DE GRUYTER OPEN

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 48 (61) 2016

DOI: 10.1515/slgr-2016-0061

Michał Klincewicz
Jagiellonian University

ARTIFICIAL INTELLIGENCE AS A MEANS TO MORAL ENHANCEMENT

Abstract. This paper critically assesses the possibility of moral enhancement with ambient intelligence technologies and artificial intelligence presented in Savulescu and Maslen (2015). The main problem with their proposal is that it is not robust enough to play a normative role in users' behavior. A more promising approach, and the one presented in the paper, relies on an artificial moral reasoning engine, which is designed to present its users with moral arguments grounded in first-order normative theories, such as Kantianism or utilitarianism, that reason-responsive people can be persuaded by. This proposal can play a normative role and it is also a more promising avenue towards moral enhancement. It is more promising because such a system can be designed to take advantage of the sometimes undue trust that people put in automated technologies. We could therefore expect a well-designed moral reasoner system to be able to persuade people that may not be persuaded by similar arguments from other people. So, all things considered, there is hope in artificial intelligence for moral enhancement, but not in artificial intelligence that relies solely on ambient intelligence technologies.

Keywords: moral enhancement, artificial intelligence, rhetoric, ambient intelligence.

1. The moral lag problem and moral enhancement

In a notorious opening section to the *Nicomachean Ethics*, Aristotle tells us that only those already brought up well and in possession of the moral virtues can engage in the study of ethics (*Nicomachean Ethics* 1095b4–6). If Aristotle is right, children who are not educated and socialized in the right way early in their development will inevitably lead a crippled moral life, if not be outright vicious. Whatever one thinks of Aristotle's views, it is difficult to resist at least the spirit, if not the letter of that grim assessment. There is a lot of evidence that bad behavior and bad environments correlate and even evidence that bad environments lead to significant neuroanatomical changes (Glenn and Raine 2014).

ISBN 978-83-7431-512-8 ISSN 0860-150X

The explanation of this state of affairs is not to be found exclusively in the way that children are brought up. Even with the right development, conflicts between moral judgments and rationality broadly construed are common. As McIntyre points out, "in the presence of powerful contrary inclinations that bring about a failure to be resolute, resisting rationalization and remaining clearheaded about one's reasons to act can constitute a modest accomplishment" (MacIntyre 2006, 311). Summing up, Aristotle's view about engaging in the study of ethics and in leading a moral life stays pertinent at least in spirit: it is difficult to be moral.

There are reasons to think that leading a moral life is even more difficult today than in Aristotle's time. Many contemporary societies face rapid technological advance and moral practice is not catching up. Some urge that the moral psychology that we inherited from our ancestors leaves us unprepared to confront moral issues raised by the advent of computers, biotechnology, and new forms of medicine. We can be disposed to morally ignore anonymous, large groups of people and care only about those closest to us (Meyer, Masten et al. 2012). We can be apathetic towards enormous amounts of suffering and put too much emphasis on the relatively lower suffering of individuals who may be physically or socially closer to us (Slovic 2010).

Let us call the sum of these problems the Moral Lag Problem. The Moral Lag Problem is a shorthand name for all the things that cause us to be not as moral as we could or should be. Persson and Savulescu, among others, see this situation not only as unfortunate, but as an existential threat to our species (Persson and Savulescu 2011). They argue that this particular set of moral shortcomings will inevitably lead to calamities like global ecological disaster or apocalyptic wars and possibly even destroy life on Earth (Rees 2004).

What may not be immediately apparent is that technological advance may also hold a solution to the Moral Lag Problem. Technology may be used to change the environment in a way that makes it easier for people to act morally. Technology may also be used to change moral agents themselves in a way that makes it easier for them to act morally – this is the idea that Parsson and Savulescu take up in their proposal for wholesale moral bioenhancement.

Solutions to the Moral Lag Problem that focus on changing the environment come in many forms, most of which aim to dampen or even eliminate the effects of environmental stimuli that may lead to morally compromising behavior (Van den Hoven, Lokhorst et al. 2012). For example, street lights that make people less aggressive or restaurant advertisements that

offer morally relevant information, such as whether the food is "fair trade" or locally produced. Whatever in the environment causes people to be less moral than they could be – new technologies, foibles of their psychology, or vicious socialization – may have less of an effect with such measures in place.

Solutions that focus on changing the agent are as diverse (Persson and Savulescu 2013). Pharmaceuticals, such as serotonin re-uptake inhibitors (Crockett 2014), modafanil (de Sio, Faulmüller et al. 2014), and oxytocin (De Dreu 2012) all have been proposed as possible interventions that would lead to moral improvement in the people that take them. There are also possible direct neural interventions into relevant brain areas (Baertschi 2014). Some have also proposed genetic manipulation (DeGrazia 2013; Harris 2010). Finally, there are the more old-fashioned methods, such as moral education (Harris 2013).

A seldom considered but promising solution to the Moral Lag Problem lies with rapid technological advances in computing technology and artificial intelligence (AI) in particular. Savulescu and Maslen have argued that it can be a route to moral enhancement (Savulescu and Maslen 2015). In their view, "the moral AI would monitor physical and environmental factors that affect moral decision-making, would identify and make agents aware of their biases, and would advise agents on the right course of action, based on the agent's moral values" (Savulescu and Maslen 2015, 80). Along similar lines, Borenstein and Arkin have given concrete examples of the way in which an AI implemented in a robot can serve as a moral nudger (Borenstein and Arkin 2016).

This paper examines some of the available engineering options that can achieve this aim and concludes that the most promising approach will borrow from extant work in machine ethics and be coupled with extant engineering solutions in AI, such as those proposed by Savulescu and Maslen among others. This is a unique new possibility for changing the agent morally and presents the main original contribution of the present work.

Section 2 critically examines the ways in which a moral AI could be implemented and tasked with moral enhancement using extant technologies, and focuses on the work of Savulescu and Maslen. Section 3 demonstrates how a better solution may be found in machine ethics coupled with AI and lays out the idea of an AI adviser that can generate moral arguments based on standard first-order normative theories.[1] Section 4.1 considers an objection to this proposal, which is the charge that such an AI module would be an unacceptable form of moral paternalism, and then Section 4.2 answers that objection.

2. Moral enhancement with artificial intelligence

Savulescu and Maslen's (2015) proposal for a moral AI has two important features. First, it is a complex of subsystems: a moral environment monitor, which provides morally-relevant information about one's surroundings; a moral prompter, which informs the user about morally important situations or events in the environment; a moral adviser, which nudges the user to what they should do; and a moral observer, which observes the behavior of the user (Savulescu and Maslen 2015, 84).

Second, the functioning of these systems is to some extent determined by exogenous interference from the user. For example, the user can set up which values the various subsystems should prioritize over others. It may be useful to quote them here at length to make it clear what it is that they have in mind:

> Before offering advice in the first instance, the AI would ask the agent to indicate which of a long list of morally significant values or principles he holds and is guided by. Importantly, he is also asked to assign a weight (between 0 and 1) to each value. Thus, an agent who cared very strongly about not harming others (non-malevolence), a bit about not breaking the law (legality) and not at all about protecting the environment (environmental protection) might assign them weights of 1, 0.5 and 0, respectively. We suggest a non-exhaustive list of possible values to include:
> - Autonomy (of others – e.g. not being paternalistic)
> - Benevolence (helping others)
> - Non-malevolence (not harming others)
> - Justice/fairness
> - Legality
> - Environmental protection
> - Family/significant relationships
> - Fulfilling duties/commitments/promises
> - Maximising net utility (making sure overall benefits outweigh overall costs)
>
> For any given scenario, the AI would compute the extent to which the courses of action open to the agent would uphold or compromise these values (fully uphold value = 1; fully compromise value = −1), amplifying or diminishing based on the weight indicated by the agent. The AI would then use these weighed values to suggest the best course of action (Savulescu and Maslen 2015, 88).

These settings can also be adjusted by the system to reflect what the user does. For example, if the user acts in a way that prioritizes Legality over Justice/fairness, the system would adjust its settings to reflect that in its future operation.

It may also be important to emphasize that Savulescu and Maslen propose an AI in its weak version. Weak AI, as opposed to strong AI, does not aim at human-level cognition in a computer, but merely at using the techniques in the field of AI to generate better software. Strong moral AI would presumably aim at the level of moral agency or at least a human-level capacity for moral reasoning. Weak moral AI "gathers, computes and updates data to assist human agents with their moral decision-making" (Savulescu and Maslen 2015, 84).

Their proposal for a new type of moral enhancement is for a moral AI that would be profiled to pick-up, process, and present *morally* relevant information to the human user. This proposal has the virtue of relying on technologies that have all been either developed or are in the process of being developed as a part of ongoing research programs into ambient technologies and affective computing (Picard 2000). However, it faces the problem of being too tied up with the values of its users and not being able to provide a solution to the Moral Lag Problem.

This becomes apparent when we look into the details of the functioning of the proposal's subsystems. For one, the purportedly morally relevant information that the moral environment monitor provides is likely to be morally relevant only for particular individuals in particular situations, if at all. Most morally relevant information about food may not be all that important to someone who is desperately hungry, on a special diet, or a nutritionist.

Similarly with the information provided by the moral prompter, which could, for example, "prompt agents faced with justice/fairness decisions to mitigate gender biases by making themselves gender-blind where possible" (Savulescu and Maslen 2015, 88). Information about gender bias may not be important to someone that cares about it a great deal already.

If the agent has overcome these biases or does not feel that overcoming them is a part of their moral values, then the AI's advice becomes altogether useless. On the other hand, if the agent cares about these values, then it is not clear what the AI adds that their own reflective attitude to gender-bias would not. In order to have the moral prompter prompt in a morally relevant manner, it would have to have an insight into the moral values of the person beyond whatever platitudes it may have received in the process of its creation or it would simply need to have the right moral values, whatever those are.

Finally, the moral adviser module, which seems to be the heart of the moral AI, may be informed:

> that high levels of testosterone and sleep deprivation are factors that make individuals more likely to make utilitarian decisions (to over-value maximizing net utility), [and thus] warn the agent that, for this reason, his current assessment of the advised course of action may not be consonant with his more enduring moral principles (Savulescu and Maslen 2015, 88–9).

But the moral adviser module would not be giving good advice if the agent's more enduring moral principles are wrong or if the agent is biased. It is also possible that in some situations over-valuing net utility is the morally right thing to do and the moral adviser would actually impede rather than enhance moral behavior if it gave advice against it.

In sum, the key problem with the proposal is that all of the component parts of the moral AI are systematically tied up to the agent's own moral values and these values might be based on morally compromising biases and beliefs. In response, the authors may stress the value of moral pluralism, which is not the same as relativism, and point out that "common human morality, whilst not always in agreement on finer points, does require some objective standards" (Savulescu and Maslen 2015, 91). Pluralism avoids paternalism and preserves the agent's autonomy.

Moral relativism is the view that moral values are relative and thus mind-dependent properties. Relativism has significant theoretical and empirical problems that are outside of the scope of this paper to discuss. One of the most troubling is that any value, including the most abhorrent, can end up being moral if relativism is true. Given this, if Savulescu and Maslen's proposal implies relativism, it would put it at a disadvantage.

Moral pluralism, on the other hand, is the view that moral values are diverse, which suggests that there is no one single ethical or metaethical theory that is correct, but that we may use some or all of them to reach morally justified conclusions. This view is compatible with moral realism and the view that moral properties are mind-independent. Given what we know about the proposed AI, it is clear that it functions in a morally pluralistic way – each user can select their own settings for values. That, however, is not the same as saying that any values would do, so the proposal is not relativistic.

This response goes some way in answering the worry, but it opens the proposal to another objection, namely, that the moral AI they propose will not lead to moral enhancement. Savulescu and Maslen urge that not only does the proposed moral AI preserve the agent's moral autonomy, it in fact enhances it by prompting the agent to reflect on and assent to moral values and principles, and by equipping the agent to be more successfully guided by the values and principles he endorses (Savulescu and Maslen 2015, 91). But

this is simply not enough to face up to the Moral Lag Problem in a serious way. To do that, the moral AI would have to play not only an advisory or facilitative role, but also a normative one.

Imagine a situation in which a hypothetical person, a white American male named John, could report that he just saw a police officer verbally abusing a Chinese person. Minimally, we would expect John to report the police officer's behavior to someone, so the police officer could be reprimanded. But John will not do even this, because of his bias against Chinese people. John has been socialized to have this bias in his childhood by racist parents, friends, and neighbors.

John is then hooked up with a moral AI with all of its different subsystems, as presented by Savulescu and Maslen, and faced with an identical situation. Could we expect John's behavior to be any different than the first time? That depends on why it is that John did not report the abuse the first time.

If John's not reporting was due to an unconscious bias against Chinese people, then there is a chance that he may act differently now that the moral AI gives him more information about the situation at hand. However, we have good reason to believe that John's not reporting was due to reasons John knows about and which are grounded in his bias against Chinese people. If this latter possibility is what actually happened, then the moral AI is unlikely to change how John acts.

To see this consider how the different modules function. The moral environment monitor in the AI may tell John that the person is Chinese and that that is morally relevant. John, being a racist, would probably recognize this as helpful, in that it helped him recognize that he should not report the police officer's abuse. This is how racism works.

The moral prompter may also tell John that this is a situation in which he is faced with a moral dilemma that involves reacting appropriately to an observed transgression of laws by someone else. John may find it helpful to know that this is a situation where laws are broken by the police officer, but he would also be compelled by his racist views to not do anything about it. Again, this is how racism works.

In fact, we have a good reason to think John would be compelled by his racist values to rationalize the police officer's transgression away. It would be more than a modest accomplishment, to use MacIntyre's phrase, if John resisted rationalization and remained clearheaded about his reasons for acting. It is easy to find examples of such rationalization at work. For example, in Duluth, Minnesota in 1920 a lynch mob was prompted by police officers that lynching African Americans for purportedly assaulting a white woman

is a transgression of US laws (Fedo 2016). Nonetheless, that did not stop the lynching.

The moral adviser module would also ultimately fail to change John's behavior in the critical moment. From the list of values provided by Savulescu and Maslen, presumably the ones relevant to John's situation would be justice/fairness and benevolence. For the sake of argument, let us say that John cares a great deal about these and assigns them maximum weight in the moral AI. This would make the moral adviser AI "compute the extent to which the courses of action open to the agent would uphold or compromise these values" (Savulescu and Maslen 2015, 88). The moral adviser may therefore advise John that his values of justice/fairness and benevolence recommend a course of action that includes reporting the police officer's abuse.

However, it is highly unlikely that John would act differently given this advice from the AI. It is well-known that racists typically dehumanize or devalue those against whom they are prejudiced and rationalize conflicts between their racism and their other values by "othering" (Dominelli 1998). This involves putting emphasis on the differences of the other person and then drawing the conclusion that these differences are essential, morally relevant properties, which somehow justify treating the other differently than they would have been otherwise.

Like any other racist, John would likely do something like this when confronted with the hypothetical situation, even if advised by his moral AI adviser not to do so. We could also expect that if the AI attempts to change the weights assigned to John's values of justice and benevolence to compensate for John's failure to act, John would protest. In his eyes, John cares about justice and benevolence a great deal, but Chinese people just do not count in certain situations, because they are "other" in some way. Given what we know about racist psychology, we have no reason to think that John would act any different if he was prompted or advised by the moral AI.

The racist John example easily generalizes to other situations where people may act immorally due to the sorts of things that are at the heart of the Moral Lag Problem. For example, persons disposed to violent confrontation when challenged by a lifetime in an environment where that kind of behavior is encouraged are unlikely to change when prompted by an AI that they are about to get into a confrontation. If informed by a system that monitors their autonomous nervous system that they may be getting angry they are more likely to take it as a suggestion to get ready for a fight rather than reflect on the moral dimension of their decisions or behavior.

Being prompted to reflect on information from the environment or moral values and principles one endorses is not sufficient to lead one to act morally in situations where doing the morally right thing conflicts with these values. So while Savulescu and Maslen's proposal that the AI be systematically tied to the agent's own moral values avoids the problem of paternalism and preserves autonomy, it also precludes the AI from playing a normative role. And this means that the system they propose is not robust enough to achieve even modest moral enhancement and certainly not on a scale that would be sufficient to overcome the Moral Lag Problem. We have good reason to expect that people enhanced with this AI will go on doing whatever they were doing.

3. Machine ethics and moral enhancement

The more promising alternative for changing behavior by an AI lies in machine ethics. Machine ethics, in the sense relevant to the present discussion, is an interdisciplinary project that has as its aim the creation of a computer program that can engage in moral reasoning (Anderson and Anderson 2011). Moral reasoning leads a moral agent to make moral decisions based on relevant reasons, broadly construed. Machine ethics aims to implement that in a computer.

A computer program that can engage in moral reasoning could play a normative role in a way in which the moral AI discussed in the previous section could not. First of all, such a system can give suggestions on how to behave in concrete situations. Secondly, and perhaps more importantly, such a system could give reasons that a rational person could be persuaded by. Thirdly, it can, if prompted to do so, give answers to first-order normative questions, such as "should I report this to the authorities?" with a definite "yes" or "no" and then also provide reasons in support of that answer.

Now let us suppose that racist John is hooked up to an AI system that can do all three of these things: give suggestions, give reasons, and give answers to first-order normative questions. So, in addition to being informed about the moral environment, prompted about the moral dimension of a police officer abusing a Chinese person, being advised about this being a situation that involves breaking laws, and so on, John is also being explicitly informed about what he should do, morally speaking, and why. For example, the AI would suggest to John that he should report this abuse to some authorities or, if he is willing to go beyond what is minimally morally required of him, that he tell the police officer to stop, because that would maximize happiness or something like that.

Racist John would of course likely ignore this sort of advice. But if he does, the AI would be able to give him reasons why he should not ignore it. For example, the AI could generate an argument that starts with some uncontroversial premises and leads to the conclusion that not reporting the police officer's abuse is immoral. This argument may take as premises Kantian deontological theory, utilitarianism, Rawls' theory of justice as fairness, or some other normative theory. It could choose one of these or all of them at once, perhaps depending on the moral philosophy closest to John's, or perhaps on some other non-arbitrary reason that informed the engineers responsible for designing the system.

Racist John is still likely to resist even these AI-generated arguments. However, if John is in any way reason-responsive and cares about morality, he is forced to confront the conflict that they generate. This is because John's racist ideology, like any other racist ideology, is, at least arguably, incompatible with ethical theories like Kant's, utilitarianism, and so on. So John has to respond to this conflict either by deciding to act on the reasons provided by the AI or by intentionally rejecting these reasons in favor of his racist beliefs. John could also accept both, even though they are contradictory, but he would still at least be made aware that they are in conflict. Either way, in this situation John cannot simply rationalize away the immorality of what is happening in front of him by "othering" or something else like it.

All things considered, John is still likely to favor his racist beliefs over the reasons provided by the AI or resign himself to having beliefs in contradictory things. But there are two important differences between his state of mind now that he has been advised by the AI that can engage in moral reasoning and his state of mind after being advised, informed, and prompted by the AI proposed by Savulescu and Maslen. First, now John is aware of the conflict between his racism and ethical theory. This gives him a genuine opportunity to change his mind in light of reasons that he can rationally accept. Second, there is a chance, however small, that John will trust the AI more than his racist beliefs and act on the moral reasons provided by the AI.

The second possibility, which is that John would listen to the AI over his own racist beliefs, is particularly promising. There is empirical evidence that strongly suggests that people can be successfully persuaded to change their behavior by appropriately designed technologies (Hamari, Koivisto et al. 2014). Given a wealth of evidence of technology for behavior change, we have some reason to hope that a similar effect can be used in cases such as racist John's.

If the agent-computer relationship is appropriately groomed, people can trust computers very much in the way that they can trust each other (Muir 1987, Nickel 2013). Agent-computer trust can be especially high when it comes to automation, sometimes even to the detriment of human agents, who may end up trusting an automated system when not appropriate (Parasuraman, Molloy et al. 1993). This means that agent-computer trust can be managed to some extent and match or even outrun levels of agent-agent trust (Lee and See 2004).

What may explain the difference between agent-agent and agent-computer trust is that one of the operative psychological mechanisms for dealing with uncertainty in agent-agent interactions is affect, while in agent-computer interactions, the operative psychological mechanism is confidence (de Vries 2004). Confidence in the performance of an artifact, such as a computer program or an automated airplane pilot, often depends on non-social factors such as past performance (Carlson, Desai et al. 2014, Sauer, Chavaillaz et al. 2015). Agent-agent trust, on the other hand, engages many disparate mechanisms, such as social cognition and emotions, which renders it subject to a number of biases (Emerson 1976).

The sometimes undeserving trust that people put in technology, especially automated technology that is designed to generate confidence in the user, can be exploited in an appropriately designed moral reasoning AI to persuade people like racist John to change their behavior. While John is unlikely to listen to reasons and arguments given by another person, extant evidence suggests that he may well listen to reasons and arguments from a machine. Machine advice may override John's distrust in people advice. So we have a reason to think that a system that can advise and give reasons would be more successful in changing behavior than the kind of system proposed by Savulescu and Maslen.

This system would have to take advantage of the best research on human-computer interaction and the psychological mechanisms of agent-computer trust to work. This is in addition to the daunting engineering challenges that face the project of creating an artificial moral reasoner, which could supply people with reasons and arguments. All of this implies a massive amount of research, which may never be carried out. Nonetheless, the potential for addressing the Moral Lag Problem is there.

Generalizing the proposed solution with adequate attention paid to customization and user experience is a promising avenue for overcoming dispositions to behave immorally on a large scale. This is because a moral AI can play a normative role and provide reasons that rational, reason-responsive people can be persuaded by. While the advice from the moral AI proposed

here is not likely to change the behavior of groups bound on doing harm, like the one in 1920 Duluth, Minnesota, it does give a promising alternative to pharmaceuticals and neural intervention for individualized moral enhancement.

4.1. Paternalism and autonomy: objection

The main problem with the moral AI proposed by Savulescu and Maslen lies in its systematic connection with the values of its user. This connection makes it difficult for the AI to play a normative role necessary to seriously address the Moral Lag Problem. The virtue of the AI having a connection to the values of its user is that it allows for moral pluralism, which in turn avoids the problem of moral paternalism. Paternalism

> is the interference of a state or an individual with another person, against their will, and defended or motivated by a claim that the person interfered with will be better off or protected from harm. The issue of paternalism arises with respect to restrictions by the law such as anti-drug legislation, the compulsory wearing of seatbelts, and in medical contexts by the withholding of relevant information concerning a patient's condition by physicians (Dworkin 2016).

Similarly in the context of moral AI, the issue of paternalism arises because the moral AI interferes with the person that uses it on the assumption that it would be better for them to be interfered with in that way.

Whether paternalism is problematic depends on the nature of the interference with the person. As Gerald Dworkin points out,

> in the case of automobile seat belts, for example, the restriction is trivial in nature, interferes not at all with the use or enjoyment of the activity, and does, I am assuming, considerably reduce a high risk of serious injury. Whereas, for example, making mountain-climbing illegal completely prevents a person from engaging in an activity which may play an important role in his life and his conception of the person he is (Dworkin 1972, 83).

Dworkin observes that mountain-climbing can play an important role in the conception of the kind of person one is. One can conceive of themselves as a mountain-climber or a passionate hobbyist of mountain-climbing. Interfering with that conception constitutes paternalistic intereference that undermines autonomy.

Dworkin's observation about activities, such as mountain-climbing, can be extended to beliefs. Some beliefs can play an important role in one's life and how one conceives of themselves. Good examples of such beliefs are religious beliefs or beliefs about values, which can for some people even be

central to that conception. Sadly, among these core beliefs one can sometimes find racist beliefs, such as John's. Racism can be a value that people hold central to the conception of who they are.

This connection between some beliefs and one's conception of the kind of person one is creates a problem for paternalistic interference. Paternalistic interference that has as its aim a change in beliefs may undermine that conception and violate autonomy (Shiffrin 2000). Some philosophers argue that even if interference of this kind is motivated by the need for moral enhancement, everyone should be given the freedom to fall (Harris 2011, DeGrazia 2013). Racist John should have the freedom to hold racist beliefs.

Left unanswered the objection eliminates the motivation for having an AI that can engage in moral reasoning and play a normative role in human behavior. If the moral AI proposed in Section 3 is paternalistic in a way that infringes on autonomy, then it is not an improvement on Savulescu and Maslen's proposal, but a step in the wrong direction.

4.2. Paternalism and autonomy: reply

The paternalism objection can be answered by stressing that while the moral reasoner AI does interfere in the conception of oneself, it does so in a way that does not undermine autonomy. The relevant distinction here is between methods of paternalistic interference in the conception of oneself that *coerce* and those that *persuade*. Coercion strips agents of their say in the constitution of their conception of themselves, thus violating autonomy. Persuasion leaves it up to them to decide, which means that agents undergoing paternalistic interference of this kind in some sense consent to the change. The moral reasoner AI presented in Section 3 falls into the latter category.

Persuasion with sophistical arguments has a bad reputation in philosophy and has been criticized by Plato (*Republic* 382d-e, *Gorgias* 500a–503c), Thomas Hobbes (*De Cive* 2.12.11–12), and Immanuel Kant (*Critique of Judgment*, sec. 53, 328), among many others. This tradition suggests that persuasion of this kind is at best an art and at worst it undermines autonomy and appeals to the most base aspects of human psychology to bamboozle unsuspecting interlocutors. If this is what the moral AI reasoner does as well, then the paternalism objection indeed hits its mark.

But there is also a tradition that springs from Aristotle's *Rhetoric*, which urges that persuasion with non-scientific arguments is not only compatible with personal autonomy, but that it plays an important role in moral and political discourse (Perelman and Olbrechts-Tyteca 1969, Dow 2015). Aristotle agrees with Plato that rhetoric can be used to take advantage of people – sophistry can be a type of intellectual coercion – but he also sees

rhetoric as the best way to appeal to people that are not persuaded by scientific argument (*Rhetoric* 1355a24–26, 1356a20–33).

In this latter tradition we can distinguish cases where autonomy is respected, in which case we have rational persuasion from cases where autonomy is not respected and we have rational manipulation. Rational manipulation occurs when:

> one might introduce reasons into a person's deliberations in order to play on her neuroses (or simple, individual preoccupations) so as to impede her deliberation and control the likelihood of a certain outcome to her deliberation. One might inundate someone with lots of relevant information, doing so precisely in order to overwhelm her and hinder her deliberation. One might present someone with evidence with the aim of emphasizing certain relevant information and facts, while intentionally neglecting to mention other information one acknowledges as relevant (Tsai 2014, 89–90).

In rational manipulation reasons are given, but the process of rational deliberation is short-circuited in the listener by some rhetorical device.

Rational persuasion can still infringe on autonomy in very special cases. First, it can reveal a negative attitude towards one's interlocutor. Second, it can preclude an opportunity to engage in rational deliberation (Tsai 2014, 91–92). Neither necessarily applies to the moral reasoner AI presented here.

The moral reasoner AI does not have a negative attitude towards its user, because it does not have any attitudes at all. It can be interpreted to have such attitudes, if its user takes the intentional stance towards it and simply attributes beliefs and desires in order to predict and understand its behavior (Dennett 1981). But that would be no different than taking an intentional stance towards a thermostat. If John interprets a thermostat as having a negative attitude towards him, that should not be a reason to think that John's autonomy is infringed upon. This is because such an attribution does not have the same moral significance as an attribution of a negative attitude to another person.

The other exception also does not apply. The moral AI does not preclude opportunities for rational deliberation. On the contrary, it invites its user to engage in rational deliberation that they may not have gone through otherwise. Racist John and people like him are just the kind of people that Aristotle's conception of persuasion by argument is intended for. This is precisely because rational persuasion can appeal to them in ways that other arguments may not. As it does its work, the moral AI reasoner respects John's autonomy, even if it attempts to persuade him to change his racist beliefs. The objection from paternalism does not apply.[2]

5. Conclusion

AI can provide some hope for dealing with the Moral Lag Problem, but only if it can play a normative role and change how people behave. To do that, the AI has to be able to provide moral reasons that a rational, reason-responsive person could be persuaded by. Machine ethics is the only extant research program that has as its aim modeling human-level moral reasoning in a computer. So an AI that can help with the Moral Lag Problem will involve machine ethics and a system that can provide reasons that a rational agent can be persuaded by.

This argument first challenges Savulescu and Maslen's (2015) proposal and then makes room for a new and potentially more effective use of AI for moral enhancement. The sketch of such a system that is provided here constitutes the positive contribution that the paper makes. This AI system will generate persuasive arguments using algorithms that formalize moral reasoning based on first-order normative theories, such as utilitarianism or Kantianism. Contrary to the critic, this system would not be paternalistic in a way that would undermine autonomy. It would provide a way to reach people that may otherwise be immune to what Aristotle calls scientific argument.

Acknowledgment
Work on this paper was financed by the Polish National Science Centre (NCN) SONATA 9 Grant, PSP: K/PBD/000139 under decision UMO-2015/17/D/HS1/01705.

NOTES

[1] It should be noted that the technology discussed here faces formidable technical and theoretical obstacles. Furthermore, even if these obstacles are overcome there remains the question whether its development and/or use is morally acceptable. Discussion of these issues is outside the scope of the present paper.

[2] Presumably, devices such as the moral reasoner AI imagined here would be used in very specific conditions. Before anyone uses them, they would first have to make a decision to use it, which implies at least tacit consent to being advised in this way. This is yet another reason to think that the moral reasoner AI is engaging in rational persuasion and not rational manipulation.

REFERENCES

Anderson, M. and S. L. Anderson (2011). *Machine ethics*, Cambridge University Press.

Baertschi, B. (2014). "Neuromodulation in the service of moral enhancement." *Brain topography* **27**(1): 63–71.

Borenstein, J. and R. Arkin (2016). "Robotic nudges: the ethics of engineering a more socially just human being." *Science and engineering ethics* **22**(1): 31–46.

Carlson, M. S., et al. (2014). "Identifying factors that influence trust in automated cars and medical diagnosis systems." in *AAAI Symposium on The Intersection of Robust Intelligence and Trust in Autonomous Systems*.

Crockett, M. J. (2014). "Moral bioenhancement: a neuroscientific perspective." *Journal of medical ethics* **40**(6): 370–371.

De Dreu, C. K. (2012). "Oxytocin modulates cooperation within and competition between groups: an integrative review and research agenda." *Hormones and behavior* **61**(3): 419–428.

de Sio, F. S., et al. (2014). "How cognitive enhancement can change our duties." Frontiers in systems neuroscience 8: 131.

de Vries, P. W. (2004). *Trust in systems: effects of direct and indirect information*, Technische Universiteit Eindhoven.

DeGrazia, D. (2013). "Moral enhancement, freedom, and what we (should) value in moral behaviour." *Journal of medical ethics*: medethics-2012–101157.

Dennett, D. C. (1981). "True Believers: The Intentional Stance and Why It Works," in A.F. Heath, ed., *Scientific Explanation: Papers Based on Herbert Spencer Lectures Given in the University of Oxford*. Oxford: Clarendon Press: 53–75.

Dominelli, L. (1998). "Multiculturalism, anti-racism and social work in Europe," in eds. C. Williams, H. Soydan and M. R. D. Johnson, *Social Work and Minorities*. London: Routledge: 36–57.

Dow, J. (2015). *Passions and Persuasion in Aristotle's Rhetoric*, Oxford University Press, USA.

Dworkin, G. (1972). "Paternalism." *The Monist*: 64–84.

Dworkin, G. (2016). "Paternalism." *Stanford Encyclopedia of Philosophy*. from http://plato.stanford.edu/entries/paternalism.

Emerson, R. M. (1976). "Social exchange theory." *Annual review of sociology*: 335–362.

Fedo, M. (2016). *The lynchings in Duluth*, Minnesota Historical Society Press.

Glenn, A. L. and A. Raine (2014). "Neurocriminology: implications for the punishment, prediction and prevention of criminal behaviour." *Nature Reviews Neuroscience* **15**(1): 54–63.

Hamari, J., et al. (2014). *Do persuasive technologies persuade? – a review of empirical studies*. International Conference on Persuasive Technology, Springer.

Harris, J. (2010). *Enhancing evolution: The ethical case for making better people*, Princeton University Press.

Harris, J. (2011). "Moral enhancement and freedom." *Bioethics* **25**(2): 102–111.

Harris, J. (2013). "'Ethics is for bad guys!' Putting the 'moral' into moral enhancement." *Bioethics* **27**(3): 169–173.

Hobbes, T. (2004). *De cive*, Kessinger Publishing.

Kant, I. (1987). *Critique of judgment*, Hackett Publishing.

Lee, J. D. and K. A. See (2004). "Trust in automation: Designing for appropriate reliance." *Human Factors: The Journal of the Human Factors and Ergonomics Society* **46**(1): 50–80.

MacIntyre, A. (2006). *Ethics and Politics: Volume 2: Selected Essays*, Cambridge University Press.

Meyer, M. L., et al. (2012). "Empathy for the social suffering of friends and strangers recruits distinct patterns of brain activation." *Social cognitive and affective neuroscience*: nss019.

Muir, B. M. (1987). "Trust between humans and machines, and the design of decision aids." *International Journal of Man-Machine Studies* **27**(5–6): 527–539.

Nickel, P. J. (2013). Trust in technological systems. *Norms in technology*, Springer: 223–237.

Parasuraman, R., et al. (1993). "Performance consequences of automation-induced 'complacency'." *The International Journal of Aviation Psychology* **3**(1): 1–23.

Perelman, C. and Olbrechts-Tyteca, L. (1969). *The New Rhetoric: A Treatise on Argumentation*, University of Notre Dame Press, Notre Dame.

Persson, I. and Savulescu, J. (2011). "Unfit for the future? Human nature, scientific progress, and the need for moral enhancement." In *Enhancing human capabilities*, ed. J. Savulescu, R. ter Meulen, and G. Kahane. Oxford: Wiley-Blackwell: 486–500.

Picard, R. W. (2000). *Affective Computing*, MIT Press.

Plato (1997). *Plato: complete works*. Indianapolis, Hackett.

Rowe, C. J., & Broadie, S. (2002). *Nicomachean ethics*. Oxford University Press, USA.

Sauer, J., et al. (2015). "Experience of automation failures in training: effects on trust, automation bias, complacency and performance." *Ergonomics*: 1–14.

Savulescu, J. and H. Maslen (2015). Moral Enhancement and Artificial Intelligence: Moral AI? *Beyond Artificial Intelligence*, Springer: 79–95.

Shiffrin, S. V. (2000). "Paternalism, unconscionability doctrine, and accommodation." *Philosophy & Public Affairs* **29**(3): 205–250.

Slovic, P. (2010). If I look at the mass I will never act: Psychic numbing and genocide. *Emotions and risky technologies*, Springer: 37–59.

Tsai, G. (2014). "Rational persuasion as paternalism." *Philosophy & Public Affairs* **42**(1): 78–112.

Van den Hoven, J., et al. (2012). "Engineering and the problem of moral overload." *Science and engineering ethics* **18**(1): 143–155.