Leikas, Jaana; Koivisto, Raija A.; Gotcheva, Nadezhda

Article

Ethical framework for designing autonomous intelligent systems

Journal of Open Innovation: Technology, Market, and Complexity

Provided in Cooperation with:
Society of Open Innovation: Technology, Market, and Complexity (SOItmC)

Suggested Citation: Leikas, Jaana; Koivisto, Raija A.; Gotcheva, Nadezhda (2019) : Ethical framework for designing autonomous intelligent systems, Journal of Open Innovation: Technology, Market, and Complexity, ISSN 2199-8531, MDPI, Basel, Vol. 5, Iss. 1, pp. 1-12, https://doi.org/10.3390/joitmc5010018

This Version is available at:
https://hdl.handle.net/10419/241309

Journal of Open Innovation: Technology, Market, and Complexity

MDPI

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 utilitarism 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 rightsBeneficence: "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-Decembe-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/~{}ear/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, Fenland, 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*; Zelta, 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*; Zelta, 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.