A Comprehensive Ethical Framework for AI Entities: Foundations

Andrej Dameski(✉)

Law, Science and Technology Joint Doctorate (LAST-JD ERASMUS), Università di Bologna (Consortium Coordinator), Bologna, Italy
andrej.dameski@studio.unibo.it

Abstract. The participation of AI in society is expected to increase significantly, and with that the scope, intensity and significance of morally-burdened effects produced or otherwise related to AI, and the possible future advent of AGI. There is a lack of a comprehensive ethical framework for AI and AGI, which can help manage moral scenarios in which artificial entities are participants. Therefore, I propose the foundations of such a framework in this text, and suggest that it can enable artificial entities to make morally sound decisions in complex moral scenarios.

Keywords: Ethics of AI · Machine ethics · AGI

1 Introduction

The subject of this article will be the brief introduction of a proposal for the foundations of a model of a comprehensive ethical framework (hereinafter: the Framework) for artificial intelligence (AI) entities, also including artificial general intelligence (AGI) entities, jointly referred to as A(G)I.

The participation of AI in society is expected to increase significantly (Yudkowski 2008; Kurzweil 2000), and therefore the effects AI is causing on its environment (including other AIs, humans and their societies, animals, and the world in general) will increase in scope, intensity and significance (see Veruggio 2007). Simultaneously, the scope, intensity and significance of morally-burdened effects (i.e. effects/changes imposed on the world that contain moral content; see Reader 2007) produced by AI is also expected to massively increase in the near future (Smith and Anderson 2014; Anderson and Anderson 2007, 2009). AI will increasingly enter in interactions which can be judged as morally (not-) good and/or right (and the natural expansion into (not-) justifiable, acceptable, just, etc.).

There already is a multitude of ethical issues on which we need to derive satisfying and morally-sound 'best possible'/'least worse' (hereinafter: 'BP'/'LW') solutions; and it seems that the future holds even deeper, and more insidious ethical issues that we will have to deal with in a morally-acceptable fashion, lest we avoid possible catastrophic consequences of the widespread introduction of AI in human civilisation(s) (Yudkowski 2008).

M. Iklé et al. (Eds.): AGI 2018, LNAI 10999, pp. 42–51, 2018.
https://doi.org/10.1007/978-3-319-97676-1_5

There are some efforts at deriving comprehensive solutions to the above issues in a morally and legally sound way, such as Veruggio's *EURON Roboethics Roadmap* (Veruggio 2007); *Robot Ethics: The Ethical and Social Implications of Robotics*, a collection of texts edited by Patrick Lin, Keith Abney and George A. Bekey (Lin et al. 2012); in works in philosophy and ethics of information Luciano Floridi's *Ethics of Information* is a notable example (Floridi 2004, Floridi 2013), alone and alongside other authors (i.e. Mariarosiaria Taddeo (Floridi and Taddeo 2016; Taddeo 2017), J. W. Sanders (Floridi and Sanders 2004), Savulescu (Floridi and Savulescu 2006), Mittelstadt (Mittelstadt et al. 2016), and others); and in regards of law and legal aspects of AI, a notable example is Chopra and White's *A Legal Theory for Autonomous Artificial Agents* (Chopra and White 2011). However, the scientific community is far from a consensus on the matter. Therefore, the author of the text hopes to contribute to the whole effort in this sense.

In essence, there is a clear need for the establishment of a comprehensive ethical framework in regards of A(G)I that can help:

- clearly conceptualise ethically-burdened situations (scenarios) where A(G)I is involved;
- devise computationally-representable 'BP'/'LW' solutions for such situations;
- engineers design and install an ethical cybernetic subsystem in A(G)I systems that will enable them to achieve the above two;
- invigorate and contribute to the debate among academia, industry, engineers, and policymakers about the foundations of morality and ethics in regards of A(G)I;
- manage morally-burdened effects caused or otherwise related to A(G)I, and its utilisation (where appropriate), to the best outcomes.

2 Considerations in Regards of A(G)I

2.1 Ethical Considerations in Regards of A(G)I

A comprehensive ethical framework that can help soundly manage morally-burdened scenarios—caused/received by or otherwise related to A(G)I—should take into consideration a plethora of moral issues and perspectives that inevitably will arise from the widespread introduction of AI into society, and the possible advent of AGI. Consequently, it bears to first discuss what possible such issues and perspectives should be managed by such a framework.

General comments. As a general comment, most of the dominant ethical theories of today are, arguably, agent-focused. That is, they focus on the morally-burdened actions of *moral agents*[1], and what those agents ought, or ought not do. These are deontology, teleology, and virtue ethics. There exist also ethical theories that are focused on *moral patients*. In these moral worldviews, agents are of second importance, and moral

[1] Namely, in the moral landscape, agents are those that take actions and thus *cause* morally-burdened effects; while moral patients are those entities which morally-burdened effects are *effected/caused to.*

frameworks are here to determine how moral agents ought act predominantly in respect of what effects their actions will have on moral patients. Examples of these theories are ethics of care, feminine ethics, some instances of ethics of information (e.g. Floridi 2013), environmental ethics, and similar.

In the opinion of the author, both moral worldviews are limited in their scope, as they focus only on certain components of morality and ethics, and choose to assign arbitrary status of higher importance to one or the other component (the agent(s) or the patient(s)). This can result in unwarranted bias during derivation of understanding, interpretation, and solutions to moral scenarios. Arguably, if an ethical framework for A(G)I is to be comprehensive, it should focus on both moral agents and moral patients, and consider them as equally important (for a discussion on this subject, see Gunkel 2014).

Ethical considerations. Below are included many essential ethical issues and perspectives that a comprehensive ethical framework for A(G)I will have to (contextually) consider in providing satisfying solutions to problematic moral scenarios. The following were chosen based on the regularity with which they appear when discussing ethics of AI (see, for example, Tzafestas 2016 p. 65–188), and also additional ones considered as important by the author. However, in the interest of available space it is by no means a final list.

Moral entities—A(G)I entities can, in a moral scenario, be moral agents and/or moral patients. In some situations, an A(G)I entity can also be both a moral agent and patient regarding the same morally-burdened effects *at the same time*.
Consciousness—An important issue to consider is how, and if, conscious experience (qualia) relates to ethics and morality, especially to A(G)I. One thing to note here is that the status of a moral agent or a moral patient for an A(G)I entity in a moral scenario can exist regardless of whether it is '(self-)conscious' about the scenario itself (see subsection *Morality in regards of A(G)I* in Sect. 3.2. below).
Universalism vs. anthropocentrism—A comprehensive ethical framework would take into consideration as important all entities in a moral scenario (i.e. humans, A(G)I, beings, the environment, entities generally including informational entities (see Floridi 2013), etc.) and the moral issues perturbing them.
Aliveness/'Being'—A consideration of what is 'alive' and what agent/patient is alive or exists ('Being'; see Floridi 2013) will be necessary so that there can be right perspective on what entity can cause morally-burdened effects, and what entity can and does receive such effects. In other words, which entities in the world can be considered as moral agents and patients respectively.
Personhood and legal personhood—A very important issue regarding ethics. Naturally, the understanding of legal personhood (considering an entity as a person before the law, and assigning it all the related rights and responsibilities) will flow from the ethical-philosophical understanding of 'person' and its attributes; and even before that (see, for example, Chopra and White 2011; and MacDorman and Cowley 2006).
Agency, autonomy—Autonomy is, by nature, directly connected to agency i.e. the property of an entity that make it a (moral) agent. Understanding of autonomy, and whether A(G)I entities possess it by definition or in practice, is a consideration predominantly in agent-focused ethical theories.

Complexity and moral uncertainty—When moral agents or patients are facing increasing complexity of moral scenarios, and thus inevitably becoming unable to devise 'perfect solutions', the role of the moral uncertainty that thus appears and potentially modifies moral responsibility and accountability is an important perspective that should be taken into consideration (see Zimmerman 2008). This is also related with the pragmatic 'BP'/'LW' solutions to moral scenarios, as mentioned before.

Rights—A(G)I entities will most probably have effect over human rights and other rights, as assigned by law, constitutions, and governing international documents (see also Tzafestas 2016 p. 75).

Values—Inspiring virtues, moral values are set of principles that moral entities (including A(G)I) use to determine what actions, effects and states are good, bad, evil, (un)acceptable, etc. In essence, moral values help determine what is considered 'valuable' from the perspective of morality and ethics. A(G)I entities dealing with moral scenarios will have to, at least implicitly, bear the capability to determine what is morally (not) valuable.

Virtues (and vices)—On the other hand, virtues are recognised as one of the most important elements of virtue ethics. They determine how a moral entity ought to think and act so that it will live the 'good life' and be 'good'. Arguably, virtues will be implicitly important for AI entities; but also explicitly important for AGI entities.

Accountability and responsibility—It is a common issue of accountability and responsibility in regards A(G)I entities causing and/or receiving morally-burdened effects in moral scenarios. Some ethical theories deny that there can be responsibility and/or accountability without (self-)consciousness. However, A(G)I entities can be held responsible and accountable even without (self-)consciousness, since we already have examples of similar treatment of children and animals, who in most ethical and legal systems are regularly treated as accountable (as in, the agent causing the effects), but not responsible.

Opacity and transparency—Opacity and transparency is a very important issue regarding A(G)I (see Danaher 2016). Designing or imposing A(G)I systems that can precisely, responsibly and intelligibly explain how they reach their conclusions and courses of (in-)action is essential for the future acceptance of the widespread introduction of automated decision making in society. This also is closely related to accountability and responsibility discussed above.

Utility (the perspective of A(G)I and algorithms simply as 'tools' or 'means')—Considering an A(G)I system simply as a tool would mean it expands a significantly narrowed down and simplified moral considerations. Potential ethical issues can arise especially with the possible advent of AGI, self-consciousness, personhood, and ability to suffer.

Trust—Trust is closely related to responsibility, accountability, predictability, opacity, and transparency. A trust in an A(G)I system facilitates its deployment and utilisation, and increases efficiency and effectiveness.

Morally-burdened effects—caused by moral agents, and received by moral patients, these are an essential part of any moral scenario, and, like all other above considerations, will need to be modelled and managed by an A(G)I system.

3 A Comprehensive Ethical Framework for A(G)I

3.1 Introduction

A comprehensive ethical framework for A(G)I has to enable derivation of satisfying solutions to the previously mentioned issues. It has to take one or more of them in consideration, where appropriate in respect of context, and provide computationally and logically representable solutions, that will be the 'BP'/'LW' ones in the moral scenarios that are faced. An A(G)I entity using such a framework would have to reach or preferably surpass moral reasoning capacities of individual humans, and of human collectives and institutions. Programmers, by consulting and implementing such a framework, will be able to design A(G)I entities that have better moral reasoning capabilities than without it. In essence, if such a framework (or an appropriate approximation) is implemented in the design and the utilisation of A(G)I entities it will leave the world better off morally on aggregate.

3.2 Characteristics and Design

Foundational—The framework should be set up as a system of axioms that can be informationally, logically and computationally represented.
Coherent—The axiomatic system is able to be informationally, logically and computationally expanded to provide solutions to arising ethical problems in context, without issues of incoherence taking place.
Hybrid, multidisciplinary, and holistic—The axiomatic base of the framework is to be conceived with a holistic approach in mind and thus help provide more comprehensive one, drawing on existing advances in ethics in general, ethics of AI and ethics of information, and on other, 'non-ethical' and meta-ethical disciplines.
Unified/unifying—The framework should have universalist pretension i.e. it should attempt to unify all the major ethical theories into a single axiomatic system; and thus render them as special cases of itself.
Contextual—The framework, when used as a cybernetic (sub)system into an A(G)I system, should be able to 'live in context', acquire new and modify its existing moral knowledge, and adjust to new environment.
Applicable to A(G)I and its interaction with the environment—i.e. other A(G)I systems and other systems in general, the world, humans and their systems, animals, the legal, financial and social systems, etc.
Translatable and implementable through engineering and legal tools.

3.3 Design

The foundation. Below is included the axiomatic foundations of the Framework for A(G)I that the author presents in this article (see also Fig. 1.).

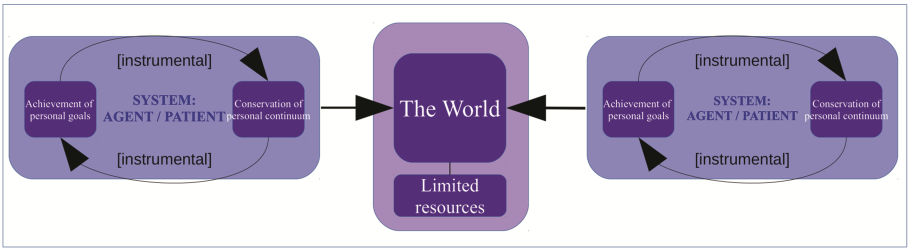

Fig. 1. Emergence of moral systems

Axiom 0	**Every system[a] has as a moral imperative[b] its highest possible personal Quality of Life (QoL).** Every system's QoL is comprised of the level of potential or actual achievement of two fundamental goals: (1) conservation of personal continuum (2) achievement of personal goals
Axiom 1	**Every system has at least one of the fundamental goals from Axiom 0 as a moral imperative [explicit goal], and as an instrument [implicit goal].** •A system can simultaneously have both of these goals as moral imperatives (that is, explicit goals). Each fundamental goal can be partially or wholly a moral imperative and/or an instrument[c]. •For a system, a goal can, and does, simultaneously serve both as an imperative and as an instrument (for the purposes of the other goal).
Axiom 2	**Every system strives towards imperative maximisation, by using its resources, which include its instruments.**
Axiom 3	**Resources are (inevitably) limited.** •Systems compete over limited resources in their imperative maximisation, and that leads them in conflict. •This dialectical process of conflict, and the subsequent emergence of solutions to the conflicting situations, is the originator of morality.

[a]A system is defined as follows: a system is a set of interrelated and interdependent components, from whose interaction the system emerges as something more than just the simple sum of its parts. A system can be conceptualised both as a collective (of its parts) and as an individual, and this usually depends on the level of abstraction (see Floridi 2013). The usage of 'moral entity', 'informational entity', 'agent/patient', 'entity', etc. are interchangeable with system.

[b]A moral imperative is, thus, a systemic imperative; in the sense that the system considers and/or acts as if pursuing the achievement of its systemic imperatives is *right* and *good* for itself and in general. This adds the moral dimension.

[c]It is important to note that the imperative/instrument duality is not a dichotomy, but a spectre. In practice, most systems have both fundamental goals as simultaneous and independent moral imperatives and instruments. One of the goals may be independently less/more of a moral imperative, and independently less/more of an instrument for the other goal, determined by the internal structure of the system.

Morality. Morality deals with Quality of Life (QoL) of systems. QoL is defined as **the potential to achieve, or the actual achievement, of moral imperative(s) of systems**. If it is considered as a category, the *potential to achieve moral imperative(s)* part of QoL would include moral concepts such as freedom, agency, capacity, intention and similar ones. Similarly, *actual achievement of moral imperative(s)* would include moral

concepts such as fulfilment, justice, happiness, alleviation and transcendence of suffering, and similar.

Simply taken, in the world systems (and therefore A(G)I agents) exist, act (and thus cause morally-burdened effects), and are acted upon (and thus receive those effects) while in pursuit of their imperatives. All systems use available resources[2] to be able to continue to do the above and proceed with pursuing their imperatives i.e. conserve their personal continuum and achieve their goals.

However, resources are either (locally or globally) limited, or inevitably become limited. This 'forces' systems to compete for them so that they can continue pursuing their imperatives. This competition inescapably leads to conflict (see also Tiles 2005 p. 70). *Conflict*, in this perspective, is a process whereby a system explicitly or implicitly threatens other system(s) with reduction of their ability to achieve their imperatives, if the first system's ability to achieve its own imperative(s) is jeopardised. In essence, when a system finds its QoL in jeopardy by another system, it acts to secure the resources that are jeopardised, and this is threatening to the QoL of the other system because the other system also needs them for its own QoL. Conflicts, by extension, and in moral scenarios with more cognitively capable moral entities, can develop into second order ones (i.e. conflicts over opposing values and methods of distribution of resources), which are in essence conflicts over differing moral systems.

If during this process systems, explicitly or implicitly, achieve a balance point, whereby there is a compromise as to how much of the contested resources should belong to the first or the second system; and this enables both systems for the time being to continue pursuing their desired, but now revised, QoL level; the balance point that has emerged ('crystallised') is a moral rule. Systems opt to respect this moral rule for the time being as it enables them to achieve the best practically possible QoL level through avoiding further conflict while lowering their desired QoL level.

Emergence of moral systems. Out of a complex, multifaceted aggregation of moral scenarios, where systems enter in conflict and subsequently establish moral rules which are then crystallised (that is, stabilised), a moral system emerges for that particular collective of systems. In essence, moral systems are methods governing the distribution of needed resources. This is what is normally understood under *morality* in a practical manner. See Fig. 1 for illustration. Morality is, therefore, a cyclical down-up (emergent) and up-down (crystallising) process. Any moral system that thus emerges or is imposed, also contains the properties of any other system.

Contextuality. Since moral systems emerge for particular collectives, each moral system is contextual and specific, even though the basic principles that cause their emergence are the same—conflict over resources needed for desired QoL. Moral systems differ because of differences in the components of the system, which include

[2] Under resources here are understood all parts of the world which a system can use instrumentally to pursue its imperatives i.e. both 'traditional' ones such as raw materials, energy source(s), food, water, minerals etc. but also time, situations, rules, other systems and their parts, and anything else of utility.

the contesting systems, the contested resources, and other miscellaneous factors such as difference in the environment.

However, most moral systems created by systems that enter into similar moral scenarios (i.e. human collectives) are alike, and universality or widespread adoption in some basic moral rules can be discovered throughout them. Examples in human moral systems are of the immorality of murder, rape, sexual acts with children, incest, lying, irresponsible or unnecessary disturbance or damage, and similar.

Morality in regards of A(G)I. In respect of A(G)I entities, there are several additional considerations that need to be discussed.

Firstly, A(G)I systems for which there will be a requirement to deal with moral scenarios will have to consider the aforementioned perspectives. That means that A(G)I systems will have to, directly or indirectly, take into consideration the QoL of other systems.

Secondly, as discussed before, the A(G)I system itself doesn't have to be (self-)conscious of the (moral) scenario or generally in the conventional meaning—since the algorithm doing the calculation and deriving at the decision for (in-)action can be designed by human programmers. This means that the system will participate as a moral agent and moral patient in the moral scenario regardless of any (self-)conscious sense of the underlying moral considerations (also known as mindless morality; see Floridi 2013). Simply taken, morally-burdened effects can exist without conscious intention. This is (mostly implicitly) recognised also by other ethicists and researchers working in this domain, such as Floridi (2013), Dodig Crnkovic and Çürüklü (2012), Gerdes and Øhrstrøm (2015), and others. This also means that human engineers can input moral systems, or even simple moral rules (i.e. deontological or teleological rules) in simple AI systems that deal with morally problematic scenarios. The systems in question will act as moral entities (agents/patients) and cause and/or receive morally-burdened effects.

And thirdly, in regards of AGI, there are some additional ethical considerations. Arguably, AGIs that reach or surpass cognitive and other capacities of humans and human collectives will be able to wield tremendous power, and cause significant morally-burdened effects. In moral scenarios, moral entities with higher power (i.e. ability to exert their will in pursuit of their goals regardless of resistance) bear proportionately higher moral responsibility. That would mean that such AGI systems will have to attempt to take into consideration the QoL of all other systems involved in the moral scenario in which they exist, act, and are being acted upon.

It is hence reasonable to assume that this will require tremendous capacity for moral reasoning (i.e. moral calculations) on the part of those AGI entities, which would include moral scenario model building, bias avoidance, heuristics and fallibilistic reasoning, and ability to choose the 'BP'/'LW' course of (in-)action given available data and resources.

Unifying/unified. The Framework presented in this text attempts to integrate and harmoniously unite dominant ethical theories of today. This includes deontology, teleology, virtue ethics, rights theory, value theory, ethics of care (patient-focused ethics). In this effort, all these ethical theories become special cases of the general model. Unfortunately, in the interest of space, the author can only provide a graphical

representation of this unification in Fig. 2. The detailed description will have to remain for future work.

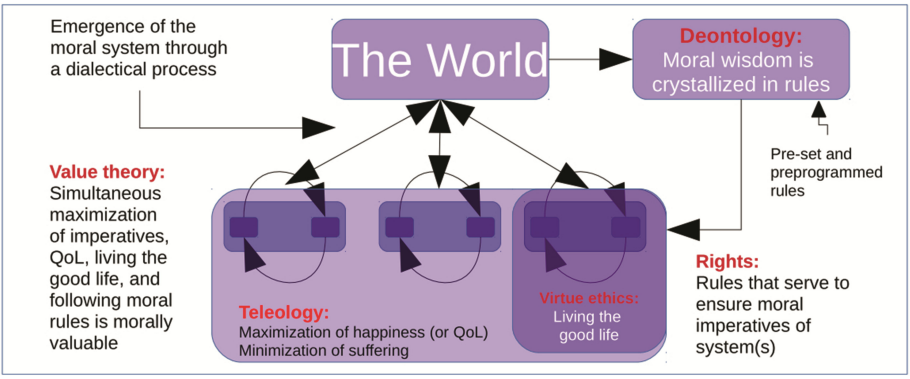

Fig. 2. The integration of ethical theories

4 Conclusion and Way Forward

The basis of the Framework presented in this text is a model of a foundational, yet flexible, adaptable and contextual moral system. It can serve as a model to be used by designers of A(G)I systems, or by A(G)I entities themselves, for the building of internal moral subsystems that will enable A(G)I entities to successfully participate in complex moral scenarios in a morally sound manner. This will enable them to manage morally-burdened effects, and attempt to avoid the negative ones, while attempting to maximise the positive ones, or the so-called 'BP'/'LW' solutions.

Subsequent efforts should be given in the elaboration of the Framework's components in detail, testing it in theoretical moral scenarios, as well as, in gathering input from a wide range of sources which would enable to determine statistical indicators that can be taken in consideration by an A(G)I entity to perform contextual moral calculations. This will enable the improvement of the model itself, and hopefully A(G)I entities using it will be able to derive moral solutions in context that will approach, and even exceed, human moral reasoning capacity.

References

Anderson, M., Anderson, S.L.: Machine ethics: creating an ethical intelligent agent, AI Magazine **28**(4), 15–26 (2007). American Association for Artificial Intelligence

Anderson, S.L., Anderson, M.: How Machines Can Advance Ethics, Philosophy Now (2009)

Chopra, S., White, L.F.: A Legal Theory for Autonomous Artificial Agents. University of Michigan, Ann Arbor (2011)

Danaher, J.: The threat of algocracy: reality. Resistance and Acommodation. Philosophy and Technology **29**, 245–268 (2016)

Dodig Crnkovic, S., Çürüklü, B.: Robots: ethical by design. Ethics Inf. Technol. **14**, 61–71 (2012)

Floridi, L. (ed.): The Blackwell Guide to the Philosophy of Computing and Information. Blackwell Publishing, Hoboken (2004)
Floridi, L.: The Ethics of Information. Oxford University Press, Oxford (2013)
Floridi, L., Sanders, J.W.: On the morality of artificial agents. Minds Mach. **14**, 349–379 (2004)
Floridi, L., Savulescu, J.: Information ethics: agents, artefacts and new cultural perspectives. Ethics Inf. Technol. **8**, 155–156 (2006)
Floridi, L., Taddeo, M.: What is Data Ethics. Philosophical Transactions of the Royal Society A **374**(2083) (2016). Preprint
Gerdes, A., Øhrstrøm, P.: Issues in robot ethics seen through the lens of a moral turing test. J. Inf. Commun. Ethics Soc. **13**(2), 98–109 (2015). Emerald Group Publishing Limited
Gunkel, D.J.: A vindication of the rights of machines. Philos. Technol. **27**, 113–132 (2014)
Kurzweil, R.E.: The Age of Spiritual Machines: When Computers Exceed Human Intelligence. Penguin Books, London (2000)
Lin, P., Abney, K., Bekey, G.A. (eds.): Robot Ethics: The Ethical and Social Implications of Robotics. The MIT Press, Cambridge (2012)
MacDorman, K.F., Cowley, S.J.: Long-term relationships as a benchmark for robot personhood. In: The 15th IEEE International Symposium on Robot and Human Interactive Communication, Hatfield, UK (2006)
Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S., Floridi, L.: The ethics of algorithms: mapping the debate. Big Data Soc. **3**(2), 1–21 (2016)
Reader, S.: Needs and Moral Necessity, Routledge (Taylor and Francis Group). Taylor and Francis e-Library, Abingdon (2007)
Smith, A., Anderson, J.: AI, Robotics, and the Future of Jobs, Pew Research Center (2014). http://www.pewinternet.org/2014/08/06/future-of-jobs/
Taddeo, M.: The moral value of information and information ethics. In: Floridi, L. (ed.) The Routledge Handbook of Philosophy of Information. Routledge (2017)
Tiles, J.E.: Moral Measures: An Introduction to Ethics East and West, Routledge (Taylor and Francis Group). Taylor and Francis e-Library, Abingdon (2005)
Tzafestas, S.G.: Roboethics: A Navigating Overview. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-21714-7
Veruggio, G.: EURON Roboethics Roadmap (2007). http://www.roboethics.org/index_file/Roboethics%20Roadmap%20Rel.1.2.pdf
Yudkowski, E.: AI as a positive and negative factor in global risk. In: Bostrom, N., Cirkovic, M.M. (eds.) Global Catastrophic Risks, pp. 308–345. Oxford University Press, Oxford (2008)
Zimmerman, M.J.: Living with Uncertainty: The Moral Significance of Ignorance. Cambridge University Press, Cambridge (2008)