Author's manuscript.

Appeared in: M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (eds.). 2018. *Envisioning Robots in Society –Power, Politics, and Public Space, Proceedings of Robophilosophy 2018 / TRANSOR 2018*, Series; Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam.
February 14–17, 2018, University of Vienna, Austria

Making Metaethics Work for AI: Realism and Anti-Realism

Michał KLINCEWICZ[a,1] and Lily FRANK[b]
[a] *Department of Cognitive Science, Institute of Philosophy, Jagiellonian University, Kraków, Poland*
[b] *Philosophy and Ethics, Technical University of Eindhoven, Eindhoven, Netherlands*

Abstract. Engineering an artificial intelligence to play an advisory role in morally charged decision making will inevitably introduce meta-ethical positions into the design. Some of these positions, by informing the design and operation of the AI, will introduce risks. This paper offers an analysis of these potential risks along the realism/anti-realism dimension in metaethics and reveals that realism poses greater risks, but, on the other hand, anti-realism undermines the motivation for engineering a moral AI in the first place.

Keywords. Artificial intelligence, metaethics, risk, engineering

1. Introduction

This paper offers an analysis of the practical consequences of design and use of artificial intelligence in light of current state of art in philosophical metaethics. It focuses on AI that will be tasked with making explicitly moral decisions, serving as moral advisers, or contributing to morally charged decision making processes [1-5]. Some examples of practical use of these technologies include help in sentencing of convicted felons; mediating discussions in ethical discussion in hospitals and physicians' offices; or risk-benefit calculations in counter-terrorism operations.

AI designed to play a role in morally charged human decision-making processes and activities should aim to at least approximate human moral psychology that underlies moral decisions, albeit in a way that improves on the typical foibles that come with it. If a full replication of human moral psychology was required, then it would be unclear why a moral AI should come into existence in the first place. More consultation with equally fallible humans would be enough. On the other hand, if an AI issued moral advice that greatly diverged from what we would expect from human moral psychology, its advice would not be judged as relevant or as having to do with human concerns.

Elsewhere we argued that the metaethical assumptions that engineers make when designing an AI that in some way approximates human moral psychology will determine the engineering challenges they will face [6]. Engineering and design decisions for such an AI will always be informed by tacit or explicit metaethical assumptions, including but not limited to: nature of moral judgment, characterization of moral motivation, the existence of mind-independent moral properties, status of moral epistemology, and what

[1] Michał Klincewicz, Institute of Philosophy, Grodzka 52, Krakow 33-044, Poland. E-mail: michal.klincewicz@uj.edu.pl

Author's manuscript.

Appeared in: M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (eds.). 2018. *Envisioning Robots in Society –Power, Politics, and Public Space, Proceedings of Robophilosophy 2018 / TRANSOR 2018*, Series; Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam.
February 14–17, 2018, University of Vienna, Austria

differentiates the moral domain of knowledge from other domains. The analysis offered here takes into account those engineering challenges and offers a further analysis of risks. The dimension of difference we focus on is realism/anti-realism about moral properties.

2. Moral Epistemology

Realism in metaethics is primarily the ontological view that moral properties or moral facts exist and are mind-independent. This also means that the way the world is, including the way human beings are, determines what actions are morally right and wrong, independently of human aims or interests. Many versions of realism include a view about semantics of sentences that contain moral terms. Moral sentences, according to realism, express propositions that can be true or false and moral terms refer to moral properties.

Alternatively, anti-realism is the view that moral properties or facts do not exist, or at least they do not exist in a mind-independent way. This view may also reject the semantic position that sentences with moral content can be true or false or can refer to moral properties. This means that, at least to some extent, human beliefs about right and wrong are the truth makers for statements containing moral content.

If moral realism is true and combined with a naturalistic position on morality, then moral facts could be thought to be akin to facts in science, which means that moral knowledge demands the kind of effort that scientific knowledge requires. Cornell realism is a prominent example of naturalist realism, which assumes that "ethical facts and properties are exhaustively constituted by natural ones" [7, p. 550]. According to naturalist realism, moral facts and properties are constituted by or supervene on basic physical facts and properties, in a way that psychological, sociological, or historical facts are constituted by or supervene on basic physical facts and properties. An AI designed to serve as a moral advisor based on naturalist realism will therefore function like a moral microscope, revealing parts of the world previously unavailable to the naked eye. The moral AI may therefore have access to moral truths that unaided capacities for moral sensitivity and reasoning would not.

A realist view assumes that the world may be full of moral facts to which we currently have no access, but could at some point in the future discover. Naturalist realism could liken these to facts about particle physics, which were obscure to people living just a hundred years ago. This analogy brings into relief an important potential risk for a realist AI: unexpected decisions, policies, or behavior of the AI may conflict with *prima facie* moral norms, and we may have no obvious way to decide whether to follow the advice of the AI or these *prima facie* moral norms. When it comes to science we have some evidence of scientific method leading to discovery of new scientific facts. In the case of morality, there is much more debate about reliable, appropriate, and justified methods of gaining moral knowledge and discovering and clarifying the moral facts. The risk here is that the realist assumptions brought to bear on AI design get the moral facts wrong and we are compelled to follow the advice given by the AI.

This is not to say that there are no defensible methods in the study of ethics to which the realist can help themselves. Some Cornell realists, such as David Brink [8], defend a coherentist moral epistemology based on wide reflective equilibrium, in which "our reasoning in the sciences as well as in ethics involves the continuing accommodation of empirical information to a body of more theoretical views already tentatively in place, making mutual adjustments to achieve the best overall fit...among our views of different

Author's manuscript.

Appeared in: M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (eds.). 2018. *Envisioning Robots in Society –Power, Politics, and Public Space, Proceedings of Robophilosophy 2018 / TRANSOR 2018*, Series; Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam.
February 14–17, 2018, University of Vienna, Austria

levels of generality" [9, p. 102]. However, methods in moral epistemology cannot be verified in the same ways that scientific methods often can. It is harder to independently verify whether the predictions moral theories make are correct or not.

If a scientific theory posits the existence of a molecule with certain properties we can conduct an experiment to see whether or not those properties manifest under appropriate circumstances. Such experiments are more controversial and harder to conceive in ethics. Furthermore, foundationalist moral epistemologists would likely demur that our system of moral beliefs should be based on a small number of basic foundational beliefs, which provide support for the rest and not on a wide equilibrium. Engineers designing an AI on realist assumptions will face the burden of justifying which of these realist epistemologies is right. If they opt for foundationalism, they will face further risks that would come from their decision about which moral norms the rest of the moral structure will be built upon.

If the engineer opts for moral realism, they should aim to create an AI that is better than we are at discovering the moral facts. The problem is that we do not have such a method at hand. It is not a good idea to build an AI as if a method of discovery of moral facts existed. It would be as if an engineer built a device for discovering facts about molecules without some falsifiable empirical hypothesis about the nature of molecules. Such an AI would likely routinely give the wrong answer to first-order moral questions. While this may be somewhat innocuous with a device for detecting molecules, it is risky in the moral case. Giving the wrong answer to a first-order moral question such as: *is it moral give this person a liver transplant as opposed to this one?* could not only get the moral facts wrong, it could lead to outright immoral behavior that leaves us worse off than we would have been without advice from the moral AI. This does not mean that anti-realism more accurately describes the moral phenomenon or experience. Rather, when we consider these risks, anti-realism seems to fare better than realism at mitigating risks that are a consequence of no known method of discovering moral facts.

3. Biases

The main source of risk for anti-realism are the inherent risks and uncertainties that anti-realism generates on its own, outside of it being engineered into an AI. The first of these is that anti-realist assumptions can give credence to the view that morality is arbitrary and that anything goes in the moral domain. Importing this problem to an AI would exacerbate it and likely result in distrust in the moral advice the AI gives.

Secondly, importing anti-realism into an AI would make it difficult to distinguish between moral judgments and moral biases, which are responsible for a wide range of moral illusions. Anti-realism in general deals better with this problem than realism since it makes it in principle difficult to make such a distinction. Realism, on the other hand, assumes that we can make such a distinction and thus can lead to situations where moral illusions masquerade as genuine moral judgments. However, when this problem is brought into an AI that is presumed to not have biases, a similar masquerading can take place.

Sinnott-Armstrong persuasively argues that our moral intuitions are subject to a range of biases, including framing effects [10]. In the context of risky choices framing refers to the phenomenon that when people are presented with two options that are in fact identical but described in such a way that in one case the losses are emphasized and

Author's manuscript.

Appeared in: M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (eds.). 2018. *Envisioning Robots in Society –Power, Politics, and Public Space, Proceedings of Robophilosophy 2018 / TRANSOR 2018*, Series; Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam.
February 14–17, 2018, University of Vienna, Austria

in the other case the benefits are emphasized, people avoid risks when it comes to gains and are more willing to take risks when it comes to avoiding losses [11].

One famous example that illustrates a framing bias is the "Chinese disease" experiment [12]. Participants of that experiment were presented with two sets of choices. In the first they read the following vignette:

Imagine that the US is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows.

> **Program A**: If Program A is adopted, 200 people will be saved.
> **Program B**: If Program B is adopted, there is 1/3 probability that 600 people will be saved, and 2/3 probability that no people will be saved.

Even though the expected value of these choices is the same, participants consistently prefer option A because it presents a certainty of saving 200 lives. In the second part of the experiment participants were presented with the same vignette and the following two choices:

> **Program C**: If Program C is adopted 400 people will die.
> **Program D**: If Program D is adopted there is 1/3 probability that nobody will die, and 2/3 probability that 600 people will die.

In this case participants chose option D, they preferred risk taking over the certainty of 400 deaths. Even though options A and C, and options B and D are the same, participants choices did not reflect this.

Sinott-Armstrong, among many others, conclude that the influence of biases on moral decisions considerably undermines the epistemic value we attach to any of our moral intuitions. This presents a problem for both realism and anti-realism. As already mentioned, this is a very serious problem for a realist moral AI. If we aim to create AI that are moral decision-making aides, we want to make sure we do not introduce our biases from the very beginning and then impart realist metaphysics into the device.

Another class of biases in the moral domain are fairness effects, which have to do with people's willingness to accept risks. In short, "people accept risks more readily if the risk distribution is perceived as fair" [13]. These findings come from experiments done by Keller and Sarin [14] and Sjöberg [15]. Probability neglect is another cognitive bias often discussed in the context of public policy making and public fears surrounding very bad, but very unlikely outcomes [16-18]. This body of research suggests that "when a hazard stirs strong emotions, most people will pay an amount to avoid it that varies little even with extreme differences in the starting probability...when the probability of loss is very low, people will tilt toward excess action. They will favor precautionary steps even if those steps are not justified by any plausible analysis of expected utility" [19, p. 116].

Sunstein points out, for example, that the public fear of flying after the September 11th, 2001 terrorist attack was vastly out of proportion to the likelihood of being killed by terrorists. In cases where the outcome is emotionally dreadful (other cases include nuclear meltdowns, toxic waste, carcinogenic foods, and bad reactions to vaccination) excessive public policy actions tend to be taken and yet people will remain insensitive to

Author's manuscript.

Appeared in: M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (eds.). 2018. *Envisioning Robots in Society –Power, Politics, and Public Space, Proceedings of Robophilosophy 2018 / TRANSOR 2018*, Series; Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam.
February 14–17, 2018, University of Vienna, Austria

the ever-diminishing risk of the negative event [20]. The opposite tends to be true when a highly emotionally desirable reward is present. When we play the lottery, we are able to discount the exceeding small probability of winning and prefer taking disproportionately bigger gambles for larger gains [21]. The inverse relationship between risk and benefit is especially relevant in the field of new technologies: "Activities or technologies that are judged high in risk tend to be judged low in benefit and vice-versa" [22]. Slovic et al. 1999 experiments on toxicologists suggest that this relationship is mediated by the level of negative or positive affect the subject doing to the risk assessment experiences regarding the technology or policy in question [23].

Framing and fairness biases pose less of a problem for anti-realism than for realism in general. If there are no moral facts or properties in the world that exist in a mind independent way then there is little discernible difference between moral beliefs and moral biases. An anti-realist may be fine with that. An AI that builds biases in is therefore not going to make matters worse. It will merely be giving out advise. This speaks strongly for the engineer to deliberately opt for anti-realism while constructing or designing a moral adviser AI. There are other risks, however, that make this option problematic.

4. Problems for Anti-Realist AI

If the realist AI functions ideally like a moral microscope, an anti-realist AI functions like an automated food critic. It issues sophisticated expert opinions, which are ultimately matters of individual taste. Depending on the version of anti-realism involved, this AI could even contradict itself. Whoever uses it may therefore get the impression that right and wrong are also matters of taste. This creates a risk of the moral AI being an example of moral judgments being arbitrary opinions that are verified by our intuitions. In order to avoid this, special care would have to be taken to ensure that the average user of the AI is sophisticated enough to distinguish anti-realism from arbitrariness.

Programming an AI with anti-realist assumptions in mind generates further problems that do not affect realist AI: an existential problem and a practical problem. The existential problem has to do with the added value of creating an AI that can serve as a moral adviser or decision aid, given the assumption that moral facts and properties are merely reflections of human mental states and that what constitutes a judgment about right or wrong, good or bad may be subjective. In other words, it is not clear why we should build an anti-realist moral AI at all.

The practical problem is best stated a conditional: if we think the anti-realist AI can add value, engineers will have to imbue the AI with some foundational moral values in a non-arbitrary way. This latter possibility is likely to reify the moral values accepted in the place and time of the AI's creation or the personal idiosyncrasies of its designers. Neither possibility is likely to result in a situation where the AI morally advises anyone; it merely convinces people to moral positions they already hold.

The practical problem generates a risk that can be attributed to the development of many, if not all, new technologies, namely, reifying norms that are not desirable. Technology inevitably mediates values [24], but it can also change them, reify the value reflected in the technology. For example, wheelchair ramps on government buildings embody the value of egalitarian access to civic life (reflecting our values). Dating apps like Tindr encourage short-term relationships, even when its users may be looking for long-term relationships (changing values) [25]. Moral AI that is specifically designed to

Author's manuscript.

Appeared in: M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (eds.). 2018. *Envisioning Robots in Society –Power, Politics, and Public Space, Proceedings of Robophilosophy 2018 / TRANSOR 2018*, Series; Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam.
February 14–17, 2018, University of Vienna, Austria

aid in moral reasoning shares in this special risk, because its explicit purpose and function is to issue moral guidance, thus changing values.

In this context, biases, such as those discussed in section 2, raise alarms about our ability to rationally engage in the very project of evaluating the benefits of AI as a moral adviser. It could be that the engineers or society at large falls into one or more of these biases. In turn, we can imagine such an AI being taken seriously by medical boards, politicians, and other people in positions of power, and then agreeing to the AI moral advice, even though it is grounded in mere biases. This is a similar, but importantly different risk that faces the realist AI. For a realist AI this risk exists at the level of moral epistemology and first order moral facts: it may be getting them wrong. The anti-realist AI adviser has this problem by design, so the problem is ultimately practical in that they concern the usefulness of the device in moral decision-making.

5. Summary Analysis and Recommendations

Realism poses serious risks that come from epistemological difficulties regarding moral facts. Anti-realism, on the other hand, seems ill-suited to the task since it undermines the motivation to engineer a moral AI in the first place. In the following table we summarize the analysis from section 1, 2, and 3. For the sake of clarity we mention the most salient risks for each position in a row.

Table 1. Summary of risks of engineering a moral AI associated with realism and anti-realism

Realism	Anti-realism
If realism is false, risk is low because an anti-realist metaethics is *prima facie* compatible with many sets of first order normative propositions.[2]	If realism is true but we do not have the correct moral epistemology to rely on, then the realist AI may yield false moral recommendations, bad advice, misleading its users, and telling them to do potentially immoral things.
If realism is true but we do not have the correct moral epistemology to rely on, then the realist AI may yield false moral recommendations, bad advice, misleading its users, and telling them to do potentially immoral things. This is the most troubling possibility, since the moral adviser would leave its users potentially worse off than before.[3]	If anti-realism is true, it becomes unclear what the added value of a moral AI adviser is. The advice it gives may as well be disregarded as one moral view among many, which are not mutually exclusive. Users of the AI have no reason to take its advice over their own moral intuitions.[4]

We considered three paradigmatic cases of the limitations to human risk perception and assessment, such as the fairness effect, framing effects, the inverse relationship between risk and benefit, and probability neglect, which are well documented in the decision theory and psychological literature. We also speculated about the potential risks involved in creating a moral adviser AI with realist and anti-realist assumptions. From

[2] Arguably this is also true for realism. As a metaethical position it is neutral with respect to say, Kantianism or Utilitarianism. But this kind of neutrality requires that we have a suitable moral epistemology. For the anti-realism comparing first order normative theories or individual beliefs can only be done based on non-normative practical concerns, see Prinz [26] on moral progress, for example.

[3] Of course, if the epistemology is partly accurate, that is it gets the moral recommendations right some of the time, then this particular risk would be mitigated.

[4] The exception being that the AI could give empirical information relevant to moral decisions (e.g. what the likely consequences of an action might be).

Author's manuscript.

Appeared in: M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (eds.). 2018. *Envisioning Robots in Society –Power, Politics, and Public Space, Proceedings of Robophilosophy 2018 / TRANSOR 2018*, Series; Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam.
February 14–17, 2018, University of Vienna, Austria

our analysis we conclude that at the present time anti-realism generates fewer serious risks than realism, since there is no chance of it ever giving moral advice that purports to be getting the moral facts true, while in fact doing the opposite. It merely advises, as a food critic might, leaving it ultimately up to its users' intuitions to decide. However, if an anti-realist moral adviser is built, it becomes unclear what its ultimate value may be. People being advised by it may as well ignore its advice, just as people that ignore a food-critic's advice often do. Ultimately, it is their own taste that matters. This puts in question the very project of engineering a moral adviser AI in the first place, at least in its anti-realist version. Whatever human beings achieve by engaging in dialogue and using some version of reflective equilibrium is likely to be just as morally authoritative, if not more, than whatever may come as a result of the moral AI.

References

[1] J. Savulescu and H. Maslen. Moral Enhancement and Artificial Intelligence: Moral AI? In J. Romportl, E. Zackova, J. Kelemen (eds), *Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics* **9**, 79-95. Springer, Cham, 2015.
[2] M. Klincewicz. Artificial Intelligence as a Means to Moral Enhancement. *Studies in Logic, Grammar and Rhetoric* **48** (2016), 171-187.
[3] J. Borenstein and R. C. Arkin. Robotic Nudges: The Ethics of Engineering a More Socially Just Human Being. *Science and Engineering Ethics* **22 (1)** (2016), 31-46.
[4] J. Borenstein and R. C. Arkin. Nudging for Good: Robots and the Ethical Appropriateness of Nurturing Empathy and Charitable Behavior. *AI & Society* (2016), 1-9.
[5] A. Giubilini and J. Savulescu. The Artificial Moral Advisor: The 'Ideal Observer' Meets Artificial Intelligence. *Philosophy & Technology* (2017), 1-20.
[6] L. Frank and M. Klincewicz. Metaethics in Context of Engineering Ethical and Moral Systems. In *2016 AAAI Spring Symposium Series* (2016), 208-213.
[7] W. J. FitzPatrick. Recent Work on Ethical Realism. *Analysis* **69 (4)** (2009), 746-60.
[8] D. O. Brink. *Moral realism and the foundations of ethics.* Cambridge University Press, Cambridge, 1989.
[9] N. Sturgeon. Ethical Naturalism. In David Copp (ed) *Oxford Handbook of Ethical Theory*. Oxford University Press, Oxford, 2006.
[10] W. Sinnott-Armstrong, Walter. *Moral Skepticisms.* Oxford University Press, Oxford, 2006.
[11] D. Kahneman and A. Tversky. Choices, Values, and Frames. *American Psychologist* **39 (4)** (1984), 341-50.
[12] A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science* **211(4481)** (1981), 453-458.
[13] L. Sjöberg and B-M. Drottz-Sjöberg. Fairness, risk and risk tolerance in the siting of a nuclear waste repository. *Journal of risk research* **4(1)** (2001), 75-101.
[14] L. R. Keller and R. K. Sarin. Equity in social risk: Some empirical observations. *Risk Analysis* **8(1)** (1988), 135-146.
[15] L. Sjöberg. *Risk and society: Studies in risk generation and reaction to risks, Vol 3*. Taylor & Francis, 1987.
[16] Y. Rottenstreich and C. K. Hsee. Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological science* **12(3)** (2001), 185-190.
[17] C. R. Sunstein. Probability neglect: Emotions, worst cases, and law. *The Yale Law Journal* **112(1)** (2002), 61-107.
[18] P. M. Sandman, N. D. Weinstein, and W. K. Hallman. Communications to reduce risk underestimation and overestimation. *Risk Decision and Policy* **3(2)** (1998), 93-108.
[19] R. Zeckhauser and C. R. Sunstein. Dreadful possibilities, neglected probabilities. In M. Kerjan and P. Slovic (eds) *The irrational economist: making decisions in a dangerous world*. pp. 116-24. Public Affairs Press, 2010.
[20] C. R. Sunstein. Terrorism and probability neglect. *Journal of Risk and Uncertainty*, **26(2-3)** (2003), 121-136.
[21] R. Dobelli. *The art of thinking clearly: better thinking, better decisions.* Hachette, UK, 2013.
[22] A. S. Alhakami and P. Slovic. A psychological study of the inverse relationship between perceived risk and perceived benefit. *Risk analysis* **14(6)** (1994), 1085-1096.

Author's manuscript.

Appeared in: M. Coeckelbergh J. Loh, M. Funk, J. Seibt, M. Nørskov (eds.). 2018. *Envisioning Robots in Society –Power, Politics, and Public Space, Proceedings of Robophilosophy 2018 / TRANSOR 2018*, Series; Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam.
February 14–17, 2018, University of Vienna, Austria

[23] P. Slovic. Trust, emotion, sex, politics, and science: Surveying the risk-assessment battlefield. *Risk analysis* **19(4)** (1999), 689-701.
[24] P-P. Verbeek. *Moralizing technology: Understanding and designing the morality of things.* University of Chicago Press, Chicago, 2011.
[25] L. Frank and M. Klincewicz. Swiping Left on the Quantified Relationship: Exploring the Potential Soft Impacts. *The American Journal of Bioethics* **18(2)**, (2018), 27-28.
[26] J. Prinz. *The Emotional Construction of Morals*. Oxford University Press, New York, 2007.