Mutual Enhancement between Science and Religion

with Fraser Watts, "Mutual Enhancement between Science and Religion: In the Footsteps of the Epiphany Philosophers"; William H. Beharrell, "Transformation and the Waking Body: A Return to Truth via Our Bodies"; Marius Dorobantu and Yorick Wilks, "Moral Orthoses: A New Approach to Human and Machine Ethics"; Galen Watts, "Religion, Science, and Disenchantment in Late Modernity"; and Rowan Williams, "Epiphany Philosophers: Afterword."

MORAL ORTHOSES: A NEW APPROACH TO HUMAN AND MACHINE ETHICS

by Marius Dorobantu and Yorick Wilks

Abstract. Machines are increasingly involved in decisions with ethical implications, which require ethical explanations. Current machine learning algorithms are ethically inscrutable, but not in a way very different from human behavior. This article looks at the role of rationality and reasoning in traditional ethical thought and in artificial intelligence, emphasizing the need for some explainability of actions. It then explores Neil Lawrence's embodiment factor as an insightful way of looking at the differences between human and machine intelligence, connecting it to the theological understanding of embodiment, relationality, and personhood. Finally, it proposes the notion of artificial moral orthoses, which could provide ethical explanations for both artificial and human agents, as a more promising unifying approach to human and machine ethics.

Keywords: artificial companions; artificial intelligence; embodiment; ethics; explainable AI; David Hume; Neil Lawrence; machine learning; relationality; theology

> There are no moral phenomena at all, only a moral interpretation of phenomena.
>
> Friedrich Nietzsche

Marius Dorobantu is a PhD candidate in Theology at the University of Strasbourg, Strasbourg, France; e-mail: marius.dorobantu@gmail.com. Yorick Wilks is emeritus Professor of Artificial Intelligence at the University of Sheffield, Oxford, UK; e-mail: ywilks@ihmc.us.

Ethical issues associated with the emerging field of artificial intelligence (AI) are one of the most pressing concerns society faces at the moment. Many questions require an answer, ranging from the purely philosophical to the very practical. Navigating through the interdisciplinary dialogue between computer science, philosophy, psychology, neuroscience, and theology, this article explores questions related to the topic of ethical explanation in human and machine decision making: what would it mean for a machine to be ethical? How different are we really from machines in terms of the current inscrutability of behavior? How rational is decision making in humans and machines? Can we locate the differences between human brains and the machine learning (ML) algorithms that learn to model the world? How do concepts like embodiment and relationality influence decision making, seen from various points of view as theology and information theory? Could an artificial moral orthosis help explain not only AI decisions, but also human ones?

An orthosis is a concept we owe to Kenneth Ford and his co-authors (2015), a notion we shall take to be a kind of artificial companion (Wilks 2010) to explain and help us understand the ethical behavior of humans and machines. We shall want to contrast this explanation function with a more conventional machine ethics concerned with the processes and programs that drive machine behavior and whose ethical properties are of interest to us. Medically, an orthosis is an externally applied device designed and fitted to the body to aid, say, rehabilitation, and contrasted with a prosthesis, which replaces a missing part, like a foot or leg. Here, it will mean an explanatory software agent associated with a human or machine agent.

When it comes to the difficulty inherent in any idea of AI ethics, the problem of "AI alignment" is illustrative, in spite of, or perhaps precisely because of, its extreme nature: how can an AI system be programmed from the beginning in such a fashion that, even if it becomes radically more intelligent than humans, its goals would still be aligned with ours? In other words, how can we make sure it would still care about us, or at least would not try to wipe us out? This problem is both difficult, and urgent, and no one has currently any idea how to solve it (Yudkowsky 2016). Whatever the solution, it will certainly be more complex than Isaac Asimov's three laws of robotics (Asimov 1950, 40). The sci-fi author brilliantly proved in several novels and short stories, known as his *Robot* series, how inefficient the three laws can be, in spite of their apparent logical suppleness.

But even if we accept the remote possibility that the alignment problem could be technically solved—that is, programmers could design AI that would forever remain faithful to a set of values, regardless of its level of (super)intelligence—there still remain the questions of which values and whose values? This points us to the fundamental problem of human ethics: we do not have a basic set of ethical principles that everyone agrees upon.

Aligning the ethics of an AI system with our own presumes that we are sure what our own ethical principles are. Yet there is no general agreement as to whether there are universal cross-cultural and cross-temporal principles or outcomes in concrete cases, in spite of ethical theories to the contrary, and monuments like the Universal Declaration of Human Rights. It was shown recently (Awad et al. 2018, 61) that Japanese and Western citizens disagree significantly as to whether an automated car should, in an emergency, seek to save an old pedestrian or a child.

Moreover, our understanding of what is good seems to be rapidly evolving over time, in the sense that practices that used to be reasonable two or three decades ago are no longer acceptable today, and vice versa. Should we miraculously agree now upon a common set of ethical principles, it is doubtful that we would still hold on to them in 2050. This problem is best illustrated by the thought experiment of imagining any of the previous human generations inventing AI and forever endowing it with its values, be it ancient Mediterranean honor and shame, or nineteenth-century racial slavery.

The point we wish to make is that human and machine ethics are equally problematic. In what follows, we will explore how the two are similar and ways in which they differ. We seek to contrast and compare the problems determining what ethical reasoning an AI system and humans are using in a given case, and to show that that determination may be more similar than appears at first glance. Finally, we will also propose a more promising unified approach toward human and machine ethics, which would take advantage of the complementarity of the two types of intelligence.

We would like to reconsider AI and ethics from a new starting point, or at least a new emphasis, given that much recent discussion has degenerated into little more than rehearsing codes of practice, of the kind that litters technical companies' publications. Elsewhere, Philippa Foot's "trolleyology" (2002, 85), the ethical discussion device that asks whether a vehicle should, for example, kill an old man or five children, and which originated as a teaching tool for ethics, has become dominant in discussions of the ethics of automated cars. But it has not led to any decisions about which to kill in any concrete case, even though it served to highlight the real problems an automated vehicle will face which, as we noted above, will include strong cultural differences across the world.

John Gray's *Straw Dogs* (2002) had an influence on our own thinking about these issues and we would like to draw out some consequences for how we see ethical machines and ourselves. We start with the old issue of the transparency of human and machine reasoning processes, and to ask what is our access to them. The point we want to reach in this article is to reintroduce the notion of orthosis into ethical explanation in humans and machines.

The Inscrutability of Human and Machine Action

Gray's starting point is that professional discussions of ethical decision making have little or nothing to do with how humans or animals actually seem to act. He believes they act simply like machines, and he means that in a positive sense, rather in the way Lao Tzu described the wise man not making choices but seeing a situation and acting rightly. In other words, humans and animals, for Gray, do not calculate ethical rules or consequences before acting, as the ethics textbooks tend to assume. He may be right about the conscious processes of humans in action, but his position is also circular: since it is clear that humans do not act randomly, then there must be some causal explanation of what they do. We do not know how we speak or see but the job of AI over fifty years has been to model these functions and to suggest possible mechanisms that would produce roughly the outputs we do. This is the gap that Jonathan Zittrain (2019) has recently called *intellectual debt*, and which we shall return to below: the gap between knowledge that things work and knowledge of *how* they do it, in the way we knew aspirin "worked" for a century before scientists told us how it dulled pain.

When one says humans do not act randomly, one remembers that there was a school of thought that celebrated irrationality as a positive virtue, of the sort expressed by the French literary notion of the *acte gratuit,* the act that was free simply because it had no rational basis at all. This was typified by the character in the novel by Gide (1914), who pushes a total stranger out of a train for no reason at all. In the particular case of ethical explanations of actions, the legal system exists in part at least to give exactly such explanations. It not only decides guilt and punishes, but explains (bad) actions, in terms of motives and desires: what Daniel Dennett (1971) has called "folk psychology," but one that seems to serve our civilization pretty well. We can barely imagine social life without this prop, even if it is all in some sense a fiction, as Gray claims to believe. Gray and Dennett have in common a downgrading of the role of consciousness as the theatre of our own, possibly self-serving, explanations of our own actions. For them the real action is all elsewhere, and inaccessible to us, a sentiment consistent with David Hume's famous declaration that reason "is the slave of the passions" ([1738] 2007, 266).

But we believe Gray is right to remind us that the true explanations of human action, whatever and wherever they may be, are as opaque as is much machine decision making, most obviously modern systems driven by ML. That this point is not yet generally appreciated can be seen from a recent influential book (Eubanks 2018, 168), where the author writes: "I find the philosophy that sees human beings as unknowable black boxes and machines as transparent, deeply troubling." And something of that same pre-ML assumption about humans and machines was present in

Donald Michie's observation in the 1960s that car drivers would prefer traffic lights to a policeman on point duty (what an ancient occupation that now seems!). Michie argued that the traffic light—called a "robot" in some English dialects to this day—could be trusted to be fair and essentially transparent though a policeman could not.

Within the current technical world, it is now a standard observation that humans may be unhappy with ML systems, regardless of the usefulness of their decisions in practice, if they cannot understand them. U.S. government agencies have recently funded just such explanatory methods (e.g., The DARPA XAI [Explainable AI] project; see Mueller et al. 2019). Similarly, the European Commission has legislated a demand (Order GDPR 2016/2679) that deployed ML systems must explain their decisions. It has done this even though no one at the moment knows how to perform this systematically, which reveals something about the technology–politics interface. One could say that much of the conscious explanations we give of our own actions are our own internal XAI, or rather XB, explainable brains. But it is important to remember that traditional ethical thought, like AI reasoning itself, assumed such reasoning to be both transparent and broadly correct.

Mechanized Reasoning as the Core of Traditional AI and Its Effect on Ethical Thinking in AI

The traditional discussion of ethics within AI (e.g., Akoudas et al. 2005) is often taken straight from the mainstream philosophy of ethics (which is to say the views of Immanuel Kant or John Stuart Mill depending on one's taste) and is one of seeing machine ethics as calculations from rules or consequence summation. These two traditional ethical approaches have now slipped somewhat into the intellectual background because, like Foot's trolley world, they have decided nothing in crucial cases.

Meanwhile, technical advances, such as automated cars or medical robotics or diagnostics, may well be based on ML and neural networks whose actions will need explaining, perhaps in courts, just like those of humans.

It is important to emphasize in all this that those two main ethical traditions both appeal to calculation, logical or arithmetical, as their basis, which is why they have appealed for so long to the computationally minded. But, and this is crucial, these are not real calculations that are ever carried out, and real values are never in fact assigned to possible outcomes in such discussions, even though, in the real world, automated machines do, of course, make practical decisions every day.

Much discussion of ethical issues in AI is inhibited, in our view, by the basic assumptions about the role of rationality and reasoning in humans and AI, the very views that Gray set out to demolish. These rationality assumptions, that rational thinking follows the rules of logic, and such rules are the basis of ethical decisions, reappear quite naturally in almost all AI discussions of ethical behavior in machines.

What is usually called GOFAI (good old fashioned AI), from its early origins in mechanical theorem proving, continued the great philosophical tradition of Gottfried Wilhelm Leibniz, who believed all matters—not just factual but ethical and religious—could be reduced to calculation: "justice follows certain rules of equality and of proportion [which are] no less founded in the immutable nature of things, and in the divine ideas, than are the principles of arithmetic and of geometry" (Leibniz 1988, 71).

For Leibniz, God was rational as was creation, and reason ruled supreme in this and all possible worlds, evil notwithstanding. His motivations were almost exactly that of the founding AI-ers, and they could imagine no other basis to AI. Yorick Wilks (1973) was among those who questioned this fixation, and argued that logic was an implausible foundation for the structure and understanding of human language. All this was long before the current rise of ML weakened the appeal of the old logic paradigm in AI (Dorobantu 2019, 6). In psychology, there have been many related findings (e.g., right back to Wason and Johnson-Laird 1972): namely, that it is almost certain that humans do not in general carry out logical operations when reasoning, but they do so from concrete exemplars that they are familiar with. This same skepticism can be found in Hume in the eighteenth century, as cited by Wilks above: "And if [ideas about facts] are apt, without extreme care, to fall into obscurity and confusion, the inferences are always much shorter in these disquisitions, and the intermediate steps much fewer than in the [deductive] sciences" (Hume [1751] 1907, 60–61).

Hume, like Leibniz, had in mind everyday reasoning and common sense though with an utterly opposed result, but his words apply equally to moral reasoning, where he also argued that that was grounded in sentiment or feeling and not in principles of reason, which was rather "the slave of the passions" as he put it, rather than its master.

Ethical Explanation in Humans and Machines

Drew McDermott (2008) makes the following important distinction: "The term 'machine ethics' actually has two rather different possible meanings. It could mean 'the attempt to duplicate or mimic what in people are classified as ethical decisions,' or 'the modeling of the reasoning processes people use (or idealized people might use) in reaching ethical conclusions.'" We shall call the former the ethical decision-making problem by an agent, which we take to be the core sense of "machine ethics," and the latter the ethical explanation problem, which is the focus of this article and the phenomenon that we are proposing an orthosis for, both for human and machine actions of ethical relevance. The original use of this term “machine ethics” is normally credited to M. Mitchell Waldrop (1987), to capture the ethical rules that might bind an AI computer’s actions, the original version of Asimov’s laws of robotics (1950), and the first sense of the term for McDermott.

The latter notion, of ethical *explanation*, is the basis of the suggestion of this article that we should consider the central ethical task of AI as the provision of explanatory orthoses for both humans and machines, since the underlying behavior of both is opaque in a way that mainstream discussion refuses to recognize.

Much of this claim is hardly novel with regard to opacity and its problems. Eugene Charniak, twenty years ago, at the start of the era of revived ML, wrote that he did not want to deal with ML systems if he could not understand how they achieved decisions, no matter how good their results (Charniak 1996). The opacity of human functioning can be both “upward” and “downward,” from microstructure to overall purpose and vice versa. Even if we were given “brain code,” it has been almost an axiom of much cognitive science that we could not determine what a person is actually doing or thinking, just as we cannot determine what a computer is actually doing from its machine code or its circuits. If we think of that as opacity from the bottom up, from knowledge of individual neurons or circuits to a machine’s real purpose, then, by contrast, Sigmund Freud and Dennett, in their very different ways, argued the opacity of human mental functioning from the top down, as it were: that conscious introspection was no guide to our real motives and processes. Similarly, with computers, their external behavior, including their language if any, can give no guide to what their actual underlying learned networks or machine code are doing.

More recently, Nick Bostrom and Eliezer Yudkowsky (2014) argued that, to be considered ethical, machines must be programmed with comprehensible rules if we are to tolerate them among us, so that we can understand them and why they do what they do. This is very much in the spirit of Charniak’s plea many years before, and refers not to the explanation of machine action but to the representation of the process that drives the action itself. Yet, if machines that take decisions are based on ML algorithms, as many now are, it is not clear that such transparency is or will be available. As we try to provide scientific explanations, in parallel with the machines as it were, those will not be the only possible explanations of the phenomena and behavior they model. There will always be alternative explanations of any behavior—courtroom drama rests largely on that fact. And that could be something quite new and orthosis-like added alongside whatever a machine is actually programmed with. It seems clear that, in the current generation at least, ML systems will not be programmed the way Yudkowsky and Bostrom (and Charniak) have demanded, and they might not be able to perform as successfully as they do if they were programmed in the transparent manner demanded.

An interesting footnote to "machine inscrutability" is that Michie also argued, forty years ago, that a major future function of AI would be to keep in operation large software programs, perhaps in critical social roles like air traffic control, which were so old that all documentation had been lost and they were effectively uneditable and inscrutable, though still apparently reliable. Yet they could not be trusted in the roles they had because they were not understood and might one day fail disastrously, and yet they were often too large and expensive to replace from scratch.

The existence of such large but inscrutable programs in the public domain gave rise to the drive for proofs of software correctness, and Michie suggested, not wholly seriously, that in the meantime a major role of AI might be to wrap around such programs and stop them doing anything disastrous, if their decisions seemed out of line and dangerous. Yet the wraparounds might still not actually be understanding the basic programs themselves, while they presumably would be wholly transparent in their own functioning.

Things have not gone that way, partly perhaps because of the inscrutability of the recent programs, though in a different way from the earlier ones, not from age and loss of documentation but from deliberate ML design. One can see in Michie's metaphor of wrapping code something like a rational cortex wrapped round, and attempting to control, the function of our deep inaccessible, instinctual, and inarticulate "crocodile" brain in our brain stem. We noted earlier Zittrain's argument (2019) that there has always been "intellectual debt" in science from things that work, though we know not how, like aspirin, which science pays off when explanations are subsequently given. He might have added that we ourselves are perhaps the supreme example of that, in our millennia of ignorance as to our own functioning, physiologically and psychologically. Yet, even as those new explanations of behavior have been developed, we have retained and refined the "folk psychology" of motive, desire, and responsibility in both our courts and our everyday life, because we seem as a species totally unwilling to give up such notions and resort to alternatives in terms of chemistry, upbringing, and brain structure.

Human versus Machine Intelligence: The Advantage in Our "Handicap"

Judea Pearl (2018) has recently entered this debate and argued that what ML systems based on big data lack is a clear concept of causation, as opposed to an association between data sets. Ethical argument, he suggests, requires a notion of causation which current ML systems cannot provide, which weakens them scientifically, he argues, and makes them ineligible as ethical decision makers. This brings the traditional discussion of the Humean notion of cause and its relation to *mere* association right back into central focus.

This last point brings us to one of the crucial questions of the current debate on machine ethics, and indeed of the more general field of intelligence, artificial and human: what is the fundamental difference between humans and machines in terms of cognitive abilities and decision making? The question can be reformulated in the following way: since ML is based on neural networks that claim to mimic, even though in a simplified fashion, the biochemical processes of the human brain, how is it that an AI algorithm needs millions of examples to learn to recognize a pattern, although the human brain can learn it from a handful of exemplars? We know intuitively that there must be a fundamental difference between the two systems, otherwise the order of magnitude of the difference in efficiency could not be properly accounted for.

A possible beginning of an answer to this question has been put by computer scientist Neil Lawrence (2017), who speaks of the *embodiment factor*, as the ratio between how much information an entity can communicate as opposed to compute. Although humans and computers may be approximately similar in their computing capability, computers can communicate information at a much higher speed. Lawrence's speculations lead, as we shall show, to an explanation of the origin of consciousness and the seat of rationality, one quite different from Julian Jaynes's (1976) classic account linking consciousness to a human talking-to-itself in an identifiable historical epoch.

The comparison in computing power is between an estimation of how much would be needed to fully simulate a human brain, approximately 1 exaflop (Ananthanarayanan et al. 2009), and the current level of the most powerful supercomputers. As of today, the supercomputer Summit stands at 0.2 exaflops, but Aurora and Frontier, set to be launched in 2021, will reach 1 and 1.5 exaflops, respectively (Vincent 2019). A desktop computer usually sits at a much lower figure, of around 10 gigaflops.

The communication comparison is done in terms of an entity's ability to share information. For computers, it refers to the amount of binary information they are able to transmit, with or without a wire, in 1 second. For humans, it is the sum of all our information output, both verbal and nonverbal. However, the concept can be broadened to any entity that communicates information. If a simple intelligence does not communicate explicitly, it still shares information through its observable behavior (Lawrence 2017, 3). Although computers can communicate information at around 1 gigabit/s, humans are only able to reach a maximum of 60 bit/s, which is about 100 million times slower. (Lawrence 2017, 4). This, combined with the humans' superior computing power, leads to astronomical differences in the embodiment factor, with humans being somewhere between 10^{10} and 10^{15} times more limited than computers in their communication, leading Lawrence to call humans "locked-in intelligences."

There seems to be a striking correlation between this huge difference in embodiment, on the one hand, and the equally large difference in learning efficiency, on the other. Lawrence makes a convincing case that the two may be causally related. In his view, it is precisely because we are and have always been so limited in our ability to communicate with each other that we were evolutionarily pressured to develop our particular type of intelligence, which relies heavily on understanding and modeling the world, the other agents (alias the theory of mind), and ourselves (self-perception, self-consciousness). Computers, on the other hand, experience this limitation at a much lesser degree, hence their poorer modeling of the environment.

The human brain has to be extremely parsimonious in its output, given that it can communicate only a tiny amount of information. Nonetheless, it has huge computational resources available, which enable it to run countless simulations of how different actions and communication strategies could play out, allowing it to maximize the efficiency of its output. Having accurate and reliable models is thus critical for optimal communication. One has to understand one's environment, which must include a good model of how others might receive the message. For this, not only is it necessary to be able to simulate the minds of others, but the simulation must go as far as picturing what the other agents might think of oneself. In other words, the brain has to simulate others' simulation of itself. Lawrence compares this hugely complex and continuous process of modeling the world with an internalized film production crew, which ceaselessly plays out various plots and scenarios. In his words, "each of us is a director" (2017, 5). An example of a formal model of such a modeling of other minds has already been set out (Wilks and Ballim, 1990), which has been implemented from time to time as the VIEWGEN system of agents' beliefs.

Computers can deploy their power more efficiently than humans, since they can communicate almost as fast as they compute, by transmitting full data, as it were. This is exactly what enables current ML algorithms to perform at superhuman levels at a variety of tasks. But it is this same advantage, combined with the high availability of training data, that makes ML "lazy" at modeling the world and other agents. Why would an ML algorithm devote time and computing power building models, creatively playing with them and constantly updating them, if the job required from it—usually pattern recognition—can be achieved through simpler repetitive and exhaustive algorithms? Why would these metaphorical AI go-karts venture off-road if they can "monotonously complete laps of their information processing circuits extremely efficiently"? (Lawrence 2017, 7).

The answer may be that it is precisely in the effort-wasting "mental space off-road" that intelligences evolve the crucial capacities of contextualization and explanatory understanding of the world. Human intellectual processes may be messy and clumsy by comparison but they produce the kind of behaviors that can be considered ethical in McDermott's second sense, even though currently inscrutable. This is the reason why the kind of explainable AI demanded by the U.S. government, the European Commission, Bostrom, Yudkowsky, or Charniak is not likely to be developed within the current ML paradigms.

Embodiment, Ethics, Relationality, and Personhood

Lawrence takes the argument further, suggesting that the emergence of consciousness could be caused by our embodiment, namely by our brain's inability to fully communicate its mental state (2017, 7). In the same way that we construct tiny internal simulated versions of every agent we interact with, we also seem to do the same thing for ourselves. We create a model of who we think we are, based mainly on how others relate to us, and continue to test, update, and refine that model for the whole of our lives. Lawrence is thus among those (Wilks 1984; Parisi 2007) who have identified the emergence of consciousness and its explicit inferences with the ability to talk to oneself, first in humans and now to be sought in machines.

This idealized and simplified constructed version of who we think we are can easily be identified in several dual-cognitive models. It has many similarities with Daniel Kahneman's (2011) system 2, or Jonathan Haidt's (2006) elephant rider. From the ethics point of view, this "fictitious me" constructed by my brain may well be the source of both self-consciousness and the generations of explanations of actions, among which will be ethical explanations. This is exactly the kind of task that we would like AI to become better at in order to fulfill its role of moral orthosis.

At a more fundamental level, what Lawrence's embodiment argument suggests is a confirmation of something that has been developing in the philosophy of personhood and in theological anthropology, namely, the so-called relational turn. Although substance used to be traditionally seen as taking precedence over relation, modern philosophy, at least starting from Kant and Hegel, has reversed the order of the two (Shults 2003, 11–32).

In the philosophy of mind, the constructionist approach of the early 1990s challenged the idea that the fundamental units of society are the individuals, and that the relationships between them are mere by-products (Gergen 1991, 156). On the contrary, it argued, minds are not fixed essences, but they are "being built from the symbolic resources of cultures by means of participation in human relationships" (Clocksin 2002, 10). A more developed description of Lawrence's embodiment factor could work in complementarity with constructionism, by providing a plausible support for this view in neuroscience and information theory.

Although the AI field may be largely technical and may try to keep as much as possible out of philosophical debates, in practice that might prove more difficult than it sounds. William Clocksin (2002, 11) points out, for example, that the preconstructionist view of the mind has fostered in AI a subfield called intelligent agent or autonomous agent research, which departed precisely from the premise of the individual as the center of knowledge and possessor of rationality. Furthermore, Noreen Herzfeld (2005) shows an analogy between trends in the history of AI and the history of interpretation of the theological concept of being in the image of God (*imago Dei*). She draws a parallel between how theologians have shifted their interpretation of *imago Dei* from a substantive, to a functional, and finally to a relational one, and how the field of AI has evolved from a substantive approach (symbolic AI) to a functional definition in the late 1980s. More interestingly, she correctly predicts a second turn, to relational AI, toward machines that could learn from interaction with humans and that could perhaps finally pass the Turing Test.

The embodiment factor argued by Neil Lawrence shares an even deeper intuition with the relational turn in theological anthropology. In his seminal 1990 book *The Call to Personhood*, Alistair McFadyen argues for an understanding of the notions of self and personhood as largely emergent from one's relationships with others: "The understanding of oneself as a continuous point of identity ('me') in an extensive range of relations, evidenced in self-referential and self-indexical use of 'I,' is not the result of some private, inward experience of one's self. It is, rather, the result of others indexing and referring to 'you' in this way. This is a communication of their experience and expectation of 'you' as a unified and continuous subject of communication, a 'self'" (1990, 95).

Not only does McFadyen propose that the self is a construction, and that it is directly caused by relationships, but he also makes the theological argument for embodiment. Although for Lawrence embodiment is defined as the ratio between computation and communication, in theology the body is the medium through which communication and relationship with others is possible, and at times it is the communication itself (1990, 89). William Beharrell has taken this point even further, in another article from this set, showing a shared recognition between various religious and medical traditions, such as Christianity, Buddhism, or Tibetan, Chinese, and Indian medicine, that "attention to embodied experience is significant for self-representation."

The two keywords that stand out from this brief interdisciplinary dialogue between computer science, information theory, neuroscience, philosophy, ethics and theology are *embodiment* and *relationality*. If we were to realize intelligent programs that could help us understand them and ourselves, as moral orthoses, a preliminary conclusion is that both embodiment and relationality should play an important part in their development. Embodiment can mean intentional limitations on their bandwidth communication and/or their access to large collections of data. But it can also mean actually having a body, be it one made of silicon, through which they could relate to us and with other artificial agents in ways more complex and more productive than through a chat bot.

Artificial Moral Orthoses

The orthosis suggestion above, which might bring all parties together, is that of an external explanatory system, using an ontology of rules, causes, and outcomes. It might come to function in parallel with inscrutable brains and ML systems and provide possible explanations of why they act as they do, rather in the way the DARPA XAI project wishes to create. Looking at the discussion on embodiment above, it might sound more promising to develop the orthosis-type of explanatory AI with a different methodology than what is currently used in ML.

The problem for any explanatory orthosis, as for scientific reasoning in general, is to find the best explanation. One could say the court system, at the heart of our civilization, is exactly that social orthosis for deviant behaviors: it finds the best explanation for such human behavior, and perhaps in the future for machine behavior. It may all, as Gray sometimes suggests, be a gigantic fiction but we can hardly imagine society without it. Elsewhere (Wilks 2010), the notion of a Companion has been developed and implemented: an agent permanently attached to a human and which gains the maximum possible knowledge about its human "owner" via dialogue over an extended period of years.

This notion amplifies that of the orthosis in a natural way, in that the Companion, so envisaged, would in principle be exactly the agent holding all the relevant information about the habits, preferences, tastes, choices and history of a human whose acts were under scrutiny, and which would supply the data needed to make inferences about his or her basis of action. It might plausibly contain self-revelations (or confessions) by an "owner" that could be crucial to ethical explanations of that person's actions. Indeed, one can imagine a person consulting their own ethical orthosis/Companion, as a form of therapy, in an effort to understand why they had acted as they did.

Machine Ethics: AI Machines as Ethical Actors?

We assume here that machine ethics—a machine acting so that ethical principles can be involved in its actions, in McDermott's first sense of the term—is in principle possible, in addition to the explanatory orthosis.

Saying that involves not accepting James Moor's (2006) view that only humans are *in principle* ethical agents, even though that is true at the time of writing. He writes: "Some might say that only humans should make such decisions, but if (and of course this is a big assumption) computer decision-making could routinely save more lives in such situations than human decision-making, we might have a good ethical basis for letting computers make the decisions" (Moor 2006, 18).

This is surely right; continuing to seek an ethical machine, even for pragmatic reasons, relies in part on a machine not-having self-interest, as in Michie's defense of traffic lights (over policemen) that assumed their lack of partiality to particular drivers. This lack, in a machine, is the opposite of McDermott's view that a machine, precisely because of its lack of self-interest, cannot make ethical decisions. Michael and Susan Leigh Anderson (2010) claimed that McDermott's view was an odd account of ethical dilemmas about a best outcome between alternatives, rather than having no self-interest. McDermott is almost certainly wrong about this, and his view cannot be squared with any classical ethical theory, such as Kant's, which would rule out self-interest by definition.

A more promising idea in McDermott's article is that "the machine must be tempted to do the wrong thing, and some machines must succumb to temptation, for the machine to know that it is making an ethical decision at all." To count as ethical, this implies, a decision must be between alternative courses of action. In a similar vein, we argued (Wilks and Ballim 1990) that necessary condition for a machine having a belief—as opposed to simply processing data—was that it should be able to compare two world states and decide which to believe, in the sense that an ATM never does when handing out cash and so cannot be said to be either having a belief or making an ethical decision. Our case below that a machine could in principle have beliefs (and indeed make ethical decisions about others) as well as beliefs about the beliefs of others, rests on a model like that of the point-of-view VIEWGEN system (Wilks and Ballim 1990). On this view, intelligent behavior is closely connected to the consideration of alternatives, not only in belief and action but also as to meanings, when we interpret, for example, metaphors.

It has become increasingly clear in recent times that emotional, affective behavior—the understanding and display of emotion in language and behavior—is a crucial part of intelligent behavior and cognition, even though it was an ignored, even forbidden, sideline in AI until the 1990s. It was given support by the pioneering work in psychology by Stacy Marsella et al. (2010), who showed how ubiquitous emotion is even in human relationships with laptops, a theme extended more widely to human affective relations with machines by David Levy (2007) and others. In what follows, we shall assume that a Companion-like orthosis will need to understand and display a range of human emotional behaviors.

Theological Considerations

Last, but not least, there is one more question that needs to be asked: would the use of such moral orthoses be in any way theologically problematic? In other words, should we even attempt to understand more about ourselves through the deployment of artificial explanatory companions? The answer to such a question cannot be simple. We will, however, try to approach it from two different angles: the use of technology, in general, as complementary to our own experience in understanding ourselves, and the universal character and theological limitations of human search for meaning.

First, if one were to understand the artificial orthosis as a sort of alien intelligence or oracle that would magically provide us with mystical information about ourselves and the world, then the theological concern would be legitimate. However, the moral orthosis and AI in general would be anything but that, and its ontology as an artifact of human creativity, that is, technology, cannot be stressed enough. And as technology, it would not be the first one that complements and augments our power to explain our actions to ourselves and to others. Writing, for one, enables just that. Writing a journal is nowadays one of the most frequent recommendations in the personal development literature, precisely because the mere exercise of writing one's thoughts on paper or on a keyboard can lead to unexpected revelations about the deeper motivations of one's behavior. Photography and cinematography are yet another example. Seeing pictures and recordings of ourselves can significantly contribute to re-shaping the internal mental model each of us has of himself or herself, which in turn plays a crucial role in how we explain our actions. The artificial Companions would just continue in this technological tradition, providing more valuable information that could help us overcome our cognitive biases.

Second, the limitations of any artificial moral orthosis have to be acknowledged. The human thirst for meaning and explanation is universal, but it is doubtful that AI will ever completely satisfy that thirst. Theologically, we must acknowledge that in the fallen current state of humanity, any such attempt would be futile. The only orthosis that can provide such ultimate explanations in the pre-eschatological stage of history is the Holy Spirit, as it is clear from Jesus' promise to the apostles (John 14:26), and from the fulfillment of that promise at the Pentecost (Acts 2). Artificial moral orthoses will thus be circumscribed by the same limitations imposed by the fallen human condition. As long as our perception of them will not conflate into something more than that, we can assume that we are theologically safe from any form of idolatry.

Finally, to assume that there will always be something about humans that will escape explanation even from superintelligent artificial observers is not "lazy apophaticism," as Fraser Watts puts it in a different article from this set. It is rather, we might say, very much in the spirit of the Epiphany

Philosophers to accept and celebrate the human ultimate inscrutability, as suggested by Rowan Williams in this same set of articles. Theologically, this inscrutability could stem from humans' fundamental status as beings created in the image of God, and who are destined to discover their purpose and fulfill their destiny only in relationship with God.

Conclusion

A main argument of the article was that an ethical machine is a real and serious possibility, in that machines undoubtedly take decisions already with ethical implications, and that these require ethical explanation in just the way humans' actions do. But such machine decision making may well not be based on the traditional core-AI perspective, in which rationality is central, but may be based on quasi-inscrutable ML processes and models of sentiment and emotion that may be quantitative in form. We argued that both human and machine actions, inscrutable to their own agents or not, will still require explanation, and that an ethical orthosis might provide such explanations in both cases. Developing such orthoses might require a different approach than current ML, one that should take more seriously the concepts of embodiment and relationality. The orthoses would function as artificial Companion agents to be associated with human and machine actors, with their embodiment of emotion simulations, and performing computations over the beliefs, goals, and points of view of other agents. These explanations might well embody not only reasoning but also be closer to ethical accounts based in moral sentiment or emotion (MacIntyre 1985) in the Humean tradition of the primacy of sentiment over reason in this area.

Acknowledgments

We are indebted to comments and criticisms from Selmer Bringsjord, Clark Glymour, Noel Sharkey, Fraser Watts, John Tait, and Eugene Charniak. The errors are, as always, all our own.

A version of this article was presented at the annual meeting of the Epiphany Philosophers held at Magdalene College, Cambridge, UK, on January 9, 2019.

References

Akoudas, Konstantine, Selmer Bringsjord, and Paul Bello. 2005. "Towards Ethical Robots via Mechanized Deontic Logic." In *AAAI Fall Symposium on Machine Ethics*, edited by Geert-Jan M Kruijff and Fiora Pirri, 17–23. Menlo Park, CA: The AAAI Press.

Ananthanarayanan, Rajagopal, Steven K. Esser, Horst D. Simon, and Dharmendra S. Modha. 2009. "The Cat Is Out of the Bag: Cortical Simulations with 10^9 Neurons, 10^{13} Synapses." *Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*—-sc '09, Article no. 63. https://doi.org/10.1145/1654059.1654124

Anderson, Michael, and Susan Leigh Anderson. 2010. "Robot Be Good." *Scientific American* 303:72–77.

Asimov, Isaac. 1950. "Runaround." In *I, Robot*, edited by Isaac Asimov, 25–45. New York, NY: Doubleday.

Awad, Edmond, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefo, and Iyad Rahwan. 2018. "The Moral Machine Experiment." *Nature* 563:59–64.

Bostrom, Nick, and Eliezer Yudkowsky. 2014. "The Ethics of Artificial Intelligence." In *The Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William M. Ramsey, 316–34. Cambridge, UK: Cambridge University Press.

Charniak, Eugene. 1996. *Statistical Language Learning*. Cambridge, MA: Bradford Books.

Clocksin, William F. 2002. "Artificial Intelligence and Theological Anthropology." Report on Theological Anthropology, Faith and Order Commission, World Council of Churches FO 33. Grand-Saconnex, Switzerland: World Council of Churches.

Dennett, Daniel. 1971. "Intentional Systems." *The Journal of Philosophy* 68:87–106.

Dorobantu, Marius. 2019. "Recent Advances in Artificial Intelligence (AI) and Some of the Issues in the Theology and AI Dialogue." *ESSSAT News and Reviews* 29:4–17.

Eubanks, Virginia. 2018. *Automating Inequality*. New York, NY: Macmillan.

Foot, Philippa. 2002. *Moral Dilemmas*. Oxford, UK: Clarendon Press.

Ford, Kenneth M., Patrick J. Hayes, Clark Glymour, and James Allen. 2015. "Cognitive Orthoses: Towards Human-centered AI." *AI Magazine* 36:5–8.

Gergen, Kenneth J. 1991. *The Saturated Self: Dilemmas of Identity in Contemporary Life*. New York, NY: Basic Books.

Gide, André. 1914. *Les caves du Vatican*. Paris, France: Editions de la nouvelle revue.

Gray, John. 2002. *Straw Dogs*. London, UK: Granta Books.

Haidt, Jonathan. 2006. *The Happiness Hypothesis: Finding Modern Truth in Ancient Wisdom*. New York, NY: Basic Books.

Herzfeld, Noreen. 2005. "Co-creator or co-Creator? The Problem with Artificial Intelligence." In *Creative Creatures: Values and Ethical Issues in Theology, Science and Technology*, edited by Ulf Görman, Willem Drees, and Hubert Meisinger, 45–52. London, UK: T&T Clark.

Hume, David. (1751) 1907. *An Enquiry Concerning the Principles of Morals*. London, UK: Longman, Green, and Co.

———. (1738) 2007. *A Treatise of Human Nature*. Oxford, UK: Clarendon Press.

Jaynes, Julian. 1976. *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Boston, MA: Houghton Mifflin.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Lawrence, Neil. 2017. "Living Together: Mind and Machine Intelligence." Available at https://arxiv.org/abs/1705.07996.

Leibniz, Gottfried Wilhelm. 1988. "Opinion on the Principles of Pufendorf." In *Leibniz: Political Writings*, edited by Patrick Riley, 64–76. Cambridge, UK: Cambridge University Press.

Levy, David. 2007. *Love and Sex with Robots*. New York, NY: Harper Collins.

MacIntyre, Alasdair. 1985. *After Virtue* (2nd ed.). London, UK: Duckworth.

Marsella, Stacy, Jonathan Gratch, and Paulo Petta. 2010. "Computational Models of Emotion." In *A Blueprint for Affective Computing: A Sourcebook*, edited by Klaus R. Scherer, Tanja Bänziger, and Etienne B. Roesch, 21–46. Oxford, UK: Oxford University Press.

McDermott, Drew. 2008. "Why Ethics Is a High Hurdle for AI." *Proceedings of the North American Conference on Computers and Philosophy (NACAP)*, Bloomington, IN.

McFadyen, Alistair. 1990. *The Call to Personhood: A Christian Theory of the Individual in Social Relationships*. Cambridge, UK: Cambridge University Press.

Moor, James H. 2006. "The Nature, Importance, and Difficulty of Machine Ethics." *IEEE Intelligent Systems*, 21:18–21.

Mueller, Shane T., Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. "Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI." DARPA XAI Program. Available at https://arxiv.org/pdf/1902.01876.pdf.

Parisi, Domenico. 2007. "Mental Robotics." In *Artificial Consciousness*, edited by Antonio Chella and Riccardo Manzotti, 191–211. Exeter, UK: Imprint Academic.

Pearl, Judea. 2018. *The Book of Why: The New Science of Cause and Effect*. New York, NY: Basic Books.

Shults, F. Leron. 2003. *Reforming Theological Anthropology: After the Philosophical Turn to Relationality*. Grand Rapids, MI: William B. Eerdmans.

Vincent, James. 2019. "World's Fastest Supercomputer Will Be Built by AMD and Cray for US Government." *The Verge*. Available at https://www.theverge.com/2019/5/7/18535078/worlds-fastest-exascale-supercomputer-frontier-amd-cray-doe-oak-ridge-national-laboratory.

Waldrop, M. Mitchell. 1987. "A Question of Responsibility." *AI Magazine* 8:29–39.

Wason, Peter Cathcart, and Philip Johnson-Laird. 1972. *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.

Wilks, Yorick. 1973. "Understanding without Proofs." In *Proceedings of the 3rd International Joint Conference on Artificial Intelligence*, 270–77. San Francisco, CA: Morgan Kaufmann Publishers.

———. 1984. "Machines and Consciousness." In *Minds, Machines and Evolution*, edited by Christopher Hookway, 105–29. Cambridge, UK: Cambridge University Press.

———, ed. 2010. *Artificial Companions*. Amsterdam, The Netherlands: John Benjamins.

Wilks, Yorick, and Afzal Ballim. 1990. "Liability and Consent." In *Law, Computers and Artificial Intelligence*, edited by Ajit Narayanan and Mervyn Bennun. Norwood, NJ: Ablex.

Yudkowsky, Eliezer. 2016. *The AI Alignment Problem: Why It's Hard and Where to Start*. Machine Intelligence Research Institute (MIRI) website. Available at https://intelligence.org/files/AlignmentHardStart.pdf.

Zittrain, Jonathan. 2019. "The Hidden Costs of Automated Thinking." *The New Yorker*, July 23.