Journal of Future Robot Life 1 (2020) 9–16
DOI 10.3233/FRL-190107
IOS Press



# A moral companion: A new approach to AI and ethics

Yorick Wilks *
*Florida Institute of Human and Machine, USA*

*"There are no moral phenomena at all, only a moral interpretation of phenomena"*

*Nietzsche*

## 1. INTRODUCTION

I would like to reconsider AI and ethics from a new starting point, or at least a new emphasis, since much recent discussion has degenerated into little more than rehearsing codes of practice, of the kind that litters technical companies' publications. Elsewhere, Phillipa Foot's "trolleyology" (1978), the ethical discussion device that asks whether a vehicle should, for example, kill a fat 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.

John Gray's *Straw Dogs* (2002) had an influence on my own thinking about those issues and I would like to draw out some consequences for how we see ethical machines and ourselves. I start with the old issue of the transparency of human and machine reasoning processes, as to ask what is our access to them. The point I want to reach in this brief paper is to reintroduce the notion of orthosis into ethical explanation in humans and machines. An orthosis is a concept I owe to Ken Ford and his co-authors (2015), a notion I shall take to be a kind of artificial Companion (Wilks, 2010) to explain and help us understand the ethical behavior of humans and machines. I shall want to contrast this explanation function with a more conventional machine ethics concerned with the processes and programs that drive machine behavior 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.

## 2. 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

*Florida Institute of Human and Machine, Cognition 15, SE Osceola, Ocala FL 34471, USA. E-mail: ywilks@ihmc.us.

(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 text books tend to assume, so neither should machines, he might have added, but didn't. He may be right about the conscious processes of humans in action, but his position is also circular: humans do not act randomly so 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 suggest possible mechanisms that would produce roughly the outputs we do.

When one says humans do not act randomly, there was of course 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, and 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 such explanations. It not only decides guilt and punishes but explains (bad) actions, in terms of motives and desires: what 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.

I believe Gray is right to remind us that human action is as opaque as is much machine decision making, most obviously modern systems driven by machine learning (ML). That thi point is not yet generally appreciated can be seen from a recent influential book (Eubanks, 2018, p. 168) where the author writes: "I find the philosophy that sees human beings as unknowable black boxes and machines as transparent, deeply troubling." Yet, something of exactly 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. US government agencies have recently funded just such explanatory methods (e.g. The DARPA XAI project, see Hoffman et al., 2018). Similarly, the European Commission has legislated a demand (Order GDPR 2016/2679) that deployed machine learning 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. But it is important to remember that traditional ethical thought, like AI reasoning itself, assumed such reasoning to be transparent.

## 3. TRADITIONAL ETHICAL THOUGHT AND ITS ROLE IN AI DISCUSSION

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 Kant or Mill depending on your 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 decided nothing in crucial cases.

Meanwhile, technical advance has shown that technological developments, 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 emphasise 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, automata do, of course, make real decisions every day.

## 4. MECHANISED REASONING AS THE CORE OF TRADITIONAL AI AND ITS EFFECT ON ETHICAL THINKING IN AI

Much discussion of ethical issues in AI is inhibited, in my 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 – transfer naturally to speculations on what a machine competent to take ethical decisions or to reason ethically would be like.

What is often called "core AI" sprang from mechanical theorem proving: the automation of deduction, a dream going back to Leibniz. For him, deduction was of divine inspiration and all matters, ethical, mathematical and practical could be settled by the appropriate calculations. As he put it: ". . . , 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, p. 71).

Reason ruled supreme for him, not only in mathematics but in ethics, politics and metaphysics, and since this world was demonstrably the best of all possible worlds, so the very basis of creation was both ethical and rational. Leaving aside this extra metaphysical and theological bonus, his program is not too far from that of core AI-ers, for whom the principles of logic play an essential role in our description of the world, not only in science but in everyday life.

I raised doubts years ago about this focus in AI, finding it inappropriate for the description of how our language and reasoning in everyday life actually function (Wilks, 1973), and long before the current rise of ML weakened the appeal of the old logic paradigm in AI. 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 perform very few processes by anything like deduction, as opposed to various heuristics and reasonings from individual cases. In that early critique I cited the words of Hume: "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, pp. 60–61).

When citing those words I intend their relevance to be to the modeling of common sense beliefs and knowledge and how we should model reasoning about everyday life in AI. But their relevance is equally to moral reasoning which Hume also did not believe to be deductively founded, not only because of the well-known non-inferability of "ought" statements from "is" statements, but more because of his belief that ethics was founded in sentiment and that reason was rather "the slave of the passions" as he put it, rather than its master.

## 5. ORTHOSES AND ETHICAL EXPLANATION IN HUMANS AND MACHINES

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.' "I'll call the former the ethical-decision making problem by an agent, which I take to be the core sense of "machine ethics" and the latter the ethical explanation problem, which is the focus of this paper and the phenomenon that I am proposing the orthosis for, both for human and machine actions of ethical relevance. The original use of this term "Machine Ethics" is normally credited to (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.

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

Much of this claim is hardly novel as regards opacity and its problems: Charniak twenty years ago, at the start of the revived ML era, 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 function can be both "upward" and "downward", from microstructure to overall purpose. Even if we were given "brain code", it has been almost an axiom of much Cognitive Science that we cannot determine what a machine is actually doing from its machine code or its firing 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, 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.

More recently, Bostrom and 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 point many years ago, and refers not to the explanation of machine action but to 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 will be available, as we noted, unless something quite new and orthosis-like is added alongside whatever it is they are 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 way 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 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, but 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.

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, 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.

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, 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.

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: that 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), I have developed, and implemented, the notion of a Companion: 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, as a form of therapy consulting their own ethical orthosis/Companion in an effort to understand why they had acted as they did.

## 6. 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. To do this requires not accepting Moor's (2006) claim that only humans are full ethical agents, even though this undoubtedly reflects present reality. The only issue then is whether that must always be so, which is to say: is the machine ethics project in principle possible? 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. (p. 18)"

I am sure this is correct: the considerations that lead one to continue looking in such a direction include the lack of self-interest of a machine, in the sense of Michie's traffic lights mentioned earlier, and the ability to consider a wide range of possibilities and outcomes (assuming them to be relevant) that a person might not know or forget.

Michie's emphasis was on a machine's lack of self-interest, which is the opposite of McDermott's view that a machine cannot make ethical decisions precisely because of its lack of self-interest. The Andersons (2010) have criticized this view of McDermott as giving an odd account of ethical dilemmas, which are normally about the choice of a best outcome between alternatives, rather than having no self-interest in an outcome. No classic ethical theory requires an agent to have a self-interest in an outcome, and in a Kantian rule-based system it would be excluded from consideration. Even a Humean sentiment-based ethics does not seem to require self-interest, and one's sentiments are, as is well known often at variance with one's "interests". Here, surely, McDermott is wrong both from any classical ethical perspective, and from any view seeking to go beyond those. Another highly interesting idea in McDermott's paper 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." He expands this point to argue that an ethical decision, to count as ethical, must be between alternative courses of action that it considers and compares. In that sense an ATM machine is never making an ethical decision, whether it gives one money or takes one's card back.

This is a very attractive idea, and I argued some time ago in (Wilks & Ballim, 1990) and elsewhere that an AI-based necessary condition for a machine having a belief – as opposed to simply acting on the basis of data – should be that it could compare two possible states of the world (which would normally include models of the beliefs of others). The basis of the system was computing or generating points of view, and there is a clear continuity of notions here, and the possibility of building into a future ethical machine a point-of-view engine capable of beliefs as a condition for it taking an ethical decision in McDermott's sense. The common theme here is that intelligent behavior may be intimately connected to the comparison of alternatives, and in a range of cognitive phenomena, rather than in a straight computation from unquestioned data,

The expression of verbal emotion and sympathy by computers, once an eccentric sideline in AI, has now progressed substantially and become a major subject in its own right, boosted as it has been by supporting research in psychology (Marsella et al., 2010), including substantial evidence of the ability of humans to establish emotional relationships with a wide range of non-human entities and mechanisms (Levy, 2007). We can thus consider, at least as a hypothesis, that effective conversational machine devices (what I termed Companions above) may be able to offer some simulacrum of emotion and apparent understanding of human pains and other emotional states, so that they might rise to that level of human mutual affection that is largely a mixture of politeness and other linguistic behaviors, and so need not rest on human computer substrate commonality at all. I shall assume a Companion, forming an integral part of the proposed ethical orthosis, would indeed have access to such simulations of emotion.

## 7. CONCLUSION

The paper has argued that an ethical machine is a real 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 view in which rationality is central, but may be based on quasi-inscrutable machine learning processes and models of sentiment and emotion that may be quantitative in form. I 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, working in tandem with 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.

**ACKNOWLEDGEMENTS**

I am 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 mine.

**REFERENCES**

Akoudas, K., Bringsjord, S. & Bello, P. (2005). Towards ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*.

Anderson, M. & Anderson, S. (2010). Robot Be Good. *Scientific American*.

Asimov, I. (1950). Runaround. In *Robot*. New York: Doubleday.

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In *The Handbook of Artificial Intelligence*. Cambridge: Cambridge UP.

Charniak, E. (1996). *Statistical Language Learning*. Cambridge, MA: Bradford Books.

Dennett, D. (1971). Intentional systems. *The Journal of Philosophy*, *68*(4), 87–106. doi:10.2307/2025382.

Eubanks, V. (2018). *Automating Inequality*. Macmillan: New York.

Foot, P. (1978). *Moral Dilemmas*. Oxford: Clarendon Press.

Ford, K., Hayes, P.J., Glymour, C. & Allen, J. (2015). Cognitive orthoses: towards human-centered AI. *AI Magazine*, Winter.

Gide, A. (1914). Les caves du Vatican. Editions de la nouvelle revue: Paris.

Gray, J. (2002). *Straw Dogs*. London: Granta Books.

Hoffman, R., Klein, G. & Mueller, S. (2018). *Literature Review and Integration of Key Ideas for Explainable AI*. Final Report on DARPA XAI Program.

Hume, D. (1751/1907). An Enquiry Concerning the Principles of Morals. David Hume, Essays Moral, Political, and Literary edited with preliminary dissertations and notes by T.H. Green and T.H. Grose, Longmans, Green: London.

Leibniz, G.W. (1988). Opinion on the principles of pufendorf. In P. Riley (Ed.), *Political Writings* (2nd ed., p. 71). Cambridge: Cambridge University Press.

Levy, D. (2007). *Love and Sex with Robots*. New York: Harper Collins.

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

Marsella, S., Gratch, J. & Petta, P. (2010). Computational models of emotion. In K.R. Scherer, T. Bänziger and E. Roesch (Eds.), *A Blueprint for an Affectively Competent Agent: Crossfertilization Between Emotion Psychology, Affective Neuroscience, and Affective Computing*. Oxford: Oxford University Press.

McDermott, D. (2008). Why ethics is a high hurdle for AI. In *Proc. North American Conference on Computers and Philosophy*, Bloomington, Indiana: NACAP.

Moor, J.H. (2006). The nature, importance and difficulty of machine ethics. *IEEE Intelligent Systems.*

Pearl, J. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Waldrop, M.M. (1987). A question of responsibility. *AI Magazine.*

Wason, P. & Johnson-Laird, P. (1972). *Psychology of Reasoning: Structure and Content*. Cambridge, MA: Harvard University Press.

Wilks, Y. (1973). Understanding without proofs. In *Proc. of the International Conference on Artificial Intelligence*, Stanford, CA.

Wilks, Y. (Ed.) (2010). *Artificial Companions*. Amsterdam: J. Benjamins.

Wilks, Y. & Ballim, A. (1990). Liability and consent. In A. Narayanan and M. Bennun (Eds.), *Law, Computers and Artificial Intelligence*. Norwood, NJ: Ablex.