Moral Orthoses: A New Approach to Human and Machine Ethics

Yorick Wilks

■ *I argue that both human and machine actions are more opaque than is generally realized and that the actions of both require explanation that an ethical orthosis might provide as aspects of artificial Companions for both human and machine actors. These explanations might well be closer to ethical accounts based on moral sentiment or emotion in the tradition of the primacy of sentiment over reason in this area of human and machine action.*

Nietzche once said that "there are no moral phenomena at all, only a moral interpretation of phenomena." That insight has implications for how we should see ethical machines and ourselves. The political philosopher John Gray has argued (2002) that we have little or no insight into how we take decisions, moral or otherwise, and a great deal of modern psychology agrees with him.

With the rise of machine learning as the core AI paradigm, we are getting used to the idea that we do not know how our programs make decisions either; hence, the rise of research in XAI, explainable AI, and the DARPA program to provide that. The European Commission has legislated a demand (Order GDPR 2016/2679) specifying that deployed machine learning systems must explain their decisions. The commission has done this even though no one knows how to provide what they are requiring. What would follow if we and machines are in roughly the same position with respect to the transparency of our ethical decision-making?

I want to reintroduce the notion of orthosis into ethical explanation: medically, an orthosis is an externally applied device designed and fitted to the body to aid rehabilitation, and usually 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. Could such an orthosis explain our own ethical behavior to us, as well as that of machines?

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). For Gray, we do not calculate ethical rules or consequences before acting, as the ethics text books tend to assume — and so neither should machines, he might have added. 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 for what they do. We can barely imagine legal and social life without this prop, even if it is all a fiction. That Gray's point is not yet generally accepted can be seen from a recent influential book in which the author writes: "I find the philosophy that sees human beings as unknowable black boxes and machines as transparent deeply troubling" (Eubanks 2018, 168)

It is important to remember that traditional ethical thought, like AI reasoning itself, assumed ethical reasoning to be one of calculations from rules or the summation of consequences. Ethical traditions appeal to calculation, logical or arithmetical, as their basis, which is why they have appealed for so long to the computationally minded. But these are not real calculations that are ever carried out, and real values are never assigned to possible outcomes in such discussions, even though, in the real world, automated systems like cars have to make real decisions every day.

Drew McDermott 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.' " (2008, 2). The latter is the ethical explanation problem, and I suggest we should consider as one central task of AI the provision of explanatory orthoses for both humans and machines, since the underlying behavior of both is opaque.

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. Yet, if machines that take decisions are based on ML algorithms, it is not clear that such unambiguous transparency will be available. There will always be alternative explanations of any behavior, and courtroom drama rests on that fact, unless, that is, something quite new and orthosis-like is added alongside whatever it is they are programmed with, which might be just another way of advocating for XAI.

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 datasets. Ethical argument, he suggests, requires a notion of causation that current ML systems cannot provide, which weakens them scientifically and makes them ineligible as ethical decision makers.

What might bring all parties together is this concept of orthosis: an external explanatory system, using an ontology of rules, causes, and outcomes, that might come to function in parallel with inscrutable brains and ML systems and provide possible explanations of why they act as they do.

Elsewhere (2010) I have developed, and implemented, the notion of a Companion: a personal web agent that is permanently associated with a human, that gains the maximum possible knowledge about its human "owner" via dialogue over an extended period of years, and that has been designed to handle the vast quantity of personal and public data that increasingly we cannot.

Such a Companion might be very like an orthosis to supply the data needed to make inferences about one's basis of action and might also contain self-revelations (or confessions) by an "owner" that could be crucial to ethical explanation. 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. Recent Senate SCOTUS hearings might have profited from such a device.

I have argued that both human and machine actions will require explanation and that an ethical orthosis might provide such explanations, in both cases, as aspects of artificial Companions for both human and machine actors. These explanations might well be closer to ethical accounts based on moral sentiment or emotion (MacIntyre 1985) in the tradition of the primacy of sentiment over reason in this area of human and machine action.

Acknowledgments

I am indebted to Ken Ford, Director of Florida IHMC, for the notion of orthosis, and to comments and criticisms from Selmer Bringsjord, Clark Glymour, Noel Sharkey, Fraser Watts, John Tait, Margaret Boden, and Eugene Charniak. The errors are, as always, all mine.

References

Bostrom, N., and Yudkowsky, E. 2014. The Ethics of Artificial Intelligence. In *The Handbook of Artificial Intelligence*. Cambridge, UK: Cambridge University Press.

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

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

MacIntyre, A. 1985. *After Virtue*. London: Duckworth.

McDermott, D. 2008. Why Ethics Is a High Hurdle for AI. Paper presented at the North American Conference on Computers and Philosophy. Bloomington, IN, July 10–12.

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

Wilks, Y., ed. 2010. *Artificial Companions*. Amsterdam, Netherlands: John Benjamins.

Yorick Wilks is a senior research scientist at the Florida Institute of Human and Machine Cognition and a professor of artificial intelligence at the University of Sheffield. He is a fellow of the AAAI and ACM.