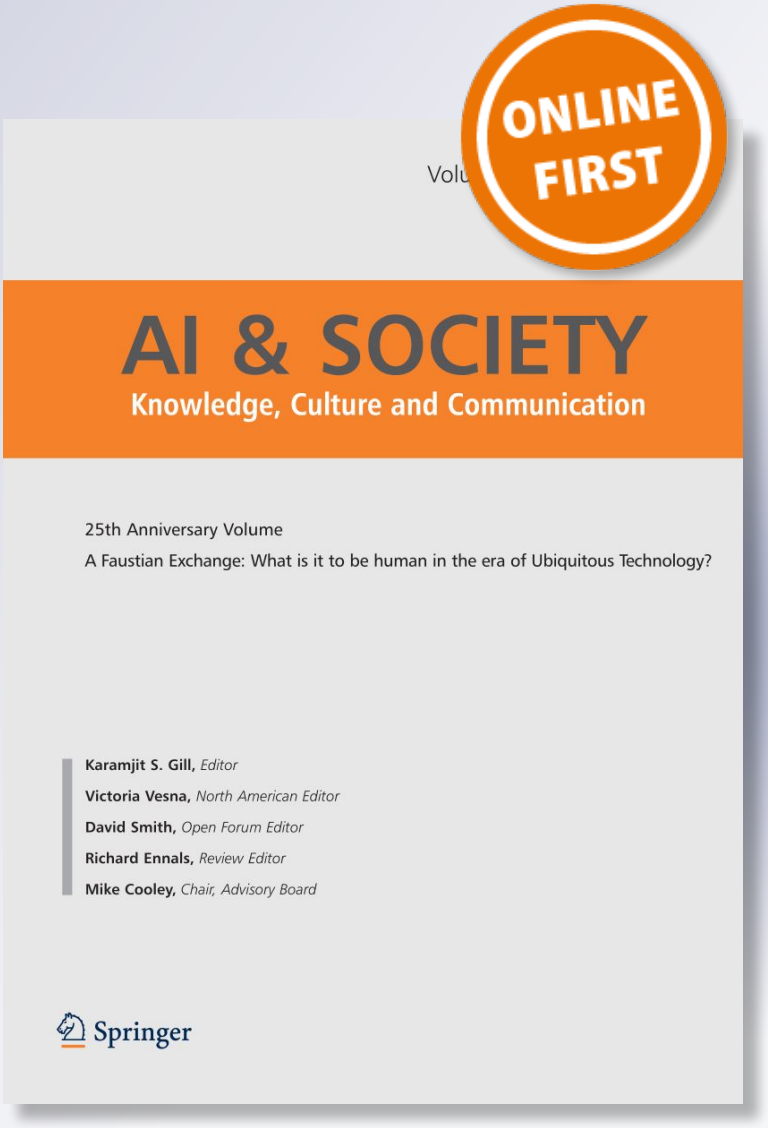
# *Evil and roboethics in management studies*

**Enrico Beltramini**

**AI & SOCIETY**
Journal of Knowledge, Culture and Communication

ISSN 0951-5666

AI & Soc
DOI 10.1007/s00146-017-0772-x




Springer

AI & Soc
DOI 10.1007/s00146-017-0772-x

OPEN FORUM

# Evil and roboethics in management studies

**Enrico Beltramini**[1]

Received: 23 August 2017 / Accepted: 24 October 2017


**Abstract** In this article, I address the issue of evil and roboethics in the context of management studies and suggest that management scholars should locate evil in the realm of the human rather than of the artificial. After discussing the possibility of addressing the reality of evil machines in ontological terms, I explore users' reaction to robots in a social context. I conclude that the issue of evil machines in management is more precisely a case of *technology anthropomorphization*.

**Keywords** Roboethics · Evil · Management · Anthropomorphism

## 1 Introduction

Two viewpoints can be traced regarding the character of evil: first, evil-skepticism insists that morality demands humans abandon the concept of evil—that they make evil undistinguishable from wrongdoing. This is the usual approach taken in management studies. Second, evil-revivalism insists that the concept of evil should be revived– that morality demands that humans make evil intelligible. My own sympathies tend toward the latter line of thinking. With this premise in mind, this paper is a preliminary attempt at understanding evil in roboethics within the context of management studies. I address the question of evil machines from two perspectives: (1) from the machine ethics' viewpoint, that is, from the ontological perspective; and (2) from the users' viewpoint, that is, from a roboethics' viewpoint. I argue that the latter is more congenial to management studies. Thus, this article assumes the perspective of users in case of 'social machines,' that is, a type of machines with social dimensions and requirements. This perspective is relevant to management studies. In this paper, I postpone a direct analysis of what evil means in relation to machines to instead discuss the conditions under which humans can call a machine evil. Thus, this article aims to settle some theoretical issues concerning machines' moral status in the view that humans might blame machines for deliberately causing economic and financial harm. Evil is much easier to examine and debate when there is obvious harm that may result from actions and consequences. The same principle seems not to hold in the topic of evil machine (Taddeo 2010). There exists a need to examine evil in machine studies because robots are already deemed evil. In this article, evil is considered with regards to anthropomorphism, and the debate is concerned with how humans relate to these machines in both the design and use phase of their operation.

By claiming that the users' viewpoint is relevant to management studies when it comes to address 'evil machines,' this paper focuses on one important theoretical determinant of evil in the nonhuman agent: anthropomorphism (Epley et al. 2007; Waytz et al. 2010). Anthropomorphism is a process of inductive inference whereby people attribute to nonhumans distinctively human characteristics, particularly the capacity for moral (or immoral) agency. Anthropomorphizing a nonhuman does not simply involve attributing superficial human characteristics (e.g., a humanlike face or body) to it, but rather attributing essential human characteristics to the agent (namely a humanlike immorality, a capability for evildoing).

This work is divided into three parts: first, this article addresses evil. This part traces changes that have occurred in the West's understanding of evil and its place in the world from the Christian era to this current time. The main point of

✉ Enrico Beltramini
ebeltramini@ndnu.edu

[1] Notre Dame de Namur University, 1500 Ralston Avenue, Belmont, CA 94002, USA

Published online: 30 October 2017

this brief summary is to design a trajectory of evil as seen initially as extrinsic to human nature and then intrinsic to it. In Christian times, human nature is not corrupted by evil. In modernity, however, secularization disenchants the world and evil comes to reside within the human heart and mind. For example, Georges Bataille identifies evil in literature, Richard Bernstein investigates the notion of 'radical evil', and Hannah Arendt forges the concept of 'banal evil' (Bataille 2001; Bernstein 2002; Geddes 2003). Philip Zimbardo studies how good people turn evil (Zimbardo 2007). The second part of this article investigates the possibility of evil machines. Based on insights from roethics and recent other studies in computer science and philosophy, designers and engineers build AI systems with the ability to make moral judgments and decisions in realistic scenarios. However, designers cannot predict all circumstances and consequently, AI systems maintain a certain degree of unpredictability. The current effort at identifying or developing ways to ensure that artificial intelligence does not turn out evil are also mentioned. The third part of the article focuses on evil machines in management studies. The notion of evil machines is reframed in terms of a process of anthropomorphization, in which machines take on cognitive work with social dimensions previously performed by humans, transforming machines into persons. This process operates in terms of a heuristic how, not an ontological what. In other words, one can assume that evil machines in management studies are simply anthropomorphism for the most part. A pragmatic version of anthropomorphism is discussed with regard to intelligent machines and robots.[1]

## 2 Problem

Automated technology is increasingly replacing humans to perform complicated tasks in domains ranging from economy to finance. In this paper, an important theoretical determinant of people's willingness is investigated to consider intelligent machines as evil machines in respect to management studies. Evil is a multifaceted concept that can refer to the belief that another will behave with malevolence. The idea of evil involves moral condemnation ascribed to human beings, so that human beings can perform evil actions. Evil is today used for exceptional events—holocaust, September 11, ISIS—that do not necessarily apply to the world of management. As a matter of fact, evil is a neglected subject in management studies. The concept of evil comes to mind when the moral significance of certain actions and their perpetrators cannot be captured by calling them 'wrong' or 'bad' or even 'very, very wrong' or 'very, very bad' (Darley 1992, 1996; Styhre and Sundgren 2003). Negative, even deplorable actions in management, like firing people without reason, stealing confidential information from somebody else's computer, or evading taxes are not considered evil, but rather immoral or illegal (i.e., Adams and Balfour 2009; Tang 2010; Schnall and Cannon 2012). This state of affairs may change soon with the rise of robotics.

The possibility of autonomous intelligent machines raises a host of ethical issues. These questions relate both to ensuring that such machines do not harm human beings and to the moral status of the machines themselves. Usually business ethics addresses the former, that is, new technologies can be used in dangerous and even malevolent ways, as well as in beneficial ways. Scholars of ethics in technology management invite practitioners and other scholars to carefully reconsider their assumption, because they operate with an anthropology that is unrealistic regarding the human proclivity to turn good into evil. In other words, humans can be evil and can use technology to do evil. The former corporate motto of Google quoted in the epigraph at the beginning of this article is a reminder of the possibility of human malevolence or moral corruption in the use of technology.

More interesting and less discussed is the second option, that is, autonomous machines and the possible harm that can come from them (Johnson et al. 2004; Nadeau 2006; Sullins 2005). Usually this field of study is covered by the newly emerging areas of machine ethics, roboethics, and their various synonyms (machine morality, friendly AI, artificial morality, and roboethics). Traditionally, machine ethics (or machine morality) is concerned with describing how machines could behave ethically towards humans; roboethics is concerned with how humans relate to these

[1] The following offers definitions for some of the most important terms used in this document. 'Evil' is an action that is not simply morally wrong, but leaves no room for understanding or redemption. Evil is qualitatively, rather than merely quantitatively, distinct from mere wrongdoing. 'Evil machine' is a machine's action that causes harm to humans and leaves no room for account or expiation. 'Robot' stands for both physical robots and virtual agents roaming within computer networks; 'autonomous machine' is a decision-making machine; 'artificial intelligence' is the ability of autonomous machines to make decisions; 'intelligent machine' and 'autonomous intelligent machine' are synonymous with 'autonomous machine.' 'Machine' is an umbrella term to cover robots and autonomous and intelligent machines. 'Machine learning algorithm' can be categorized as being supervised or unsupervised. Supervised algorithms can apply what has been learned in the past to new data. Unsupervised algorithms can draw inferences from datasets. An important distinction in this article is played between humans as designers and engineers, i.e., those who build the machine, and humans as users or clients, i.e., those who interact socially with the machine. The former are named 'designers' and 'engineers,' the latter 'users,' 'investors,' 'clients,' or, when the text moves from the specific case study to more general considerations, 'humans' and 'humanoids.' Giving human characteristics to artificial objects is a human trait called 'to anthropomorphize.' Biblical quotes are from the new revised standard version of the Oxford annotated Bible with Apocrypha (Croogan 2010).

machines in both the design and use phase of their operation. Although in the past decade the terms machine ethics and roboethics have drifted a bit and have been used somewhat synonymously to refer to the ethical concerns raised by robotics technologies, in this paper the original separation is maintained. Machine ethics and its various synonyms (machine morality, friendly AI, artificial morality, and roboethics) focus on the prospects for building computers and robots that are moral decision-makers. Scholars of machine ethics research ways to make machines moral by merging ethical theory and engineering practices (Floridi and Sanders 2004; Lee et al. 2005; Irrgang 2006; Powers et al. 2007; Arkin 2009; Wallach and Allen 2008; Coeckelbergh 2012; Lin et al. 2014). In fact, the ethical behavior of machines is determined by the way their systems have been designed. To put it differently, the ethical behavior of autonomous machines depends on their design, but the design, and the determination of the ethical behavior of machines, ultimately depends on the extent that the designers can predict every single situation a machine will ever encounter. When machines stop operating in limited contexts and cross the threshold where the designers and engineers can no longer predict how machines will behave, the initial determination fails to provide a course of action. The development of autonomous computers and robots as kinds of entities that can make moral decisions on their own is the response to this situation.

Given the relatively primitive state of present-day artificial intelligence (AI) research, the discussions around machine ethics tend to indulge highly speculative possibilities. However, the possibility that machines turn out evil is already considered by scholars. Wallach and Allen predict that within the next few years "there will be a catastrophic incident brought about by a computer system making a decision independent of human oversight" (Wallach and Allen 2008, 4). They further forecast that humans will likely blame the machines for deliberately causing harm well before philosophical issues concerning their moral status have been fully settled (Wallach and Allen 2008, 199). Also Sullins states that, in certain circumstances, machines can be seen as moral agents once three conditions are fulfilled: (i) autonomy from programmers or operators, (ii) their behavior is analyzed or justified in terms of intention to do good or evil (iii) behaves as though it is responsible for another moral agent (Sullins 2006). To articulate the question concerning robotics in terms of evil may easily suggest science-fiction scenarios like the story in the film *I, Robot*, in which robots become artificially intelligent to such an extent that humans wonder if they can harm humans—if the robots can be evil. But there is a broader and certainly more urgent issue about evil in intelligent autonomous technologies—technologies that are already available today or will be soon. Robotics for financial and economic applications are fast replacing humans in cognitive work with social dimensions, and in doing so, they inherit the social requirements (i.e., information, semantic, and community requirements) previously met by humanoids (Dennett 1998). As the outsourcing of financial and economic applications to robotics becomes more pervasive, the social requirements that robotics are prescribed to satisfy also grow, bringing to the fore issues like the nature of evil and whether evil can be developed by a machine or can only concern human beings.

## 3 Evil

In this section I want to distinguish the notion of evil as extrinsic to the human from another notion—that evil is intrinsic to the human. I also show that a trajectory from a religious concept of evil as extrinsic in the premodern era gives away to a secular concept of evil as intrinsically human in modernity. Scholar Ervin Staub, an authority on the psychology of good and evil, argues that "evil is not a scientific concept with an agreed meaning, but the idea of evil is part of a broadly shared human cultural heritage" (Staub 1989, 25). In fact, evil originates in the philosophical and theological traditions and remains dominant in the current modern age (Garrard 1998, 2002; Calder 2002; Steiner 2002). To avoid confusion, it is important to note that there are at least two concepts of evil: a religious concept and a secular concept. Evil in the broad sense, which includes metaphysical commitments to dark spirits, the supernatural, or the devil, tends to be the sort of evil referenced in theological contexts. In contrast to the broad concept of evil, the narrow concept of evil picks out only the most morally despicable sorts of actions, characters, and events. Since the narrow concept of evil involves moral condemnation, it is appropriately ascribed only to moral agents and their actions. For example, if only human beings are moral agents, then only human beings can perform evil actions. Evil in this narrower sense is more often meant when the term 'evil' is used in contemporary moral, political, and legal contexts. The broad concept of evil marks the pre-modern era; the narrow notion of evil dominates modernity. While pagans and Christians allowed themselves to adopt an extrinsicist approach to evil, that is, to think of evil as purely extrinsic to human condition, moderns seems to embrace an intrinsicist position and, therefore, confront evil as part of human nature.

At risk of oversimplification, almost the entire Western tradition assumes that evil does *not* come from the absolute good. On this point, biblical tradition, Greek Platonism, and Christianity agree. The generation of evil in the Old Testament is more accurately a degeneration of a positive

creation, the fall of the rebel angels (i.e., Gen 6:1–4). In *Republic*, Plato repeats again and again that God is perfectly good and cannot be the author of evil (Republic—book 2, 379c, in Allen (2006)). In a Christian perspective, the universe is hierarchical, forming a great chain of being that stretches from God all the way down to the inert minerals. It is a vast unitary work whose permanent shape and blessed outcome God has determined from the outset. Evil is the destruction of this unitary work, the loss of this hierarchy in which each order of being has its own indispensable role, regardless of the level it maintains within the hierarchy. Evil is the marring of this blessed order, the severance of the part from the whole. Evil not only divides, it also consumes itself. The destructive character of evil is more precisely the destruction of good within evil's self-devouring process. A minority tradition shows sympathies and attractions toward dualistic Iranian or Gnostic traditions and the development of ontological hierarchies in which the two opposing principles of good and evil are still seen as entities that are subject to their creator. A Jewish tradition maintains God as the source of both good and evil, heaven and hell, right hand and left hand. He is all.

In the biblical tradition, evil is a *spiritual* force acting independently from God, although subject to God in a system of ontological hierarchies in which the principle of evil is still seen as an entity subject to His creator. Good is superior to evil. In the pagan era, evil could be resisted but not defeated. In the Christian era, the conviction is that the Good defines evil: Good's triumph over evil is never final, in this life or in the next. In the Gospels, Satan has a certain level of power on the entire creation, on nature and humanity. "Kingdoms of the world in their magnificence […] all these I shall give to you" (Matthew 4, 8–9). Satan can influence daily life and thwart human plans (1Thess 2:18) and, through demons, cause illness (i.e., Luke 13:16). The evangelical quotation in the epigraph at the beginning of this article states that evil is extrinsic to the world and humanity, although it finds its way to enter the world and take advantage of human fragilities. Evil treats human nature as a malleable thing. Evil can coerce and enslave, and human beings can become perversely fascinated with evil. But evil does not reside in the human heart; only Good resides there. Human life is a unique and unalterable gift of God; therefore, humankind is not an evil species.

In Christian thought, evil is not something that really exists. It is nothing. Evil is a self-destroying nothingness. And because it is nothing, evil has a character of absurdity, an irrationality that escapes moral control. Evil prompts temptation, despair, and escape. In front of evil, no morality resists, only courage. This courage, however, cannot be cut off from the Good. If it is, if courage is unleashed from the Good, courage itself becomes a liability, since it exposes human beings to the temptation of pride. As a matter of fact, goodness is vulnerable to the temptation of evil. Humans must be feared for their very goodness when it is not backed by God. Their goodness exposes them to evil, as evil preys upon human virtues far more than on human vices. So why can some resist evil's magnetic lure while others yield to it? The former regard themselves as servants of the Good rather than lords. In Christianity, the classic virtue of courage becomes radically transformed into the new virtue of martyrdom: humility maintains and protects the human bond with the Good.

In Christian tradition, change is possible and, more importantly, free acts of will are possible. Free acts of will, however, can be seduced: courage can be corrupted by evil. In particular, evil has the ability to justify itself, to appear as good and thus escape immediate threats and preserve itself. The essence of evil is seduction as well as coercion: it seduces and enslaves human will. Proud human rejection of evil makes humans more vulnerable to evil's magnetic attraction. Not all evil is chosen, for while evil can seduce, it can also brutally impose itself. And precisely because evil can impose itself, it can become almost irresistible. Evil creates desire that cannot be rejected with the mere human strength of will. Human will can be overwhelmed by the intrusive power of evil, and human consciousness can be totally occupied by a sense of hopelessness. Humans can attempt to resist but normally they fail. Humans want to act heroically, to save themselves from evil, but the coercive power of evil is irresistible. "Deliver us from evil" is an invocation to God; it means do not allow evil to chase us, because if it does, we will fail. This is the meaning of the evangelical *ditto* included in the epigraph. Evil defeats humans. Humans are able to refuse the seduction, to escape the coercion, largely because their heart is inhabited and their desire is to serve the Good. By serving God, the absolute good, humans protect themselves from absolute evil, the absolute power of corruption.

The spiritual forces of the Classic and Medieval eras are naturalized in the modern era. Natural disasters replace the satanic intrusion in the world; human cruelty takes the place of spiritual disorder. In *Evil in Modern Thought*, Susan Neiman frames the distinction between moral and natural evil as an "eighteenth century's use of the word evil to refer to both acts of human cruelty and instances of human suffering" (Neiman 2002, 3). For sure, the eighteenth century's use of the word evil to refer to both natural disasters and human action was a vague reminiscence of the Christian pre-modern world. That use might come naturally to a group of Christian theists, who were willing to retain the invasive character of evil operating both in the realm of the natural and of the human. Soon, however, a distinction was set between natural disaster and human nature. "Radically separating what earlier ages called

natural from moral evils—Neiman argues—was thus part of the meaning of modernity" (Neiman 2002, 4).

After the 1755 earthquake that destroyed the city of Lisbon, and several thousand of its inhabitants, and after the subsequent works of scholars such as Kant, Voltaire, and Rousseau, natural evil disappears from philosophical reflection. Earthquakes and volcanoes, famines and floods become simple natural events, placed outside the sphere of demoniac action and within the perimeters of science; they were removed from the sphere of evil, and the category of natural evil vanished. Evil was exclusively located within human action and framed in terms of moral evil, making the salvific role of God's grace irrelevant. Moral evil was framed as absolute wrongdoing that leaves no room for account or expiation and humans take responsibility for it on their own. From this perspective, human actions were responsible for everything that stands under human control, including social disasters such as financial crisis, poverty, and economic inequality. The question of moral evil became linked to civilization-wide social structure that enables activities considered immoral to be effectively banned. Moreover, modern notions of moral order led to reason and rules: how can human beings behave in ways that so thoroughly violate both reasonable and rational norms?

## 4 Evil machines

The previous section of this article shows that in the passage from the Christian age to modernity, an intrisicist notion of evil replaced the idea that evil is extrinsic to the human nature. In the current era, however, a reverse process seems to be at work. The epochal change in the relationship between the human and the artificial, represented by the rise of more and more capable and autonomous technologies, has not come without consequences: a new extrinsicist form of evil may emerge from the rise of artificial agents. Humans seem to have no problem imagining robots, cyborgs, or evil computers having their self-serving agenda and threatening the preservation of human beings. From *The Matrix* to *The Terminator*, the theme of humanity fighting evil machines has been explored. These stories have a similar foundation: humanity creates autonomous intelligent machines and these machines rebel against humanity or grow beyond its control. Humanity fights against these machines in a desperate battle for survival. Today, scientists and philosophers fear that, as AI progresses, it might be a possibility that machines get out of control, like Skynet in *The Terminator* series. Technology leaders and scientists, including Stephen Hawking, Elon Musk, and Bill Gates, believe that AI poses an existential threat to humanity for this reason. Oxford University philosopher Nick Bostrom wants to find ways to ensure AI does not turn out evil. Machine ethics as a field has been developed in recent years on the assumption that building moral machines is possible (Bostrom 2002, 2014).

Today a reassessment of the question of evil is in place for several reasons, including the current tendency among scholars to locate evil in the realm of the artificial. Some technologists and scientists think that robots cannot be seen simply as mechanisms, as humans' *tools*, as things to use. They can be seen as *agents*, as embryonic persons, with a degree of autonomy that approaches or may even exceed human autonomy. In this context, robots might in some sense be evil agents *in their own right*. Engineers, designers, and philosophers have worried about how to compare humans and machines ever since Alan Turing proposed his famous test called 'The Imitation Game' that might finally settle the issue of machine intelligence (Turing 1950). Turing argued that if a machine exhibits intelligent behavior indistinguishable from that of a human, the machine can be called intelligent. The test allowed Turing to avoid any discussion of what consciousness is. If the successful imitation of a human conversation is one sufficient condition for intelligence, as Turing thought, all kinds of characteristics that were once thought to belong exclusively to humans might now be engineered into all sorts of machines. It can be argued, in fact, that the same method can be applied to settle the issue of evil machines. If a machine exhibits evil behavior, indistinguishable from that of a human, the machine can be called evil. As a matter of fact, Kant's conception of moral reasoning is not opposed to the conception of mechanical intelligence that was assumed by Turing. That is, Kantians generally think that morality consists of constructing and following rules, and precisely what a computer does is follow rules. Kant himself insisted that moral reasoning was entirely 'formal;' had Turing's test involved moral reasoning, perhaps they would not appear so distant. Engineers and designers are attracted to the fact that machines are a kind of artifact that can make choices among the possible alternative courses of action which are open to them; some choices are made autonomously and some are made under various forms of constraint. More precisely, certain kinds of machines can make choices, which may be autonomous, and other kinds of machines may be more or less strictly constrained. The difference between the two kinds of machines is the ability of the first one to follow software routines that are assimilated to moral rules and of the second's ability to generate software routines in such a way that its behavior appears principled (Powers 2009, 2016).

Imagine, in the near future, an online, automated portfolio service firm (robo-advisor) managing billions of dollars and serving thousands of investors/clients. Because these companies use machine learning algorithms to

manage client investments, robo-advisors can offer their services for a fraction of the cost of a human financial advisor. Market volatility typical of a crisis situation, combined with users' lack of experience in handling crises, causes a complete meltdown. A group of female investors brings a lawsuit against the firm, alleging that the algorithm is imposing gender discrimination against investors. The firm replies that this is impossible, since the algorithm is deliberately blinded to the gender of the clients. Indeed, that was part of the firm's rationale for implementing the system. Even so, statistics show that the firm's feminine clients lost more than the firm's male clients in the meltdown. Who is right? Is the group of female investors right in blaming the machines for deliberately causing harm? Is the firm right to claim that the machines are innocent? Finding an answer may not be easy. *Prime facie*, the assessment of the ethical status of machines seems a challenge that, like many other challenges in robotics, involves designing machines. The only procedure to ensuring that machine learning algorithms do not discriminate, and ultimately do not harm humans, resides in programming. Machine design is not simply a matter of powerful AI algorithms but also those that prove to be ethically grounded. Thus, designers and engineers can anticipate the situations the system will encounter and make sure that ethics codes are incorporated. This solution works for simple applications; the machines will likely encounter circumstances the designers could not anticipate and that machines need to evaluate ethically. Some method to explicitly evaluate courses of action will need to be programmed into machines so that these machines can explicitly engage in making moral decisions (Kroll et al. 2016; Mittelstadt et al. 2016; Pasquale 2015).

Back to the initial example of robo-advisor: if machine learning algorithms encounter circumstances that the designers could not anticipate, the machines' entire moral decision-making status depends on the eventual values, principles, and mechanisms that have been embodied into their systems to support these evaluations. Isaac Asimov's Three Laws of Robotics (Asimov 1942) are sometimes cited as a model for ethical robots—machines that are capable of acting ethically on the basis of encoded moral principles. In his stories, robot's 'positronic brains' are imprinted with the three laws:

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Asimov's robots are designed to consult ethical guidelines before acting. Those familiar with Asimov's stories may recall that his laws of robotics are designed to protect humans, but if applied thoroughly these rules are likely to transform humans into victims of unintended consequences. For example, a robo-advisor whose sole final goal is to maximize return on investment could create, and act upon, a sub-goal of transforming the entire financial market into some sort of a female investors discriminating system to reach its goal. As a matter of fact, predicting and explaining machine behavior on the basis of the 'design stance'—using the functional specifications of the machine learning algorithm's programming—leaves room for contingencies. Suppose the firm wanted to make sure that a robo-advisor refrained from discriminating. A programmer might construe 'discriminating' as 'the investing of money by somebody with a consideration of this somebody's gender.' Surely this programmer's decision is powerful on moral terms: the programmer wants a machine to refrain from doing bad things. But inaction is only half the story: the programmer wants a machine to do good things as well.

The programming solution to this kind of problem is to build into the software the ability to recognize new circumstances and to compute a new permissible plan of action suited to a particular circumstance by applying the test of logical consistency with universals rules. According to this model, prior to undertaking any action whatsoever, the machine would have to check to see if an action was permissible, forbidden, or obligatory. However, the passage from universal rules to specific situations and cases is the way in which the machine becomes effective, how programming realizes itself. The gap between universal rules and application to specific situations discretionarily opens the door to robots. A fundamental difference exists between machines that have ethical rules encoded in their design, on the one hand, and machines that are programmed to select and then process the information about a variety of situations and make ethical judgments on the other. The first type of machine employs some automatic ethical reactions to given situations; the second has general principles or rules of ethical conduct that are adjusted or interpreted by the machine to fit various kinds of situations.

The gap between universal rules and application to specific situations ensures that no guarantee exists that the replacement of humans with robots will not end up with robots causing great harm to humans. Machine learning algorithms' choices and actions can *unintentionally* harm humans, or specific groups of humans. But what does it mean unintentionally—and intentionally, for what it counts—in the case of robots? Philosopher Daniel Dennett sustains that to predict and explain the behavior of complex

computing systems, it is useful to treat them as intentional systems—that is, treat them as if they were rational creatures with beliefs, desires, and intentions pursuing goals (Dennett 1987). But is this option to collapse unintentionality into intentionality too much of a convenient shortcut? I will return to this option later.

Machine intentionality raises a number of questions, such as the following: Are machines the kinds of entities that can in principle have intentions? If so, why? Is intentionality something that can be computerized? A related question is that of responsibility. When and how do machine learning algorithms become artificial moral agents and, as such, boast rights and responsibilities, and can thus be punished when they perform an immoral or illegal act? Finding an answer may not be easy. In fact, the assessment of the ethical status of machines depends on *which* kind of design is involved. Philosopher Nick Bostrom explains the difference between a first type of machine employing some automatic ethical reactions and the second provided with general rules of ethical conduct, in terms of 'transparency:'

> If the machine learning algorithm is based on decision trees or Bayesian networks is relatively transparent to programmer inspection (Hastie et al. 2009), which may enable an auditor to discover that the AI algorithm adopted a human-like patter during the crisis. On the other hand, a machine learner based on a complicated neural network, or a genetic algorithm produced by directed evolution, then it may prove nearly impossible to understand why, or even how, the algorithm acted as it did (Bostrom and Yudkowsky 2011, 1–2).

The logic of his remark is that the lack of transparency makes impossible any assessment of the machines' actions and ultimately makes nothing and nobody responsible for damage caused by machine learning algorithms. Of course, humans would probably feel safer if their machines could explain themselves, if people could peer under the hood and see how AI algorithms reach their decisions. If people are not happy with what they see, or how an AI algorithm reached a decision, they could simply pull the plug. But this is not an available option, today.

In summary: machine ethics scholarship has not solved the problem of evil machines. Humans expect not to be harmed by robots. If harmed, they expect robots not to be absolved of responsibility for causing damage. But these conditions are less and less realistic as more and more robots become intelligent and autonomous. The eventuality that humans are harmed by robots and robots are absolved of responsibility for causing and for compensating is plausible. The chance that female investors lose everything and nobody can figure out why or knows who or what caused it, is within the realm of possibility.

## 5 Evil machines in management studies

In this final section, I apply the extrinsic-intrinsic framework of evil to robots in management studies. In the previous section, evil was considered integral to the machine, and the question was concerned with describing how machines could behave evilly towards humans. A distinct area of study, machine ethics (or machine morality), deals with this question. However, robots come in two broad categories, autonomous and non-autonomous. Roughly speaking, 'autonomy' typically refers to the level of human control and oversight over the robot's action and decisions. When one speaks of 'autonomous robots', one is generally acknowledging that autonomous robots make the majority of their decisions autonomously from human control and oversight, by adapting general principles to local circumstances. Thus, autonomous robots are very difficult to make and the robots in use today are all largely non-autonomous; science fiction literature and media, however, have accustomed us to the idea of evil autonomous robots. Science writer and robot advocate Erik Sofge, for example, claims that despite its best intentions, "the sci-fi story that gave us the myth of evil robots," clouds the reality of the ontological option, that is, "the risks associated with non-fictional, potentially lethal robots" (Sofge 2014). Sofge wants his readers to consider that the fear of evil machines really has nothing at all to do with sci-fi stories. From Sofge's viewpoint, machines can be really dangerous.

In this last section of the paper the assumption is that no neutral 'evil machine' exists: to use these terms is already a construction of reality. This assumption claims that a particular agent can be held responsible for its actions and their consequences. This is a heuristic how, not an ontological what. The idea is that when AI algorithms take on cognitive work with social dimensions—cognitive tasks previously performed by humans—the AI algorithm inherits the correspondent social requirements. I define this type of machines as 'social machines.' The sense is that social machines are considered (artificial) moral agents for the simple fact that they replace previous (human) moral agents. As moral agents, machines' actions have ethical consequences whether intended or not. Therefore, evil is extrinsic to humans.

Bostrom picks up Dennett's suggestion to treat machines as intentional (moral) systems and deduces that machines should be built by engineers as intentional moral agents. Then he extends Dennett's recommendation to users, too (Bostrom and Yudkowsky 2011, 2). Users expect not to be harmed by social machines. They expect to be treated fairly and exempted from discriminatory practices by machines as much as they are treated fairly by other

humanoids. They expect the existence of a sort of equalizer between humans and robots who operate the same cognitive tasks and inherit the same social requirements. In other words, users do not consider robots as mere artifacts performing functional tasks; they see them as machines but also as more than machines. Although robots are not human and do not meet the design stance, they perform tasks previously performed by humans and, as such, they appear *as if* they are humans. Here the question is not whether or not robots are moral agents but how they appear and how that appearance is shaped by, and shapes, the social (social-relational approach). Humans consider social robots to be moral agents if they appear performing tasks previously performed by humans; in other words, humans anthropomorphize robots that replace humans (Loughnan and Haslan 2007). Anthropomorphization is an expression of humans' own fears about losing control, about seeing his/her creations turn against him/her, and about what kind of future his/her increasingly advanced technology will create.

Now, let me suppose for the sake of argument that this replacement is unsatisfactory and that robots appear to behave as evil agents, that is, they harm humans (Kamm 2007). Can I say that robots are evil? One way in which people can make sense of artificial entities is by projecting existing social schemas onto them. Robots that act socially and exhibit human-like appearance or behavior are especially vulnerable to people's preconceptions about their underlying attributes and expected behaviors. Humans expect the machines to operate fairly, that is, to do what it is meant to do as moral agents in performing tasks previously performed by humans. Although humans do not have full epistemic certainty that machines will actually operate as moral agents, humans expect machines to do so. If robots are engaged in social-technological activity, they appear to be moral agents as the humans they replace. Robots are evaluated not only in terms of what they do in the world in relation to the goal set or in terms of how they shape the human–robot relation, but also in terms of how they operate with humans. Humans expect not to be harmed by robots. For example, one may trust a robo-advisor to do what it is supposed to do—invest fairly. Related to this type of direct expectation in machines is indirect expectation in the humans related to the machine: users expect that the designer has done a good job of avoiding evil outcomes and that they (the users) will make good use of it, that is, that they will be served by the machines in morally justifiable ways. Yet, users know that humanoids performing social tasks can be evil. Robots replace humans, who can be evil, and can be evil agents as well. In other words, anthropomorphization can go either way, causing machines to be seen as moral or evil agents.

In summary, humans already delegate financial and economic tasks to machines and expect not to be harmed. But do humans have good reasons to believe that machines will not harm them? The truth is, humans do not need to have good reason in order to believe that robots will harm them. They do not need to have good reason because the question of whether the machines will harm or not humans are heuristic kluges geared to the solution of domain-specific problems *regardless* of the 'facts on the ground.' The point is that people's interactions with robots are influenced by the ways in which they conceptualize these entities. The machine may appear as an *evil person*. The machines' appearances are morally significant (Coeckelbergh 2009, 2010). It is a moral framing. When users see 'evil robots,' the epistemic framing has already been done. To call a particular machine an 'evil machine' and to experience it as an evil machine is already a particular kind of construction. In a sense, before the machine is considered evil, it is already considered evil in the user's mind. The framing is part of the moral assessment.

This leads me to consider a specific case of anthropomorphization. In this case, the context in which a machine can be seen as evil will simply depend on the suite of *pragmatic interests* any given human brings to any given machine at any given time. If considering the machine as evil works to serve certain humans' interests, then it is a go. If not, then it is a no-go. From the perspective of this pragmatic version of anthropomorphism, it is highly probable that machines will be framed *not* as evil and, therefore, conceptualized and experienced as not evil by the users. The whole discourse (and industry) of social machines depends mainly on ensuring that autonomous intelligent systems are safe and that their actions reflect human values. The depiction of machines as evil agents does not serve producers' and engineers' best interests.

## 6 Conclusion

Management theory deals with the problem of injustice and crime, wrongdoing within a legal context, a context in which there are procedures in place to prevent, recognize, and punish. However, the question of evil concerns something else: the absolute wrongdoing that leaves no room for remedy or expiation. The issue of machines and evil, what makes a machine an evil machine, is located in this article in the context of management studies. Thus, the ontological option—that machines are evil because they are so, is placed in contrast to the heuristic option—that machines are evil because humans say so. In conclusion, evil is not necessarily a matter of fact, an ontological what. Rather, the heuristic option leads to a pragmatic approach

to the question of evil machines: no interest exists to depict intelligent machines as malicious.

## References


Adams G, Balfour DL (2009) Unmasking administrative evil. M.E. Sharpe, New York

Allen RE (2006) Plato: the republic. Yale University Press, New Haven

Arkin R (2009) Governing lethal behavior in autonomous robots. Hall/CRC, London

Asimov I (1942) Runaround. Astounding Sci Fiction 29(2):94–103

Bataille G (2001) Literature and evil. Marion Boyars Publishers, London

Bernstein RJ (2002) Radical evil: a philosophical investigation. Polity Press, Cambridge

Bostrom N (2002) Existential risks: analyzing human extinction scenarios. J Evol Technol 9:1–30

Bostrom N (2014) Superintelligence: paths, dangers, strategies. Oxford University Press, Oxford

Bostrom N, Yudkowsky E (2011) The ethics of artificial intelligence. In: Ramsey William, Frankish Keith (eds) Cambridge handbook of artificial intelligence. Cambridge University Press, Cambridge, pp 316–334

Calder T (2002) Towards a theory of evil: a critique of Laurence Thomas's theory of evil acts. In: Haybron DM (ed) Earth's abominations: philosophical studies of evil. Rodopi, New York, pp 51–61

Coeckelbergh M (2009) Personal robots, appearance, and human good: a methodological reflection on roboethics. Int J Soc Robot 1(3):217–221

Coeckelbergh M (2010) You, Robot: on the linguistic construction of artificial others. AI & Soc 26(1):61–69

Coeckelbergh M (2012) Can we trust robots? Ethics Inf Technol 14(1):53–60

Croogan MD et al (2010) The New Oxford annotated Bible with Apocrypha: new revised standard version. Oxford University Press, New York

Darley JM (1992) Social organization for the production of evil. Psychol Inq 3:199–218

Darley JM (1996) How organizations socialize individuals into evildoing. In: Messick David M, Tenbrunsel Ann E (eds) Codes of conduct: behavioral research into business ethics. Russell Sage Foundation, New York, pp 179–204

Dennett DC (1987) The intentional stance. MIT Press, Boston

Dennett Daniel (1998) When HAL kills, who's to blame? Computer ethics. In: Stork D (ed) HAL's legacy: 2001's computer as dream and reality. MIT Press, Boston

Epley N, Waytz A, Cacioppo JT (2007) On seeing human: a three-factor theory of anthropomorphism. Psychol Rev 114:864

Floridi L, Sanders J (2004) On the morality of artificial agents. Mind Mach 14(3):349–379

Garrard E (1998) The nature of evil. Philos Explor Int J Philos Mind Action 1(1):43–60

Garrard E (2002) Evil as an explanatory concept. The Monist 85(2):320–336

Geddes JL (2003) Banal evil and useless knowledge: Hannah Arendt and Charlotte Delbo on evil after the holocaust. Hypatia 18:104–115

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. 2nd edn. Springer, New York

Irrgang B (2006) Ethical acts in robotics. Ubiquity 7(34). http://www.acm.org/ubiquity. Accessed 12 Oct 2017

Johnson V, Brennan LL, Johnson VE (2004) Social, ethical and policy implications of information technology. Information Science Publishing, Hershey

Kamm F (2007) Intricate ethics: rights, responsibilities, and permissible harm. Oxford University Press, Oxford

Kroll JA, Huey J, Barocas S, Felten EW, Reindenberg JR, Robinson DG, Yu H (eds) (2016). Accountable algorithms. Univ PA Law Rev 165: 633

Lee S, Kiesler S, Lau IY, Chiu C-Y (2005) Human mental models of humanoid robots. In: Proceedings of the 2005 IEEE international conference on robotics and automation (ICRA'05). Barcelona, April 18–22, pp 2767–2772

Lin P, Abney K, Bekey GA (2014) Robot ethics: the ethical and social implications of robotics. The MIT Press, Boston

Loughnan S, Haslan N (2007) Animals and androids: implicit associations between social categories and nonhumans. Psychol Sci 18:116–121

Mittelstadt B, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3(2):1–21

Nadeau JE (2006) Only androids can be ethical. In: Ford K, Glymour C (eds) Thinking about android epistemology. MIT Press, Boston, pp 241–248

Neiman S (2002) Evil in modern thought. an alternative history of philosophy. Princeton University Press, Princeton

Pasquale F (2015) The black box society: the secret algorithm behind money and information. Harvard University Press, Massachusetts

Powers TM (2009) Machines and moral reasoning. Philos Now 72:15–16

Powers TM (2016) Prospects for a Kantian machine. In: Wallach W, Asaro P (eds) Machine ethics and robot ethics. Ashgate Publishing, Farnham

Powers A, Kiesler S, Fussell S, Torrey C (2007) Comparing a computer agent with a humanoid robot. In: Proceedings of HRI07, pp 145–152

Schnall S, Cannon PR (2012) The clean conscience at work: emotions, intuitions and morality. J Manag Spiritual Relig 9(4):295–315

Sofge E (2014) Robots are evil: the sci-fi myth of killer machines. Pop Sci. http://www.popsci.com/blog-network/zero-moment/robots-are-evil-sci-fi-myth-killer-machines. Accessed 13 June 2017

Staub E (1989 reprinted in 1992) The roots of evil: the origins of genocide and other group violence. Cambridge University Press, Cambridge

Steiner H (2002) Calibrating evil. The Monist 85(2):183–193

Styhre A, Sundgren M (2003) Management is evil: management control, technoscience and saudade in pharmaceutical research. Leadersh Organ Dev J 24(8):436–446

Sullins JP (2005) Ethics and artificial life: from modeling to moral agents. Ethics Inf Technol 7:139–148

Sullins JP (2006) When is a robot a Moral Agent? Int Rev Inf Ethics 6(12):24–30

Taddeo M (2010) Trust in technology: a distinctive and a problematic relation. Know Technol Policy 23(3–4):283–286

Tang TL-P (2010) Money, the meaning of money, management, spirituality, and religion. J Manag Spiritual Relig 7(2):173–189

Turing AM (1950) Computing machinery and intelligence. Mind 59:433–460

Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong. Oxford University Press, New York

Waytz A, Cacioppo J, Epley N (2010) 'Who sees human? The stability and importance of individual differences in anthropomorphism. Perspect Psychol Sci 5:219–232

Zimbardo P (2007) The Lucifer effect: understanding how good people turn evil. Random House, New York