Artificial intelligence and social responsibility: the case of the artificial intelligence strategies in the United States, Russia, and China

Anton Saveliev and Denis Zhurenkov
Philosophy of science and technology, Institute of Philosophy, Russian Academy of Sciences, Moscow, Russian Federation

Received 31 January 2020
Revised 3 September 2020
29 September 2020
Accepted 19 October 2020

Abstract

Purpose – The purpose of this paper is to review and analyze how the development and utilization of artificial intelligence (AI) technologies for social responsibility are defined in the national AI strategies of the USA, Russia and China.

Design/methodology/approach – The notion of responsibility concerning AI is currently not legally defined by any country in the world. The authors of this research are going to use the methodology, based on Luciano Floridi's Unified framework of five principles for AI in society, to determine how social responsibility is implemented in the AI strategies of the USA, Russia and China.

Findings – All three strategies for the development of AI in the USA, Russia and China, as evaluated in the paper, contain some or other components aimed at achieving public responsibility and responsible use of AI. The Unified framework of five principles for AI in society, developed by L. Floridi, can be used as a viable assessment tool to determine at least in general terms how social responsibility is implied and implemented in national strategic documents in the field of AI. However, authors of the paper call for further development in the field of mutually recognizable ethical models for socially beneficial AI.

Practical implications – This study allows us to better understand the linkages, overlaps and differences between modern philosophy of information, AI-ethics, social responsibility and government regulation. The analysis provided in this paper can serve as a basic blueprint for future attempts to define how social responsibility is understood and implied by government decision-makers.

Originality/value – The analysis provided in the paper, however general and empirical it may be, is a first-time example of how the Unified framework of five principles for AI in society can be applied as an assessment tool to determine social responsibility in AI-related official documents.

Keywords Artificial intelligence, Social responsibility, National strategies, Ethics, Social responsibility, National strategies

Paper type Research paper

1. Introduction

Artificial intelligence (AI) is often understood as a double-edged sword of modern science – a technology whose destructive potential is virtually unlimited without proper guidance from its creator. For example, such prominent scientist and public intellectuals as Stephen Hawking and Martin Rees, as well as innovations entrepreneur Elon Musk and AI-researcher Stuart Russel have quite eloquently

The Authors would like to thank Dr. Igor Perko, Dr. Francesco Caputo and Dr. Teodora Ivanusa for their helpful advice on various technical and philosophical issues, examined in this paper. The Authors, however, bear full responsibility for the Paper.

0368-492X
DOI 10.1108/K-01-2020-0060

spoken about the disruptive power the AI possesses, mentioning above all the risks of total annihilation of mankind, should strong AI-technology go rogue, or fall into wrong hands. Public perception of AI-related risks spreads even further, calling for collective action.

In 2009, leading AI-researchers (Horvitz and Selman, 2009) voiced their rising concern during the Asilomar conference, which eventually led to the signing of an open letter (Future of Life Institute, 2015), and the creation of Asilomar AI Principles (Future of Life Institute, 2017) – an assortment of 23 guidelines, outlining AI developmental issues and ethics for the development of beneficial AI. On the other hand, potential dangers of weaponized AI-applications were openly discussed at a leading AI conference, the Association for the Advancement of Artificial Intelligence (AAAI) 2015, and at a workshop on AI and ethics of the same conference. The international community (UNESCO, 2018; Floridi and Cowls, 2019) also shared a certain mistrust of AI, but in general agreed that it is difficult to develop and formalize any clear guidelines to avoid major risks: the AI technology is too complex.

All this indicates that the AI is no longer considered a purely technological phenomenon, but rather a social one. While the possible impact of AI-related technologies on modern society is yet unknown, researchers, enthusiasts and political leaders are still struggling to define a framework, which will ensure the utilization of AI for the benefit of social responsibility. With the growing interest in AI, more and more countries are creating their national strategies to develop and harness this technological marvel of the 21st century. These pioneering documents often seem vague and uncertain, but still offer considerable insight on how AI will affect people's lives, safety and well-being, should governments in technologically advanced countries fully embrace this technology. This fact alone is worthy of thoughtful review and analysis, to find out, how these strategies define (if define at all) the use of AI for social responsibility. For this purpose, the authors have chosen official strategic documents in the field of AI development and implementation of the three technologically advanced countries – the USA, Russia and China. These documents will be examined through the methodology, based on the *Unified framework of five principles for AI in society*. Italian philosopher and researcher Luciano Floridi put forward the idea of basic ethical principles of AI, using his concept of *Philosophy of Information*. These principles now form the basis of the EU Ethics guidelines for trustworthy AI. In this paper, we are going to assess, how the Unified framework of five principles for AI in society can be applied for a better understanding of social responsibility's role in AI-related official documents. The choice of Luciano Floridi's theoretical contribution as a basis for this research methodology is justified by the fact that it allows us to consider AI as an autonomous agent, and thus enables us, at least as a part of the theoretical assumption, to use the concept of "responsibility" with respect to AI.

1.1 Conceptual background

Currently, the ethical and social impact of AI can be roughly divided into two distinct areas of study – the human perspective and the AI perspective. This distinction is not dictated by the primordial nature of AI a possibly alienated (in logic and behavior) entity, but rather by our incomplete understanding of its role. The human perspective is of paramount importance for us and will be briefly reviewed in this section.

John P. Sullins in his works argues that a state of moral agency can be ascribed to AI, especially if this AI fulfills a social role. In this case, it must possess a duty of care, which is only possible if it is a moral agent (Sullins, 2006). Sean O'Heigeartaigh defines the perils of moral outsourcing in connection to AI: human biases are numerous but they are relatively easy to grasp and to correct. A machine error, on the contrary, can lead to unthinkable

catastrophic consequences (O'Heigeartaigh, 2013). Even more so, certain researchers believe, that moral outsourcing cannot be chastised even by the means of strict algorithmic accountability (Pavaloiu and Kose, 2017). In this regard, P. Boddington points out that virtue ethics and Kantian morality do not allow for the outsourcing of moral judgments, since another machine, however, sophisticated, would be unable to do the right things for the right reasons and in the right manner (Boddington, 2017). On the other hand, some scientists advocate the delegation of morality to computer systems. Jos de Mul argues that computer systems can act as important moderators of human agency, enhancing our moral autonomy (de Mul, 2010). Jeffrey K. Gurney explores AI ethics through the framework of a crash-optimization algorithm, in which the algorithm author allows the autonomous machine to decide who and what to hit under certain conditions, dictated by classical moral dilemmas (Shopping Cart Problem, Motorcycle Problem, The Car Problem, etc.) (Gurney, 2016).

Issues of socially responsible application of AI are addressed by J. Borenstein and Y. Pearson. Using a capabilities approach analysis, they argue, that robots can indeed enhance the caregiving process and empower caretakers with more freedom in their life (Borenstein and Pearson, 2010). Sharkey (2008), however, points out the dangers of this approach: caretakers can harbor possibly unhealthy attachment to inanimate entities: i.e. elderly, left in the exclusive care of machines, would be deprived of the human contact. J. Kofas proclaims that AI will significantly impact the sense of identity and community in society, by undermining the community culture and widening inequality in society. He argues, that AI and robotics will enhance living standards, but the most valued applications of these technologies will be available exclusively to wealthy classes (Kofas, 2017).

The ethical and social shortcomings of algorithms (which are the main pillars of AI) are addressed by Friedler *et al.* who examine the meaning of a "fair algorithm" on the basis of philosophy and computer science. Researchers provide a mathematical definition of "fairness", and demonstrate that fairness in output depends on the interactions between the construct space, observed space and the decision space of the algorithm (Friedler *et al.*, 2016). Anderson and Sharrock point out differences between algorithms and social actors: the former are bound by mathematical instructions, the latter can exercise discretion. Yet they believe that algorithms can be relied on to make satisfactory ethical decisions (Anderson and Sherrock, 2013).

Speaking about the general design perspectives of AI ethics, Pavaloiu and Kose remind us about data interpretation bias, found in the top-down approach: e.g. the AI might indirectly infer if an individual is depressed based on social media data and this can affect future employment of this individual (Pavaloiu and Köse, 2017). Allen *et.al* state that the top-down approach permits the architect to tailor the capacity of the system in design. At the same time, there can be a possible conflict between the pre-encoded rules and the constant pressure to predict and compute outcomes for every action. To prevent this conflict we are going to need a universal standard for AI ethics since computers are still limited in their computational capacity and thus will struggle to find the right outcome for every complex ethical dilemma. Meanwhile, the bottom-up approach can be based on the rewards and punishments system, in which AI can be compared to a child brought up in a rough neighborhood. In conclusion, researchers call for a hybrid model which, while framing broad governing rules, would provide scope for learning through experience (Allen *et al.*, 2005).

There are certain developments in specialized applicable standards, such as for the protection of personal data (Heisenberg, 2005; Albrecht, 2016) or in robotics (Schlenoff, 2015; Noorman and Johnson, March 2014), as well as in the research on the legal personality of AI (Gadzhiev, 2018; Arkhipov and Naumov, 2017; Morkhat, 2018; Cerka *et al.*, 2017; Morkhat, 2017). In this regard, we would like to point out the works of the Russian Research Center for

the Problems of Regulation of Robotics and AI (Arkhipov and Naumov, 2017), which is drafting legislation in the field of robotics and AI (http://robopravo.ru).

While the above cited studies do indicate a growing interest on behalf of an academic community, one is forced to agree that a general and universally recognized legal or ethical system of AI-related risks control does not exist (Darnault, *et al.* August 2019). Nonetheless, this issue does not stop technologically advanced nations from creating and formalizing their own AI-development and utilization strategies, aimed to better adapt the AI technologies for the benefits of security and economic prosperity. Despite being state-mandated, these strategies are not restrictive, nor clearly defined, focusing on all AI-related issues and matters at once. In the absence of any specific AI-control laws of national or international origin, even these vague documents are still a great source of insight. AI is no longer just a promising technological gimmick, instead, it is perceived as a social, political, and history-changing phenomenon, easily comparable to the atomic bomb, the infamous brainchild of 20th century physics.

In addition, just like with the nuclear fissure, society is both scared of its power and eager to harness it in. This paper explores how responsibility to society, as well as any means to facilitate social responsibility, is defined (if defined at all) in national AI development strategies in three technologically advanced countries – the USA, Russia and China. To achieve this task, American, Russian and Chinese strategies will be reviewed within the combined framework of existing voluntary applicable standards in the field of AI and the theoretical basis of modern philosophy of information. As authors, we must state, that the choice of these three countries for our analysis is not justified by their ethical contribution to the seemingly different domains of social responsibility and AI. On the contrary, our decision was dictated by the interest in the development of AI technologies in these countries on the part of government authorities and major market players, which found its reflection in official documents.

USA and China are currently leading the AI race in terms of technological advance and market application, while Russia is but following in their steps, trying to emulate successful approaches set by both countries. While not possessing access to autonomous robotic systems, Russia lags in R&D and spending on AI technologies as well. The Russian government's future investment in AI research is unknown, Brookings Institution report estimates that the country spends approximately $12.5m a year on AI research, putting it far behind China's plan to invest $150bn through 2030. The US Department of Defense alone spends $7.4bn annually on unclassified research and development on AI and related fields (Brookings Institution, 2018). Nonetheless, despite these disparities, Russia is still capable to become a local leader in applied AI technologies due to relatively advanced IT-infrastructure and great fundamental scientific potential.

2. Theory and methodology

AI is an area of information technology that deals with the development of intelligent systems, i.e. systems possessing capabilities that are traditionally associated with the human mind – language understanding, learning, ability to argue, solve problems, etc. Moreover, the terminology of AI systems also includes the following concepts:

- Narrow AI, which has been available for decades and is still widely used today. Narrow AI systems are complex software programs that can execute discrete 'intelligent' tasks such as recognizing objects or people from images, translating language, or playing games. Narrow AI systems can execute complex calculations, but they are limited by the boundaries of their task, operating environment and specified programming (SIPRI, 2019).

- Autonomous AI, which provides the ability of the system to function for a long time without the participation of an operator.
- Adaptive AI, which implies the ability of the system to adapt to new conditions, acquiring knowledge that is not embedded in the creation.
- Artificial General Intelligence (AGI), which is characterized by such high adaptability that the system possessing it can be used in a variety of activities with appropriate training, both independent and directed (with the help of an instructor).
- Strong AI (Strong AI), Human-Level AI (Human-Level AI) with a level of adaptability comparable to the human level.
- Super-human AI (Super-human AI), which implies even higher adaptability and speed of learning.

Most of the aforementioned paradigms still satisfy the definition, proposed by John McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon in seminal Proposal for the Dartmouth Summer Research Project on AI (McCarthy *et al.*, 2006 [1955]):

> For the present purpose the artificial intelligence problem is taken to be that of making a machine behave in ways that would be called intelligent if a human were so behaving.

However, this classic definition is viewed by some researchers as counterfactual. Italian philosopher Luciano Floridi, whose theoretical findings are extensively used in this paper, states that:

> Just because a dishwasher cleans the dishes as well as (or even better than) I do does not mean that it cleans them like I do, or needs any intelligence to achieve its task (Floridi and Cowls, 2019).

Instead, Floridi provides his broad definition, which is nonetheless still compatible with the classical one, and conceptualizes AI as:

> [. . .] a growing resource of interactive, autonomous, and often self-learning agency [. . .], that can deal with tasks that would otherwise require human intelligence and intervention to be performed successfully. In short, AI is defined based on engineered outcomes and actions and so, in what follows, we shall treat AI as *a reservoir of smart agency on tap*[. . .] (Floridi, 2019).

This definition is highly relevant for this research, as it enables the use of "responsibility" concerning AI applications. Stanford Encyclopedia of Philosophy declares, that "In very general terms, an agent is a being with the capacity to act, and agency denotes the exercise or manifestation of this capacity" (Stanford Encyclopedia of Philosophy, 2019). Therefore, *if we embrace Floridi's definition of AI, we are allowed to empower AI and robotic systems with a limited ability to act autonomously* and even independently at least in the imagination of a scientist exploring the hypothetical reality. *This assumption is a cornerstone of this research*, which argues that responsibility (social or other) cannot be achieved without the ability to act without autonomy or independence. It is true that among the society of homo sapiens species there are individuals capable of independent actions, but wholly or partially devoid of responsibility for them, yet the legislative systems, currently employed in the world, do not possess the status of a responsible individual incapable of independent action.

Social responsibility is an ethical theory in which individuals are accountable for fulfilling their civic duty, and the actions of an individual must benefit the whole of society. ISO 26000, the international standard for social responsibility declares, that there must be a balance between economic growth and the welfare of society and the environment. If this equilibrium is maintained, then social responsibility is accomplished, and it will guide us towards greater/social good. It is a rather simple consequentialist approach, which leads us to a complex

dilemma. Consequentialism-based ethical constructs mandate us to increase the probability of good consequences, but they fail to define what these potential consequences are. In other words, the exact meaning of "socially responsible" and "good" depends solely on what the agent in question considers as such. However, such an approach could cause dangerous, if not life-threatening, problems when applied to AI. What if an AI system makes a decision, which is not universally recognized as "good" in nature by society? Should a self-driving car prioritize potentially expensive damage to public property and environment over the life of a single human in a dangerous situation? And what about "augmented intelligence" systems which involve an extensive amount of networked human-to-machine and machine-to-machine communications? What happens if such a system gets tampered with to produce potentially harmful effects?

These questions cannot be answered until we define the ontological and ethical status of objects (or subjects?), which possess or operate AI. L. Floridi's Philosophy of information defines *information object* as any and every part of animate and inanimate nature, everything that existed, exists or will exist (may exist), everything that somehow or other participates in communication interactions, ensures the emergence, transfer, or preservation of knowledge, that is, ensures the existence of the *infosphere*. The modern world is filled with information objects (robots, programs, engineering artifacts, etc.), which cannot be called alive, but they have high value and importance in the system of information exchange. In this regard, L. Floridi proclaims the highest value of information and information objects as the basis for new information ethics. Everything that is information and contributes to its multiplication has value: artifacts, distant stars, past and future generations, ancient civilizations, and modern computer programs. In this case, the absolute evil is any destruction of information and impoverishment of the infosphere (*entropy*, both in L. Floridi's and C. Shannon's terminology) (Floridi, 2002; Shannon, 1948).

In this regard, L. Floridi proclaims that information and information objects, inhabiting the infosphere, are to be considered as the ultimate value of macro-ethics (Floridi, 2008). Evil is everything that causes or promotes entropy, good is everything that stimulates the growth of the infosphere. A moral act is evaluated in terms of the well-being of the infosphere, and it can be performed even by an inanimate information agent. For example, a computer program that protects data and makes calculations in this sphere does a positive thing. Destruction of organized and coherent data and destruction of the entire information system is considered as a negative action. This philosophical view of the ontological and ethical status of information objects allows us to overcome the anthropocentric bias, without renouncing our human and biological origins, and to establish a valid connection with a great heritage of known and proven ethical systems. For this purpose, L. Floridi creates a concept of "inforg" – an informationally embodied organism, entity made up of information, that exists in the infosphere. These organisms are made up of information, thus they differ from natural agents with whom they coexist. This concept is quite similar to Norbert Wiener's view of organisms as entities defined by patterns of persisting "Shannon information" – information placed into the physical realm to be manipulated by the laws of nature and science (Wiener, 1961). This means that inforgs consist of matter, energy, and Shannon information. The concept of inforg in relation to the information is not only philosophical and cybernetical but biological as well. Recent experiments show, that it is possible to use DNA as data storage, meaning that binary helix can be encoded to hold binary information (Goldman *et al.*, 2013), which in its own turn reinforces the concept of living organisms as "persisting patterns of Shannon information encoded within an ever-changing flux of matter-energy" (Bynum, 2010). Speaking of persisting patterns, one must keep in mind that the Shannon information within an inforg contains the identity of said

organism. Human identity is not defined by matter, but rather by encoded patterns of Shannon information within the body. Our bodies are adaptive organisms, forever changing in response to their environment, but our identity persists through time and (to certain extent) governs this constant adaption. Thus, we are able to manipulate Shannon information in the metaphysical realm to influence (and govern) the physical one. This alludes that information ethics is crucial in a viable systems approach to the solution of the growing number of challenges brought forth by the digital age and information revolution.

However, what are the key principles of information ethics in relation to AI and society? If we are already living in the Universe, defined by information on both physical and metaphysical levels, does it mean that there is a possible underlying connection between these planes of our existence? In his joint article with Josh Cowls "A Unified Framework of Five Principles for AI in Society" L. Floridi analyzed six high-profile initiatives established in the interest of ethical and socially beneficial AI, including:

- The Asilomar AI Principles, developed under the auspices of the Future of Life Institute in 2017 (Future of Life Institute, 2017).
- The Montreal Declaration for Responsible AI developed under the auspices of the University of Montreal in 2017 (Montreal University, 2017).
- The General Principles offered in the second version of Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. This document is a joined crowdsourced effort of 250 global scientific and academic leaders under the patronage of the Institute of Electrical and Electronics Engineers (IEEE) (IEEE, 2017).
- The Ethical Principles offered in the Statement on AI, Robotics and 'Autonomous' Systems, published by the European Commission's European Group on Ethics in Science and New Technologies, in March 2018 (European Commission, 2018).
- The 'five overarching principles for an AI code' offered in the UK House of Lords AI Committee's report (House of Lords, 2018).
- The Tenets of the Partnership on AI, a multi-stakeholder organization consisting of academics, researchers, civil society organizations, companies building and utilizing AI technology, and other groups (Partnership on AI, 2018).

Floridi and Cowls have reviewed all these documents and found that together they yield 47 basic principles of how AI can be used in a socially beneficial way. According to the concept, embraced by the aforementioned researchers, all these principles have a high degree of cohesion with the four core principles commonly used in bioethics: *beneficence*, *non-maleficence*, *autonomy*, and *justice* (Floridi and Cowls, 2019; Beauchamp, 2012). According to Floridi, bioethics is the one that most closely resembles information ethics in dealing ecologically with new forms of agents, patients and environments (Floridi, 2013). However, Floridi and Cowls argue that a new principle is needed in addition: *the principle explicability*, which is "understood as incorporating both intelligibility (for non-experts, e.g. patients or business customers, and experts, e.g. product designers or engineers) and accountability" (Floridi and Cowls, 2019). Thus, the complete Unified framework of five principles for AI in society (hereinafter referred to as framework), according to Floridi, would look like this:

(1) *Beneficence: promoting well-being, preserving dignity, and sustaining the planet.* This principle, found to some extent in all six initiatives, firmly underlines the central importance of promoting the well-being of people and the planet with AI

technologies, continued prospering for mankind and the preservation of a socially-responsible environment for future generations;

(2) *Non-maleficence: privacy, security, and 'capability caution'*. This principle cautions against various negative consequences of overusing or misusing AI technologies, especially when dealing with personal privacy and militarized AI applications. However it is still unclear what the exact nature of the possible AI-maleficence would look like: are we dealing with the people developing AI, or the technology itself.

(3) *Autonomy: the power to decide (to decide)*. "When we adopt AI and its smart agency, we willingly cede some of our decision-making power to technological artifacts", states Floridi. This important principle speaks for a balance between the decision-making power people retain for themselves and the power they delegate to artificial agents. Humans should retain the power to decide which decisions to take: exercising the freedom to choose where necessary, and ceding it in cases where overriding reasons, argues Floridi.

(4) *Justice: promoting prosperity, preserving solidarity, avoiding unfairness*. While justice may seem as a rather broad concept, all AI enthusiasts and thinkers agree that elimination of unfair discrimination and the need for shared prosperity should stand in the center of AI use.

(5) *Explicability: enabling other principles through intelligibility and accountability*. Simply put, this principle must answer the question: is humanity the patient, receiving the 'treatment' of AI, the doctor prescribing it, or possibly both. So the explicability principle really should incorporate both the meaning of how AI works and the meaning of who is responsible for the way it works. Floridi argues, that this principle complements the others: for AI to be beneficent and non-maleficent, we must be able to understand the responsibility or harm it is doing to society, and in which ways (Floridi and Cowls, 2019).

At first glance, the framework developed by Floridi is sufficient to analyze any national strategy in the field of AI: such documents are full of broad and declarative statements of general "do responsibly" nature. However, the question is to what extent these ethical principles (borrowed from bioethics) are appropriate for assessing the social responsibility in AI utilization. Sure, Floridi's framework was adopted by the EU and is now an integral part of AI Ethics Guidelines of the European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG), but maybe there is a deeper connection between the seemingly different worlds of silicon and carbon, and not just our desire to use old laws for new "trespassing"? Our modern reality is inexorably hybrid and not only in terms of various "interreality systems" we use constantly – from advanced heads-up displays in planes and even cars to QR-code scanners in smartphones. Virtual tools, coupled with their real-world counterparts, inevitably lose their "virtuality" if only in the eyes of the common user, who becomes increasingly reliant (if not dependent) on them, just like he relies on caffeine and painkillers in everyday life, and probably even more so. In this respect, we can hypothesize that AI is no longer simply a tool for us to wield, nor is it a lifestyle application, but a completely new dimension to the informatisized reality we live in. And as much as we would like to present this process as something completely new to society, we might just discover instead that it is not so novel in fact. In the modern world, there is a saying: "The human being is the sum of social relations" that can be paraphrased as "The human being is the sum of information and its transmission channels". Human lifestyle as *biosocial* creatures is very much hybrid in nature: we strive to create our own metaphysical reality through

culture and thinking and we cannot fully if ever at all, leave the inherent reality of nature. As a result, what we get is an entanglement, or co-production of nature/culture, and not simply modern social and cultural phenomena cast in biological metaphors (Rabinow, 1996). And the more we embed AI into our society and culture, the more it becomes entangled into the surrounding nature as well as the nature of humanity as species and information agents.

Ultimately, we can agree with Floridi and propose the working hypothesis that bioethics and the ethics of information and AI can indeed overlap and complement each other. Thus, the Unified framework of five principles for AI in society *can be temporarily adopted as a potentially viable assessment tool* for how social responsibility is *implied and implemented* in AI-development strategies. To make this task a little bit easier, the authors would like to propose the following visualization of the chosen assessment tool (Figure 1).

The diagram presented above is a visualization of the analysis of AI development strategies in this article. For lack of a more precise methodology, we will empirically assess the positions of each of the considered strategies with the Unified framework, where the blue diagram corresponds to each case of the exact/relative resemblance of a particular principle (*beneficence, non-maleficence, autonomy, and justice*), and the orange diagram – to how *explicit* each of the principles is represented (e.g. how well it is supported, what are the means to implement it, etc.). The scale of the blue and orange diagrams is measured according to the following grade system (Table 1).

The final verdict will be based on a simple semantic analysis, where the framework grade units, chosen above, may or may not encounter their semantic representation in the analyzed text. The authors of this paper do not aim to determine the exact frequency of these encounters. Instead, they will try to examine a basic semantic correlation between the Unified framework provisions and those of the analyzed documents. This methodological perspective may appear to be too vague, but the authors fear that any other more specialized approach could be too narrow for the documents which appear to be vague in the letter as well as in spirit.

3. National artificial intelligence strategies and social responsibility

3.1 The USA

The USA could be considered as an active leader in terms of AI technologies. It is worth to mention, that the majority of global companies in the AI industry are concentrated there – over 2,000 US-based companies are leading the way in terms of AI start-ups – 1,393 (over 54% of the global market) compared to 389 in China (about 15%). Nevertheless, despite this unquestioned market leadership, the country still struggles to establish an active legal regulation in the field of AI. The closest thing the USA has to the national strategy is the Executive Order on AI,

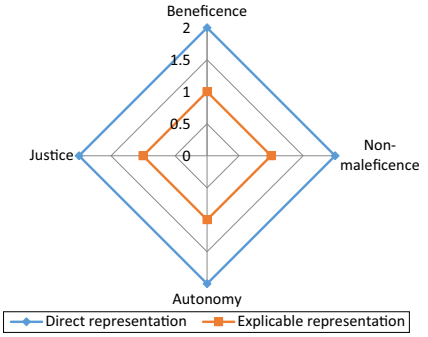

Figure 1.
Assessment tool for the framework of five principles for AI in society

Grade	Basic principle representation (Blue diagram)	Explicability of the principles (Orange diagram)
0	The principle is not mentioned	No supporting means mentioned and no further representation
1	The principle is mentioned at least once in the document	Supporting means are mentioned, but are purely declarative
2	The principle is detailed in at least one of the document's strategic goals	Supporting means are detailed and are backed by legal initiatives
3	The principle is represented in numerous strategic goals of the document	Supporting means are well detailed, are backed by legal initiatives and have assigned government-assisted programs

Table 1.
Grade system for the framework of five principles for AI in society

signed by President D. Trump on February 11, 2019 (Executive Office of the President, 2019). The document announced the American AI Initiative – the USA' national strategy on AI. This strategy is a concerted effort to promote and protect US AI technology and innovation. The Initiative implements a whole-of-government strategy in collaboration and engagement with the private sector, academia, the public and like-minded international partners. It directs the Federal government to pursue five pillars for advancing AI:

(1) promote sustained AI R&D investment;

(2) unleash Federal AI resources;

(3) remove barriers to AI innovation;

(4) empower the American worker with AI-focused education, and training opportunities; and

(5) promote an international environment that is supportive of American AI innovation and its responsible use.

The USA is also actively leveraging AI to help the Federal government work smarter in its processes and services.

Executive Order on AI is further detailed and supported by the National AI R&D Strategic Plan of 2019, which establishes the following eight strategic priorities:

(1) make long-term investments in AI research. Prioritize investments in the next generation of AI that will drive discovery and insight and enable the USA to remain a world leader in AI;

(2) develop effective methods for human-AI collaboration. Increase understanding of how to create AI systems that effectively complement and augment human capabilities;

(3) understand and address the ethical, legal, and societal implications of AI. Research AI systems that incorporate ethical, legal, and societal concerns through technical mechanisms;

(4) ensure the safety and security of AI systems. Advance knowledge of how to design AI systems that are reliable, dependable, safe, and trustworthy;

(5) develop shared public datasets and environments for AI training and testing. Develop and enable access to high-quality datasets and environments, as well as to testing and training resources;

(6) measure and evaluate AI technologies through standards and benchmarks. Develop a broad spectrum of evaluative techniques for AI, including technical standards and benchmarks;

(7) better understand the national AI R&D workforce needs. Improve opportunities for R&D workforce development to strategically foster an AI-ready workforce; and

(8) expand public-private partnerships to accelerate advances in AI. Promote opportunities for sustained investment in AI R&D and for transitioning advances into practical capabilities, in collaboration with academia, industry, international partners, and other non-Federal entities.

Each of the aforementioned priorities is a complete strategy in itself, with corresponding R&D programs and goals (United States Government, 2019).

A unified framework of five principles is rather easily applicable to the US AI strategy:

- *Beneficence (1)* is perhaps the most represented principle here, as all eight strategic priorities are mentioned to be potentially beneficial to "nearly all aspects of society, including the economy, healthcare, security, the law, transportation, even technology itself." The benefits to the society are described to be numerous, with enhanced quality-of-life through the use of AI in disaster recovery and medical diagnostics, new workplaces on th0065 market, etc. On the basic scale, the beneficence principle in the US AI strategy can be assigned grade "3". However, on the explicability scale, this principle is not very well represented at all. There are no legal initiatives and mentioned priorities do not define specific research agendas for Federal agency investments. Nonetheless, they still provide an expectation for the overall portfolio for US Federal AI R&D investments, thus grade "1.5" can be assigned (Figure 2).
- *Non-maleficence* (2) is indirectly mentioned in goal 3. Understand and Address the Ethical, Legal, and Societal Implications of AI, which calls for ethical and secure AI deployment to prevent all possible harm to humans. Besides, non-maleficence is

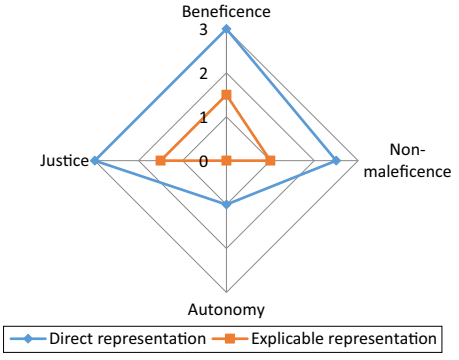

Figure 2.
The framework of the US AI strategy principles

indirectly implied in goal 4. Ensure the Safety and Security of AI Systems. It is stated, that AI should be "safe by design", where safety is not the only concern of system designers, but also is considered throughout the AI system lifecycle. On the basic scale we can assign grade "2.5", but on the explicability scale only grade "1" is possible since all supporting means are declarative.

- *Autonomy (3)* principle is represented in strategic goal 2. Develop Effective Methods for Human-AI Collaboration, where it is stated that the deployment of AI systems should be considered "as a design option for operators who wish to decide whether they want to use them or not". Grade 1 can be safely assigned on the basic scale, but the total lack of supporting means and further legal representation means only grade "0" on the explicability scale.
- *Justice (4)* is well represented in the US strategy both in the meaning of just and fair AI use (extensively detailed in goal 3. Understand and Address the Ethical, Legal, and Societal Implications of AI) and "social justice" as well (goal 7. Better Understand the National AI R&D Workforce Needs). The latter calls for the need to support American researchers in the field of AI, as well as undergraduate students. On the basic scale grade "3" is assigned, but only grade "1.5" on the explicability scale: numerous supporting means are declared, albeit without any legal and financial backing.

3.2 Russia

In the past 10 years in Russia, some 1400 AI scientific projects were carried out. Yet most of them (1200) were non-profit, as the private sector did not show much interest in the development and use of AI. Hence, AI was mostly driven by state and state-owned businesses. In the last 10 years, Russia allocated about $343m for R&D in AI. In comparison, the US state budget allocates about $200m annually for research in AI.

The Russian National AI Development Strategy was published on the 10 of October 2019. Until this date, the country lacked clear strategic documents, outlining AI development, use, and priorities. Russian Strategy mostly focuses on civilian AI applications (academia, science, healthcare, infrastructure, etc.), where new technologies can provide fast if not an instant benefit to the public sector.

Russian AI Development Strategy can be divided into six blocks, each corresponding to the certain national priority field:

(1) *Legal and ethical framework*: Establish clear legal foundations for creating and running industrial and academic centers for interdisciplinary applied research; determine the distribution of responsibility between owners, developers, and suppliers of data for AI use; clarify the regulation of the circulation of results of intellectual activity using AI; create national standards for testing, certifying, and confirming compliance for AI systems.

(2) *Scientific and academic community*: Create a new methodology for calculating a researcher's contribution to scientific development; finance AI researchers via contests and competitions; increase Russian-published articles on AI research; establish transparent methods for research and computer modeling.

(3) *Data regulation*: Create online repositories to collect, store, and process scientific data, including for training AI algorithms; provide Russian citizens with tools to manage their personal data and anonymize personal data posted in the public domain.

(4) *Hardware and software development*: Develop AI-ready microprocessors that run faster and use less electricity; simplify the licensing of intellectual rights to AI development software; create centers for collective AI solutions testing.

(5) *Education*: Use AI to monitor student performance to raise graduation rates; create a training system for AI professionals in universities and academia; create unified AI open learning platform for all educational levels; establish a secondary AI vocational education system; train engineers to allow the creation of domestic AI hardware.

(6) *Healthcare*: Use AI applications to conduct preventive examinations and to minimize defects and errors in diagnostic and invasive procedures; integrate AI tools into the Russian Unified State Health Information System; create National data bio-bank for big data analysis.

The document presents Russia's officially recognized definition of AI as "a set of technological solutions that makes it possible to simulate human cognitive functions [. . .] as well as to obtain results during the performance of specific tasks that are at least comparable to the results of human intellectual activity. This set of technological solutions shall consist of information and communications infrastructure, software [. . .], and data-handling procedures and services" (The President of the Russian Federation, 2019). The strategy stresses the strategic importance of AI as a prerequisite for Russia's entry into the group of economic world leaders as well as the country's technological independence and competitiveness. Even though Russia is not currently considered a leader in the realm of AI, the document states that Russia has the potential to "[become] an international leader in the development and use of artificial intelligence technologies".

This disparity is rather evident in the view of the Unified framework assessment tool:

- *Beneficence (1)* is a well-represented principle in the Russian AI strategy. Both Education and Scientific community priority fields call for new funding grants and programs in AI-training to enhance the overall positive impact of AI on the Russian system of science and education. The healthcare priority field states that AI could be vital for early diagnostics and the Russian healthcare system can increase its efficiency via extensive AI applications, thus significantly raising at least some of the aspects of people's life quality. On the basic scale, the beneficence principle in the Russian AI strategy can be assigned grade "3". On the explicability scale, this principle is represented rather poorly – "1". There are no legal initiatives and all mentioned priority fields lack project agendas to support their noble goals (Figure 3).
- *Non-maleficence* (2) is not only extensively mentioned but also declared among the main principles of the strategy. Legal and ethical priority field calls for the formulation of the appropriate ethical standards in the field of AI and ethical rules for human interaction with AI to prevent potentially maleficent use of AI technologies and ensure their safety for humans. On the basic scale we can assign grade "3", but on the explicability scale only grade "1" is possible: Russian strategy rather eagerly declares support for the *non-maleficence principle*, but does not provide any details on how it will be implemented in the form of government-backed programs and initiatives.
- *The autonomy (3)* principle is represented poorly. The strategy in review does not mention any decide-to-delegate models for AI users, simply stating that such delegation is possible except decisions that might infringe upon the rights and legitimate interests of individuals. We assign 1 on the basic scale and "0" grade on explicability scale.

- *Justice (4)* is fairly represented in the strategy. The legal and ethical priority field states that AI use must not infringe human rights in any way as well as be transparent and non-discriminative. Other priority fields also have some indirect meaning of 'just' AI utilization. Overall, equal access to the benefits of AI is clearly stated, however, there are not any means and legal initiatives to back this declaration. We assign" 2" on the basic scale and "0" grade on explicability scale.

The evident lack of explicable support means makes the Russian National AI Strategy looking overly optimistic. The document assumes that Russian society as well as public organizations and public enterprises will readily accept the benefits of AI applications and will fully cooperate with the Strategy's priorities. Also, Russian AI strategy presumes (and perhaps in a false way) that relevant Russian federal and regional government institutes will facilitate smooth implementation of the strategy, fully collaborating with existing public and private AI research projects.

3.3 China

Currently, Chinese companies are the most popular among international tech-investors: in 2018, Chinese Sense Time ($1,200m), UBTech Robotics ($820m), Megvii Technology ($600m), YITU Technology ($300m) and US Dataminer ($391m), CrowdStrike ($200m) and Pony.ai ($214m) became the leaders in terms of investment in the AI market. The purpose of AI development in China is to become a "scientific and technical superpower", a world center of innovations in the field of AI (the leader in all AI fields) with an active commercial AI industry and a leader in the creation of industrial robots.

China is one of the most active players in the field of AI, and this applies not only to technology and market infrastructure development but also to regulation. At present, China has the most extensive system of legislative acts and state plans containing priorities for the development of AI-technologies, but this diversity is significantly hindered by the vague and overly declarative nature of these documents. The need for the development of AI systems is outlined in the 13th Five-Year Plan for Economic and Social Development of the People's Republic of China 2016–2020. According to this document, the Chinese authorities expect economic growth of at least 6.5% during the five-year period 2016–2020, which is expected to be achieved through "breakthrough programs" in the field of robotics, innovative technologies for various purposes, and cybersecurity systems using AI. There is also a need to develop robotic systems for industrial, service, medical, and other purposes, as well as to promote the use of AI in the national economy, regardless of the industry (Central Committee of the Communist Party of China, 2016).

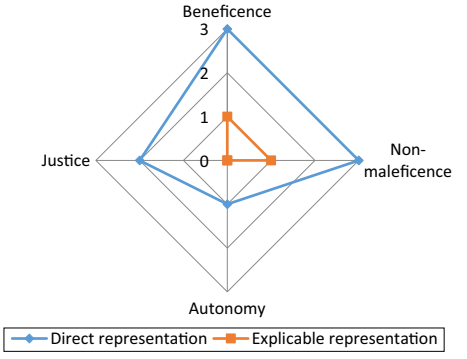

Figure 3.
The framework of the Russian AI strategy principles

However, the most important document declaring China's desire to become a world leader in the development of AI is the Plan for the Development of New Generation AI Technologies from 2017 (State Council of the People's Republic of China, 2017). The Plan outlines a three-step strategy, which the following achievements:

(1) By 2020, China will be one of the world's most advanced levels of AI technologies. By 2020 AI should have become an important driver of economic growth with the value of core AI industries to reach 150bn RMB, while that of AI-related industries to reach 1tn RMB. Chinese enterprises should possess world-leading competitiveness in a series of technology breakthroughs like in intelligent big data, intelligent cross-media, swarm intelligence, hybrid enhanced intelligence, and indigenous intelligent systems.

(2) By 2025, China is to become the world's leading country in the field of some AI technologies and their applications, and also achieve breakthroughs in fundamental AI theories. AI should be extensively applied in areas such as smart manufacturing, smart healthcare, smart cities, smart agriculture, and national defense, with the value of core AI industries to reach 400bn RMB, while that of AI-related industries to reach 5tn RMB. The year 2025 will also mark the preliminary establishment of a new legal framework for AI, including standards, safety assessments, and supervision.

(3) By 2030, China is to become the world's leading level in all AI theories, technologies, and applications. Also, the country will be the global center for AI technologies and AI economy. The core industry value will exceed 1tn RMB, AI-related industries value – 10tn RMB. AI will have deep integration in social governance, national security, and defense.

The Plan indicates six key tasks to implement to fulfill the above objectives, specifically relating to:

(1) the establishment of an open and cooperative AI technology innovation system;

(2) the cultivation of a high-end and highly-efficient intelligent economy;

(3) the building of a safe and convenient intelligent society;

(4) the strengthening of civil-military integration in the field of AI;

(5) the establishment of a ubiquitous, safe and efficient intelligent infrastructure system; and

(6) forward-looking layout of a new generation of AI-related major projects.

Despite its detail and consistency, the Chinese strategic plan is perhaps the hardest document to analyze through a unified framework. Of course, it is declarative no less than the reviewed US and Russian strategies; however, the concentration of its provisions on economic and technological advantages brings it in line with business plans, rather than with government strategic documents:

- *Beneficence (1)* is a principle that is rarely mentioned but strongly implied in the Chinese strategic plan. Practically, all strategic plan's priorities and tasks are meant to bring economic prosperity to the country for years to come as well as to establish a new quality of life. That includes the application of innovative AI throughout education, health care, pension, and other urgent needs involving people's livelihood.

There are several key social services in the plan, which should embrace AI technologies via extensive government support to become more beneficial and effective, namely Intelligent Education, Intelligent Medical Care, and Intelligent Health and Elder Care Systems. Still, the plan does not provide any means to achieve this goal. There are no programs or projects assigned, as well as no legal initiatives to back them. On the basic scale grade, "3" can be assigned. On the explicability scale – only "1" (Figure 4);

- *Non-maleficence* (2) is represented in the strategic plan's guarantee measure 1. Develop laws, regulations, and ethical norms that promote the development of AI, which calls for safe and ethical AI use. For this purpose, the document advocates for the development of a legal foundation for AI as well as for research on AI behavior science. Nonetheless, the strategy fails to expand on these matters and provides no implementation measures. On the basic scale we can assign grade "1", but on the explicability scale only grade "0" is possible.
- *The autonomy (3)* principle is not represented at all. The strategy in review does not mention any decide-to-delegate models for AI users. It is stated that AI can enhance the decision-making process, but any measures to balance AI-assisted decisions are not mentioned or implied. We assign" 0" on both scales.
- *Justice (4)* is represented mostly in the meaning of a traceable and accountable AI deployment (guarantee measure 1. Develop laws, regulations, and ethical norms that promote the development of AI) and ethical and moral multi-level judgment structure and human-computer collaboration ethical framework. No further clarification is provided of these definitions, as well as no supporting means. We assign" 1" on the basic scale and "0" grade on explicability scale.

4. Conclusions and future directions for research

The Unified framework of five principles for AI in society has proven to be more than just a coherent theoretical architecture for securing positive social outcomes from AI technology, but also a potentially viable assessment tool to define how certain strategic documents in the field define and implement these outcomes. Although social responsibility and ethics in AI may seem as highly theoretical aspects, they still have a strong practical dimension, which is unfortunately rarely recognized. At present, there are many voluntary ethical standards of social responsibility accepted by AI, but they have not yet found broad recognition in national strategic documents, dealing with AI matters. It often happens that national strategies in the field of AI seek to address all possible issues at once, but in reality, without

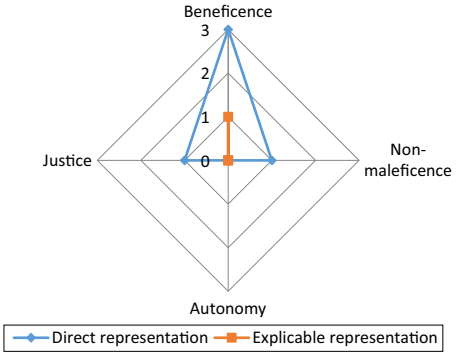

Figure 4.
The framework of the Chinese AI strategy principles

a proper ethical framework (of both enabling and constraining nature) they address none at all. This is particularly evident if we compare thee documents with one another. A cross-country comparison diagram of social responsibility principles in AI is presented below (Figure 5). Composite explicability principle grade here is calculated as the arithmetic average of the country's basic principles explicability grades.

The analyzed strategies of the US, Russia, and China may lean towards some or other declared basic principles of the social responsibility, but they fail to explicit them through mutual cross-priority implementation and accompanying support measures. This evident disproportion only stresses the significance of the *explicability principle*, introduced by L. Floridi in the Unified framework. Without this enabling principle's proper manifestation, each strategy in the field of AI becomes just a collection of good wishes. The authors of this paper argue that an effective national AI strategy should focus not only on declarations but provide a complex framework of enabling activities and priorities. This holistic approach will allow the nation to unlock the full potential of AI technologies in an ethical, transparent, and safe way.

As the authors of this paper would like to propose, the convergence of the ethics of AI with bioethics is not only permissible but desirable. Van Rensselaer Potter's original concept of bioethics as an interdisciplinary area of knowledge (Potter, 1994), integrating values of society with the values of science, can become a blueprint for a similar approach in AI. Therefore, it seems relevant to consider further similarities between the prospective model of the ethics of AI and the accumulated wealth of theoretical, methodological, and practical knowledge in bioethics.

Also, it seems relevant to broaden our understanding of the Philosophy of Information in regards to systems science. If we are going to embrace the crucial concept of Philosophy of Information, which declares every entity in the Universe as an information object, we must discover the principles governing the information systems on both physical and metaphysical levels. And it is the metaphysical level which is of paramount importance since it allows us to manipulate Shannon information to govern information systems, ourselves included. On this journey, we must keep in mind that the only coherent metaphysics is that which intends to build bridges between and unify elements of scientific theories (Ross *et al.*, 2009).

5. Theoretical and practical implications

This research allows us to better understand the linkages, overlaps, and differences between modern philosophy of information, AI-ethics, social responsibility, and government regulation. The analysis provided in this paper can serve as a basic blueprint for future attempts to define

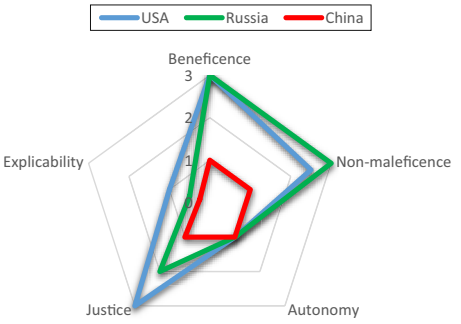

Figure 5.
Cross-country comparison diagram of social responsibility principles in AI

how social responsibility is understood and implied by government decision-makers. The experience of the USA, Russia and China in this field can be used in a variety of ways to analyze and improve other countries' AI development strategies. Certain provisions of the work may be useful for government agencies and can be applied in developing new administrative and legal instruments to support the use and development of AI.

References

Albrecht, J.P. (2016), "How the GDPR will change the world", *European Data Protection Law Review*, Vol. 1, pp. 287-289.

Allen, C., Smit, I. and Wallach, W. (2005), "Artificial morality: top-down, bottom-up, and hybrid approaches", *Ethics and Information Technology*, Vol. 7 No. 3, pp. 149-155.

Anderson, R.J. and Sherrock, W.W. (2013), "Ethical algorithms: a brief comment on an extensive", available at: www.sharrockandanderson.co.uk/wp-content/ (accessed 03 September 2020).

Arkhipov, V.V. and Naumov, V. (2017a), "Artificial intelligence and autonomous devices in legal. *Tr. SPIIRAN*, Vol. 6 No. 47.

Arkhipov, V.V. and Naumov, V. (2017b), "On some issues of the theoretical basis for the development. *Law*, Vol. 5 No. 157.

Beauchamp, T.L. and Childress, J.F. (2012), *Principles of Biomedical Ethics*, Oxford University Press Oxford.

Boddington, P. (2017), *Towards a Code of Ethics for Artificial Intelligence*, 1st ed., Springer New York.

Borenstein, J. and Pearson, Y. (2010), "Robot caregivers: harbingers of expanded freedom for all?", *Ethics*, Vol. 12 No. 3, pp. 277-288.

Brookings Institution (2018), "A blueprint for the future of AI", Weapons of the Weak: Russia and AI-Driven Asymmetric Warfare, available at: www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/#footnote-4 (accessed 24 August 2020).

Bynum, T.W. (2010), "The historical roots of information and computer ethics", in Floridi, L. (Ed.), *The Cambridge Handbook of Information and Computer Ethics*, Cambridge University Press, Cambridge, p. 24.

Central Committee of the Communist Party of China (2016), The 13th Five-Year Plan for Economic and Social Development of the People's Republic of China (2016–2020).

Cerka, P., Grigiene, J. and Sirbikyte, G. (2017), "Is it possible to grant legal personality to artificial intelligence", *Computer Law & Security Review*, Vol. 3 No. 5, p. 685.

Darnault, C., Parcollet, T. and Morchid, M. (2019) "Artificial intelligence: a tale of social responsibility", *Association for the Advancement of Artificial Intelligence*.

de Mul, J. (2010), "Moral machines: ICTs as mediators of human agency", *Techné: Research in Philosophy and Technology*, Vol. 3 No. 14, pp. 226-236.

Executive Office of the President (2019), *Maintaining American Leadership in Artificial Intelligence, E. O. 13859 of Feb 11, 2019*. s.l.: Executive Office of the President, Presidential Document.

European Commission (2018), European Group on Ethics in Science and New Technologies (2018, March). Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems.

Floridi, L. (2002), "What is the philosophy of information?", *Metaphilosophy*, Vol. 33 Nos 1/2, pp. 123-145.

Floridi, L. (2008), "Foundations of information ethics [electronic resource]//the handbook of information and computer ethics", available at: www.cems.uwe.ac.uk/~pchatter/2011/pepi/ (Accessed 01 September 2020).

Floridi, L. (2013), *The Ethics of Information*, Oxford: Oxford University Press.

Floridi, L. (2019), "What the near future of artificial intelligence could be", *Philosophy and Technology*, Vol. 32 No. 1, p. 1.

Floridi, L. and Cowls, J. (2019), "A unified framework of five principles for AI in society", *Harvard Data Science Review*, Vol. 1 No. 1.

Friedler, S.A. Scheidegger, C. and Venkatasubramanian, S. (2016), "On the (im) possibility of fairness", arXiv preprint, issue 1609.07236.

Future of Life Institute (2015), *An Open Letter: research Priorities for Robust and Beneficial Artificial Intelligence. s.l*, Future of Life Institute.

Future of Life Institute (2017), "Asilomar AI principles".

Gadzhiev, G.A. (2018), "Whether robot-agent is a person?", *Journal of Russian Law*, Vol. 1 No. 29.

Goldman, N., *et al.* (2013), "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA", *Nature*, Vol. 494 No. 7435, pp. 77-80.

Gurney, J.K. (2016), "Crashing into the unknown: an examination of crash-optimization algorithms", *Albany Law Review*, Vol. 183, p. 79.

Heisenberg, D. (2005), "Negotiating privacy: the European union", *The United States and Personal Data*, Boulder: Lynn Rienner Publishers.

Horvitz, E. and Selman, B. (2009), "Interim report from the panel chairs, association for the advancement of artificial intelligence".

House of Lords (2018), "House of lords artificial intelligence committee (2018, April, 16). AI in the UK: ready, willing and able".

IEEE (2017), "The IEEE initiative on ethics of autonomous and intelligent systems (2017). ethically aligned design".

Kofas, J. (2017), "Artificial intelligence: socioeconomic, political and ethical dimensions. Counter", available at: https://countercurrents.org/2017/04/artificial-intelligence-socioeconomic-political-and-ethical-dimensions/ (accessed 03 September 2020)

McCarthy, J., Minsky, M.L., Rochester, N. and Shannon, C.E. (2006/1955), "A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955", *AI Magazine, 15 December*, Vol. 27 No. 4, p. 12.

McCarthy, J., Minsky, M.L., Rochester, N. and Shannon, C.E. (2006 [1955), "A proposal for the dartmouth summer research project on artificial intelligence, August 31, 1955", *AI Magazine, 15 December*, Vol. 27 No. 4, p. 12.

Montreal University (2017), "Montreal declaration for a responsible development of artificial intelligence, s. l.: announced at the conclusion of the forum on the socially responsible development of AI".

Morkhat, P.M. (2017), *Artificial Intelligence: A Legal Perspective*, Moscow: Buki Vedi.

Morkhat, P.M. (2018), "Concerning the question about the legal personality of electronic person", *Legal Studies*, Vol. 4 No. 1.

Noorman, M. and Johnson, D.G. (2014), "Negotiating autonomy and responsibility in military robots", *Ethics and Information Technology*, Vol. 16 No. 1.

O'Heigeartaigh, S. (2013), "Would you hand over a moral decision to a machine? Why not? Moral outsourcing and artificial", available at: http://blog.practicalethics.ox.ac.uk/2013/08/would-you-hand-over-a-moral-decision-to-a-machine-why-not-moral-outsourcing-and-artificial-intelligence/ (accessed 03 September 2020).

Partnership on AI (2018), "Partnership on AI (2018). Tenets".

Pavaloiu, A. and Kose, U. (2017), "Ethical artificial intelligence–An open question", arXiv preprint, issue arXiv:1706.03021.

Pavaloiu, A. and Köse, U. (2017), "Ethical artificial Intelligence – an open question", *Journal of Multidisciplinary Developments*, No. 2, pp. 15-27.

Potter, V.R. (1994), "An essay review of – global responsibility. In", *Perspectives in Biology and Medicine*, Vol. 37 No. 4, pp. 546-550.

Rabinow, P. (1996), *Essays on the Anthropology of Reason. c1996 Xvii ed*, Princeton, N.J.: Princeton University Press.

Ross, D., Spurrett, R. and Collier, J. (2009), *Every Thing Must Go: Metaphysics Naturalized*. 1st ed. s.l, Oxford University Press.

Schlenoff, C.I. (2015), "IEEE standard ontologies for robotics and automation. IEEE std 1872-2015", 10 April, pp. 1-60.

Shannon, C.E. (1948), "A mathematical theory of communication", *Bell System Technical Journal*, Vol. 27 No. 3, pp. 379-423, 623.-656.

Sharkey, N. (2008), "The ethical frontiers of robotics", *Science*, Vol. 322 No. 5909, pp. 1800-1801.

SIPRI (2019), "The impact of artificial intelligence on strategic stability and nuclear risk", available at: www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf (accessed 24 August 2020).

Stanford Encyclopedia of Philosophy (2019), "Stanford encyclopedia of philosophy, definition of agency", available at: https://plato.stanford.edu/entries/agency/#:~:text=In%20very%20general%20terms%2C%20an,a%20standard%20theory%20of%20action. (accessed 24 August 2020).

State Council of the People's Republic of China (2017), "Next-generation artificial intelligence development plan".

Sullins, J.P. (2006), "When is a robot a moral agent?", *International Review of Information Ethics*, Vol. 6 No. 12, pp. 23-30.

The President of the Russian Federation (2019), *Decree of the President of the Russian Federation on the Development of Artificial Intelligence, s.l*, Official Portal of Russian Legal Information.

UNESCO (2018), "Towards a global code of ethics for artificial intelligence research", *The UNESCO Courier*, Vol. 3 No. July-September 2018.

United States Government. (2019), "National AI R&D strategic plan: 2019".

Wiener, N. (1961), *Cybernetics or Control and Communication in the Animal and the Machine, Reissue of the 1961 Second Edition. 2nd ed. s.l*, The MIT Press.

Further readings

Grishin Robotics (2017), "Federal act' on amendments to the civil code of the Russian Federation in terms of improving the legal relationship management in the field of robotics" (in russian)", available at: http://robopravo.ru/uploads/s/z/6/g/z6gj0wkwhv1o/file/bESvQz3Y.pdf (accessed 28 August 2020).

Kant, I. (1999), *Critique of pure reason (the cambridge edition of the works of immanuel kant)*", Translated and edited by Paul Guyer *and* Allen W. Wood. *s.l*, Cambridge University Press.

Robopapravo (2017), "The initiative for study legal aspects of robotics and artificial intelligence 'robopapravo'. [online]", available at: https://robopravo.ru/matierialy_dlia_skachivaniia#ul-id-4-35 (Accessed 28 August 2020).

Vasilyev, A.A., Ibragimov, Z.I. and Gubernatorova, E.K. (2019), "The Russian draft bill of 'the Grishin law' in terms of improving the legal regulation of relations in the field of robotics: a critical analysis", *Journal of Physics: Conference Series*, Vol. 1333, p. 052027.

Corresponding author

Anton Saveliev can be contacted at: anton.saveliev@gmail.com