Baltic J. Modern Computing, Vol. 7 (2019), No. 3, 444-453
https://doi.org/10.22364/bjmc.2019.7.3.11

# Rethinking Radical Imagination: Ethics of Artificial Intelligence

Ernesta MOLOTOKIENĖ

Klaipėda University, Herkus Mantas str. 84, Klaipėda, LT-92294, Lithuania

irama@inbox.lt

**Abstract.** The aim of this paper is to reveal the relation between the power of radical imagination and its impact on ethics of AI. The radical imagination is understood as profoundly dialogic, creative power existing only through collective, critical encounters. One of the main social and ethical challenges for radical imagination is rethinking of existence of ethical AI. The main issues the paper deals with are: 1. How might we assess whether, and in what circumstances, AIs themselves have moral status? 2. How AIs might differ from humans in certain basic respects relevant to our ethical assessment of them? 3. Is it possible to create AIs more intelligent than human, and ensuring that they use their advanced intelligence for good rather than ill? We argue that there is a strong parallel between ethical intention of radical imagination and those behind human codes of ethics for various professions: they are used to openly and transparently communicate to the outside world what are the norms and values in a particular profession, and by doing that to earn trust and acceptance from outside.

**Keywords**: Radical Imagination, Artificial Intelligence, Ethics of AIs.

## Introduction

The Oxford Handbook of the Development of Imagination defines imagination as the capacity to mentally transcend time, place, and/or circumstance (Taylor, 2013). The radical imagination is understood as profoundly dialogic, creative power existing only through collective, critical encounters. One of the main social and ethical challenges for radical imagination is rethinking of existence of ethical AI. Radical imagination in the context of AI understood as the creation of new possibilities for moral action through the modification of meanings.

The possibility of creating AI raises a host of ethical issues. These questions relate both to ensuring that AI does not harm humans and other morally relevant beings, and to the moral status of AI themselves. Much of the success of AI currently comes from a revolution in data science, specifically the use of deep learning neural networks to extract structure from data. For increasingly intelligent AI, it is vital that not just the humans comply with the code, but the AIs too. For the latter, one literally obtains a code of ethics, embedded in the AI’s program. These ethical dimensions require AI designers to act responsibly. A better idea is to incorporate ethical thinking in the design of AI systems, possibly with the use of AI technology itself. Building ethical values into AI requires three things: capacity to acquire ethical values, possibly from humans, knowledge of which human values are important and radical imagination to combine it all.

The article states that AIs with sufficiently advanced mental states, or the right kind of states, will have moral status, and some may count as persons—though perhaps persons very much unlike the sort that exist now, perhaps governed by different rules. And finally, the prospect of AIs with superhuman intelligence and superhuman abilities presents us with the extraordinary challenge of stating an algorithm that outputs super-ethical behavior. These challenges may seem visionary, but it seems predictable that we will encounter them; and they are not devoid of suggestions for present day research directions. We argue there is a strong parallel between ethical intention of radical imagination and those behind human codes of ethics for various professions: they are used to openly and transparently communicate to the outside world what are the norms and values in a particular profession, and by doing that to earn trust and acceptance from outside.

## 1. AI Challenge: Radical Imagination vs. Morality

Within philosophical ethics the term "moral imagination" has been used to refer to an excellent form of moral perception or an aspect of moral judgment, but also to the radical revision of moral understandings and the capacity to generate new possibilities for realizing moral ends or exercising virtue. It means that the radical revision of moral understandings can create new possibilities by changing what is intelligible as an expression of virtue. Changing a meaning does not mean changing everyone's mind. An agent with radical moral imagination succeeds when he achieves a more accurate understanding of the meaning of his experience or acts according to newly conceptualized criteria for virtuous action. The achievement does seem to be brought about within the agents own mind. Acting according to a radically revised meaning may not always be the right thing to do, because it could prove to have been too risky when the results are destructive to oneself or others. One may be unable to live up to the responsibilities generated by a risky choice. Recalling Aristotle, one might hold that some risks are reasonable or noble while others are unreasonable or foolish, and conclude that a person only fails to do the right thing when she takes reckless risks and shies away from noble ones. If the right thing to do is to take the right risks, the actual reception of actions need not factor into moral assessment. There is moral meaning and value in the kind of self-care that transforms the self and its possibilities that is not reducible to the value of fulfilling a duty. While the Kantian grounds the duty of self-respect in the value of ones shared rational nature that is there to be valued, taking responsibility for oneself is done for the sake of a better self and a good future, both of which do not yet exist.

When we face AI, the new radical moral dilemmas arise. Our imagination and our body are always already collaborations with human-made technology and the non-human, and we cannot speak of the radical imagination except as it emerges from a whole web of relations to the nonhuman and more-than-human world. The cultivating the radical imagination in the face of the ongoing “slow-motion apocalypse” of ecologically- and socially-destructive process is a matter of building collective and shared narratives of possibility and resistance. The radical imagination emerges out of radical practices, ways of living otherwise, of collision with AI and of cooperating differently, that reject, strain against, or seek to escape from the racist, patriarchal, heteronormative, colonial, imperial, militaristic, and fundamentalist forms of oppression that undergird our lives.

Currently, artificial intelligence can primarily be found in the area of cognitive intelligence and, to a lesser degree, in the sensorimotor area. Autonomous emotional and social intelligence, is currently not achievable artificially. Nevertheless, there has been significant debate over the past few years about whether it will be possible to categorically identify strong artificial intelligence in the future. Imagination is one of the hallmarks of human intelligence (Asma, 2017), as well as a hallmark of imagination is the ability to reason about counterfactuals (Pearl, 2009). The links between causal reasoning and radical imagination are explored from a probabilistic perspective (Walker and Gopnik, 2013). Humans seek causal explanations because they want to understand the world in simple “cause-effect” relationships. They make analogies to interpret strange worlds in terms of worlds they understand, even though such analogies are imperfect. One of the successes of machine learning, specifically deep neural networks is object recognition (Goodfellow et al., 2016).

Another hallmark of radical imagination is the ability to get curious, to seek out novel situations, and to get bored solving the same problem repeatedly. According to Sridhar Mahadevan (2018), much of the success of AI currently comes from a revolution in data science, specifically the use of deep learning neural networks to extract structure from data. In this case the main question is: can machines really be creative and moral? Can they be considered artists in their own right? If creativity is a defining characteristic of what it means to be human, how a collection of wires and transistors can be considered creative? Ultimately, humans are mere biological machines, and conversely, a thinking, dreaming computer could be considered a silicon life-form. If we can be creative, why not AI? The Harvard philosopher Sean Dorrance Kelly argued that creativity and morality is one of the defining features of human beings and can only exist within a human context (Kelly, 2019). There is no reason to claim that creativity and morality belongs to humans alone. But what can happen if AI achieves general human-level intelligence, including the empowering of imagination and/or morality?

Novel ethical questions arise because artificial minds can have very different properties from ordinary human or animal minds. In the context of radical imagination it is possible to ask how these novel properties would affect the moral status of artificialminds and what it would mean to respect the moral status of such exotic minds. Although current AI offers us few ethical issues that are not already present in the design, the approach of AI algorithms toward more humanlike thought portends predictable complications. Social roles may be filled by AI algorithms, implying new design requirements like transparency and predictability. Sufficiently general AI algorithms may no longer execute in predictable contexts, requiring new kinds of safety assurance and the engineering of artificial ethical considerations. AIs with sufficiently advanced mental states, or the right kind of states, will have moral status, and some may count as persons—though perhaps persons very much unlike the sort that exist now, perhaps governed by different rules. The prospect of AIs with superhuman intelligence and superhuman abilities presents us with the extraordinary challenge of stating an algorithm that outputs super-ethical behavior.

## 2. Philosophical Basics on Ethics of AI

According to Nicolas Cointe, Gregory Bonnet, and Olivier Boissier (Cointe et al., 2016), ethics is a normative practical philosophical discipline of how one should act towards others. Philosophical ethical dilemmas refer to situations in which any available choice leads to infringing some accepted ethical principle and yet a decision has to be made. Addressing ethical issues in AI requires that we ask questions also about how humans relate to AI, and what the implications are of replacing, supplementing or enhancing human thought, experience and action, with machines. So we need to ask questions about differences and similarities between human information processing and behavior, and those of machines, and we need to drill down and ask some fundamental questions about the relationship between ethics and agency. Answers to ethical questions depend upon answers to questions about minds and agency, including questions about responsibility and intention, and about the place of humanity in the moral universe.

The main question of AI ethics is: how we might insert ethical behavior in intelligent machines, given the range of moral uncertainty that exists. This approach focuses upon how moral reasoning might be programmed into machines themselves. Alternatively, one might understand this approach as one which includes consideration of human input, given his close examination of methods for dealing with uncertainties arising from the variety in moral theories of the different human actors involved.

On the other hand one major approach to addressing the control problem in AI is via the notion that machines themselves might be programmed to act ethically. But this putative solution has to grapple with the large range of moral uncertainty that exists—both about which general moral theory is correct, but also regarding the correct response to specific moral problems. There are uncertainties about how moral theory might be applied to practice, how we might go about testing moral theories, whether there is such a thing as moral truth, what constitutes moral justification, and so on. The general approach here seems to assume that there is a correct moral theory that we are all searching for, and that the different theories so far advocated each have some probability of being correct. This models itself rather closely on one particular account of the search for scientific truth, and will be hotly contested by many moral theorists. We can ask if methods of deciding between rival moral theories are ‘fair’, but this in itself may presuppose a shared understanding of fairness, and at times, it would appear that in advocating a decision-making procedure for choosing between moral theories, we are no longer in the realm of the moral, but of the political, where the question is not so much how to decide what the best moral answer is, per se, but how to adjudicate between the claims of those advocating for different theories.

In determining how to characterize, and how to address, the ethical questions that AI presents to us, we need to consider in great detail how the functioning and use of autonomous agents compares to the functioning of human agents. How are we to regulate AI when its very autonomy means that its behavior is hard to predict, and it is also hard to know whether this behavior should be classed as intentional or not — indeed, what would even count as intentional? Answering such questions involves not only deep understanding of AI, but of the basis for attributing intention to humans and how this impacts upon the normative assessment of actions. According Nick Bostrom and Eliezer Yudkowsky, when AI algorithms take on cognitive work with social dimensions — cognitive tasks previously performed by humans — the AI algorithm inherits the social requirements (Bostrom and Yudkowsky, 2011). It is debatable whether human intelligence is truly general — we are certainly better at some cognitive tasks than others (Hirschfeld and Gelman, 1994) — but human intelligence is surely significantly more generally applicable than nonhominid intelligence. The moral constraints to which we are subject in our dealings with contemporary AI systems are all grounded in our responsibilities to other beings, such as our fellow humans, not in any duties to the systems themselves. While it is fairly consensual that present-day AI systems lack moral status, it is unclear exactly what attributes ground moral status. One common view is that many animals have qualia and therefore have some moral status, but that only human beings have sapience, which gives them a higher moral status than non-human animals. This picture of moral status suggests that an AI system will have some moral status if it has the capacity for qualia, such as an ability to feel pain. A sentient AI system, even if it lacks language and other higher cognitive faculties, is not like a stuffed toy animal or a wind-up doll; it is more like a living animal. We argue that AI system also has sapience of a kind similar to that of a normal human adult, then it would have full moral status, equivalent to that of human beings.

These ethical dimensions require AI designers to act responsibly. There is a strong parallel between these values (and intentions) and those behind human codes of ethics for various professions: they are used to openly and transparently communicate to the outside world what are the norms and values in a particular profession, and by doing that to earn trust and acceptance from outside. For increasingly intelligent AI, it is vital that not just the humans comply with the code, but the AIs too. The idea is to incorporate ethical thinking in the design of AI systems, possibly with the use of AI technology itself. According Bostrom, building ethical values into AI requires three things: a capacity to acquire ethical values, possibly from humans, knowledge of which human values are important and radical imagination to combine it (Bostrom, 2014).

As said, much of the success of current AI comes from learning approaches, but now (and in the past) these are being criticized for their lack of explainability, and their incapability to insert and extract domain knowledge. As Martijn van Otterlo argues, ethical AI involves reasoning and learning (Otterlo, 2018). Autonomous AI systems should be designed so that their goals and behaviors can be assured to align with human values throughout their operation. Obviously, value alignment is a multi-objective optimization problem too (Vamplew et al., 2018), and does view AI as an autonomous agent which takes actions and optimizes its behavior according to a utility function. By enabling individual agents to behave ethically and judge the ethics of other agents' actions, is it enough to create a society of well coordinated and collaborative agents acting with human wellbeing as their primary concern?

## 3. Ethics in Human-AI Interactions

AI ethics is a determining factor to the extent autonomous systems are permitted to interact with humans. Therefore, research works focusing on technical approaches for enabling these systems to respect the rights of humans and only perform actions that follow acceptable ethical principles have emerged. However, existing survey papers on the topic of AI governance mostly focused on the psychological, social and legal aspects of the challenges (Arkin, 2016; Etzioni and Etzioni, 2017; Pavaloiu and Kose, 2017). They do not shed light on technical solutions to implement ethics in AI systems. Flexible incorporation of norms into AI to enable ethical user and prevent unethical use is useful since ethical bounds can be contextual and difficult to define as design time. Nevertheless, if updates are provided by people, some review mechanisms should be put in place to prevent abuse (van Riemsdijk et al., 2015). Moral decision-making by humans not only involves utilitarian considerations, but also moral rules. These rules are acquired from past example cases and are often culturally sensitive. Such rules often involve protected values, which morally forbid the commitment of certain actions regardless of consequences.

On the other hand ethics requirements are often exogenous to AI agents. Thus, there needs to be some ways to reconcile ethics requirements with the agents' endogenous subjective preferences in order to make ethically aligned decisions. According to Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R. Lesser and Qiang Yang (2018), the decision- making frameworks with ethical and moral considerations put the burden of codifying ethics on AI system developers. The information on what is morally right or wrong has to be incorporated into the AI engine during the development phase. By assuming that the majority of observed human behaviors are ethical, the proposed approach learns ethical shaping policies from available human behavior data in given application domains. The ethics shaping function rewards positive ethical decisions, punishes negative ethical decisions, and remains neutral when ethical considerations are not involved.

In AI applications which attempt to influence people's behaviors, the principles established by the Belmont Report (Bel, 1978) for behavioral sciences have been suggested to be a starting point for ensuring ethics. The principles include three key requirements: 1) people's personal autonomy should not be violated (they should be able to maintain their free will when interacting with the technology); 2) benefits brought about by the technology should outweigh risks; and 3) the benefits and risks should be distributed fairly among the users (people should not be discriminated based on their personal backgrounds such as race, gender and religion). Human centric values have been incorporated into the objective functions of recent AI-powered algorithmic crowdsourcing approaches.

According to Yilin Kang, Ariel Rosenfeld and Sarit Kraus, one of the application areas in which AI attempts to influence people's behaviors is persuasion agents (Kang et al., 2015; Rosenfeld and Kraus, 2016). The authors conducted a large-scale study to investigate human perceptions on the ethics of persuasion by an AI agent. The ethical dilemma used is the trolley scenario which involves making a participant actively cause harm to an innocent bystander by pushing him on to the train track in order to save the lives of five people. It is a consequentialist ethical outcome which requires the decisionmaker to violate a sacred value. The authors tested three persuasive strategies: 1) appealing to the participants emotionally; 2) presenting the participants with utilitarian arguments; and 3) lying. The three strategies are delivered to some participants by a person playing the role of an authority (the station master of the train station) and by a persuasion agent. The results suggested that participants hold a strong preconceived negative attitude towards the persuasion agent, and argumentation-based and lying-based persuasion strategies work better than emotional persuasion strategies. The findings did not show significant variation across genders or cultures. The study suggests that the adoption of persuasion strategies should take into account differences in individual personality, ethical attitude and expertise in the given domain.

Although emotional appeals may not be an effective persuasive strategy under ethical dilemmas, ethically appropriate emotional responses from agents can enhance human- AI interaction. According to Cristina Battaglino and Rossana Damiano (2015), an approach based on the Coping Theory (Marsella and Gratch, 2003) to allow agents to deal with strong negative emotions by changing the appraisal of the given situation was proposed. The agent assesses the ethical effects of its own actions and other agents' actions. If its own action violates a given moral value, the shame emotion is triggered which serves to lower the priority of continuing with the given action. If another agent's action violates a given moral value, the reproach emotion is triggered in the observing agent which serves to increase social distance with the given agent. The ethical decision-making process is similar to existing individual ethical decision frameworks. The triggering of emotions serves as an implicit reward for the agent and facilitates communications with humans in the loop.

Based on recent advances in AI governance techniques, it appears that most work focused on developing generalizable individual ethical decision frameworks combining rule- based and example-based approaches to resolving ethical dilemmas. In order to learn appropriate rules from examples of ethical decision-making by humans, more work on collecting data about various ethical dilemmas from people with different cultural backgrounds is required. Works on collective ethical decision-making based on multi-agent voting have also appeared, but much work is still needed to design mechanisms to represent ethical preferences by agents. How AI can act ethically when making recommendations to humans and express their ethical judgements affectively are the current foci of ethical human-AI interaction research. In addition, AI engineers need to engage more with the ethics and decision making communities.

In order for ethics to be built into AI, (Burton et al., 2017; Goldsmith and Burton, 2017) advocate that ethics should be part of the AI curricula. This is based on the observation that consequentialist ethics (or ethics based on the utilitarian analysis of possible outcomes) is most closely related to the decision-theoretic frame of mind familiar to today's AI researchers. Deontological ethics (or rule-based ethics) and virtue ethics are less familiar among AI researchers. Understanding deontological ethics can help AI researchers determine which rules are more fundamental and, therefore, should take priority in an ethical decision framework. Understanding virtue ethics, which concerns questions on whom one wishes to become, can help AI researchers frame ethical discussions in the context of changing social conditions (possibly brought on by AI technologies) and guide the incorporation of ethics into AI which shape the paths of learning. Learning materials on these different dimensions of ethics could help AI researchers understand more clearly the topic of ethical decision-making and steer the field of AI towards more emphasis on ethical interactions with humans.

## Conclusions

With AI becoming increasingly ubiquitous in our daily life, we may need to consider revising our current social contracts. Research in this area will help us establish regulations about who is responsible when things go wrong with regard to AI, and how to monitor and enforce these regulations. This research direction is inherently dynamic and interdisciplinary in nature as it must be updated with changing cultural, social, legal, philosophical and technological realities. Addressing ethical issues in AI requires that we ask questions also about how humans relate to AI, and what the implications are of replacing, supplementing or enhancing human thought, experience and action, with machines.

Radical imagination in the context of AI understood as the creation of new possibilities for moral action through the modification of meanings in the interaction with human beings, therefore building ethical values into AI requires three things: capacity to acquire ethical values, possibly from humans, knowledge of which human values are important and radical imagination to combine it all. The radical imagination emerges out of radical practices, ways of living otherwise, of collision with AI and of cooperating differently, that reject, strain against, or seek to escape from the racist, patriarchal, heteronormative, colonial, imperial, militaristic, and fundamentalist forms of oppression that undergird our lives.

There is a strong parallel between ethical intention of radical imagination and those behind human codes of ethics for various professions: they are used to openly and transparently communicate to the outside world what are the norms and values in a particular profession, and by doing that to earn trust and acceptance from outside. For increasingly intelligent AI, it is vital that not just the humans comply with the code, but the AIs too. The idea is to incorporate ethical thinking in the design of AI systems, possibly with the use of AI technology itself.

Our imagination and our body are always already collaborations with human-made technology and the non-human, and we cannot speak of the radical imagination except as it emerges from a whole web of relations to the nonhuman and more-than-human world. The cultivating the radical imagination in the face of the ongoing “slow-motion apocalypse” of ecologically- and socially-destructive process is a matter of building collective and shared narratives of possibility and resistance. Incorporating ethics into AI systems will influence human-AI interaction dynamics. By knowing that AI decisions follow ethical principles, some people may adapt their behaviors in order to take advantage of this and render the AI systems unable to achieve their design objectives.

# References

Arkin, R. C. (2016). Ethics and autonomous systems: Perils and promises. Proc. IEEE, 104(10):1779–1781.

Asma, S. (2017). The Evolution of Imagination. University of Chicago Press.

Battaglino, C., Damiano, R. (2015). Coping with moral emotions. In AAMAS, pages 1669– 1670.

Bel (1978). The Belmont report. Technical report.

Bostrom, N. (2014). Superintelligence, Oxford University Press.

Bostrom, N., Yudkowsky, E. (2011). The Ethics of Artificial Intelligence. In: Cambridge Handbook of Artificial Intelligence, eds. William Ramsey and Keith Frankish. Cambridge University Press. 1–20.

Burton, E., Goldsmith, J., Koenig, S., Kuipers, B., Mattei, N., Walsh, T. (2017). Ethical considerations in artificial intelligence courses. AI Mag., 38(2):22–34.

Cointe, N., Bonnet, G., and Boissier, O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In AAMAS. 1106–1114.

Etzioni A., Etzioni, O. (2017). Incorporating ethics into artificial intelligence. J. Ethics, 21(4):403–418, 2017.

Goldsmith J., Burton, E. (2017). Why teaching ethics to AI practitioners is important. In AAAI, pages 4836–4840.

Goodfellow, I.; Bengio, Y.; and Courville, A. C. (2016). Deep Learning. Adaptive computation and machine learning. MIT Press.

Hirschfeld, L. A., Gelman, S. A. (eds.) (1994). Mapping the Mind: Domain Specificity in Cognition and Culture, Cambridge: Cambridge University Press.

Kang, Y., Tan, A., Miao, C. (2015). An adaptive computational model for personalized persuasion. In IJCAI, pages 61–67.

Kelly, S. D. (2019) Artificial Intelligence. In the MIT Technology Review. Preprint, available at `https://www.technologyreview.com/s/612913/a-philosopher-argues-that-an-ai-can-never- be-an-artist/`

Mahadevan, S. (2018). Imagination Machines: A New Challenge for Artificial Intelligence. Preprint, available at `https://people.cs.umass.edu/~mahadeva/papers/aaai2018-imagination.pdf`

Marsella , S., Gratch, J. (2003). Modeling coping behavior in virtual humans: Don't worry, be happy. In AAMAS, pages 313–320.

Rosenfeld, A., Kraus, S. (2016). Strategical argumentative agent for human persuasion. In ECAI, pages 320–328.

Otterlo van, M. (2018). Ethics and the value(s) of Artificial Intelligence. Ethics and the value(s) of Artificial Intelligence, Nr. 3, 5/19. 206–209.

Pavaloiu A., Kose, U. (2017). Ethical artificial intelligence - an open question. J. Multidisci. Develop., 2(2):15–27.

Pearl, J. (2009). Causality: Models, Reasoning and Inference. Cambridge University Press.

Riemsdijk van, M. B., Jonker, C. M., Lesser, V. (2015). Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. In AAMAS, pages 1201–1206.

Taylor, M. (2013). Transcending time, place, and/or circumstance: An introduction. In Taylor, M., ed., Oxford Handbook of the Development of Imagination. Oxford University Press. 3–10.

Mahadevan, S. (2018). Imagination Machines: A New Challenge for Artificial Intelligence. Association for the Advancement of Artificial Intelligence. College of Information and Computer Sciences. Preprint available at `https://people.cs.umass.edu/~mahadeva/papers/aaai2018-imagination.pdf`

Vamplew, P., Dazeley, R., Foale, K., Firmin, S. and Mummery, J. (2018). Human-aligned AI is a multiobjective problem. Ethics and Information Technology 20(1). 27–40.

Walker, C., and Gopnik, A. 2013. Causality and imagination. In Taylor, M., ed., Oxford Handbook of the Development of Imagination. Oxford University Press. 342–358.

Yu, H., Shen, Z., Miao, C., Leung, C., Lesser, V. R., Yang, Q. (2018). Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18).

## Authors' information

**Ernesta Molotokienė**, Ph. D., is Chair of Department of Philosophy and Culture Studies, Faculty of the Humanities and Educational Sciences, Klaipėda University (Lithuania).

Received September 3, 2019, accepted September 25, 2019