Acta Baltica Historiae et Philosophiae Scientiarum
Vol. 8, No. 2 (Autumn 2020)
DOI : 10.11590/abhps.2020.2.04

Challenges of Aligning Artificial Intelligence with Human Values

Margit Sutrop

Department of Philosophy
University of Tartu
Jakobi 2
Tartu 50090, Estonia
Email: Margit.Sutrop@ut.ee

Abstract: As artificial intelligence (AI) systems are becoming increasingly autonomous and will soon be able to make decisions on their own about what to do, AI researchers have started to talk about the need to align AI with human values. The AI 'value alignment problem' faces two kinds of challenges—a technical and a normative one—which are interrelated. The technical challenge deals with the question of *how* to encode human values in artificial intelligence. The normative challenge is associated with two questions: "*Which* values or *whose* values should artificial intelligence align with?" My concern is that AI developers underestimate the difficulty of answering the normative question. They hope that we can easily identify the purposes we really desire and that they can focus on the design of those objectives. But how are we to decide which objectives or values to induce in AI, given that there is a plurality of values and moral principles and that our everyday life is full of moral disagreements?

In my paper I will show that although it is not realistic to reach an agreement on what we, humans, *really want* as people value different things and seek different ends, it may be possible to agree on what we *do not want* to happen, considering the possibility that intelligence, equal to our own, or even exceeding it, can be created. I will argue for pluralism (and not for relativism!) which is compatible with objectivism. In spite of the fact that there is no uniquely best solution to every moral problem, it is still possible to identify which answers are wrong. And this is where we should begin the value alignment of AI.

Keywords: *artificial intelligence (AI), artificial general intelligence (AGI), AI ethics, moral agent, moral principles, superintelligence (SAI), value, value alignment, value pluralism*

Introduction

The rapid development of artificial intelligence (AI) has enabled a wide range of beneficial applications in various contexts, from healthcare to security and governance. Fields or abilities formerly thought to be exclusively "human" are now being tackled by a machine-learning approach (Arnold *et al.*, 2017, p. 81). As AI becomes more powerful, more attention is being paid to ways of maximising its societal benefit. The growing capabilities of AI and machine learning and the development of countless new applications also pose questions of risk and trust (Sutrop, 2019). AI could be a threat to humanity either through its malicious use by humans (Brundage *et al.*, 2018), unintended consequences (Amodei *et al.*, 2016), or autonomous acts by the artificial system (Farquhar *et al.*, 2017). A joint report, *The Malicious Use of AI: Forecasting, Prevention, and Mitigation* (2018), issued by the leading AI research centres in the UK and the US (Brundage *et al.*, 2018) has recently analysed the landscape of potential security threats from malicious uses of AI technologies and explicitly stated that if adequate defences are not developed soon, AI will threaten our digital, physical and political security.

Currently, AI systems are narrowly dedicated to specific tasks such as speech, face or object recognition, spam filters or autonomous driving, and they are not capable of setting their own goals or choosing the best courses of action across domains. Researchers predict that AI will outperform humans in translating languages by 2024, writing high-school essays by 2026, driving a truck by 2027, or working as a surgeon by 2053; they believe there is a 50 per cent chance of AI outperforming humans in all tasks within 45 years and of automating all human jobs in 120 years (Grace *et al.*, 2018). As robots become increasingly autonomous and able to make decisions on their own, the risk increases that humans will lose control over AI systems. Wendell Wallach and Colin Allen (2009) have suggested that as robots take on more and more responsibility, they must be programmed with moral decision-making abilities, for the sake of our own safety.

What is AI? AI has been described in various ways. It can refer to certain human-designed systems, or to a scientific discipline that encompasses several approaches and techniques, such as machine learning, machine reasoning, and robotics. In this paper, I will discuss AI systems according to the definition of the European Commission High Level Expert Group on Artificial Intelligence as follows:

> AI systems are software (and possibly also hardware) systems designed by humans that, given a complex goal, act in the physical or digital dimension

> by perceiving their environment through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge, or processing the information derived from this data and deciding the best action(s) to take to achieve the given goal. AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions. (European Commission, 2019a)

Visionary scenarios consider the possibility that intelligence equal to our own, or even exceeding it, will already be created within the current century. An AI equipped with human-level or even greater ability, which is able to find a solution when presented with an unfamiliar task, is referred to as Artificial General Intelligence (AGI). Indeed, freed from biological constraints such as limited memory and slow biochemical processing speeds, machines may eventually become more intelligent than we are—with profound implications for us all. It has been hypothesised that if humans can create AGI at a human level of intelligence, such an innovation could give rise to higher and higher intelligence; in combination with recursive self-improvement of the AGI, this would eventually lead to Superintelligence (SAI) (Kurzweil, 2005; Tegmark, 2017). SAI is "any intellect that greatly exceeds the cognitive performance of humans in virtually all domains of interest" (Bostrom, 2014, p. 26). According to a high-level survey of expert opinion, AI specialists estimate that systems will move to superintelligence by the end of this century, and that "the chance is about one in three that this development turns out to be 'bad' or 'extremely bad' for humanity" (Müller & Bostrom, 2016). Even if we are still far from human-level AI and AI advancements may be overhyped, several research programs have been launched on risk assessment of AI systems, attempting to mitigate threats of rapid and uncontrollable AI development. The AI developers themselves worry about how to ensure that an AI does not escape its confines and cause damaging effects (Yudkowsky, 2004). Computer scientists warn, "[a]s autonomous systems become more prevalent in society, it becomes increasingly important that they robustly behave as intended" (Russell *et al.*, 2016). Ethicists stress, "[w]hen independent of the humans that created them, their true 'intelligence' is tested in terms of their status as 'moral beings'" (Iphofen & Kritikos, 2019). Since the potential threats are great, policymakers have emphasised the need to develop ethically-aligned AI which will respect our ethical principles and values. A good overview of various guidelines for ethical, rights-respecting and beneficial AI has been provided in the recent report 'Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for

AI' published by the Berkman Klein Center for *Internet & Society* at Harvard University (Fjeld *et al.*, 2020).

The research program of designing AI that conforms to human values is called **'value alignment'**. AI researchers explain this as follows: "As intelligent systems gain autonomy and capability, it becomes vital to ensure that their objectives match those of their human users; this is known as the value alignment problem" (Fisac *et al.*, 2018, p. 1). 'Value alignment' is defined as a property of an intelligent agent that allows it only to pursue goals and activities which are beneficial to humans (Soares & Fallenstein, 2014; Soares, 2015; Russell *et al.*, 2016; Arnold *et al.*, 2017; Russell, 2019). Identifying the purposes we really desire remains an open question. The opposite of 'value alignment', 'value misalignment' can be explored from two perspectives, either by focusing on the behavior of individual tech artefacts (individual level) or on the functioning of sociotechnical systems (system-level) (Osoba *et al.*, 2020, p. 332).

The AI 'value alignment' problem faces two interrelated kinds of challenges—technical and normative (Gabriel, 2020). The technical challenge deals with the question of "*how* to encode human values in artificial intelligence." The normative challenge is associated with two questions: "*which* values or *whose* values should artificial intelligence align with?" My concern is that AI developers and computer scientists underestimate the difficulty of answering the normative question. They hope that if we find out "what we want, what we really, really want" (Russell, 2015), we can create a beneficial AI which is "humble, altruistic, and committed to pursue our objectives, not theirs". But how are we to decide which objectives or values to induce in AI, given that there is a plurality of values and moral principles and that our everyday life is full of moral disagreements?

In the first part of my paper I will demonstrate how technical and normative challenges of value alignment are interrelated. I will provide a brief overview of the ways in which AI researchers have tried to solve the 'value alignment problem'. I will explain why it is problematic that in value alignment programs the word 'value' is used interchangeably with goals or preferences. In the second part of my paper I will focus on normative challenges of value alignment. Several research papers, AI guidelines and manuals discuss the need to take into account cultural differences, while also respecting universal principles. I will suggest how this apparent inconsistency can be removed.

My thesis is that although it is not realistic to reach an agreement on what we humans *really want,* given that people value different things and seek

different ends, it may be possible to agree on what we *do not want* to happen, considering the possibility that intelligence can be created that is equal to our own, or even exceeding it. I agree with Isaiah Berlin's view as expressed in 'The Pursuit of the Ideal' that even if our ends and moral principles are plural, they are not infinitely many, since they have to remain "within the human horizon" (Berlin, 2013[1947]). I will argue for pluralism (and not for relativism!) which is compatible with objectivism. Despite the fact that there is no uniquely best solution to every moral problem, it is still possible to identify which answers are wrong. And this is where we should begin the value alignment of AI.

A technical challenge: how to align AI with values?

The answer to the question how to align AI with values depends, on the one hand, on the technical possibilities and, on the other hand, on what we want AI to align with. Theories of value alignment vary on what they mean by 'value'. In the context of value alignment, Iason Gabriel (2020) notes that the notion of 'value' can serve as a place-holder for many things: AI could be designed to align with expressed or intended instructions, revealed preferences, informed preferences or desires, interest or well-being, or moral values as defined by the individual or society. In view of these different approaches, value alignment turns out to be an umbrella for various attempts to ensure that general artificial intelligence will remain beneficial to humans.

Through what means do AI researchers and designers hope to reach value alignment? There seem to be two basic—albeit opposite—approaches, either feeding the artificial agent the right principles or applying (cooperative inverse) reinforcement learning. Both approaches—feeding the artificial intelligent with principles and reinforcement learning—have their proponents and critics. With progress in the area of machine learning, the idea of training a system on data (either supervised or unsupervised) has drawn increasing attention. Thus there seems to be a turn away from traditional machine ethics with its top-down articulation of rules, norms or models of behaviour and toward learning values either through imitation or by identifying the preferences of humans. Nevertheless, some critics doubt that beneficial AI can be achieved in this way, arguing that AI can be truly beneficial only if it can become a moral agent which acts on ethical principles or has moral virtues. Iason Gabriel (2020) has recently argued for a principle-based approach. His view is that instead of looking for

"true" moral principles for AI, the real task is "to identify fair principles for alignment that receive reflective endorsement despite widespread variation in people's moral beliefs" (Gabriel, 2020).

The idea behind **reinforcement learning** is that desirability of any state sequence can be expressed as a sum of immediate rewards associated with each state in the sequence. The objectives that the reinforcement learning agent is supposed to follow can be defined by specifying reward function. In the case of inverse reinforcement learning (IRL), an AI system infers the underlying utility function of an agent by observing its behavior. Hadfield-Menell *et al*. (2016) argue that the value alignment problem can only be solved through cooperative inverse reinforcement learning (CIRL) which they describe as "a cooperative, partial-information game with two agents, human and robot; both are rewarded according to the human's reward function, but the robot does not initially know what this is."

Mark O. Riedl and Brent Harrison (2016) have argued that given potentially infinite undesirable outcomes in an open world, if values cannot be easily enumerated by human programmers, they can be learned by **reinforcement**. They hypothesise that an artificial intelligence that can read and understand stories can learn the values tacitly held by the culture from which the stories originate. Riedl and Harrison (2016) describe their preliminary work on using stories to generate a value-aligned reward signal for reinforcement learning agents that preempts psychotic-appearing behavior. They argue,

> Value alignment in a reinforcement learning agent theoretically can be achieved by providing the agent with a reward signal that encourages it to solve a given problem and discourages it from performing any actions that would be considered harmful to humans. A value-aligned reward signal will reward the agent for doing what a human would do in the same situations when following social and cultural norms (for some set of values in a given society and culture) and penalise the agent if it performs actions otherwise. A reinforcement learning agent will learn that it cannot maximise reward over time unless it conforms to the norms of the culture that produced the reward signal. (Riedl & Harrison, 2016)

How can a value-aligned reward signal be learned? Riedl and Harrison's (2016) answer is that, "[a] value-aligned reward signal is produced from the crowdsourced stories in a two-stage process". When humans tell stories to other humans, they skip over many details and events. By crowdsourcing the narratives,

one can get better coverage of all steps involved in a situation and extract the most reliable pattern of events. Riedl and Harrison describe the reinforcement learning process as follows: in the first step, the plot graph learning process "aligns" the crowdsourced example narratives to extract the most reliable patterns of events; in the second step, the plot graph is translated into a "trajectory tree" which is used to produce a reward signal. If the reinforcement learning agent performs an action that is a successor of the current node in the trajectory tree, it receives a reward; if it fails, it receives a small punishment. The aim is to make the reinforcement learning agent solve some problem in the most human-like fashion. The problem is that in many stories humans violate social and cultural norms. The authors' hope is that the most typical human behaviour that emerges from crowdsourced stories is compliant to the norms.

The proposal to achieve value alignment through reinforcement learning has been criticised by Thomas Arnold, Daniel Kasenberg and Matthias Scheutz (2017) as ethically inadequate and insufficient to guide artificial agents in decision making. They point out a number of problematic areas associated with reinforcement learning, including data bias, generalisation issues, and the adequacy of reward functions to represent temporally complex norms (Arnold *et al.*, 2017, p. 81). Arnold *et al.* (2017) claim that even if there is agreement that top-down approaches of giving AI systems ethics would be intractable, IRL on its own cannot solve the problem and train such an agent to be moral. They point out that ethical behavior depends not only on the acts themselves, but on intentions, reasons, norms, and counterfactuals. The question IRL faces is what kind of model could truly learn ethical practices.

Alternatively, value alignment of AI can take the form of **imitation learning**. In their article entitled 'The virtuous machine—old ethics for new technology' Nicolas Berberich and Klaus Diepold (2018) draw inspiration from Aristotle to demonstrate that virtue ethics can provide a solution to the value alignment problem. Instead of focusing on the formulation of duties and maxims (as deontologists do) or on the identification of desirable and maximisable consequences of actions (i.e., happiness), Berberich and Diepold argue that one should concentrate on moral actions performed by virtuous moral agents. "With respect to machines this changes the primary question from 'By what kind of algorithm can a machine choose the right action?' to 'How can we build a machine that, owing to its constitution, acts appropriately in arbitrary situations?'" (Berberich & Diepold, 2018, p. 4). The article concludes by stating,

1. Virtue ethics fits nicely with modern artificial intelligence research and is a promising moral theory.
2. Taking the virtue ethics route to building moral machines allows for a much broader approach than simple decision-theoretic judgement of possible actions. Instead, it takes other cognitive functions into account like attention, emotions, learning and actions. (Berberich & Diepold, 2018, p. 23)

Berberich and Diepold believe that since humans have for centuries learned virtues from imitating the behaviour of virtuous moral agents, virtue theory should be the guiding moral theory for building truly moral machines.

They point out that since virtues are an integral part of one's character, the AI would not have the desire of changing its virtue of temperance (Berberich & Diepold, 2018, pp. 23–24).

The difficulty in this argument is that virtue ethics has a hard time giving reasons for actions, and this means one cannot claim responsibility.

The most influential approach to the problem of AI value alignment has been provided by Stuart Russell, whose ideas are summarised in his recent book, *Human Compatible: AI and the Problem of Control* (2019). Russell's (2019) central idea is that in order to maintain control, we have to design beneficial AI which is "humble, altruistic, and committed to pursue *our* objectives, not theirs". In order to guide AI researchers and developers in thinking about how to create beneficial AI systems, Stuart Russell formulates three principles:

1. The machine's only objective is to maximise the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior. (Russell, 2019, pp. 172–173)

Russell believes that there is plenty of data about human actions (what has been written, filmed or observed) and about attitudes to these actions. In other words, rather than having a detailed ethical taxonomy programmed into them, AI systems should infer human values by observing and emulating our behavior.

It is noteworthy that in his earlier writings and interviews, Russell has spoken about 'human values' which have now been replaced by the expression 'human preferences'. In an interview from 2015, Russell stated:

> To the extent that human values are shared, machines can and should share what they learn about human values. […] by assigning very broad priors over what human values might be, and by making the AI system risk-averse, it ought to be possible to induce exactly the behavior one would want: before taking any serious action affecting the world, the machines engage in an extended conversation with us and an extended exploration of our literature and history to find out *what we want, what we really, really want.* (Russell, 2015)

In his 2019 book, Russell replaces the expression 'human values' with 'human preferences'. He also provides explanations as to why he has dropped the word 'value'. Russell points out that since the word 'value' is often understood as 'moral value', use of this term can be confusing. He does not want to give an impression of attempting to solve moral dilemmas like stops on a trolley route. In his view, such an approach is not worth considering, as in true dilemmas there are good arguments for both sides, and artificial agents cannot cause more harm than humans even if they take a wrong decision. Russell's concern is how to avoid the worst catastrophes and to ensure that we never lose control of machines more intelligent than we are.

In addition, Russell insists that one should not equate value alignment with the wish to instill in machines a single ideal value system which may raise the question whose values should be encoded or who has been given the right to align AI with such values. He says that he uses the word 'value' as a technical term, "synonymous with utility, which measures the degree of desirability of anything from pizza to paradise" (Russell, 2019, p. 178). He also explains what he means by 'preferences': these cover everything one might care about. Furthermore, in order to be beneficial, the purely altruistic machine should aim to maximise the realisation of human preferences. Instead of "putting in values" or preferences, the machines "should learn to predict better, for each person, which life that person would prefer, all the while being aware that the predictions are highly uncertain and incomplete" (Russell, 2019, p. 178).

Russell believes that, in principle, an AI could learn billions of different predictive preference models. With respect to the issue of how the machine should trade off the preferences of multiple humans, Russell (2019, p. 174) refers to the utilitarian principle of "the greatest happiness for the greatest numbers". The crucial question is how the machine might learn about human preferences. His answer is, again, very simple: by observing human choices which reveal information about human preferences. In Russell's opinion, the real complications arise because humans

are not perfectly rational and the machine must take such an imperfection into account (Russell, 2019, p. 177).

It turns out that, in essence, what Russell means by 'value' is in reality a 'preference'. However, the same problem crops up with the concept of 'preference' as it did with the concept of 'value'. Both have multiple meanings, and their use in different disciplines is highly variable. In its most general meaning, a value is that which is worthy, that which is worth having, getting or doing, or that which possesses some property or properties that make it so. Values are action-guiding; they are taken into account when making decisions and when planning activities. In the most general terms, preferences are evaluations; they concern matters of value. Preferences are *subjective* in that the evaluation is typically attributed to an agent. Russell uses the term as *subjective comparative evaluation*, of the form 'Agent *A* prefers *X* to *Y*'. What one prefers is not necessarily something objectively valuable, but rather something subjectively valuable from one's personal point of view.

But this is where the problem begins. We can imagine the situation where the designer of a self-driving car asks potential customers whether they would prefer a car which will protect the persons riding in it or those on the street. Or, even worse, whether they would like a car which violates traffic rules, ignores traffic lights and brings the owner quickly and safely to the destination. Should such preferences be fulfilled by artificial intelligence? By leaving aside moral values and concentrating only on human preferences, it is difficult to understand how situations in which AI might fulfil the preferences of immoral agents can be avoided.

As there are more non-democratic countries than democratic ones in the world, we should be careful not to allow AI to be programmed to do something devastating. Somewhere there will always be efforts underway to create an AI that will be destructive, for example, for military purposes. Some experts have questioned the use of robots in military combat, especially if such machines were to be given some degree of autonomous functions, e.g., being able to independently choose targets to attack with weapons.

What is the solution? AI should be programmed not to fulfil immoral preferences and refuse to do immoral actions but they should be able to distinguish between moral and immoral actions. Thus, besides values we should also provide them with the capacity of moral deliberation to decide when a certain moral principle applies, and to be able to rank values in different contexts. It seems dangerous

to let AI become an autonomous decision-maker without its becoming a moral agent. To be a moral agent entails far more criteria than adopting ethical values. Amitai Etzioni and Oren Etzioni (2018) note that to be a moral agent requires a certain set of attributes:

> (a) Self-consciousness. If the agent is not aware of itself in any given situation, and of the alternative courses that might be followed, then no moral decisions can be rendered. (b) The agent must be aware that she can affect the situation. (c) The agent must be able to understand the moral principles to be employed in arriving at a particular moral choice. (d) The agent must have a motive to act morally. This involves having passions, as otherwise moral preferences are merely intellectual preferences with nothing to fuel the moral action. Some scholars also add that a will or intention is required… (Etzioni & Etzioni, 2018, pp. 239–240).

This argumentation indicates that if we want to avoid the scenario that AI can satisfy both moral and immoral preferences, we need to integrate ethics into AI, not only align it with values. The research programme of AI value alignment should ideally be interdisciplinary, besides computer scientists also involving philosophers, lawyers, economists, and cognitive scientists, psychologists and pedagogical scientists. Its success will depend on how much the AI designers will be able to integrate the knowledge provided by these disciplines in the design of AI. So far, discussions on value alignment have not actively included ethicists. As a result, there is a danger that AI ethics will undergo simplification, reduction or dogmatism. In several working groups which aim to secure beneficial AI, ethics has been boiled down to law or reduced to some specific principle (fairness, transparency, equality or accountability) without providing any reasons why one or another principle should be prioritised. The analysis of the policy documents on ethical AI (Fjeld *et al.*, 2020) showed that the most common governance regime suggested for dealing with AI is international human rights law—64 per cent of all documents contained a reference to human rights.

Annette Zimmermann and Bendert Zevenbergen (2019) have recently issued warnings concerning seven traps into which AI ethics can fall: reductionism, oversimplification, the relativism trap, value-alignment trap, dichotomy; the myopia trap; and the trap of the rule of law. For the purposes of this paper, it is especially interesting how the authors present the relativism trap and the value-alignment trap as two unacceptable options. The relativism trap is associated with the tendency to conclude that since pervasive moral disagreements exist, ethics is relative. Should we argue that relativism is wrong, we risk falling into

the value-alignment trap according to which there must be one morally right answer. If we cannot reach an agreement, we will have failed. I am quite inclined to agree with the authors' view that AI ethics should welcome value pluralism without collapsing into extreme value relativism. We should accept that it may not always be possible to determine "the one right answer", thereby avoiding moral disagreement. However, it is often possible to clarify that at least some paths of action are clearly wrong and some paths of action are comparatively better.

The normative challenge: which values should AI align with?

In order to answer the normative question 'Which values should AI align with?' one must first answer the question of whether values are objective or subjective; universal or culturally relative. Are there absolute values or *prima facie* values? Several regulatory bodies that write guidelines about ethical AI stress that one has to be sensitive to cultural differences while at the same time respecting universal principles.

Thus the *Ethics Guidelines for Trustworthy AI* (European Commission, 2019b) assert that "AI systems should respect the plurality of values and choices of individuals", thereafter claiming that "certain fundamental rights and principles, such as human dignity, are absolute and cannot be subject to a balancing exercise" (European Commission, 2019b, p. 13). In order to make it possible for European citizens to reap the benefits of AI, it needs to be "aligned with our foundational values of respect for human rights, democracy, and rule of law" (European Commission, 2019b, p. 4). IBM's manual entitled *Everyday Ethics for Artificial Intelligence* states: "Care is required to ensure sensitivity to a wide range of cultural norms and values. As daunting as it may seem to take value systems into account, the common core of universal principles is that they are a cooperative phenomenon." (IBM, 2019)

Another important document, *Ethically Aligned Design. A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, published under the auspices of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems poses the rhetorical question: "If machines engage in human communities as autonomous agents, then those agents will be expected to follow the community's social and moral norms. A necessary step in enabling machines

to do so is to identify these norms. But whose norms?" (IEEE, 2019) Again, it seems that in spite of stressing the variety of social and moral norms, this document emphasises that values-based design of AI should respect universal ethical principles and internationally recognised human rights. Indeed, the document states that

> […] whether our ethical practices are Western (e.g., Aristotelian, Kantian), Eastern (e.g., Shinto, School of Mo, Confucian), African (e.g., Ubuntu), or from another tradition, honoring holistic definitions of societal prosperity is essential versus pursuing one-dimensional goals of increased productivity or gross domestic product (GDP). Autonomous and intelligent systems should prioritize and have as their goal the explicit honoring of our inalienable fundamental rights and dignity as well as the increase of human flourishing and environmental sustainability. (IEEE, 2019, p. 1)

It seems to me that all these documents point in the same direction—namely, that we should accept plurality of values while at the same time prioritising certain specific values, and doing both of these things without necessarily providing explanations of how pluralism of values and objectivity of ethics are compatible. In the following I will try to explain how pluralism can accommodate both persistent moral disagreements and the objectivity of ethics.

Pluralism differs from both relativism and subjectivism. Unlike the relativist, who believes that rightness depends on the culture or society, and unlike the subjectivist, who believes that rightness depends on the individual, the pluralist holds that in some cases the question of what is right lacks a determinate answer. Rightness is indeterminate, as there is irreducibility of values. As shown by Susan Wolf in the article 'Two levels of pluralism' (1992), this irreducibility can happen on two levels of decision: on the level of values and on the level of moral systems. Wolf says that on the first level we **constrain moral judgements within a moral system**, i.e., we have to decide which value, for example, equality or liberty, is more important and we have to admit that there is no right answer to this question. On the second level, **we constrain moral assessments of the systems themselves**. Here we have to decide which moral system, with its different understanding of moral rightness, we should prefer (for example, Kantianism or utilitarianism or virtue ethics). Wolf (1992) shows that there is a possibility that after a careful assessment it may turn out that these value systems are all incommensurably good, and living according to any of these moral systems will lead to a morally good life.

Also Isaiah Berlin supports pluralism which is compatible with objectivism. Pluralism means for him that "there are many different ends that men may seek and still be rational, fully men, capable of understanding each other" (Berlin 2013[1947], p. 11). In fact, we do have shared values which form bridges between people and make intercommunication between cultures in time and space even possible. "But our values are ours, and theirs are theirs" (Berlin 2013[1947], p. 11). We might criticise the values of other cultures, but we have to understand where they come from, what it means to live the way they do. Here Berlin emphasises people's empathy and capacity for imagination, their need to put themselves in the shoes of a representative of another era or culture.

Pluralism in Berlin's view does not exclude a world of objective values or ends that men pursue for their own sake. Ends and moral principles are plural, but they are not infinitely many, since they have to remain "within the human horizon" (Berlin 2013[1947], p. 12). We should be able to distinguish between objectively good and bad life forms and condemn slavery, sexism, racism, or Nazism. Given that there are people who argue from an egoistic position of power, with the intention of preserving their preferred state or that some lack imagination and empathy for what it would mean to be in another's situation (in cases such as slavery, Nazism, female circumcision, racial or gender hatred, disparagement of minorities, abuse of animals), we need to have the opportunity to say that such things are unacceptable. This remains so even if we favor value pluralism.

We can get a better idea of how pluralism can be compatible with objectivity by reading John Kekes's book *The Morality of Pluralism* (1993). Kekes makes a distinction between primary and secondary values. Primary values are the same for everyone, for they constitute the minimum requirements of good lives and are derivable from the facts of human nature. For example, we share the same physiological and psychological needs and capacities. Thus, the same things, goods pertaining to the self, intimacy, and social order are needed by everybody to live a good life. At the same time, the identity of secondary values depends on variable social and personal circumstances. For example, whereas killing is universally considered to be evil, what sort of killing counts as murder may be dependent on context. Kekes explains that secondary values are contingent on primary ones; either they are particular forms of expression of these values (for instance, the primary value of intimacy will take the form of certain sexual practices) or they are genuinely new goods, the realisation of which depends on the presence of primary values (for instance, in order to derive pleasure from creative participation in art or science, one's elementary needs for food and shelter need to be satisfied)

(Kekes, 1993, pp. 42–43). Primary and secondary values can be either moral or non-moral. While moral values (both primary and secondary) affect others, non-moral values (both primary and secondary) affect the agent or occur naturally. Primary values (both moral and non-moral) are universal requirements of all good lives, while secondary values count as benefits or harms because a tradition or a conception of a good life makes them so (Kekes, 1993, p. 45).

Kekes explains that pluralists differ from relativists in holding primary values to be objectively justifiable by referring to our **common nature**. He admits that even monists must not reject plurality of values, provided there is some overriding value which allows authoritative ranking of the plural values. Kekes points out that there is a great variety of monist conceptions of what the overriding value or some principle of ranking values may be: the utilitarian ideal of the greatest happiness for the greatest number of people, the welfare economists' goal of preference satisfaction, the Kantian principle of the categorical imperative, the contractarian list of fundamental human rights, or something else (Kekes, 1993, p. 47). Thus, the main difference between monists and pluralists is that for the former there is one overriding value or principle, while for the latter values are manifold and contextual.

Kekes explains that we usually aim to construct a conception of the good life out of the primary and secondary values that we ourselves favor and that our tradition supplies. Although he argues that for pluralists there is no authoritative way of settling the value conflict, there still seems to be at least one overriding principle for comparing and ranking different values. To my mind, by arguing that primary values originate in our common human nature, Kekes is also suggesting an authoritative overriding principle which helps to compare and rank values. For instance, he proposes that in value conflicts one should try to reduce secondary values to primary values (which have their source in human nature), by making it explicit that in spite of all differences, the fact that situatedness in various historically conditioned traditions and individual life-situations has led to differentiation in our conceptions of a good life, we continue to have similar human needs and capacities. Although Kekes himself has not recognised that his own appeal to common human nature is monistic in a similar way as the appeal to human rights framework, categorical imperative or general happiness principle, it is still possible that he ultimately speaks for pluralism and not monism. I think that as long as one accepts that there can be a plurality of ultimate values, all of which can be equally justified, one is a pluralist. Thus, indeed, pluralism really is compatible with objectivism.

Conclusions

I have tried to show that it is good that AI researchers have identified the importance of ensuring that AI will remain beneficial and that therefore AI should align with human values. However, since the term 'value' only functions as a placeholder and in reality one could limit oneself to making AI fulfil human preferences, the problem may arise that AI will not necessarily act morally, but potentially also satisfy immoral preferences. Thus, in order to secure ethical AI one should compel it to follow moral principles or obtain moral values.

In the course of the argument the question emerged whether there are universal values or whether AI should reckon with cultural differences. I hope I have shown that value pluralism can accommodate both moral disagreements caused by diversity of values and the objectivity of ethics, accepting that there are no uniquely correct answers to moral questions. Moral ambiguity does not necessarily imply relativism. Accepting plurality does not mean that anything goes. We can identify wrong answers as long as we share our expectations of morality. Indeed, the aim of morality is to regulate our life in society, enabling us to live together peacefully and allowing human beings to satisfy their various needs: physiological, psychological, social, cultural and intellectual.

Thus, value alignment should begin with seeking agreement on negative values—what we do not want to lose and what we do not want to happen if it becomes possible to create artificial intelligence with intellectual capacities equal to or beyond our own. Our efforts to define what is immoral should keep in mind that morality's function is to satisfy the basic human needs which we all share.

Value plurality and objectivity can be joined if we accept that there are many moral principles by which one can justify and rank moral values. Different moral theories can prefer different ultimate principles. One can live a moral life according to all of these. Nevertheless, there are fewer choices about what is morally wrong, and thus we should seek agreement on what is clearly immoral.

Acknowledgements

The research done for this paper was supported by Kone Fellowship at the Helsinki Collegium for Advanced Studies at the University of Helsinki and by the Centre of Excellence in Estonian Studies (European Union, European Regional Development Fund).

The paper profited from the discussion following the oral presentations at the 15th Estonian Philosophy Conference in Tallinn in August 2019 and at the research seminar of the Helsinki Collegium for Advanced Studies in November 2020. I am very grateful to Dr. Tiina Kirss for her careful editing and help with English expression and to Dr. Mari-Liisa Parder for her help with references. Also, I would like to thank Dr Peeter Müürsepp for his encouragement to write this paper as well as Kait Tamm and Piret Frey for careful editing.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J. & Mané, D. (2016), 'Concrete problems in AI safety,' *ArXiv*, 25 July, v2. Retrieved from https://arxiv.org/abs/1606.06565 [accessed 2 Dec 2020]

Arnold, T.; Kasenberg, D. & Scheutz, M. (2017), 'Value alignment or misalignment—what will keep systems accountable?' The AAAI -17 Workshop on AI, Ethics, and Society, WS-17-02. Retrieved from https://aaai.org/ocs/index.php/WS/AAAIW17/paper/view/15216/14648 [accessed 2 Dec 2020]

Berberich, N. & Diepold, K. (2018), 'The virtuous machine—old ethics for new technology?' *ArXiv*, pp. 1–25. Retrieved from http://arxiv.org/abs/1806.10322 [accessed 2 Dec 2020]

Berlin, I. (2013[1947]), 'The Pursuit of the Ideal,' in H. Hardy (ed.) *The Crooked Timber of Humanity*, Princeton & Oxford: Princeton University Press.

Bostrom, N. (2014), *Superintelligence: Paths, Dangers, Strategies*, Oxford: Oxford University Press.

Brundage, M.; Avin, S. *et al.* (2018), *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.* Retrieved from https://maliciousaireport.com/ [accessed 2 Dec 2020]

Etzioni, A. & Etzioni, O. (2018), 'Incorporating ethics into artificial intelligence,' in A. Etzioni (ed.) *Happiness is the Wrong Metric: A Liberal Communitarian Response to Populism*, Library of Public Policy and Public Administration, 11, Cham: Springer, pp. 235–252. Retrieved from https://www.springer.com/gp/book/9783319696225 [accessed 2 Dec 2020]

EU Commission (2019a), *A Definition of AI: Main Capabilities and Disciplines.* Retrieved from https://www.aepd.es/media/docs/ai-definition.pdf [accessed 2 Dec 2020]

EU Commission (2019b), *Ethics Guidelines for Trustworthy AI*. Retrieved from https://ec.europa.eu/futurium/en/ai-alliance-consultation [accessed 2 Dec 2020]

Farquhar, S.; Halstead, J.; Cotton-Barratt, O.; Schubert, S.; Belfield, H. & Snyder-Beattie, A. (2017), *Existential Risk Diplomacy and Governance*, Global Priorities Project. Retrieved from https://www.fhi.ox.ac.uk/wp-content/uploads/Existential-Risks-2017-01-23.pdf [accessed 2 Dec 2020]

Fisac, J. F.; Gates, M. A.; Hamrick, J. B.; Liu, C.; Hadfield-Menell, D.; Palaniappan, M.; Malik, D.; Sastry S. S.; Griffiths, T. L. & Dragan A. D. (2018), 'Pragmatic-pedagogic value alignment,' in N. Amato, G. Hager, S. Thomas & M. Torres-Torriti (eds.) *Robotics Research: Springer Proceedings in Advanced Robotics*, vol. 10, Cham: Springer, pp. 49–57. https://doi.org/10.1007/978-3-030-28619-4_7

Fjeld, J.; Achten, N.; Hilligoss, H.; Nagy, A. C. & Srikumar, M. (2020), 'Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI,' *Berkman Klein Center Research Publication*, no. 2020-1, pp. 1–71. https://doi.org/10.2139/ssrn.3518482

Gabriel, I. (2020), 'Artificial intelligence, values, and alignment,' *Minds and Machines*, vol. 30, pp. 411–437. https://doi.org/10.1007/s11023-020-09539-2

Grace, K.; Salvatier, J.; Dafoe, A.; Zhang, B. & Evans, O. (2018), 'When will AI exceed human performance? Evidence from AI experts,' *ArXiv*. Retrieved from https://arxiv.org/abs/1705.08807 [accessed 2 Dec 2020]

Hadfield-Menell, D.; Dragan, A.; Abbeel, P. & Russell, S. (2016), 'Cooperative inverse reinforcement learning,' *ArXiv*. Retrieved from https://arxiv.org/abs/1606.03137 [accessed 2 Dec 2020]

IBM (2019), *Everyday Ethics for Artificial Intelligence*, IBM Corp. Retrieved from https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf [accessed 2 Dec 2020]

IEEE (2019), *Ethically Aligned Design. A Vision for Prioritizing Human well-being with autonomous and intelligent systems*, The Institute of Electrical and Electronics Engineers, Incorporated. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf [accessed 2 Dec 2020]

Iphofen, R. & Kritikos, M. (2019), 'Regulating artificial intelligence and robotics: ethics by design in a digital society,' *Contemporary Social Science*, vol. 2041, pp. 1–15. https://doi.org/10.1080/21582041.2018.1563803.

Kekes, J. (1993), *The Morality of Pluralism*, Princeton, NJ: Princeton University Press.

Kurzweil, R. (2005), *The Singularity Is Near*, New York: Viking.

Müller, V. & Bostrom, N. (2016), 'Future progress in artificial intelligence: a survey of expert opinion,' in V. Müller (ed.) *Fundamental Issues of Artificial Intelligence*, Synthese Library, 376, Cham: Springer, pp. 553–571.

Osoba, O.; Boudreaux, B. P. & Yeung, D. (2020), 'Steps towards value-aligned systems,' *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 332–336. https://doi.org/10.1145/3375627.3375872

Riedl, M. & Harrison, B. (2016), 'Using stories to teach human values to artificial agents,' The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence, 'AI, Ethics, and Society', February 12–13, 2016, Technical Report, WS-16-02.

Russell, S. (2015), 'Will they make us better people?' 2015: What do you think about machines that think? *Edge*. Retrieved from https://www.edge.org/response-detail/26157 [accessed 2 Dec 2020]

Russell, S. (2017), 'Provably beneficial artificial intelligence,' in *The Next Step: Exponential Life, BBVA OpenMind*. Retrieved from https://people.eecs.berkeley.edu/~russell/papers/russell-bbvabook17-pbai.pdf [accessed 2 Dec 2020]

Russell, S. (2019), *Human Compatible. AI and the Problem of Control*, London: Allen Lane, Penguin Books.

Russell, S.; Dewey, D. & Tegmark, M. (2016), 'Research priorities for robust and beneficial artificial intelligence,' *AI Magazine*, vol. 36, no. 4, pp. 94–105.

Soares, N. (2015), *The Value Learning Problem*, Technical Report, 2015-4, Machine Intelligence Research Institute.

Soares, N. & Fallenstein, B. (2014), *Aligning Superintelligence with Human Interests: A Technical Research Agenda*, Technical Report, 2014-8, Machine Intelligence Research Institute.

Sutrop, M. (2019), 'Should we trust artificial intelligence?' *Trames*, vol. 23, No. 4, pp. 499-522.

Taddeo, M. (2010a), 'Modelling trust in artificial agents. A first step towards the analysis of e-trust,' *Minds and Machines*, vol. 20, no. 2, pp. 243–257.

Taddeo, M. (2010b), 'Trust in technology: A distinctive and a problematic relation'. *Knowledge, Technology and Policy, vol. 23*, nos. 3–4, pp. 283–286.

Tegmark, M. (2017), *Life 3.0: Being Human in the Age of Artificial Intelligence*, London: Allen Lane.

Wallach, W. & Allen, C. (2009), *Moral Machines: Teaching Robots Right From Wrong*, Oxford: Oxford University Press.

Wolf, S. (1992), 'Two levels of pluralism,' *Ethics*, vol. 102, no. 4, pp. 785–798.

Yudkowsky, E. (2004), *Coherent Extrapolated Volition*, San Francisco, CA: The Singularity Institute. Retrieved from https://intelligence.org/files/CEV.pdf [accessed 2 Dec 2020]

Zimmermann, A. & Zevenbergen, B. (2019), 'AI ethics: seven traps,' *Freedom to Tinker.* Retrieved from https://freedom-to-tinker.com/2019/03/25/ai-ethics-seven-traps/ [accessed 2 Dec 2020]

Margit Sutrop is a professor for practical philosophy at the Institute of Philosophy and Semiotics, University of Tartu and a Kone fellow at the Helsinki Collegium for Advanced Studies at the University of Helsinki.