to demonstrate compliance with the GDPR, i.e., the data controller should implement measures that meet the principles of data protection by design and data protection by default. Furthermore, the data controller is responsible to implement effective measures and it must be able to demonstrate the compliance of processing activities even if the processing is carried out by a data processor on behalf of the controller.

Article 25 states that Data Protection Impact Assessments have to be conducted when specific risks occur to the rights and freedoms of data subjects, and Articles 37–39 state that Data Protection Officers have to ensure compliance within organizations. In the case of non-compliance of these three main rules, strict penalties apply, starting with 20 Mio € and reaching up to 4% of the company's global turnover. In addition to that, one should keep in mind that there is no grace period, i.e., the GDPR is in full effect since 25 May 2018.

6. Findings and Discussion

Delivery robots as part of the last-mile B2C-distribution raise currently considerable attention and represent a growing business sector that is driven by traditional logistics service providers, but also by a number of start-ups that are located all around the globe. The existing devices are still in the test phase, and the considered cases show that the main area of operation is in food, flower, and grocery business where the robots are charged and unloaded by humans. The case of Starship Technologies reveals that the delivery robots can be considered as cyber-physical systems (CPS), since they are self-organised, self-optimized, and internet-linked—but they are autonomous only up to 90%, i.e., full self-organization is still a future issue [3].

Other important features of Industry 4.0 publications are related to internet-based linked machine-to-machine-communication and interaction as well as the ability to get integrated into cross-company processes safeguarding the capability to operate in a networked manufacturing and logistics environment [3–5]. Here, research shows that M2M-technologies are today only partly realized—e.g., in the RoboVan solution, where the van acts as a hub that communicates with the delivery robots as feeders. But, a self-guided and M2M-based organization of delivery robots that integrate themselves into the full supply chain and realize autonomously the last-mile of the delivery without media discontinuity, i.e., without intervening of human work-force, is still to come. A benchmark for such a system is the pilot project “AMATRAK” at ISL Bremen, which realized a self-guided container transportation system, where containers are able to choose and book suitable and optimal transportations means, according to their own needs [2].

The competitive advantage of autonomous delivery robots as compared to other delivery modes is the low cost of less than 1€ per unit and delivery, which makes them up to 15 times cheaper than traditional delivery services. Their limited delivery radius, together with the fact that land-based delivery robots have to share the sidewalk with pedestrians and other traffic, make their preferred area of operation suburbs and low-density traffic areas, which makes them a delivery service mainly supplementing the existing ones.

Another important aspect which has not been in the focus of discussion in robot-friendly circles is the question to which degree society would in fact be ready to accept an excessive use of delivery robots. The shared use of sidewalks between delivery robots and pedestrians cause already today considerable acceptance problems in some places, which are expressed in different legal frame conditions, depending on the location. Some of them can seriously endanger the business model of delivery robots: Whereas, e.g., Estonia already has adapted its traffic laws for the shared use of space for humans and robots (see reform act on Estonian traffic act from 14 June 2017 on amendments of Section 2 of the same act) [36], other countries are still hesitating. A closer look to the USA reveals that not all parts of society welcome sharing sidewalks with robots by nature. Currently, a number of United States (USA) States allow for robots to participate in the traffic and adapted accordingly their state traffic laws. At the same time, within some States certain cities or municipalities formulated their own traffic law concerning robots, which makes the USA a much diversified legal patchwork with changing and partly contradicting laws for robot operations in traffic. Recently, the case of San Francisco's anti-robot laws gained extensive media coverage when they banned autonomous delivery devices from most sidewalks entirely and permitted them only in low-foot traffic zones [37]: While in those places where few specimen roaming around on selected cities' walkways today are well respected and are observed with curiosity, the public perception is starting to deteriorate in some cities with a higher “population” of delivery robots—for

instance, in San Francisco the local government passed in early December 2017 strict regulations on delivery robots, capping permissions for robots “at three per company, and nine total at any given time for the entire city. The robots will now only be allowed to operate within certain industrial neighbourhoods, on streets with 6 ft-wide sidewalks, and must be accompanied by a human chaperone at all times” [38]. The city reacted this way to protests by a “coalition of residents, pedestrian advocates, and activists for seniors and people with disabilities”, claiming that “sidewalks are not playgrounds for the new remote controlled toys of the clever to make money and eliminate jobs” [38].

Besides this ongoing regulatory discussion around the world, another important frame condition is dedicated to the allowed weight of delivery drones. Both US States Virginia and Idaho allow for robots to operate autonomously, but, in Virginia, the law allows for land-based robots to operate up to a weight of less than 50 pounds, in Idaho the legal weight limit is 80 pounds. These discussed examples point out that delivery robots face a scattered landscape of legal regulations even within individual nations, which makes their operation decisively more difficult.

Finally, the new EU data protection regulation formulates new challenges for the development and operation of delivery robots. The considered cases disclose that, until now, data protection issues are not ranging in the top of agenda of the delivery robot world. But, since the new European General Data Protection Regulation took effect on 25 May 2018, a huge set of data necessary to operate a delivery robot have to be considered as personal data that are not only locally processed in the robot, but are also transferred and stored via internet links. Consequently, the applicable new data protection rules have to be taken into account in the design of the robots. Although partly, there may be a legitimate interest of the user/controller for collecting and processing in order to prevent harm to bystanders and to maintain integrity of the robot, interviews with the management and developers [22] have shown that there is little to no awareness that any collection of personal data from any human individual in the vicinity of the moving robots beyond the absolute necessary (PbD) requires the explicit consent from the respective individual, which is practically impossible to obtain, meaning that the GDPR is generally violated by delivery robots that are collecting this information.

This research show that the new EU General Data Protection Regulation requires a higher level of attention in the whole sector of delivery robots.

7. Conclusions

Delivery robots by design seem to provide the “missing link” between wholesale logistics and the consumer, and expectations that they will considerably contribute to solve the last-mile-problem in near future are well-founded. The current technical solutions are realizing only partly the Industry 4.0 concepts, but a closer view to funding and growth indicators reveal that the whole robot sector is highly dynamic and it represents a strongly growing market for the upcoming years, especially against the background of ecologically friendly logistics (e.g., via green transport corridors) and the combination of delivery robots and artificial intelligence [25,39,40]. Anyway, excessive enthusiasm falsifies the perception of delivery robots when it comes to implementation of these technologies into existing legal frameworks.

This paper intended to shed some light on two aspects where major challenges will arise in future (and partly even today), being strict liability for accidents that are caused by delivery robots under traffic law and considerable penalties in case of violations of GDPR requirements by delivery robots’ data collection and transmission mechanisms—risks an entrepreneur deciding in favour of making use of delivery robots may not have taken into account so far.

Another, less legal aspect that may impede the future success of delivery robots as business model is the question how much society—and municipal governments—will indeed welcome an excessive use of pedestrian walkways by delivery robots. The related legal framework which evolves around the sector of delivery robots represents a patchwork of different rules on national, regional and municipality level, making it complicated to realize the competitive advantage of the business model of the delivery robots for the last-mile.

While it has to be conceded that all great novel technologies have so far faced an initial hype together with massive initial protests before they were broadly accepted later, an interested entrepreneur has to be aware that some perseverance may be required once this business model is chosen, i.e., that legal restrictions in close future could rather hinder than facilitate the use of delivery robots. The research gives an empirically validated insight in the current developments in the sector of delivery robots, but by taking into account the high dynamic and innovative character of the whole sector, the picture can only give an actual snapshot of the evolution of delivery robots.

Author Contributions: Conceptualization, Investigation, Methodology, Validation, Writing—original draft, T.H. and G.P. The authors contributed equally.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Prause, G.; Atari, S. On sustainable production networks for Industry 4.0. *Int. J. Entrep. Sustain. Issues* **2017**, *4*, 421–431. [CrossRef]
2. Prause, G. Sustainable business models and structures for industry 4.0. *J. Secur. Sustain. Issues* **2015**, *5*, 159–169. [CrossRef]
3. Kagermann, H.; Wahlster, W.; Helbig, J. *Recommendations for Implementing the Strategic Initiative INDUSTRY 4.0*; National Academy of Science and Engineering: Berlin, Germany, 2013.
4. Bauer, W.; Schlund, S.; Marrenbach, D.; Ganschar, O. *Industry 4.0—Volkswirtschaftliches Potenzial für Deutschland*; BITKOM: Berlin, Germany, 2014. (In German)
5. Brettel, M.; Friederichsen, N.; Keller, M.; Rosenberg, M. How Virtualization, Decentralization and Network Building Change the Manufacturing Landscape: An Industry 4.0 Perspective. *Int. J. Mech. Ind. Sci. Eng.* **2014**, *8*, 37–44.
6. Belussi, F.; Sedita, S. Managing the fragmented value chain of global business: Exploitative and explorative offshoring toward emerging market economies. In *The Past, Present and Future of International Business & Management, Advances in International Management*; Devinney, T., Pedersen, T., Tihanyi, L., Eds.; Emerald Group Publishing Limited: Bingley, UK, 2010; Volume 23, pp. 399–429.
7. Dujin, A.; Geissler, C.; Horstkötter, D. *Industry 4.0: The New Industrial Revolution*; Roland Berger Strategy Consultants: Munich, Germany, 2014.
8. Prause, G. E-Residency: A business platform for Industry 4.0? *Entrep. Sustain. Issues* **2016**, *33*, 216–227. [CrossRef]
9. Kersten, W.; Seiter, M.; von See, B.; Hackius, N.; Maurer, T. *Trends und Strategien in Logistik und Supply Chain Management*; Bundesvereinigung Logistik (BVL): Bremen, Germany, 2017; ISBN 978-3-87154-607-5.
10. Punakivi, M.; Yrjölä, H.; Holmström, J. Solving the last-mile issue: Reception box or de-livery box? *Int. J. Phys. Distrib. Logist. Manag.* **2001**, *31*, 427–439. [CrossRef]
11. Laseter, T.; Houston, P.; Chung, A.; Byrne, S.; Turner, M.; Devendran, A. The last-mile to nowhere. *Strategy Bus.* **2000**, *20*, 40–49.
12. Rodrigue, J.-P.; Comtois, C.; Slack, B. The “Last-Mile” in Freight Distribution. In *The Geography of Transport Systems*, 2nd ed.; Routledge: Abingdon-on-Thames, UK, 2009; p. 212, ISBN 978-0-415-48323-0.
13. Lee, H.; Whang, S. Winning the Last-mile of E-Commerce. *MIT Sloan Manag. Rev.* **2001**, *42*, 54–62.
14. Song, L.; Cherrett, T.; McLeod, F.; Guan, W. Addressing the Last-mile Problem: Transport Impacts of Collection and Delivery Points. *Trans. Res. Rec.* **2009**, 9–18. [CrossRef]
15. Boyer, K.K.; Prud’homme, A.M.; Chung, M. The last-mile Challenge: Evaluating the effects of customer density and delivery window patterns. *J. Bus. Logist.* **2009**, *30*, 185–201. [CrossRef]
16. Cha, I.; Shah, Y.; Schmidt, A.; Leicher, A.; Meyerstein, M. Trust in M2M Communication—Addressing New Security Threats. *IEEE Veh. Technol. Mag.* **2009**, *4*, 69–75. [CrossRef]
17. Wu, G.; Talwar, S.; Johnsson, K.; Himayat, N.; Johnson, K. M2M: From Mobile to Embedded Internet. *IEEE Commun. Mag.* **2011**, *49*, 36–43.
18. Zhang, Y.; Yu, R.; Xie, S.; Yao, W.; Xiao, Y.; Guizani, M. Home M2M Networks: Architectures, Standards and QoS Improvement. *IEEE Commun. Mag.* **2011**, *49*, 44–52. [CrossRef]

19. Maurer, M.; Gerdes, C.; Lenz, B.; Winner, H. *Autonomes Fahren—Technische, Rechtliche und Gesellschaftliche Aspekte*; Springer: Berlin, Germany, 2015; ISBN 978-3-662-45853-2. (In German) [CrossRef]
20. SBS. *Technological Disruption and Innovation in Last-Mile Delivery—White Paper*; Stanford Business School, Stanford University: Stanford, CA, USA, 2016.
21. Olaniyi, E.O.; Reidolf, M. Organisational Innovation Strategies in the Context of smart Specialisation. *J. Secur. Sustain. Issues* **2015**, *5*, 213–227. [CrossRef]
22. Starship. Data Protection? We Need a Feasible Balance between Business and Privacy. Available online: <https://www.european-business.com/starship-technologies/interviews/data-protection-we-need-a-feasible-balance-between-business-and-privacy/> (accessed on 19 May 2018).
23. CBinsight. The Robotics Startup Funding Landscape Broken Down in One Infographic. Available online: <https://www.cbinsights.com/research/robotics-deals-consumer-enterprise-medical/> (accessed on 19 May 2018).
24. International Data Corporation. *Worldwide Semiannual Robotics and Drones Spending Guide*; International Data Corporation: Framingham, MA, USA, 2017.
25. Basu, S.; Omotubora, A.; Beeson, M.; Fox, C. Legal Framework for Small Autonomous Agricultural Robots. Available online: <https://link.springer.com/content/pdf/10.1007%2Fs00146-018-0846-4.pdf> (accessed on 10 July 2018).
26. Blaxter, L.; Hughes, C.; Tight, M. *How to Research*, 3rd ed.; McGraw-Hill Education: London, UK, 2006; p. 672.
27. Seeck, S. *Erfolgsfaktor Logistik*; Springer: Berlin, Germany, 2010. (In German)
28. Heinla, A. An Interview with Ahti Heinla of Starship Technologies. 27 July 2017. Available online: <https://www.tharsus.co.uk/an-interview-with-the-genius-behind-starship-technologies-ahiti-heinla/> (accessed on 19 May 2018).
29. Straßenverkehrs-Ordnung. Available online: https://www.gesetze-im-internet.de/stvo_2013/ (accessed on 10 July 2018).
30. Verordnung über die Zulassung von Personen zum Straßenverkehr (Fahrerlaubnis-Verordnung—FeV). Available online: https://www.gesetze-im-internet.de/fev_2010/BJNR198000010.html (accessed on 10 July 2018).
31. Road Traffic Convention of 1958. Available online: <http://www.unece.org/fileadmin/DAM/trans/conventn/crt1968e.pdf> (accessed on 19 May 2018).
32. See Report of the Sixty-Eighth Session of the Working Party on Road Traffic Safety, Annex 1. Available online: <http://www.unece.org/fileadmin/DAM/trans/doc/2014/wp1/ECE-TRANS-WP1-145e.pdf> (accessed on 19 May 2018).
33. Voigt, P.; Bussche, A.V.D. *General Data Protection Regulation*; Springer: Berlin, Germany, 2016; p. 679.
34. Directive 95/46/EC on the Protection of Individuals with Regard to the Processing of Personal Data (Data Protection Directive). Available online: <http://www.wipo.int/wipolex/en/details.jsp?id=13580> (accessed on 10 July 2018).
35. EUGDPR 2018, Data Protection in the EU. Available online: <https://www.eugdpr.org/> (accessed on 19 April 2018).
36. Estonian Traffic Act. Available online: <https://www.riigiteataja.ee/en/eli/516022016004/consolide> (accessed on 10 July 2018).
37. Wired. San Francisco Just Put the Brakes on Delivery Robots. Available online: <https://www.wired.com/story/san-francisco-just-put-the-brakes-on-delivery-robots/> (accessed on 19 May 2018).
38. Wong, J. San Francisco Sours on Rampant Delivery Robots: Not Every Innovation Is Great. Available online: <https://www.theguardian.com/us-news/2017/dec/10/san-francisco-delivery-robots-laws> (accessed on 19 May 2018).
39. Prause, G.; Hoffmann, T. Cooperative Business Structures for Green Transport Corridors. *Balt. J. Eur. Stud.* **2017**, *7*, 3–27. [CrossRef]
40. Rault, R.; Trentesaux, D. Artificial Intelligence, Autonomous Systems and Robotics: Legal Innovations. In *Service Orientation in Holonic and Multi-Agent Manufacturing*; Borangiu, T., Trentesaux, D., Thomas, A., Cardin, O., Eds.; Springer: Berlin, Germany, 2018; Volume 762.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

MDPI
St. Alban-Anlage 66
4052 Basel
Switzerland
Tel. +41 61 683 77 34
Fax +41 61 302 89 18
www.mdpi.com

MDPI Books Editorial Office
E-mail: books@mdpi.com
www.mdpi.com/books



We are currently witnessing a revolution in production systems that comes as a consequence of the massification of automation and of the presence of autonomous robots and AI (artificial intelligence) in all productive sectors, in the field of war and security, and at all social levels. This book deals with several issues and problems within the field of Robotics, Artificial Intelligence and Law. In fact, it attempts to address the "dangerous liaisons" between them and their implications for modern Societies. Its structure is designed to function as a work and consultation tool for the students of the Master of Law of the UMSNH in the seminars on Law and New Technologies and Robots, A.I, Human Rights and Transhumanism.