




Artificial Intelligence as a Socratic Assistant for Moral Enhancement

Francisco Lara · Jan Deckers 

Received: 3 October 2018 / Accepted: 19 February 2019 / Published online: 26 February 2019
© The Author(s) 2019

Abstract The moral enhancement of human beings is a constant theme in the history of humanity. Today, faced with the threats of a new, globalised world, concern over this matter is more pressing. For this reason, the use of biotechnology to make human beings more moral has been considered. However, this approach is dangerous and very controversial. The purpose of this article is to argue that the use of another new technology, AI, would be preferable to achieve this goal. Whilst several proposals have been made on how to use AI for moral enhancement, we present an alternative that we argue to be superior to other proposals that have been developed.

Keywords Moral enhancement · Artificial intelligence · Technology ethics · Socratic ethics

Introduction

Human societies have always tried to make their citizens more moral. Many have also sought to do so on their own initiative. Methods that have traditionally been used to improve appropriate behaviour include religion,

propaganda, literature, ethical reflection and education. However, these traditional methods of moral enhancement are often slow and ineffective. Questionable behaviours include lying, corruption, racism, murder, and paedophilia. These behaviours continue to exist despite efforts at all levels to prevent them. The ineffectiveness of such efforts derives, to some extent, from the problems associated with influencing behavioural patterns that are influenced greatly by our biology. Many of us are biologically predisposed to have limited cognition and to have a limited level of altruism. The influence of biological conditioning has been corroborated with experiments that show that the degree of altruistic predisposition is very similar in identical twins and not in fraternal twins who do not share the same genes [1, 2].

Thanks to the advance of the neurological sciences and biotechnology, we are now able to influence the moral deliberation and behaviour of individuals by either influencing or intervening directly in their biology. These interventions include the use of various substances, such as oxytocin and serotonin, as well as of various techniques, including transcranial magnetic stimulation and the provision of neurofeedback.¹ The aim of these interventions would be to promote trust in others and to foster the desire to collaborate. Some authors have argued that voluntary implementation of these interventions would be morally permissible or morally desirable [4, 5]. Others have argued that such interventions are urgent and that they should be

F. Lara
Department of Philosophy I, Facultad de Filosofía y Letras,
Edificio de Psicología, Campus de Cartuja, University of Granada,
18071 Granada, Spain
e-mail: flara@ugr.es

J. Deckers (✉)
School of Medical Education, Faculty of Medical Sciences,
Newcastle University, Newcastle-upon-Tyne NE2 4HH, UK
e-mail: jan.deckers@ncl.ac.uk

¹ A brief survey of these and some other interventions is provided in Savulescu & Maslen [3].

compulsory if we are to survive as a species. For Persson & Savulescu this is so because our natural morality is unable to face the inescapable and serious challenges of a society that is radically different from the society that evolutionarily shaped our morality [6].²

However, these biological interventions have been criticised for their possible adverse effects on health, for their threats to autonomy and personal identity, for being morally counterproductive by favouring, at least if they are optional, the immoral behaviour of the free-rider, and for their potential to foster exclusive empathy [8]. In addition, these proposals for moral bio-enhancement are very simplistic. Morality is not only about motivation. It does not consist exclusively in trusting others and being willing to collaborate. As trusting those who are untrustworthy and cooperating with those who wish to take advantage may not be good, for example, morality also requires interpreting situations, deliberating properly and, above all, having reasons to act.

Therefore, it is worth considering other forms of moral enhancement offered by this new technologically advanced world. In this article, we will consider one that has not yet been studied much, given the scarce bibliography about it, one that we will call ‘moral artificial intelligence enhancement (henceforth: AIenhancement)’. Our goal is to present a new proposal on how artificial intelligence could make individuals more moral, more rapidly and successfully than traditional methods, and with fewer risks and controversies than bio-enhancement. Risks and controversies would be decreased as AIenhancement does not aim to alter the biological causes of our motivations, but merely to improve our moral capacities without the use of biotechnological interventions.

In order to do so, we will: firstly, characterise our new proposal by differentiating it from other proposals of moral enhancement through AI; secondly, show its feasibility by specifying some guidelines for future software that could be derived from it; and finally, try to answer the objection that this enhancement will be unsuccessful because it does not affect our motivations.

² In a similar vein, Dietrich maintains that, due to evolutionary conditions of adaptation, we are “genetically hardwired” not to be very moral. We can progress morally, he says, but there are immoral behaviours (conceived as behaviours that are harmful to others) from which, to a greater or lesser degree, we will not be able to free ourselves, given their usefulness in our evolutionary past. Such behaviours include “lying, cheating, stealing, raping, murdering, assaulting, mugging, abusing children, as well as such things as ruining the careers of competitors, negatively discriminating on the basis of sex, race, religion, sexual preference, and national origin and so forth” (p. 534) [7].

Exhaustive Enhancement

In this context, ‘artificial intelligence’ means ‘ubiquitous computing’ or ‘ambient intelligence’; that is, a system that collects information from multiple sensors and databases to process it according to its functional relevance for the system user. Artificial intelligence has already improved the quality of life of many human beings, even if it may not have improved their moral behaviour. However, it could also serve to make our lives more moral.

A first way to use AI to enhance morality derives from what we now know as ‘machine ethics’. Machine ethics is not intrinsically about human moral enhancement. It aims to turn machines into moral decision-makers.³ Dietrich is so pessimistic about the moral nature of humans and so optimistic about the possibilities of AI that he adopts an obligation to “usher in our extinction” to create a better world in which only ethical robots inhabit the Earth (p. 531) [7]. These robots would have achieved that “Copernican turn”, inaccessible to most humans by their biological conditioning, which allows them to act from the belief that each of them is not the centre of the universe (pp. 532–3) [7]. Really, he concludes, these machines would be nothing more than improved versions of *Homo sapiens*; he calls them “*Homo sapiens 2.0*” because they would be “the better angels of our nature”, without the defects of those who are less angelic (p. 536) [7].⁴

Whilst Dietrich’s proposal has not been greeted with great enthusiasm, some authors suggest that moral machines might be created to improve the morality of human decisions, rather than to make human beings redundant. Human decisions could be overridden or corrected by machines that would be better at moral

³ This aim is philosophically controversial. For example, Searle believes that making moral decisions is something that can be done only by human beings [9]. Whitby criticises it [10].

⁴ If someone holds that we have an obligation to extinguish ourselves to give way for machines that are better than us morally, it makes sense to ask oneself what the point of such an alternative world would be. Dietrich believes that this world would be given meaning by the machines themselves because they would be conscious and empathetic. He admits that we still have a lot to know about what consciousness is, but he believes that building machines with human-like intelligence will imply that these machines are conscious, as we ourselves are conscious. If we are able to build such a machine, he thinks that it could be programmed to make what he calls the “Copernican turn”: the realisation that oneself matters and that, since others are like oneself, others matter as well. Dietrich claims that the machine’s capacity to make this turn would depend partly on its capacity to have sympathy for others (p. 537) [7].

reasoning as their decisions would be characterised by a constant impartiality, consistency, and absence of emotional influences [11]. They would be free from limitations, including the irrational egocentric or group favouritism that very often leads to immoral behaviour. In addition, the machine would not get tired of the constant determination of what is right.

This way of enhancing humans, which leaves their moral decision-making entirely to machines, will be called ‘exhaustive enhancement’. In this approach, to do the right thing, we just have to obey the machine. This proposal tries to enhance moral behaviour by creating autonomous artificial agents that would have moral capabilities superior to those of humans. It would be based on a conception of morality that the system designer considers to be valid and with which they configure the system to direct human beliefs, motives, and actions.⁵ The essential aspect of this proposal is that all human participants, including the designer, would be expected to take a passive role after the original programming had been completed. We would not need to expend any psychological and/or behavioural efforts to achieve the desired enhancement apart from deciding to allow the system to make decisions for us (pp. 140–1) [13]. Human beings could then execute these decisions either by the machine controlling our behaviour directly (the direct model), for example through a brain implant that directs our actions, or indirectly through a machine or a political system sanctioning any behaviour that deviates from that which is demanded by the system (the indirect model). Whereas deviation would still be possible in the indirect model, it would be strongly discouraged. We would simply need to do as we are told.

Reasons against the Exhaustive Project

We distinguish five problems with the exhaustive project. The first is that it may be hard to set this up as the existence of pluralism may thwart the idea of finding a consensus on which theory should inform the system. The second is that, even if such a consensus could be achieved, we may still fail to set up a good system due to human or nonhuman limitations. The third is that we

doubt whether the system can be an autonomous moral agent, which is why it would not be able to make moral decisions. The fourth is that the possibility of making moral progress is excluded. The final problem is that it would mean the death of morality. We will discuss each of these problems in turn.

The first drawback of the exhaustive project is that we do not have a consensus on which ethical scheme should inform the design of the system. Should we use deontology, utilitarianism, contractualism, virtue ethics, or some other ethical theory?⁶ Regardless of which theory is chosen, it is probable that there would always be people who disagree with the system, where the choice of one theory would lead the system to direct different actions than those that would be promoted on the basis of another theory. One might argue that this need not be a problem as multiple system designers could develop multiple systems, for example deontological and utilitarian systems. Users of such systems could then simply decide which system to use. This, however, does not resolve the problem of value pluralism. It makes it worse. By surrendering our moral decision-making to machines, we would forgo the possibility of ever reaching any agreement where different systems provide different solutions to a moral issue.

A potential way in which one could try to address this problem is to ensure that the system is not programmed to make decisions on the basis of some abstract moral theory, but that it is programmed to decide on the basis of actual decisions that people have made. This is what a group of researchers of a ‘moral machine’ project at MIT intend: that the system, in this case that of a driverless car, would make moral decisions by virtue of generalizations from what people actually judge on what must be done in dilemmatic moral situations [16]. To find out what people think in moral situations that might be relevant for the development of autonomous vehicles, the research team questioned over two million people from more than two hundred countries. It found that moral judgments vary significantly between people from different cultural traditions. Whilst the authors claim that their findings imply that the “journey to consensual machine ethics is not doomed from the start”, we agree with them that they present a significant challenge towards “developing global, socially

⁵ This is what Savulescu & Maslen call “strong moral artificial intelligence” (p. 84) [6], and what Schaefer calls “direct moral enhancement” (p. 262) [12].

⁶ Although contractualism is usually known as a political theory, there are authors such as Gauthier [14] and Scanlon [15], who have developed ethical theories from the idea of the hypothetical contract.

acceptable principles for machine ethics” (pp. 63, 59) [16].⁷ They present a particularly difficult challenge for the exhaustive project as it is unclear why people should simply follow the moral decisions that are made by a system that may be perceived to make questionable decisions. Indeed, there is a significant concern that those who would design and implement an exhaustive system might foster their own ends, which may be biased by their own interests.

The second problem is human or nonhuman fallibility. System designers might make programming errors that result in them, as well as others, being directed to commit particular actions that do not flow from their values. Supporters of the exhaustive project might counter that this may be extremely unlikely. Even if this is granted, some errors might still occur due to problems that are caused by environmental factors that are outside human control, for example “a short circuit modifying some component of the system” (p. 392) [17]. Supporters of the exhaustive project might retort that this need not be a problem because the system would be an autonomous moral agent with superior moral capacities. It would therefore be able to correct any errors that were made either by their designers or that would result from any physical limitations.

This takes us to the third problem. We do not think that there is any system that can be regarded as an autonomous moral agent. Therefore, simply doing as we are told would not be advisable. Defenders of the exhaustive proposal might retort that the system could become autonomous, even if it might not be autonomous at the time that it is being designed. Like human children, who are not autonomous moral agents from the moment that they are born, they might argue that the system could also become an autonomous moral agent at some stage in its development. What this analogy ignores, however, is that children become able to reject particular assumptions as they grow older, whilst computers are never capable of doing so. Children owe this capacity to the fact that they are never machines in the first place. We do not mean to say that computers cannot be designed in such a way that they start running on different algorithms from those that they used at the outset, but that they can only do so if they are designed in that way. Many children, by contrast, will start to develop their own ethical views without ever having

been programmed to do so, and frequently in spite of their parents’ moral views. The sheer fact that some computers may be able to run on different algorithms from those that they used at the outset, however, does not imply that they are or could be moral agents.

To illustrate why we think that a computer is unable to make moral decisions, let us consider these two imaginary cases. The first case is the case of Geoff, who was brought up in a family where every Friday night they ate mussels. Geoff learned to like eating mussels and he developed an ethical theory that it was fine to eat mussels. He now finds himself on an island without a chip shop where they sell cooked mussels. However, he finds mussels on the beach and rejoices in cooking them so that he can eat them. The second case is identical to the first, with the exception that Geoff finds animals on the beach who look like mussels, but are not in fact mussels. In fact, the species has never been seen by any human being before. Geoff decides, however, to treat them in a similar manner.

A computer might simulate human reasoning as follows in the first case: Geoff approves of other people killing mussels in order to eat them. This is deduced from the principle that people are allowed to kill mussels in order to eat them. Since Geoff is also a human being, the principle of consistency demands that, if Geoff approves of other people killing mussels in order to eat them, he must also approve of doing so himself. We think that the computer will only be able to come to this decision if it has been programmed to base its moral advice on the principle of consistency. It seems implausible to us to think that the computer would be able to make the autonomous decision to value the principle of consistency. For this to be possible, the computer would need to grasp the moral importance of consistency.

In order to be able to do so, the system would need to possess not only the capacity of imitating analytic (logical) judgments, but also the capacity of making synthetic judgments.⁸ A synthetic judgment that underlies the first case is the judgment that it is good to be consistent. The reason this is a synthetic judgment is that someone who rejects that it is good to be consistent does not make a logical error. The meaning of the word good

⁷ We would like to thank an anonymous reviewer for bringing this study and this point to our attention.

⁸ The distinction between analytic and synthetic judgments is commonly made in philosophy, for example by Kant, who wrote that analytic judgments are true by definition as the predicate is contained within the subject (e.g.: “All bodies are extended.”), whereas this is not the case in synthetic judgments (e.g.: “All bodies are heavy.”) (p. A6–7) [18].

does not imply ‘to be consistent’, or vice versa. Whilst supporters of the exhaustive project might defend their project in light of this challenge by pointing out that logical conclusions that follow from well-chosen values should provide good moral advice, the second case highlights the problems associated with the exhaustive project. It presents a genuinely new situation. Whereas Geoff may rely on certain parameters to determine whether or not he would be justified in killing these animals, for example their degree of similarity to mussels or some behavioural traits, it would be wrong for Geoff to fix these parameters in advance. Doing so would imply that Geoff blinds himself to new facts that may arise in his encounters which would alter the parameters that he had hitherto considered to be significant, for example the realisation that these animals communicated with him in a way that he had never experienced with any other animals. It is unclear to us how computers, by contrast, might become sensitive to any morally relevant facts. They can only mimic synthetic judgments by relying on parameters that are “specified ahead of time” (p. 19) [19].

This point must not be misunderstood. We do not claim that a computer would not be able to identify some facts as morally salient even if they were not identified as such at the time of its design. To use an example provided by Penrose, some computers may, for example, be capable of identifying a face as a human face, even if they had never observed that particular face before, and bestow the same moral significance on that face as on other human faces (p. 18) [19]. We do claim, however, that computers lack the emotional sensitivity that is required to make moral judgments. As computers lack the emotional sensitivity to understand that animals who look like mussels may, in fact, be profoundly different from mussels, they would be expected to advise us to treat these animals as if they were mussels. One might argue that this is not a problem, as this is precisely what Geoff decides to do. This, however, misses the point. We have argued already that Geoff should not blind himself to any morally salient facts that may present themselves to him in his encounters. Imagine that the animals in question told Geoff ‘cook me’ upon the point of being put into boiling water. Geoff might be so touched by this that it might lead him to alter his intention. This kind of sensitivity is required for Geoff to operate as a moral being. When we compare this with a computer’s ‘sensitivity’, we are unaware of the existence of any computer that shows any greater

sensitivity towards the moral interests of human beings than, say, a deckchair, and it is unclear to us how an aggregate of molecules that has been assembled by human beings could ever possess such sensitivity.

To sum up this third concern, we grant that some computers may be able to simulate or mimic analytic judgments reliably and synthetic judgments less reliably (as the latter defeat logic), but we defend the view that morality depends on the ability to make synthetic judgments, for example the judgment that the logical conclusions of an axiological egotist should (not) be followed. We are not aware that an assemblage of molecules that has been put together by a human being has ever been able to make such a judgment, and we doubt whether it is even possible, as we consider emotional sensitivity to be a necessary condition for making such judgments.⁹

The fourth problem with the exhaustive project is that it stifles the possibility of moral progress. Moral judgments that are made today may no longer be acceptable at some other time. If we allow the system to decide, we would be left with a static account of morality as it would only be able to simulate, rather than to make judgments. In the nineteenth century, certain values were widespread that are no longer considered to be appropriate today, for example particular views about slavery or about the role of women. Similarly, it is likely that some views which are held dearly today will, one day, be looked upon as morally problematic. If we allow human decisions to be determined by the system, however, we would essentially forgo the possibility of moral change. To allow the possibility of moral progress, there must be moral pluralism and, consequently, dissent. Just as Mill argued about conventional morality, the morality established by the machine can never be challenged if there are no other proposals [22]. For this to occur it is necessary to question the values that determine the operation of the machine.

The fifth problem with the exhaustive project is that it would result in a loss of the sense of morality. Epistemologically, there are two possibilities: either human beings lack the capacity to know what is right and wrong, or at least some human beings do have the capacity to know what is right and wrong. If we accept

⁹ Whilst our view that no computerised system can be a moral agent is contested (see e.g. [20]), it is not without support in the literature. See for example Miller, who writes: “While robots are sensitive to physical properties, e.g. heat and light, they are not sensitive to moral properties. Accordingly, robots are not moral agents” (p. 162) [21].

the former view and accept, against our earlier contention, that machines are capable of knowing which behaviours are ethical, human beings would simply have to defer to machines. This deference would not be based on sound judgment, but on blind faith, questioning why we should adopt such faith. One might argue that blindly accepting the advice of a system undermines the very nature of morality, which depends on individuals developing their moral positions on the basis of moral reasoning, rather than on the basis of blind faith.

If we accept the second possibility, however, the possibility arises that we might nevertheless undermine our ethical capacity by following the system. The reason for this is that we would be ‘outsourcing’ moral decision-making. The adoption of such a passive role might ultimately make us less confident in making ethical decisions, which Vallor argues to result in moral de-skilling [23]. It would imply that being good at moral reasoning is not valued much. This is rightly questioned by van Wynsberghe and Robbins, who argue that, by leaving this function to machines, we deprive ourselves of what Aristotle recognised as an essential element of the good life: leading a life dominated by moral understanding, developed through practice [24].¹⁰

This shows that, even if the position that machines can be better at moral reasoning is adopted, this should not necessarily be a reason to rely on them. The direct model would entail the death of morality itself, as moral decisions would no longer be made by human beings, but by the system. Because the system would govern the behaviour of human beings, human beings would make ‘their’ decisions without practically participating in them. People would surrender their moral autonomy, rather than enhance it. A similar problem besets the indirect model. Even if we would still be able to choose against the system, the fact that any deviant behaviour would be classed as immoral would provide a strong incentive to avoid the behaviour in question. The probability that deviant behaviour would be tolerated might, in fact, be slim. This would be a particular concern for those who disagree with our contention that machines cannot be autonomous as such machines might decide to rule the roost. The concern would not be absent for those who agree with us, however, as sophisticated machines

that are good at mimicking human decision-making might be perceived to be autonomous and to be capable of making superior moral decisions. In both cases, machines might dominate us to such an extent as to severely compromise human liberty and life.

We maintain that the point is not to enhance moral conduct exclusively, but rather to enhance moral conduct because it reflects a better agent. Therefore, the deciding agent should, ultimately, always be independent of the machine. No enhancement is possible if it prevents conscious reasoning and rational deliberation, which the exhaustive proposal would prevent. Therefore, the main objective of moral AI enhancement should not be to change our behaviour, but the ways in which we make moral decisions. Rather than design an artificial mind that would regulate our behaviour, the aim should be to use the different functions of ambient intelligence (collecting, computing, and updating data) to help us to reach better decisions ourselves and, consequently, to act better.

Auxiliary Enhancement

We will now set out a new proposal that, in contrast to the exhaustive project, requires the agent’s participation. We will call it ‘auxiliary enhancement’. The agent has a less passive role here because he gives personal criteria to the machine that enable it to process information. The agent then prompts the machine to provide advice, which the agent considers before deciding. The agent does not lose their autonomy at any stage. The values with which the machine works are those of the agent. The agent’s values are a fixed benchmark with which a system is built that helps the agent to decide from that framework of values.

This is what is intended by Savulescu & Maslen and by Giubilini & Savulescu [3, 25]. In the former article, the authors envisage that the agent chooses from a broad list of values that are provided by the system and ranks their relative importance, and that the system then recommends a moral decision that accords with this value hierarchy. In the latter article, the authors suggest that different versions of the system should be available, and that the agent should choose which version they wish to adopt, based on their value system. The design of this system is inspired by Firth’s idea of morality, where a moral system is conceived as a collection of statements to which an ideal observer would react in a particular

¹⁰ Schaefer makes a similar point, arguing that Mill would have opposed the exhaustive enhancement project on the basis of the view that it would deprive the individual of the values of autonomous reasoning and personal responsibility for the effects of one’s actions [12].

way. The observer would be ideal in that they would be omniscient about non-ethical facts, omnipercipient (i.e. capable of visualizing, imagining and using all the information simultaneously), disinterested, dispassionate, and consistent, but normal (or like an average person) in all other respects (p. 321) [26]. The system envisaged by Giubilini and Savulescu (p. 7) [25] would differ from Firth's ideal observer as its moral advice would not be based on the perspective of some "absolute observer" (in the sense that Firth (p. 319) [26] gave to this term, to refer to an observer who formulates ethical judgments without containing egocentric expressions like "I like" or "I believe"), but reflect the ethical values of the particular human agent using it, within certain constraints. Giubilini and Savulescu illustrate that these constraints could, for example, be set by Catholic moral experts. Those who want to be good Catholics, but do not know exactly what principles to apply in particular circumstances could, in this way, choose to use an artificial moral advisor designed by Catholic experts, rather than some other version of the system. In this way, the authors believe that moral agents could ensure that their views accord with the Catholic ethos, thus establishing narrow reflective equilibrium, as well as balance the Catholic perspective against others, thus establishing wide reflective equilibrium.¹¹

We think that these approaches have two problems. Firstly, the role of the agent continues to be too passive. Once the agent has chosen which values they wish to adopt, the agent's only remaining decision is to decide whether or not to accept the result of the machine's deliberation. As the agent does not need to understand the rational connections between their values and the decisions that are made by the system, their moral skills may not be enhanced a great deal. If the machine stopped working, the agent's decisions might not be much better than what they were in the beginning either, as the focus is on helping the agent to make the right decisions, rather than on helping the agent to become a better moral agent.

A second problem with both approaches is that, once the values had been fixed by the moral agent, either through choosing a value hierarchy (in the former approach) or through choosing a value framework (in the latter approach), the system would merely recommend decisions that accord with those values, rather than encourage the user to question those values. The agent

may, at any time, rank the values provided by the system differently or choose to use a different version of the system, but we are concerned that they might be unlikely to do so. This is recognised by Savulescu and Maslen, who write that using the machine might encourage deference, rather than "deeper reflection" (p. 92) [3]. If it is true that people are frequently reluctant to change their moral values, they may be expected to be even more so if they think that their decision was based on the advice from a reliable machine. Whilst these approaches allow for a plurality of different perspectives, it is hard to see how they might foster wide reflective equilibrium or agreement between those who choose different (versions of) systems.

Socratic Enhancement

The approach that we envisage, by contrast, involves a constant interaction between system and agent. There should be no previous lists or systems of values from which to improve the morality of the agent. Through the constant interaction between the agent and the system, the possibility that the agent's values would be changed through their dialogue with the machine is increased. The crucial point is that the system should work to help us to reach a better decision ourselves, without committing us to any pre-designed ethical perspectives. Our decision would be determined through interacting with a system that had not been configured with previous designer or agent values.¹²

This approach therefore resembles something like Socratic help. Obviously, it would not be Socratic in the sense that the dialectical method envisaged by Plato simply recollects certain ideas that everyone has hidden within their nature because of their previous reincarnations. Nor would it be in the sense of seeking a definition of a concept, usually of some virtue. Our proposal is inspired by the role of the deliberative exchange in Socratic philosophy as an aid to develop better moral judgments. Socrates always presents himself and acts as a mere assistant who aims to refute the definitions he

¹¹ For the notion of reflective equilibrium, see for example Rawls [27].

¹² Seville & Field outline what could be an Ethical Decision Assistant that could be made available to everyone on the web [28]. The aim of this assistant would not be to give agents answers about what they should do, but to allow them to reconsider their personal values in the light of consequentialist considerations and consistency. Although they also grant an auxiliary role to AI, there is no room for the kind of machine-agent interaction that we defend here.

receives from his interlocutors. He, as Plato once points out, is like a midwife who only helps the other to give birth to his own knowledge. In our case, this knowledge would not reveal any hidden or common-sense truth, but consist in a moral judgment that was formed by applying conditions of empirical, logical and ethical rigour to one's beliefs. The agent should always have a privileged place, should always provide the first solution in a significant conflict, which is then submitted to staged scrutiny so that the machine, like Socrates, may ask relevant questions and reveal potential failures in the argumentation.

It does not seem to be appropriate to object to this proposal, as with other proposals at biological or AI moral enhancement, that it would threaten the autonomy of the enhanced agent. Rather, the intervention would increase autonomy, avoiding those (especially cognitive) limitations that prevent the agent from doing the right thing. By gathering, computerizing and updating data, the AI would enable the agent to be better prepared to make moral decisions, allowing them to find the ethical response to every dilemma for themselves.

Our proposal, therefore, has two features that differentiate it significantly from other auxiliary approaches, such as those of Savulescu & Maslen and Giubilini & Savulescu [3, 25]. The first feature has to do with the degree of participation of the agent. In those approaches, the machine deliberates for the agent; in our proposal, the degree of participation is greater as the agent deliberates in dialogue with the machine. The second feature that qualitatively differentiates our proposal from those is that the emphasis is placed on the formative role of the machine for the agent, rather than on the result. The aim is to help the agent to learn to reason ethically, rather than to help the agent to learn which actions the system deems to be compatible with particular values.

To avoid the risk of this enhancement turning into an 'exhaustive' form, the system should be programmed to be no more than an auxiliary aid for each agent, who always has the first and the final word in the process. Algorithms would be needed to avoid the machine being biased towards particular values and ethical theories. This approach therefore contrasts with the current tendency to develop prototypical computational models based on particular value theories, such as Bentham's utilitarian theory in the JEREMY program [29]; a mixture of Ross's prima facie duty theory and Rawls's reflexive equilibrium in a prototype called W.D. [29], which was then applied to the field of medical ethics

under the name MeEthEx [30]; a casuistic scheme that compares concrete dilemmas, such as the Truth-Teller; or the so-called SIROCCO program that has been applied to the field of engineering [31].

Instead, we pick up on the intention of early programs to help users rather than to give them a solution. These early programs had, above all, a pedagogical aim and, in the beginning, their specific aim was to help students to reason ethically through the presentation of practical problems. They started with videos about these problems and tried to invite the students to ethical exploration.

For example, the Ethos program, designed by Donald Searing to accompany an engineering ethics manual, provided videos and interviews on certain dilemmas to provoke students to ask new questions and to rethink their previous positions. In order to do so, it encouraged the user to make step-by-step decisions that would be recorded and then examined according to a basic scheme proposed by the program to take apart moral decisions: firstly, to delimit the problem; secondly, to outline alternatives; and finally, to evaluate each one (p. 30) [31].

Another example is an 'ethical decision-making assistant' app, created by the Markkula Center for Applied Ethics at Santa Clara University, to guide ethical decision-making. The app can be accessed via the internet or downloaded to a mobile device or computer with iTunes [32]. The user of the app must determine the degree to which certain conduct would affect possible stakeholders by considering specific ethical criteria: utility, rights, justice, common good and virtue. At the same time, the user is allowed to inform himself, thanks to certain inputs about these ethical criteria that accompany each stage of decision-making. Later, the user has to hierarchize these criteria by virtue of the relevance that he gives to them for the case in question. Finally, the app, by virtue of a score obtained from the answers given by the user, advises the user whether or not to modify the option considered.

A third example is the Dax Cowart program, a multimedia, interactive program designed to explore the ethical dimensions of the real case of Dax Cowart, a victim of an accident that resulted in serious and very painful wounds and burns. Cowart insisted on his right to stop the treatment that had been imposed on him. The key question was whether his wish should be fulfilled. The program featured videos and interviews with Dax Cowart and his doctor, lawyer, mother and healthcare

staff, allowing the user to see the problem from different perspectives. The program periodically asked the user to decide whether Dax's request to let him die should be accepted. It then presented alternative information and perspectives that could prompt users to reconsider (p. 30) [31].

A fourth example is the work by Robbins, Wallace and Puka, who initially devised a system that used web pages containing links to relevant ethical principles and theories, as well as a simple 'coach' in ethics [33]. Later on, these authors designed a much more sophisticated computational model that combined collaborative problem solving (i.e., several human subjects discussing an ethical problem), the psychological theory of planned behaviour, and the agency model of belief-wish-intention into a decision aid. It aimed to simulate different roles, including that of counsellor, group facilitator, interaction coach, and forecaster of how human subjects discuss and attempt to solve ethical dilemmas [34].

Whilst in line with these early examples of AIenhancement that focus on a constant interaction between machine and human being, we will, in the following section, outline a more ambitious system to aid ethical decision-making that might be developed thanks to advances in AI.

Criteria and Implementation of Socratic Enhancement

Taking into account the above, we would now like to propose a series of functions that should be taken into account when computer programs are designed with the aim to help moral agents to make moral judgments and to facilitate their behaviour according to such judgments.

Providing Empirical Support Although moral judgments are not fully verifiable by facts, they can be refuted when they are based on empirically refuted premises. Thanks to its access to and rapid handling of big data, an AI system has a privileged position to suggest that, sometimes, the agent's judgments may have no empirical basis and to require the agent in that case to modify their judgments to make them more truthful.

Improving Conceptual Clarity Moral judgments frequently use concepts that are not clearly defined and that can condition the validity of the judgment or the

interpretation that can be made of it. The agent is not always aware of this plurality of meanings. The AI system can warn them of this and, therefore, to remind them of the requirement to be rigorous in determining the meaning of any concepts that are used. The system can acquire this knowledge both from the big-data (crossed information coming from many dictionaries, grammar textbooks, records of use of language) and from the conclusions of experts in ethics that are taken into account when configuring the software. The use of experts is especially important because many essential concepts in ethical debates (death, guilt, person, ...) have strong normative elements that have multiple interpretations and whose knowledge is essential in the precise defence of moral positions.

Understanding Argumentative Logic Many people may agree with the view that moral judgments should be based on arguments that have to follow certain logical guidelines. For each type of argument (by generalisation, deduction, analogy, ...) there are some logical rules that the agent may not have followed either due to ignorance or due to the excessive weight given, more or less consciously, to irrelevant factors such as preconceived and biased interests or beliefs. It would therefore be of great value for moral enhancement if the computer system were designed to make the agent see the logical deficiencies of his argument. This could be done by showing them, by the formalisation of their argument, the error committed, or by using the most common repertoire of fallacies in the hope that it may make them understand why their reasoning is invalid. In addition, it would be very useful for the system to warn the agent of the need for their particular moral judgments to be ethically coherent in the light of an ultimate value criterion that gives meaning to all the agent's judgments. Agents who agree that consistency is important will agree, for example, that it is not valid to justify an action by virtue of the belief that it is derived from the ultimate criterion that it serves the happiness of all, and to demand, at the same time, an action that is justified by the principle of respect to the other, regardless of its consequences for the happiness of all those affected. In such a case, the system should prompt the agent to provide a meta-criterion that gives coherence to his judgments, which does not leave the justification of her judgments to irreconcilable reasons.

Testing whether one's Judgment May Possess Ethical Plausibility The moral agent may not only be enhanced by knowing the logic of his arguments, but also by their knowledge of the debates in normative ethics. To do this, software should be enriched through the accumulation of information that is based on the main theoretical positions on any particular issue. But we should not forget that pure logic and theoretical debate in strictly rational terms can lead to foolish and very implausible conclusions. Therefore, although the system should not be biased towards the common sense or the legal view, it could be programmed to make the agent aware of this view so that they may be prompted to think more carefully where they deviate significantly from it.

Raising Awareness of Personal Limitations AI could also be of great help alerting the agent to certain biological and environmental factors that could affect decision-making. Through monitoring the physiology, the mental states, and the environment of the agent, the system could warn of certain risks of deliberating badly. Examples of such negative influences on moral deliberation are: shortness of sleep, the time elapsed since the last meal, exhaustion, inappropriate levels of particular hormones and neurotransmitters, the presence of particular foodstuffs or psychoactive drugs that are known to have a detrimental role, and environmental factors such as heat and noise (pp. 85–6) [3].

Advising on how to Execute one's Decision The system could be of great help in advising the agent how to put into practice the moral decisions that they have reached. Some very advanced technologies have already been applied to areas of cognitive decision-making. These are software programs that quickly access a large space of digitised information and that, once certain search criteria have been added, assist human beings in areas where complicated decisions must be made [35–37]. The areas where most progress has been made in this respect are business and medicine.¹³ In the latter area the use of AI has resulted in a technology called the clinical decision-support system, which is an information technology that has been developed to improve clinical decision-making. It is being used to assist doctors in formulating diagnoses and making medical decisions.

¹³ Examples of recent software and platforms developed to improve the analytical and decision-making processes of companies are HANA (by SAP), DOMO, Apptus and Avanade [38].

The software receives information about the patient's symptoms and condition, and crosses it with medical databases and the medical history of the patient and his/her family, to conclude a diagnosis and the corresponding treatment [39–41]. This software could be a reference to help the moral agent to decide on how to apply moral judgments that he has previously adopted as valid. The software would receive information about the moral decision that has been adopted and would use it to process relevant information about how it might affect others and the environment, advising us how our behaviour would best fit our moral position.

To sum up, the system would receive, through computers, virtual reality devices or brain interfaces, information from many databases on science, linguistics, logic, and on how people think and reason morally. Moreover, it would collect information from experts in argumentation theory and ethical theory. With the help of sensors, it would also monitor the actual biology and the environment of the agent. The system would then process all this information and, using the aforementioned criteria, engage in a conversation with the agent through a virtual voice assistant. In this conversation, the system would ask a number of questions. These may include the following: *Why? And why? What makes you think that? Is this your last reason? Why do you think this is the best reason? What about this other reason? What do you mean with this word? Do you know there are other meanings? Are you aware that your assertion has no scientific basis? Are you aware that both assertions are contradictory? Are you aware that this deduction/induction/analogy... is not valid? Do you know that this is not a common value? Are you aware that your current physical condition/environment is not the best one to make an important decision? Do you know that, in these circumstances, your decision could be best executed like this?*

What about Motivation?

We believe that, with the proposed Socratic enhancement that we have sketched, human beings could overcome some of their limitations to behave morally. However, we are also aware of the proposal's limitations. The AI system will not suddenly turn us into good moral agents. This is because it cannot correct the motivational factors that prevent us from acting morally. Due to the characteristics of this type of technology (as opposed to

biomedical technologies) and to the auxiliary character of our proposal, AI may, in the first instance, only remedy our cognitive limitations. Thanks to this technology, the agent could acquire more precise, rigorous and consistent judgments about what is correct. But what good is this if the agent, because of strong emotions or a weak will, does not feel motivated by those judgments?¹⁴

This is a fair question, but the possibility that the cognitive help of AI may *indirectly* modify the motivations and emotional dispositions of the moral agent cannot be ruled out. Thanks to their ability to make better judgments, people could be sensitised to rethink some emotive positions that they may have. We refer here to the persuasive power of good arguments, the persuasion of reason, which many philosophers, from ancient Greece to modern times, have accentuated. People can change their values after thinking about them. Artificial intelligence can be an excellent tool for this. The system can make us accept new values by showing us the rational force of the arguments that support them and that we would not have seen without the help of the system. By incorporating virtual reality, the system can also persuade us to take the consequences of our actions more seriously. It may allow us to ‘experience’ the realities of (particularly) distant others more vividly, and to imagine much better how our actions and omissions might affect them. Even though these features are cognitive, it is likely that they would reduce the problem caused by weakness of the will (pp. 502–4) [28].

We think that the AI system would fulfil this persuasive role to a greater extent than human dialogists. Why? There are studies that show that our trust in machines in general, and in computers and robotic systems in particular, depends on their effectiveness rate, that is, on the statistics of their past performance and their capacity to respond appropriately to new situations [44]. Effectiveness also matters when we decide whether to trust other human beings who help us in our decision-making.

¹⁴ We are aware that the issue of moral motivation is problematic and that, for some rationalist approaches, reasons are motivating in themselves. An example of such a position is that of Harris, for whom the only goal is to enhance human beings cognitively [42]. Recent discoveries in neuroscience, however, argue that non-rational, pre-conscious factors also affect moral decision-making (see e.g. [43]). If this is the case, Harris’s exclusive focus on “rationality and education” would be deficient. For this reason, many advocates of moral enhancement believe that, for such enhancements to be successful, they should address both our ability to reflect on what is right and our motivations. See, for example, Douglas [4] and DeGrazia [5].

However, as long as people remain clear of the fact that the AI system is imbued with values that cannot be better than those possessed by human beings, their faith in human assistants is likely to be lower as human psychology makes us wary of various logical and volitional limitations that are absent in the AI system.¹⁵

Moreover, trust in AI systems could be increased if they were regularly redesigned to improve their cognitive and affective appeal [47]. With regard to the former, human beings might increase their trust in them, for example if systems provide up-to-date statistics of their past successful performance, or if the algorithms become simpler and more comprehensible, or if they show the connections between their recommendations and the goals of their users in better ways. Regarding their affective appeal, just as in human social psychology where people with similar character traits are attracted to each other, it can be predicted that software will be chosen by virtue of its expression of certain ‘personality’ traits that are similar to those of the user. Thus, virtual assistants that respond to their users with a language more in tune with their way of being and their emotional states can be expected to stimulate more trust in them [48, 49].

We believe that, if all these aspects are taken into account, the reasoning that a human being could develop with the help of a computer could influence her judgment and behaviour more than any advice received from other people. It should also be noted that in the proposal that we have presented, judgments can be even more persuasive than in other forms of AIenhancement because the system is only an efficient assistant, a midwife who helps users to give birth to a decision that is completely theirs and from which they can feel more proud precisely because of the fact that it is theirs.

Conclusions

Given our incomplete current knowledge of the biological determinants of moral behaviour and of the use of biotechnology to safely influence such determinants, it is reckless to defend moral bioenhancement, even if it were voluntary. However, the age-old human desire to be morally better must be taken very seriously in a globalised world where local decisions can have far-reaching consequences and where moral corruption

¹⁵ Muir [45] and Klineciewicz (p. 181) [46] maintain something similar.

threatens the survival of us all. This situation forces us to seek the satisfaction of that desire by means of other technologies. AI could, in principle, be a good option. Since it does not intervene directly in our biology, it can, in principle, be less dangerous and controversial.

However, we argued that it also carries risks. For the exhaustive project, these include the capitulation of human decision-making to machines that we may not understand and the negation of what makes us ethical human beings. We argued also that even some auxiliary projects that do not promote the surrendering of human decision-making, for example systems that foster decision-making on the basis of moral agents' own values, may jeopardise the development of our moral capacities if they focus too much on outcomes, thus providing insufficient opportunities for individuals to be critical of their values and of the processes by which outcomes are produced, which are essential factors for personal moral progress and for rapprochement between different individuals' positions.

We proposed a new way in which AI could help us to enhance our personal and group morality. It highlights an interactive relationship between a human agent and a computerised assistant. This assistant, in a Socratic way, should ask questions and provide relevant information to help the human agent to reach better moral judgments and realisable behavioural options that cohere with those judgments. This could be used both to exercise the cognitive skills necessary for morality and to motivate agents to behave according to what they think is right. AI could thus represent an important advance in the omnipresent problems in the history of ethics of how to know what is right and how to motivate agents to act accordingly. If people develop their arguments in dialogue with a machine that is more reliable than human dialogists and more reliable than themselves, but that leaves decision-making to human beings alone, resisting the force of reason will be much more difficult.

Acknowledgements This article was written as a part of the research project *Artificial Intelligence and Biotechnology of Moral Enhancement. Ethical Aspects* (FFI2016-79000-P), funded by the Ministry of Economy, Industry and Competitiveness of the Spanish Government. It was commenced whilst Francisco Lara was a visiting researcher in the School of Medical Education, Newcastle University, collaborating with Jan Deckers.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

Human and Animal Studies This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Wallace, B., D. Cesarini, P. Lichtenstein, and M. Johannesson. 2007. Heritability of ultimatum game responder behaviour. *Proceedings of the National Academy of Sciences* 104 (40).
- Baron-Cohen, S. 2003. *The essential difference: Men, women and the extreme male brain*. London: Penguin/Basic Books.
- Savulescu, J., & Maslen, H. 2015. Moral enhancement and artificial intelligence: Moral AI?. In J. Romportl, E. Zackova, & J. Kelemen (Eds.), *Beyond artificial intelligence. The disappearing human-machine divide* (pp. 79–95). Springer.
- Douglas, T. 2008. Moral enhancement. *Journal of Applied Philosophy* 25 (3): 228–245.
- DeGrazia, D. 2014. Moral enhancement, freedom, and what we (should) value in moral behaviour. *Journal of Medical Ethics* 40 (6): 361–368.
- Persson, I., and J. Savulescu. 2012. *Unfit for the future*. Oxford: Oxford University Press.
- Dietrich, E. 2001. Homo sapiens 2.0: Why we should build the better robots of our nature. *Journal of Experimental & Theoretical Artificial Intelligence* 13 (4): 323–328.
- Lara, F. 2017. Oxytocin, empathy and human enhancement. *Theoria* 32 (3): 367–384.
- Searle, J.R. 1994. *The rediscovery of mind*. Cambridge: MIT Press.
- Whitby, B. 2011. On computable morality. An examination of machines as moral advisors. In *Machine ethics*, ed. M. Anderson and S.L. Anderson, 138–150. Cambridge: Cambridge University Press.
- Gips, J. 1995. Towards the ethical robot. In *Android Epistemology*, ed. K.M. Ford, C. Glymour, and P. Hayes, 243–252. Cambridge: MIT Press.
- Schaefer, G.O. 2015. Direct vs. indirect moral enhancement. *Kennedy Institute of Ethics Journal* 25 (3): 261–289.
- Focquaert, F., and M. Schermer. 2015. Moral enhancement: Do means matter morally? *Neuroethics* 8 (2): 139–151.

14. Gauthier, D. 1986. *Morals by agreement*. Oxford: Oxford University Press.
15. Scanlon, T.M. 1998. *What we owe to each other*. Cambridge: The Belknap Press of Harvard University Press.
16. Awad, E., S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.F. Bonnefon, and I. Rahwan. 2018. The moral machine experiment. *Nature* 563 (7729): 59–64.
17. Yampolskiy, R.V. 2013. Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In *Philosophy and theory of artificial intelligence, SAPERE 5*, ed. V.C. Müller, 389–396. Berlin, Heidelberg: Springer.
18. Kant, I. 1998. *The critique of pure reason*. Edited and translated by P. Guyer and A.W. Wood. Cambridge: Cambridge University Press.
19. Penrose, R. 1995. *Shadows of the mind. A search for the missing science of consciousness*. London: Vintage.
20. Allen, C., G. Varner, and J. Zinser. 2000. Prolegomena to Any Future Artificial Moral Agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12 (3): 251–261.
21. Miller, S. 2018. Autonomous weapons: terminator-esque software design. In H. Prunckun (Ed.), *Cyber Weaponry* (pp. 157–169). (Advanced Sciences and Technologies for Security Applications). Cham: Springer.
22. Mill, J.S. 1859. *On liberty*, Edited by Edward Alexander. Broadview Press, 1999.
23. Vallor, S. 2015. Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. *Philosophy and Technology* 28 (1): 107–124.
24. van Wynsberghe, A., and S. Robbins. 2018. Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics*. <https://doi.org/10.1007/s11948-018-0030-8>.
25. Giubilini, A., and J. Savulescu. 2017. The artificial moral advisor. The 'ideal observer' meets artificial intelligence. *Philosophy and Technology* 31: 1–20. <https://doi.org/10.1007/s13347-017-0285-z>.
26. Firth, R. 1952. Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research* 12 (3): 317–345.
27. Rawls, J. 2001. *Justice as fairness: A restatement*. Cambridge: Harvard University Press.
28. Seville, H., and D.G. Field. 2011. What can AI do for ethics? In *Machine ethics*, ed. M. Anderson and S.L. Anderson, 499–511. Cambridge: Cambridge University Press.
29. Anderson, M., Anderson, S.L. & Armen, C. 2005a. Towards machine ethics: Implementing two action-based ethical theories, proceedings of the AAAI fall symposium on machine ethics, technical report FS-05-06. AAAI Press: 1–7.
30. Anderson, M., Anderson, S.L. & Armen, C. 2005b. MedEthEx: Toward a medical ethics advisor, proceedings of the AAAI fall symposium on caring machines: AI in elder care, technical report FS-05-02. AAAI Press: 9–16.
31. McLaren, B.M. 2006. Computational models of ethical reasoning: Challenges, initial steps, and future directions. *IEEE Intelligent Systems*, July/August: 29–37.
32. Markkula Center for Applied Ethics. 2009. A Framework for ethical decision making. <https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/a-framework-for-ethical-decision-making/>. Accessed 26 January 2019.
33. Robbins, R.W., W.A. Wallace, and B. Puka. 2004. *Supporting ethical problems solving: An exploratory investigation (proceedings of the 2004 SIGMIS conference on computer personnel research: Careers, culture, and ethics in a networked environment, 134–143)*. New York: ACM Press.
34. Robbins, R.W., and W.A. Wallace. 2007. Decision support for ethical problem solving: A multi-agent approach. *Decision Support Systems* 43 (4): 1571–1587.
35. Phillips-Wren, G., and N. Ichalkaranje, eds. 2008. *Intelligent decision making: An AI-based approach*. Berlin, Heidelberg: Springer-Verlag.
36. Talbot, P.J., and D.R. Ellis. 2015. Applications of artificial intelligence for decision-making. Multi-strategy reasoning under uncertainty. In *CreateSpace independent publishing platform*.
37. Tweedale, J.W., R. Neves-Silva, L.C. Jain, G. Phillips-Wren, J. Watada, and R.J. Howlett, eds. 2016. *Intelligent decision technology support in practice*. Springer.
38. Ruth, J. 2019. 6 Examples of AI in business intelligence applications. <https://www.techemergence.com/ai-in-business-intelligence-applications/>. Accessed 26 January 2019.
39. Haynes, R.B., and N.L. Wilczynski. 2010. Effects of computerised clinical decision support systems on practitioner performance and patient outcomes : Methods of a decision-maker-researcher partnership systematic review. *Implementation Science* 5 (12): 1–8.
40. Eberhardt, J., A. Bilchik, and A. Stojadinovic. 2012. Clinical decision support systems: Potential with pitfalls. *Journal of Surgical Oncology* 105 (5): 502–510.
41. O'Sullivan, D., P. Fraccaro, E. Carson, and P. Weller. 2014. Decision time for clinical decision support systems. *Clinical Medicine, Journal of the Royal College of Physicians of London* 14 (4): 338–341.
42. Harris, J. 2011. Moral enhancement and freedom. *Bioethics* 25 (2): 102–111.
43. Cunningham, W.A., M.K. Johnson, C.L. Raye, J.C. Gatenby, J.C. Gore, and M.R. Banaji. 2004. Separable neural components in the processing of black and white faces. *Psychological Science* 15 (12): 806–813.
44. Carlson, M.S., Desai, M., Drury, J.L., Kwak, H., & Yanco, H.A. 2014. Identifying factors that influence trust in automated cars and medical diagnosis systems, AAAI symposium on the intersection of robust intelligence and trust in autonomous systems, technical report SS-14-04. AAAI Press: 20–27.
45. 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.
46. Klineciz, M. 2016. Artificial intelligence as a means to moral enhancement. *Studies in Logic, Grammar and Rhetoric* 48 (1): 171–187.
47. 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.
48. Nass, C., and K.N. Lee. 2001. Does computer-synthesised speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7 (3): 171–181.
49. Picard, R.W. 1997. *Affective computing*. Cambridge: MIT Press.