Machine Ethics: Evolutionary Teachings

Luís Moniz Pereira

NOVA Laboratory for Computer Science and Informatics (NOVA LINCS),
Departamento de Informática, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, Portugal
lmp@fct.unl.pt

Abstract

We ponder on the teachings of human moral evolution studies for machine ethics.

Keywords: Evolutionary anthropology, utilitarianism, mutualism, contractualism, reciprocity, emergence, computational morality, Evolutionary Game Theory.

Teachings

Added dependency on cooperation makes it more competitive to cooperate well. Thus, it is advantageous to invest on shared morals in order to attract partners who will partake of mutual and balanced advantages.

This evolutionary hypothesis inspired by mutualism (Baumard 2010)—itself a form of contractualism (Ashford and Mulgan 2012)—contrasts with a number of naturalist theories of morality, which make short shrift of the importance of cognition for cooperation. For example, the theory of reciprocity, in ignoring a wider cognitive capacity to choose and attract one's partners, forbids itself from explaining evolution on the basis of a cooperation market.

Indeed, when assigning all importance to population evolutionary mechanisms, naturalist theories tend to forget the evolution of cognition in individuals. Such theories habitually start off from evolutionary mechanisms for understanding the specificity of human morals: punishment (Boyd and Richerson 1992; Sober and Wilson 1998), culture (Henrich and Boyd 2001; Sober and Wilson 1998), political alliances (Boehm 1999; Erdal et al. 1994). According to Baumard's hypothesis, morality does not emerge because humans avail themselves of new means for punishing free-riders or for recompensing cooperators, but simply because mutual help—and hence the need to find partners—becomes much more important.

In summary, it's the development of cooperation that induces the emergence of morals, and not the stabilization of morals (via punishment or culture) that promotes the development of cooperation.

Experimental results are in line with the hypothesis that the perfecting of human intuitive psychology is responsible for the emergence of morality, on the basis of an improved understanding of the mental states of others. This permits to communicate, not just to coordinate with them, and thus extend the domain cooperation, thereby leading to a disposition toward moral behaviors. For a systematic and thorough account of research into the evolutionary origins of morality, see (Krebs 2011; Bowles and Gintis 2011).

At the end of the day, one may consider three theories bearing on three different aspects of morality: the evaluation of interests for utilitarianism, the proper balance of interests for mutualism, and the discharging of obligations for the virtues principled.

A naturalistic approach to moral sense does not make the psychological level disappear to the benefit of the evolutionary one. To each its explanation level: psychology accounts for the workings of the moral sense; sociology, for the social context that activates it; and a cupola theory, for the evolution of causes that occasioned it (Sperber 1997). Moral capability is therefore a "mechanism" amongst others (Elster 1998), as are the concern for reputation, the weakness of the will, the power to reason, etc.

An approach that is at once naturalist and mutualist allows escape from these apparently opposite viewpoints: the psychological and the societal. At the level of psychological motivations, moral behavior does neither stem from egotism nor altruism. To the contrary, it aims at the mutual respect for everyone's attending interests. And, simultaneously, it obeys the logic of equity. At the evolutionary level, moral behavior is not contradictory with egotism because, in human society, it is often in our own interest to respect the interests of others. Through moral motivations, we avail ourselves of a means to reconcile the diverse individual interests. Morality vies precisely at harmonizing individual interest with the need to associate, and profit from cooperation, by adopting a logic of fairness.

The mutualist solution is not new. Contractualist philosophers have upheld it for some time. Notably, they have furnished detailed descriptions of our moral capacity (Thomson 1971; Rawls 1971). However, they never were able to explain why humans are enabled with that particular capacity: Why do our judgments seek equity? Why do we behave morally at all?

Without an explanation, the mutualist theory seems improbable: Why behave we as if an actual contract had been committed to, when in all evidence one was not?

Past and ongoing evolutionary studies, intertwining and bridging cognitive and population aspects, and both becoming supported on computational simulations, will help us find answers to that. In the process, rethinking machine ethics and its implementations.

According to (Boehm 2012), conscience and morality evolved, in the biological sense. Conscience evolved for reasons having to do with environments humans had to cope with prehistorically, and their growing ability to use group punishment to better their social and subsistence lives and create more equalized societies. His general evolutionary hypothesis is that morality began with having a conscience and that conscience evolution began with systematic but initially non-moralistic social control by groups.

This entailed punishment of individual "deviants" by bands of well-armed large-game hunters, and, like the ensuing preaching in favor of generosity, such punishment amounted to "social selection", since the social preferences of members and of groups as a whole had systematic effects on gene pools.

This punitive side of social selection adumbrates an immediate kind of "purpose", of large-brained humans actively and insightfully seeking positive social goals or avoiding social disasters arising out of conflict. No surprise the genetic consequences, even if unintended, move towards fewer tendencies for social predation and more towards social cooperation. Hence, group punishment can improve the quality of social life, and over the generations gradually shape the genotype in a similar direction.

Boehm's idea is that prehistoric humans made use of social control intensively, so that individuals who were better at inhibiting their own antisocial tendencies, by fear of punishment or by absorbing and identifying with group's rules, garnered a superior fitness. In learning to internalize rules, humankind acquired a conscience. At the beginning this stemmed from punitive social selection, having also the strong effect of suppressing free riders. A newly moralistic type of free-rider suppression helped evolve a remarkable capacity for extra-familial social generosity. That conscience gave us a primitive sense of right and wrong, which evolved the remarkable "empathy" which we are infused with today. It is a conscience that seems to be as much a Machiavellian risk calculator as a moral force that maximizes prosocial behavior, with others' interests and equity in mind, and minimizes deviance too. It is clear that "biology" and "culture" work together to render us adaptively moral.

Boehm believes the issue of selfish free riders requires further critical thought, and that selfish intimidators are a seriously neglected type of free rider. There has been too much of a single-minded focus on cheating dominating free rider theorizing. In fact, he ascertains us the more potent free riders have been alpha-type bullies, who simply take what they want. It is here his work on the evolution of hunter-gatherer egalitarianism enters, namely with its emphasis on the active and potentially quite violent policing of alpha-male social predators by their own band-level communities. Though there's a large literature on cheaters and their detection, free-rider suppression in regard to bullies has not been taken into account so far in the mathematical models that study altruism.

"For moral evolution to have been set in motion," Boehm (Boehm 2012) goes on, "more was needed than a preexisting capacity for cultural transmission. It would have helped if there were already in place a good capacity to strategize about social behavior and to calculate how to act appropriately in social situations."

In humans, the individual understanding that there exists a self in relation to others makes possible participation in moral communities. Mere self-recognition is not sufficient for a moral being with fully developed conscience, but a sense of self is a necessary first step useful in gauging the reactions of others to one's behavior and to understand their intentions. And it is especially important to realize that one can become the center of attention of a hostile group, if one's actions offend seriously its moral sensibilities. The capacity to take on the perspective of others underlies not just the ability of individuals in communities to modify their behavior and follow group imposed rules, but it also permits people acting as groups to predict and cope insightfully with the behavior of "deviants."

Social selection reduced innate dispositions to bully or cheat, and kept our conscience in place by self-inhibiting antisocial behavior. A conscience delivers us a social mirror image. A substandard conscience may generate a substandard reputation and active punishment too. A conscience supplies not just inhibitions, but serves as an early warning system that helps prudent individuals from being sanctioned.

Boehm (Boehm 2012) wraps up: "When we bring in the conscience as a highly sophisticated means of channeling behavioral tendencies so that they are expressed efficiently in terms of fitness, scenarios change radically. From within the human psyche an evolutionary conscience provided the needed self-restraint, while externally it was group sanctioning that largely took care of the dominators and cheaters. Over time, human individuals with strong free-riding tendencies—but who exercised really efficient self-control—would not have lost fitness because these predatory tendencies were so well inhibited. And if they expressed their aggression in socially acceptable ways, this in fact would have aided their fitness. That is why both free-riding genes and altruistic genes could have remained well represented and coexisting in the same gene pool."

Conclusions

For sure, we conclude, evolutionary biology and anthropology, like the cognitive sciences too (Hauser 2007; Gazzaniga 2006; Churchland 2011; Greene 2013; Tomasello 2014), have much to offer in view of rethinking machine ethics, evolutionary game theory simulations of computational morality, and functionalism to the rescue (Pereira 2016).

Acknowledgments

LMP acknowledges support from FCT/MEC NOVA LINCS PEst UID/CEC/04516/2013.

References

Ashford, E., and Mulgan, T. 2012. Contractualism. `http://plato.stanford.edu/archives/fall2012/entries/\\contractualism/`.

Baumard, N. 2010. *Comment nous sommes devenus moraux: Une histoire naturelle du bien et du mal*. Paris: Odile Jacob.

Boehm, C. 1999. *Hierarchy in the Forest: the Evolution of Egalitarian Behavior*. Cambridge, MA: Harvard University Press.

Boehm, C. 2012. *Moral Origins: the Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Bowles, S., and Gintis, H. 2011. *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press.

Boyd, R., and Richerson, P. 1992. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13(3):171–195.

Churchland, P. 2011. *Braintrust: What Neuroscience Tells Us about Morality*. Princeton, NJ: Princeton University Press.

Elster, J. 1998. A plea for mechanisms. In *Social Mechanisms: an Analytical Approach to Social Theory*. Cambridge, NY: Cambridge University Press.

Erdal, D.; Whiten, A.; Boehm, C.; and Knauft, B. 1994. On human egalitarianism: an evolutionary product of machiavellian status escalation? *Current Anthropology* 35(2):175–183.

Gazzaniga, M. S. 2006. *The Ethical Brain: The Science of Our Moral Dilemmas*. New York: Harper Perennial.

Greene, J. 2013. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York, NY: The Penguin Press HC.

Hauser, M. D. 2007. *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. London, UK: Little Brown.

Henrich, J., and Boyd, R. 2001. Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology* 208(1):78–89.

Krebs, D. L. 2011. *The Origins of Morality – An Evolutionary Account*. Oxford U. P.

Pereira, L. M. 2016. Software sans emotions but with ethical discernment. In Silva, S. G., ed., *Morality and Emotion: (Un)conscious Journey to Being*. London: Routledge.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

Sober, E., and Wilson, D. 1998. *Unto Others: the Evolution and Psychology of Unselfish Behavior*. Cambridge, MA: Harvard University Press.

Sperber, D. 1997. Individualisme méthodologique et cognitivisme. In *Cognition et sciences sociales*. Paris: Presses Universitaires de France.

Thomson, J. J. 1971. A defense of abortion. *Philosophy & Public Affairs* 1(1):47–66.

Tomasello, M. 2014. *A Natural History of Human Thinking*. Cambridge, MA: Harvard University Press.