Chapter 12
Is It Possible to Program Artificial Emotions? A Basis for Behaviours with Moral Connotation?

Abstract The fact that machines can recognize emotions, or even be programmed with something functionally similar to an emotion, does not mean that they exhibit moral behaviour. The laws defined by Isaac Asimov are of little use if a machine agent has to make decisions in complex scenarios. It must be borne in mind that morality is primarily a group phenomenon. It serves to regulate the relationship among individuals having different motivations regarding the cohesion and benefit of that group. Concomitantly, it moderates expectations about one another. It is necessary to make sure agents do not hide malevolent purposes, that they are capable of acknowledging errors and to act accordingly. One must begin somewhere, even without presently possessing a detailed knowledge of human morality, to the extent of programming ethical machines in full possession of all the functions of justification and argumentation that underlie decisions. This chapter will discuss the bringing out of a moral lexicon shareable by most cultures. The specific case of guilt and the capacity to recognize it is present in all cultures. It can be computer-simulated and can be a starting point for exploring this field.

Since we have frequently referred to terms like symbiosis, human-machinemachine interaction and cooperation, it is not surprising that we wonder which axes support the group cohesion of agents, be they humans or machines. After all, our digital partners—even if, for the time being, they do not show much introspective sense—are there to share decision-making processes with us.

Isaac Asimov, one of the most remarkable sci-fi writers, endowed his robotic characters with his celebrated Three Laws of Robotics: three red lines of moral inspiration.[1] These lines structure the moral relationship of robots with humans. Under the first law, a robot may never injure a human, or – through inaction – allow it to be harmed. The second states that robots should always obey humans, except in cases when this would conflict with the first law. Lastly, the third law endorses the need for the robot to protect its own existence, except in cases this would involve breaking the first and second laws. Later, most likely already aware of the potential, but credible, risks resulting from the technological explosion associated with the

[1] Asimov (1950).

L. M. Pereira and A. B. Lopes, *Machine Ethics*, Studies in Applied Philosophy, Epistemology and Rational Ethics 53,
https://doi.org/10.1007/978-3-030-39630-5_12

implementation of AI, he felt the need to establish a new clause, which he coined the zeroth law. Its content stipulates that a robot may not cause harm to humanity, or by inaction, allow any harm to come to it. Behind this enumeration, we imagine an author sceptical of humanity, and hopeful as to the protective capacity of artificial agents.

The restrictions presented here constitute moral precepts regulating the relationship between machines and humans, in a context in which the human is the owner and master. Because of this, they are of little use for entangled scenarios where we will not even know whether or not decision "x" or "y" was taken by humans. In these cases, we wish for a sharing of moral common sense, a distributed knowledge of rules and precepts, and an identical notion of their application. Above all, we want each decision maker to be able to spell out the arguments that supported their resolutions. In this sense, it is understood that there are unavoidable concerns related to the need to generate trust among agents. We do not yet have much experience of what is a cognitive environment filled with differentiated agent types; but we are well aware of the consequences, and the social respectability of a person who systematically does not keep his promises, or does not adequately justify his deliberations. Social cohesion requires trust and commitment. This means that artificial agents will also have to participate in these dynamics. Even if they come from diverse manufacturers, equipped with algorithms with different acting instructions, they will have to inspire confidence. In this respect, we are very much used to dealing with humans and understanding the difficulties associated with that, the problems relative to the presence of free riders, the capability of dissimulation that enables to conceal real selfishness with feigned altruism, and even the recognition that moral precepts are not coercive. Often our moral conscience allows us to identify what the good deed shall be; however, despite recognizing it as such, we do the opposite. We now have a very clear sense of how these situations break trust, and how difficult it is to regain it.

As far as machine decisions are concerned, the potential for distrust may be further amplified, either by our fears and expectations, or by the inevitable mistakes the new machines will make, which, by norm and custom, are of today being ascribed a demanding technical perfection.

On the fears side, there is not only the possibility of them getting our jobs, or even imposing required behaviours on us. On the expectation side, we do not accept that they make mistakes, at least with the same typology of human error. If an artificial driver kills a human, such event is not comparable to our disabling of a cognitive machine, or to an accident that destroys an intelligent robot. Nevertheless, legislation on artificial entities, based on a judicious autonomous agent conception, will eventually recognize robots as beings capable of moral decision, mitigating part of this difference.

In future, the moral lexicon will be common to both humans and machines, and this includes notions such as right and wrong, worthy and unworthy, cooperative and selfish, acknowledgment of guilt and apology, and so on. This will require machines to be equipped with a consciousness-like device that enables them to produce informed judgments and their associated argumentation. It is in this context that

we may question which elementary items might be shared by machines from different manufacturers, with different programming languages and, eventually, with distinct scopes, not all properly spelled out. In this sense, we may ask ourselves which categories must inevitably be present in AI development programs, and which faculties favouring moral discernment must be integrated into artificial agents.

The need and urgency for machine morality, and its respective research, has been stressed several times. As artificial agents become more sophisticated and autonomous, acting in groups within populations of both other machines and humans, this need will be increasingly justified. To understand this idea, just think about the number and diversity of robots that are being introduced into the market. Zora, a Softbank Robotics[2] creation, is not only able to accompany hospitalized adults and children, but is also an excellent helper in nursing homes, and can also be used in an educational setting to perform various functions traditionally performed by human teachers. Zora interacts with emotionally debilitated people, but it is not yet a robot capable of meeting all the resulting challenges. However, it is these challenges that delimit part of the path to be followed. That is, computational morality entails a whole research program that should be incorporated into all future AI developments.

In a first phase, it is very relevant to identify the lexicon traditionally associated with the moral context. Thus, pairs of terms such as cooperation/competition, guilt/impenitence, shame/honour can be identified as axes of the referred lexicon. Its organization in pairs of opposites expresses the possibility for agents to make choices.

Also, the need has been several times mentioned to start with small steps, which can be consequential. Regarding the study of the cognition of human individuals integrated in multiple agent groups and who frequently interact morally (they may choose to compete or cooperate with one another), it is important to understand whether the results of the research can be equally applicable to the evolution of populations of artificial agents. Or still, if the results obtained in laboratory contexts with populations of artificial agents allow for a better understanding of human morality and its "machinery".

Specifically, in relation to human morality, the answer seems to be a resounding "yes". We must always bear in mind that morality concerns groups and populations, requires cognition, and will have to evolve into an intertwining and strengthening of the relationship between nature, genetics and culture. On the other hand, evolutionary Anthropology, Psychology, and Neurology have produced new insights into the evolution of human morality. Their theories and scientific results must be considered and serve as inspiration when thinking about machine morals. Indeed, the very study of ethics and of the evolution of human morality, can now also draw on experimental means, on the theory of computation, and on robotics to represent and simulate individual—or group—moral reasoning in a multitude of circumstances. Regardless of these occasions, however, morality is to be general, making explicit a set of procedures capable of ensuring that the equal is treated as equal, and the

[2] https://www.softbankrobotics.com.

different as different. That is, it must produce rules and procedures that ensure fair and equitable treatment of agents.

Using these domains, we can better understand the morals that emerge in agent populations in a given context. In addition, human groups tend to rival each other—sometimes too much. Therefore, it is important to research moral items that are beyond cultural differences and are represented in various value systems.

It is in this context that, in the morals of groups, themes such as shame and guilt can be invoked. These are social emotions of great importance. Although both have evolved to promote cooperation, guilt and shame can be dealt with separately. Despite being able to promote acquittals and even spontaneous public confessions, guilt is an internal private phenomenon. An important point, not always properly explored, is that present guilt helps to prevent future guilt, due to the pain it might entail. It is thus a form of prospecting for the future, always very useful in games and for survival. As for shame, it has, inherently, the trait of a public performance that is personally unwanted by the agent, because it addresses his own essence, and not just his act, as is the case of guilt, and leads to shunning the social. As in the case of guilt, it can also lead to similar consequences: excuses, apology and change of behaviour. Shame, however, depends on the agent being caught, on not misleading deliberately, and on the existence of a mechanism of social reputation.

No other emotion is more directly associated with morality than guilt. If we consider the Catholic religion in terms of game theory, we know that we are born losing, with original sin. It was at its expense that Christ suffered and died to save us, which makes us doubly guilty if we choose not to play properly. Associated with guilt comes confession, the request for absolution and its respective pardon. This provides the opportunity for a reset; the game being playable over and over again. The notion of guilt is closely associated with the idea of conscience as an internal guide that tells us when an action is wrong. Furthermore, guilt is widely regarded as a fundamentally collective emotion that plays a positive prosocial role. It arises especially when there is a threat of separation or exclusion. Guilt is an unpleasant emotion, and when experienced, people try to free themselves of it: the most common coping strategies are confession, amends, self-criticism, and punishment, often self-inflicted.

We must bear in mind that, in terms of real material gain, non-compliance is the dominant strategy in many economic games: defaulters do better than cooperators, whether or not their trading partners are cooperating. This makes it rational for both parties to disobey, even if the results of mutual misconduct are worse than those of mutual cooperation. In order to mitigate this embarrassment, many Evolutionary Game Theory (EGT) theorists have argued that guilt is not only anticipatory, but mitigates this problem by promoting a cooperative attitude, adding an emotional cost to failure. Trivers (2011) speculates that coevolution has caused guilt to arise because it makes fault less attractive. People can earn materially through a defaulting strategy, but guilt makes them suffer emotionally, and this in turn leads them to cooperate. Robert Frank notes that this trend is so deeply rooted that people avoid failure, even in cases when the other party will not be a likely partner in future exchanges. Both Robert Trivers and Robert. Frank assume that the emergence of guilt is the result of evolution, selected as if it were a process of biological evolution. It is possible

that guilt came under cultural pressure as a tool to ensure that people cooperate. It is interesting to note that, alongside this moral development, nature—and later culture—has engendered a kind of arms race for the development of camouflage strategies and generation of deceptions, which has culminated in the human capacity to lie. Most likely, part of the development of the large brain that characterizes us concerns the enduring need to decode strategies that imply deception.

The evolutionary question about guilt is whether guilt is worth more than its absence with respect to the advantageous emergence of cooperation. In order to research the problem, we can explicitly introduce guilt into computational models, thereby simulating its appearance in the evolutionary scene. We can test this hypothesis through such computational models. In them, guilt is linked to the fear of recognition of malevolent intent, since infraction avoidance will have evolved from the fear of detection of the harm caused by it. The prediction of these models is that guilt will facilitate, and accelerate, the emergence of cooperation. Although its initial cost is heavier, over time the cost will eventually be recovered within the guilt-ridden population by inhibiting faulty strategies. The basic hypothesis is that, with some guilt present in the initial population, the most frequently evolutionary stable situation includes guilt and better cooperation.

In a community where there is a small number of individuals capable of feeling guilt, the models show that behaviours that derive from this variable will steadily increase their replication across the population: more than if there were no guilt. This proves two things: first, that machines must feel guilt and—second—that they must be programmed with such a capability. It also explains why the feeling of guilt in humans corresponds to a beneficial evolution/mutation in their cooperation games.

How is it implemented? Guilt is a kind of self-punishment: if I betray the other who has cooperated, I imagine a counterfactual. I.e., if some other play of mine would have been preferable, knowing already how the other would play, then I deduce something from what I won in the play, and change my behaviour. This subtracting something is the self-punishment. Our own research using EGT has computationally demonstrated the cooperative advantages of this guilt model (Pereira et al. 2017).

Under these terms, it is essential to provide cognitive machines with the capacity to recognize guilt and corresponding apologies. Guilt not only mends present situations but prevents future evils. The need to achieve this is evident as it will be its execution. Furthermore, guilt, as a structuring emotion for cohesion and confidence building among group members, is fundamental to fostering the cooperative relationships which must be our target if we are to live in a better society.

References

Asimov, I. (1950). Runaround. I, Robot (The Isaac Asimov Collection ed.). New York, NY: Doubleday.

Pereira, L. M., Lenaerts, T., Martinez-Vaquero, L. A., & Han, T. A. (2017). Social manifestation of guilt leads to stable cooperation in multi-agent systems. In S. Das et al. (Eds.), *Proceedings of the 16th conference on autonomous agents and multiagent systems* (pp. 1422–1430). May 8–12, São Paulo, Brasil.

Trivers, R. (2011). *Deceipt and self-deception: Fooling yourself the better to fool others*. London: Allen Lane.