Scientia*et***Fides** 9(1)/2021 ISSN 2300-7648 (print) / ISSN 2353-5636 (online)

Received: February 2, 2020. Accepted: May 22, 2020 DOI: http://dx.doi.org/10.12775/SetF.2021.010

The Government of Evil Machines: an Application of Romano Guardini's Thought on Technology

ENRICO BELTRAMINI
Notre Dame de Namur University
United States
ebeltramini@ndnu.edu
ORCID: 0000-0001-9704-3960

Abstract. In this article I propose a theological reflection on the philosophical assumptions behind the idea that intelligent machine can be governed through ethical protocols, which may apply either to the people who develop the machines or to the machines themselves, or both. This idea is particularly relevant in the case of machines' extreme wrongdoing, a wrongdoing that becomes an existential risk for humankind. I call this extreme wrong-doing, 'evil.' Thus, this article is a theological account on the philosophical assumptions behind the possibility of evil machines, machines that can create an existential risk for humankind, and the ethical remedies that limit that risk. After discussing these assumptions, I argue for the ineffectiveness of these ethical remedies to prevent the possibility of evil machines. The article is an application of Romano Guardini's thought on technology to evil machines.

Keywords: evil; technology; AI; ethics.

Introduction

One of the most famous and persistent myths of our time is the intelligent machine turning evil. In this article I define ‘intelligent machine’ in terms of sentient machines with general artificial intelligence (‘AI’) and ‘evil’ as extreme wrongdoing, a wrongdoing that becomes an existential risk for humankind. Thus, evil machines are machines that can create an existential risk for humankind. The possibility of evil machines (defined as above) has already been considered in the current scholarship. In his renowned book *Life 3.0. Being Human in the Age of Artificial Intelligence*, for example, MIT physicist Max Tegmark lists twelve scenarios regarding the lasting trajectory of Artificial Intelligence (AI). At least three of these scenarios assume a form of AI turned evil (to be fair, Tegmark rejects the idea of evil machine and prefer to consider the case of ‘misaligned intelligence,’ i.e., intelligence with goals misaligned with human goals). In ‘conquerors,’ AI takes control and gets rid of humankind. In ‘descendants,’ the result is the same, but the end is more graciously delivered. In ‘zookeepers,’ humans survive only to live in a cage (Tegmark 2017, 162–163). At the same time, scholars and practitioners are confident that the possibility of evil machines is avoidable through a strategy of containment, ethical limits that operate as unsurmountable deterrents to would-be evil-doers. These ethical limits should work on people (the designers and developers of machine) as well as on machines. The idea is that some ethical limits prevent intelligent machines from ignoring certain software modules that instruct them to maintain the norms of behavior that have been programmed into them. These ethical limits also prevent developers from allowing intelligent machines to ignore those software modules. In sum, scholarship addresses the problem of evil machines in terms of ethical remedies, that is, protocols that guide people dealing with the machines as well as the machines themselves and work as a rationale that restrains their use and guide the appropriate exercise of such power.

In this article I study the idea that ethics governs intelligent machines. I look at this idea theologically through the lens of the work of German theologian Romano Guardini, including *The End of the Modern World* (1998) and *Letters from Lake Como: Explorations in Technology and the Human Race* (1994). While his writings were originally published more than half a century ago, the relevance of Guardini's work on technology in the development of Catholic thought has been recently confirmed by Pope Francis (2015). Once placed into Guardini's framework, this idea that ethics governs intelligent machines reveals itself to be based on an assumption, that is, *ethos* is subordinated to *logos*. *Ethos* and *logos* are philosophical categories; they stand respectively for the embodiment of ideas in social practice (*ethos*) and the rationale behind practice (*logos*). Synonyms of *ethos* are action, practice, will, and power; synonyms of *logos* are theory, knowledge, logic, and reason. Guardini discussed the relationship between *logos* and *ethos* within a much larger theme, i.e., the transformation of culture from modernity to late (or after-) modernity (Millare 2013) In a nutshell, Guardini sees after-modernity as absolute modernity, in which technological action is a dominant force with no limits or counterparts. As a matter of fact, Guardini did not coin the expression 'absolute modernity' nor ever adopt it. Rather, it was a Swiss theologian, Hans Urs von Balthasar, who applied the sentence to summarize Guardini's specific view of after-modernity (2010, 105). Methodologically speaking, this study is a loose application of Guardini's claims about 'absolute modernity' to the problem of evil technologies.

In this article I dispute theologically the prevention of the possibility of evil machine through ethical limits, regardless of whether the latter refers to people or machines. Put simply, I argue for the ineffectiveness of a containment strategy to prevent the risk of evil technologies, defending this position by applying a specific understanding of technological culture. In this specific understanding, technological culture is one of ubiquitous technological power. In such a culture, is it possible to seriously consider a self-imposed limit, an ethical boundary, an unsurmountable principle as effective? I believe the correct answer is no. In such a reality, in fact, ethical limits raised to prevent the risk and to deflect the impact of evil technologies are ineffective; therefore, technological search for power operates unchallenged. I elaborate on this contribution and claim that in a situation of saturated technological culture, technological forces are free to be evil. To put if differently, I frame the possibility of evil as internal to the same forces that are establishing technology as a dominant factor in Western culture, so that no limit ever reveals itself as insurmountable. This is my argument.

Before proceeding further, I should define the scope of my article and clarify the terms used; I will also make explicit the limitations of my argument and the method adopted. In this article I propose a preliminary investigation on the condition of evil in a technological era. In this study, evil is defined as a force that causes human suffering and leaves no room for understanding or redemption. An evil technology, therefore, is a force that causes human suffering beyond understanding or redemption. As 'condition of evil' I mean the possibility to produce an action that is not simply morally wrong, but one that leaves no room for understanding or for redemption. I place this possibility neither in the machine nor the human in front or behind the machine, rather in the technological culture in which machines and humans operate. I define technology as a mathematically and logically driven form of *techne*, a term that stands for 'practice' or 'activity.' In this article, I consider technology in the same terms of philosopher Emanuele Severino, who, in the tradition of Heidegger, addresses technology as will of power (2016a and 2016b). Moreover, 'evil technologies' (or 'evil machines' synonymously) are technologies turned evil. 'Evil machines' stand for intelligent machines that possess the ability to make choices and therefore come with the capacity for evil. More precisely, these are machines that are free to either follow or ignore some software modules that instruct them to maintain the norms of behavior that have been programmed into them. 'Ethics' is defined in Kantian terms, that is, in terms of application of rules and normative theories, or determination of right or wrong (i.e., good and evil), while moral imperative is about doing what is right. It is in the realm of this definition of ethics that my argument stands. In this study, I neither assert that the notion of evil in a technological culture is *morally* significant, nor do I elaborate a theory of evil in a technological culture. I simply argue that the notion of evil in a technological culture is *practically* significant, with genuine theoretical interest. I do not deal with the question of the 'banality of evil,' and therefore I do not consider the question of whether an action should be considered evil only if it is intentional, or, more precisely, malicious, sadistic, or defiant. Arendt (1963), Card (2002), and Formosa (2008) think that evil actions can be banal, in that evil actions do not require evil actors. Singer (2004), Kekes (2005), and Steiner (2002) contradict Arendt, and argue that an action can be evil only if it is intentional.

This is a three-part paper. In the first part, I draft the problem I aim to address in this essay. In the second, I explain Guardini's thought on technology. In the third, I apply Guardini's framework on technology to the problem of evil technologies.

1. Problem

Broadly speaking, the recent debate over the effects of AI has been dominated by three themes. One is the threat that AI will allow governments to monitor, understand, and control their citizens far more closely than ever before. Another theme is based on the worry that a new industrial revolution will allow machines to disrupt and replace humans in every—or almost every—area of society, from transportation to the military to healthcare. The third theme refers to the way in which AI promises to reshape the world: by exceeding human intelligence and escapes human control, with possibly disastrous consequences. This is the theme of the technological singularity, the notion – introduced by scholar and science fiction author Vernor Vinge in his 1993 essay *The Coming Technological Singularity* – that an upgradable intelligent technology would enter a 'runaway reaction' of self-improvement cycles, so that the intelligent technology would continue to upgrade itself and advance technologically at an incomprehensible rate with no control on human side.

A less famous but still relevant variant on this theme is the so-called 'unfriendly AI,' according to which the problem with the singularity is not the impossibility of human control over technology, but rather the lack of shared goals between intelligent technology and humanity. In this perspective, the problem is not the increasing power of AI, but how this power is used. Tegmark is not only a physicist but also a co-founder of the Future of Life Institute (FLI). Among other reasons, FLI was created precisely to help humanity pursuing a ‘friendly AI,’ that is, an AI whose goals are aligned with humanity’s goals. In Tegmark’s view, the task for researchers, engineers, and programmers is making AI learn, adopt, and retain humanity’s goals (2017, 260). In this context, the way “to imbute a superintelligent AI with an ultimate goal that neither is undefined nor leads to the elimination of humanity,” Tegmark explains, passes through the development of an adequate forms of ethics (2017, 249–279 and 280). While Tegmark recognizes that an ethics for technological singularity is still to come, two forms of ethics have been developed to limit the effects of a pre-singularity AI self-empowerment. These two forms of ethics are: 1. machine ethics (which is concerned with the moral behavior of artificial intelligence beings); and 2. roboethics (which is concerned with the moral behavior of humans as they design, construct, and use such beings).

Scholarship on machine ethics is expanding (Wallach and Allen 2008; Lin and Abney 2011; Lin, Jenkins and Abney 2017). Recent literature in machine ethics has addressed the so-called ‘machine question,’ that is, whether a machine might be considered a legitimate moral agent that could be held responsible for decisions and actions (Gunkel 2012). At stake in this debate on machine morality is the acceptance of the machine as a moral agent, with the notion that future machines might be conscious and should be included in the moral order. On the side of robotics, the question surrounds that of ethical protocols embedded in the work of programmers and their software as the condition of friendly, non-evil technologies. To put it differently, roboethics assumes that the definition of right and wrong belongs to the people behind the machines, and machine ethics assumes that such a definition belongs to the machines themselves. Recent work in the field of robotics focuses on important ethical issues, including privacy concerns, responsibility and the delegation of decision making, transparency, bias as it arises at all stages of data science processes, and ethical practices that embed values in design and translate democratic values into practices

(Coeckelbergh 2020). Practitioners and researchers offer ways to transform modern ethical theory into practical guidelines for designers, product managers, and software engineers alike (Boddington 2017; Bowles 2018). In sum, most experts maintain the opinion that making AI safe for a specific purpose likely will be solved. Some scholars and scientists, however, show optimism for humans' future with general AI, too. General AI refers to an algorithm or set of algorithms that can perform all tasks as well as or better than humans. Narrow AI is artificial intelligence that is focused on one narrow task. With changes of small or large magnitude, those experts believe that controlling ethics through AI is going to be possible, and that no unrestricted general AI will ever be released. Other experts are more concerned, however. Less optimistic contributions come from scholars involved in the 'control problem,' a question posed by philosopher Nick Bostrom on how to limit advanced artificial intelligence while still benefiting from its use (Bostrom 2014, 127–144). Computer theorist Stuart Russell believes that AI, as it is developed currently, is on the path to becoming a mortal threat to humanity. His solution is to change the way AI is developed. Russell suggests that we can rebuild AI on a new foundation according to which machines are designed to be altruistic and committed to pursuing our objectives, not theirs. This new foundation, in Russell's view, would allow us to create machines that are provably deferential and provably beneficial as machines (Russell 2019). As it is currently built, therefore, AI is a threat for humanity. Bostrom and Russell believe that if AI surpasses humanity in general intelligence to become superintelligent, then humanity will be at the mercy of that superintelligence goodwill.

The problem in this article is neither that an adequate ethics for technological singularity has not developed yet, or whether such an adequate ethics may be developed, but rather whether any form of ethics is capable to limit, govern, or control intelligent technology in the long run. To put it differently, the problem addressed in this article is whether ethics is an effective condition to 'imbute,' i.e., govern, control, limit, a superintelligent AI. A situation where a superintelligence AI is free to pursue its goals without limits and constrains, in fact, is a necessary condition that make evil possible. The argument of this article is that any strategy that operates on the assumption that ethics governs general AI (before and especially after reaching the stage of superintelligence) is going to fail. Programmers will secure AI and make it safe, but no one can prevent someone else from modifying it so that those safeguards are altered. By 'someone else' I mean (1) humans who use AI against other humans on a massive scale, and/or (2) AI which subverts the human's control and gets free. Once the first unrestricted general AI is released, there will be no effective means of stopping its distribution and use. In dealing with the 'singularity,' the event in which a machine superintelligence (or simply 'superintelligence') exceeds human intelligence and escapes human control, several scientists, engineers, and scholars have been adamantly clear: the possibility of disastrous consequences is high. The debate is most often framed in the following way: we do not stop the progress of AI, yet we know it can cause nothing short of the annihilation of humanity. Bostrom popularizes the concept of 'existential risk,' which is the idea that superintelligence, no longer under human control, can put at risk the very existence of mankind (2002 and 2014). In the entire debate, however, there is no room for evil. Evil is replaced specifically by the notion of 'misaligned machine intelligence,' i.e., machine intelligence with goals misaligned with human goals (Tegmark 2017, 162–163). The task for researchers, engineers, and programmers is making machine intelligence learn, adopt, and retain humanity's goals (2017, 260). The way "to imbute a superintelligent AI with an ultimate goal that neither is undefined nor leads to the elimination of humanity," in the words of Tegmark, passes through the development of adequate forms of ethics (2017, 280).

Evil is not a popular topic. Scholars prefer to deal with wrongdoing. Evil-skeptics insist that morality demands that humans abandon the concept of evil. Evil-revivalists, however, insists that the concept of evil should be revived– that morality demands that humans make evil intelligible. In this paper I take a realistic stand, and I assume that evil exists in the same way that goodness, malice, or honesty exist. I have reason to believe that there really are evil actions, the worst kind of wrong actions, actions beyond redemption. With this qualification in mind, there is no room in this article for trivial evils, or excusable evils, or evil actions that people morally ought to perform (Calder 2015 and Steiner 2002). That said, the notion of evil action is vague, ambiguous, and in need of clarification. Talk about evil should be done cautiously, for good pragmatic reasons, and claims that certain actions are evil should be clarified. So, what is evil? 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" (2002, 3). The first form of evil is moral and is intrinsic to human nature (i.e., human cruelty); the second is natural and extrinsic to humanity (i.e., human suffering). In the last two centuries, philosophical reflection focused on the first form, leaving the second to the attention of science.

First in science fiction literature, then in science, philosophy, and technology studies, readers have been accustomed to address evil in technology as a case of character trait or moral property of actions (Russell 2014). With regard to the former, scholars seems to have different ideas on where evil should be actually located: in the machines or in the humans standing in front (users, clients) or operating behind (programmers, engineers) the machine (Martin 2019; Sparrow 2007; Sparrow 2016; Arkin 2009). Those who believe that technology is inherently neutral or value-free see evil in terms of human cruelty; those who believe instead that technologies have politics, that is, technological objects and systems have political qualities for they "embody specific forms of power and authority" (Winner 1986, 19), investigate the possibility of technologies as intrinsically evil. Scholars also differ with reference to the primacy of character trait over moral property of actions, or the other way around. Russell (2014, 31) claims that we ought to build an account of evil character on a prior account of evil action; Haybron (2002, 280) and Singer (2004, 190) argue for the contrary view of beginning with an account of evil character. Other authors reject the option of evil machines and argue that the option reflects cultural anxieties about robots; these scholars prefer to direct their attention to the human perception of evil machines (machines that are perceived evil) (Szollosy 2017). Not surprisingly, some simply refuse to use the word 'evil' because of what they see as its religious connotations. Others use the word ‘evil’ but make clear that it does not come with religious connotations (Martin 2019, 1).

In conclusion, in this article, to say a technology is evil is to say it (i.e., general AI before or after the singularity) puts at risk the very existence of mankind. Accordingly, the problem I like to address here can be summarized as follows: can an ethical limit be seriously considered as a concrete, practical remedy for the existential risk hypothetically presented to humanity by the singularity?

In the next two parts, the attention moves from AI in particular to technology in general, as the discourse is addressed in general terms with regard to Guardini’s thought on the culture of technology. As it will become clear at the end of the next section, Guardini’s reflection can be organized in a framework. In the second part, I apply Guardini’s thought to the problem of evil technologies.

2. Guardini’s Thought on Technology

Across the last century, Roman Catholicism has attempted to come to terms with technology. The result of this effort can be summarized in the work of two thinkers, German theologian-philosopher Romano Guardini and French theologian-sociologist Jacques Ellul. Both share the same premises, that is, the technological has replaced nature as the milieu in which human beings are required to exist (Guardini 1994, 13; Ellul 1983, 86). They reach, however, opposite conclusions: Guardini believes that technology should be evangelized, that is, assimilated into the Christian worldview; Ellul argued that technology should be challenged and rejected. Recently, Pope Francis mentioned Guardini several times in one of his most important official documents, *de facto* giving credibility and authority to Guardini’s position. To be honest, Guardini never really offered a solution to the problem of the assimilation of technology into the Christian worldview; he rather articulated an analysis of the nature of technology as a cultural phenomenon in the same line as the work of other German philosophers of his day: Karl Jaspers (1931), Oswald Spengler (1931), Ernst Cassirer (1985), Martin Heidegger

(1962), and Jürgen Habermas (1968). To understand Guardini's thought of technology as a cultural phenomenon, one must start from Guardini's view of the passage of Western civilization from modernity to absolute modernity.

In Guardini's view, modernity is the condition of the modern world. Here modernity (a word first coined by Charles Baudelaire in 1864) is not examined through just a philosophical lens but includes the cultural embodied elements. Modernity is a philosophy as well as an entire cultural package, with a complete array of philosophical and scientific foundations affecting every aspect of life, including arts, politics, economy, and society. In a nutshell, modernity is a multi-layer condition and, as such, it offers a vision of synthetic totality as a comprehensive structure of meaning. According to Guardini, modernity is gradually replaced by a more radical form of modernity, i.e., absolute modernity. The expression 'absolute modernity' refers to Arthur Rimbaud's *Une saison en enfer*, where he states that "one must be absolutely modern" ("il faut être absolument modern") (1979, 116). Being absolutely modern is being modern in an absolute way, the latter understood by Rimbaud as the condition of those who have been made free and have broken loose from any previous bond with the past. With 'absolute modernity,' Guardini means a world that is absolutely modern—modern to an absolute degree. If modernity is the condition in which the human has been severed from nature, absolutized modernity is the world in which the 'natural' has been replaced by the artificial. The end of the modern world is the entering of humanity into a world in which everything is merely factual, everything including human existence.

Being modern in an absolute way is understood by Guardini as the condition of those who have been made free and have broken loose from any previous bond with nature. Ultimately, in a condition of absolute modernity, the world is artificial, that is, the inner reality of the world is filled with norms and necessities, but at the same time there is nothing that could not be different. All things and souls fly in the void like travelers in space without any possibility of ground control. In Guardini's words, "I am who I am not by nature; rather, I am 'given' to myself" (1993, 14ff). "I am given" can probably be paraphrased as: I am an object among other objects that can be thrown with complete indifference into a cosmos operating as a mere space. A robust sense of contingency lies in that assertion. In sum, the world is not by nature, it is given. And Man is not by nature. For Guardini, the passage from modernity to absolute modernity is a transformation of culture – a transformation from a scientific culture to a technological culture. Here 'culture' stands for *Weltanschauung*, the fundamental cognitive orientation of world perception. A scientific culture is a culture in which the world is perceived through a scientific lens; analogously, a technological culture is a culture in which the world is perceived through a technological lens. The crucial point is how the transformation occurs. The shift from modernity to absolute modernity, in Guardini's opinion, is the transformation of culture through the re-articulation of the relationship between *ethos* and *logos*. Guardini claims that the primate of *logos* over *ethos*, encapsulated by the Scholastic axiom 'action follows being' (*ager esequiter esse*), then adopted in modernity with a slight adjustment, i.e., 'action follows knowledge,' has reached a dead end. In a technological culture, *ethos* drives *logos*. Thus, the movement consists in a passage from a scientific world-view with *logos* as the center of gravity, to a technological world-view centered on *ethos*.

For Guardini, the passage from modernity to absolute modernity is the shifting from scientific culture to technological culture. The crucial point is how the transformation occurs. In the following four sections, I address the question of the transformation of culture through the re-articulation of the relationship between *ethos* and *logos*. In particular, the movement from a scientific world-view with *logos* as the center of gravity to a technological world-view centered on *ethos* is considered. Thus, in the next sections a trajectory within the transformation of culture, from a culture centered on scientific knowledge to another centered on technological action, is designed. While in the previous part of the article I mentioned Guardini's emphasis on the artificial – counterpointed to natural – character of technology, in this and the next parts I focus on Guardini's identification of technology with power. Guardini focuses on technology as power in *Power and Responsibility: A Course of Action for the New Age.* While *End of the Modern World* and *Power and Responsibility* were originally translated into English and published by Regenery Press as two separate works, ISI Books has published them together in one volume.

2.1. Logos and Ethos in a Scientific Culture

In this section, the relationship between *logos* (theory) and *ethos* (practice) in scientific culture is briefly described. Two different relationships between theory and practice in scientific culture are examined through the prism of the relationship between *logos* and *ethos* in the reality of modernity. In particular, the question of the primacy of *logos* over *ethos*, and its reversal, the primacy of *ethos* over *logos,* is addressed. Let's start with the former, that is, the subordination of the *ethos* over the *logos,* of will over the knowledge. The subordination of the *ethos* over the *logos* means that every *ethos* always needs a *logos* to precede it and give it meaning. Its reversal is the subordination of the *logos* over the *ethos,* of knowledge over will. As a result of this emphasis on superiority of the will (voluntarism), *ethos* has received a primacy over *logos*, the practice over theory. In other words, the subordination of the *logos* over the *ethos* proclaims the superiority of the utility and the pragmatism of the will over knowledge and meaning. In sum, the first option (the primacy of *logos*) supports the primacy of scientific method and theoretical inquiry; the second (the primacy of *ethos*) supports the primacy of utility and pragmatism. In a scientific culture, *logos* maintains a dominant position over *ethos*.

2.2. From Scientific Logos to Technological Logos

The passage from modernity to absolute modernity, i.e., from a scientific culture to a technological culture, is a complex movement. The first step is the shift from a scientific *logos* (a *logos* within a scientific culture) to a technological *logos* (a *logos* within a technological culture). Readers may be aware that there are different types of *logos*. The case considered is that culture is transforming, moving from a culture centered on scientific knowledge to one centered on technological knowledge. This passage to a technological *logos* means that, in this era, technological rationality replaces scientific rationality. More precisely, a distinct way of reasoning, a form of calculative thinking, a calculative *logos* replaces a scientific, utilitarian *logos*. Technology has by now become the most powerful form of organization of knowledge and transformation of the world.

2.3. From Scientific Ethos to Technological Ethos

The first step is the shift from a scientific *logos* (a *logos* within a scientific culture) to a technological *logos* (a *logos* within a technological culture). However, the passage from a scientific culture to a technological culture involves the *logos* as well as the *ethos*. More precisely, the passage from one *logos* to another implies a change from one *ethos* to another. Readers may be aware that there are different types of *ethos*. In practice, the *ethos* changes as a result of the related *logos*. This is particularly true in the case of a passage to a technological *logos*. The *ethos* of science is governed by a utilitarian *logos*; the *ethos* of technological progress is governed by a calculative *logos*. Thus, the *logos* of *techne,* a calculative way of thinking, theorizing, and judging, orients and dominates over an *ethos* of technological power.

2.4. From the Logos-Ethos Relation to the Ethos-Logos Relation

I complete the description of the transformation with further comments on the relationship between *ethos* and *logos* in a technological culture. In this section of the paper, I recognize that a technological culture has subordinated *logos* to *ethos*. This subordination can be described as the dominance of doing (*ethos*) over knowing (*logos*). What ultimately matters is not knowledge, but activity. As result of the nominalist emphasis on superiority of the will in technological culture, *ethos* has received a primacy over *logos*. No matter how great the quality of knowledge, the brilliance of theory, what really matters is the energy of the volition and ultimately action. In the end, a technological *logos* is subordinated to a technological *ethos*, that is, an *ethos* of a technological culture that is all about power and driven by a thirst for unstoppable technological progress.

2.5. Science-Technology Relationship

The passage from a scientific to a technological culture changes the relationship between science and technology. In a scientific culture, technology is a means to increase scientific knowledge. In a technological culture, technology is a goal of its own. Thus, the passage from a scientific to a technological culture can be understood as an erosion of ends by means. Science must strengthen the technological means of which it makes use, but in so doing, science empowers technology and transform it in a goal. To put it differently, technology is a means that becomes necessary for increasing the power of science; because it is necessary for increasing the power of science, it is no longer a means, but rather a goal in itself. In a technological culture, technology is no longer the handmaid of the scientific forces that govern the world but is itself the power that governs the destinies of these scientific forces. The tendency of our time is that science is asked to serve the ideology of technological advancement, not the other way around: modern science tends to depend on technological knowledge.

2.6. In Summary

A common belief in these days is the primacy of knowledge over action and science over technology. Technology is a sub product of scientific knowledge, or, in the terms used above, a technological *ethos* is subordinated to a scientific *logos*. The notion of culture transformation helps make clear that the relationship between logos and ethos operates within the same culture or paradigm. Thus, the transformation from the scientific to the technological is more precisely a shift with a dominant scientific *logos* to a similarly dominant technological ethos, or more simply, from scientific knowledge to technological action. This shift from the primacy of scientific rationale to technological power implies that an *ethos* (action) of technological power, without a *logos* that precedes it and gives it meaning, is free to produce an action that is independent from knowledge. This topic is addressed in the next part of the study.

3. Application to the Problem of Evil Machines

For Guardini, the modern world is replaced by a world that is even more modern – modern in absolute terms. The movement from the modern world to the absolute modern world, or from a scientific culture to a technological culture, is the passage from a form of scientific rationality to a configuration of technological action. In the next two sections, I elaborate on the results of Guardini's reflection on technology: first, I take a closer look at technological culture. I understand technological culture as absolute becoming. Second, I identify technology as will and address evil in technological culture in terms of unlimited power. Finally, I return to the problem of AI and ethics' role in controlling it.

3.1. Technology in Technological Culture

What is an action that is independent from knowledge? In a nutshell, it means that everything is transient, ephemeral, and destined to decay. It means that everything is controvertible, deniable, and disputable. It means that action cannot be governed, therefore it cannot be constrained or limited. It means the triumph of becoming on being and the impossibility of rationalizing the logic of becoming.

The becoming, the notion that everything is (happens) *in time*, has been known to the citizens of the West living in the humanistic and scientific eras. The becoming is that nothing stays forever. Faced with the anguish of 'becoming,' the West responded in the past through the logic of the remedy, i.e., it raised the 'immutables' (God, the laws of nature, dialectics, the free market, the ethical or political laws, etc.). The immutables are entities (God, the divine laws, the laws of nature) and transcendental and permanent values (ethical, natural, etc.). The task of the immutables is to limit, constrain, and ultimately control the becoming. In the pre-scientific tradition, structures of the world were eternal and permanent, God was necessary and so were the king and the pope. Providence governed the world of people. Science came to light claiming that eternal laws, such as the law of gravity, regulate the universe, where time and space are absolute entities independent from events and Man. States have tried to control becoming through laws and soldiers, religions through God and the supernatural, philosophy through logic and spirit of criticism. All in all, the immutables stand on their necessity: to be effective in their task to constrain and control the becoming, the immutable need to be necessary.

The essential substratum of the philosophy of the last two centuries has made crystal clear that the only possible necessity is the becoming. Today, there lies the persuasion that every *thing* is contingent. This persuasion, in turn, is founded on the central trait of the philosophical thought of our time– the thought that no 'thing' exists by necessity. Here 'thing' is understood in its broader sense – the one in which, for example, a 'thing' is an animal, an intellectual state, a concept. Divine things exist by necessity, but not because there are things, but rather because there are divine (above the contingency of life). But divine things, of course, cannot exist. Non-divine things, however, insofar that they are things, do not exist by necessity. This implies the power to imagine a time in which these things that are, will be no longer. It also implies the power to imagine a time in which these things that are not, will be. In other words, by linking things to a 'when,' to a time, things that are something becomes nothing, and things that are nothing become something. In the realm of becoming, all comes and goes. In the realm of becoming, being has a birth (conceived as beginning to be by emerging from nothingness) and death (conceived as the end of being). In the realm of becoming, there is no contradiction in having a being that is not (i.e., an entity that is nothing). This oscillation of things between being and nothing (or non-being) is the contingent condition of things, their condition of non-necessity.

An example in which to think about a being that is not, is not contradictory, is historiography. The past is no longer. Insofar as it is past, this day is not and has become past. The past is no longer and, as such, it is non-being, nothingness. Yet, the past 'is.' Historians refer to the past that remains in the traces that it has left. Properly speaking, what remains in the traces is not the past, rather something that remains. The past does not remain. But, again, in historiography there is no contradiction in thinking a being that is not. What we have here is a non-absolute being combined with a non-absolute non-being—a non-necessary something with a no-necessary nothing. When it comes to things in time, the principle of non-contradiction seems not to function: it is possible to think an entity that does not exist. And it is possible because *it is an unquestionable evidence that a thing becomes*: there is a time when it is and a time when it is not. It is absolutely evident that things are in becoming, that is, they are born, exist, and then die. In other words, the being itself is ephemeral, is contingent, that is, the world is seen as existing by chance; reality has no foundation. A world of non-necessity is a contingent world. It means that the philosophy of our times has shown that nothing else, God, immutable, or eternal, is necessary. Everything is contingent. Every truth, knowledge, law or principle, value or faith, is destined to perish, replaceable, and ultimately unnecessary. But if the immutables are temporary and contingent, how can they control the becoming? How can they limit and constrain and ultimately govern the becoming when the becoming is permanent and necessary? As a matter of fact, their destiny is to become, i.e., to change, and therefore to be an integral part of the unstoppable becoming.

Both the humanist and scientific cultures share the same will, the will to know permanently the truth of the world. In the humanist culture, knowledge is produced by philosophy; in a scientific culture, knowledge is produced by science, while philosophy tends to sink into scientific knowledge. The entire body of philosophy dissolves into single sciences: psychology, logic, political science. What had been the function of philosophy in the humanistic culture has been inherited by the sciences. Both the humanist and scientific culture raise the immutables because they both inherit the traits of stability from the episteme. The episteme, or the structure of understanding, is the condition that makes knowledge possible (from episteme derives the term epistemology). It is the native essence of philosophy and science, an essence that lies beyond the ‘immutables.’ The episteme is the stable dimension of knowledge, within which are raised all the immutables of the West.

As said, *episteme* is the word with which philosophy expresses the absolute character of truth in the ancient world. More precisely, *episteme* means the staying (what is staying) despite the becoming (what is becoming). The *episteme*, which imposes itself on the becoming, ultimately suffocates becoming; thus, it is necessary that, in order for becoming to be, that episteme becomes impermanent (it is no longer *episteme* in the proper sense). Think, for example, of the sense given to the 'things,' which is in fact not constant, but according to historical epochs. It is a world in which no *episteme* (i.e., knowledge, theory, science) is permanent, in which no stable structures of knowledge are possible. The point is, the technological culture does not inherit the traits of episteme, rather of techne, a practice. According to a technological mentality, consequently, theory is not the model of action and therefore of creation. Creation is action without a preliminary model. In the same spirit, reality in a technological culture is not shaped by the form, and action is not the fruit of thought, nor the work of a precedent model. Reason is not (the only) guide of human behavior nor of the reality. Thought does not govern being, and knowledge does not precede practice. The technological culture is the very negation and destruction of the immutables.

Most of human history has been first and foremost devoted to an attempt to discover the absolute truth, the unmodifiable and incontrovertible truth. The truth is absolute when it neither depends on one's faith, or certainty of it, nor one's hopes or fear that it is the way it is, rather when it is impossible that it would be any different from the way it is. Such an attempt, however, is no longer possible in a world in which no *episteme* is permanent. When an absolute truth is denied, the consequence is not, in turn, the absolute truth of the denial, rather the practical situation to organize existence without such absolute truth (about the world). It is a world in which no definitive and undeniable truth (about the world) is possible. The abandonment of the absolute truth (except the truth that nothing is forever) and the choice of the 'becoming' form the ontological space where the forms of Western civilization, with its social and political institutions, have already moved.

Thus, ephemeral ontology, impermanent knowledge, and deniable truth, are 'things' which maintain themselves in a provisional equilibrium between being and non-being. But it is precisely because 'things' are thought of as provisional equilibrium between being and non-being that the project to dominate them, by producing and destroying things, acquires a radicalism it never before possessed. In fact, ‘produce’ means ‘to make pass from non-being into being,’ that is, to produce reality, knowledge, and truth; ‘to destroy’ means ‘to make pass from being into non-being,’ that is, to destroy reality, knowledge, and truth. The philosophy of the last two centuries has removed the way of every obstacle to action. The triumph of becoming is the rejection of traditional thought – that is, of the epistemic-metaphysical entities – and the termination of the immutables. The immutables, anticipating and controlling the becoming, *de facto* cancel the contingent character of the events. Making the immutables a form of knowledge no longer supported by truth and episteme, the philosophy of the last two centuries has made the realities of time and becoming self-evident and in need of no demonstration; consequently, philosophy opened the way to an unstoppable will of power.

3.2. Power as Unlimited Power (Evil)

A simplified version of my argument in this section could be summarized as follows: the pervasive nihilism of a civilization embracing the unquestionable belief that ‘all things pass’ leaves humanity at the mercy of an unstoppable technological power. This section is also a theoretical inquiry into ethics, with particular attention to ethics against technological domination. Can ethics save humanity?

This section benefits from the following assumptions in pursuing the inquiry: 1. ethics is a form of rationale; 2. ethics’ scope is to govern and orient action; and, 3. ethics aims to maintain the primacy of rationale over action. Can ethics govern and stabilize action, in the case that action is primarily technological progress and will of power? To put it differently, can ethics discipline action and control power in a situation of vertiginous technological advance that seems almost impossible to keep up with? I previously built the case that in a situation of vertiginous technological advancement, the proper order between reason and action is inverted; as a result, reason loses control over action and action becomes dominant over reason. I added that the nihilist substratum of the technological culture simply makes the becoming unstoppable.

It should be clear at this point that technological culture itself requires a definition. What is ‘technological culture,’ after all? It is the view of technology as a ‘project of transforming the world.’ Who does carry forward this project? The answer is, whoever understands technology as an end on its own. For brevity, it can be called a ‘technological apparatus,’ although more precisely an apparatus of technology, science, rational thought, physics, and mathematics. This is an apparatus interested in increasing the power of technology. In the previous dominant cultures – humanistic as well as scientific culture – humanist and scientific apparatus were interested in increasing ‘mankind’ and ‘knowledge,’ respectively. These previous dominant cultures overlook the authentic meaning of the technology’s project. Those who still live and see reality through the lens of humanist and scientist cultures want to make use of technology to realize their objectives. In a technological culture, however, technology has a purpose different from ‘mankind’ or ‘knowledge,’ that is, the indefinite increase of power. Science succeeded in giving mankind greater knowledge than humanism; like technology, science had a purpose different from ‘mankind.’ The same can happen with technology: although technology has a purpose different from ‘mankind,’ it can succeed in giving mankind greater power than humanism and science. With that said, technology does not take ‘mankind’ for an end in itself.

In a technological culture, the purpose of technology is the indefinite increase of power, that is, the will of power. Technology, in fact, is the will of power. Precisely because technology is will of power, technology is, after all, a specific product of Western thought. In the words of Schelling, who summarizes in this way the entire experience of the West, “Wollen ist Ursein,” the will is the primordial Being. The will of power is the will to increase one’s power over things and gods: this has always been the most profound desire of people in the West who think that power allows them to overcome pain and death. Technology’s ultimate purpose is the purpose of the supreme form of the will to power. In other words, the purpose of technology is domination. Technology is destined to pursue its purpose without limitations or counterpowers because the essential substratum of the philosophy of the last two centuries has made crystal clear that the only possible truth is the becoming. All comes and goes. All comes to life and passes away. Nothing is forever. If nothing is forever, no limits can be erected to block technology's indefinite increase of power; in fact, if a limit is not forever, it is will be overcome one day or another.

To put it differently, the project to dominate 'things' is the result of a specific understanding of the world as a place in which things are contended, that is, in which being and nothing (or non-being) contend these things between each other. The project to dominate 'things' is the result of a specific understanding of the world as a place in which the contingent condition of things, their condition of non-necessity, is necessary in order for becoming to reign. If becoming is the supreme evidence, it cannot be suffocated by enduring ontology, permanent knowledge, and incontrovertible truth. The *will* for becoming to be survives any attempt to suffocate it; precisely for this reason, it is necessary for any enduring ontology, permanent knowledge, and incontrovertible truth, within which the becoming of the world is unthinkable, to vanish. Thus, will dominates it all.

As said, the term 'culture' is used in a specific mode, synonymous with 'world-view.' This world-view, in turn, is the result of a dynamic relationship between *ethos* and *logos*. Once this insight is applied to 'technology' and 'ethics,' 'technology' stands for an orientation toward *ethos*, and 'ethics' toward *logos*. In this context, this study had advanced one claim: the shift from a scientific to a technological culture can be better understood as a shift of the relationship between *logos* and *ethos within* culture (a transformation of culture). In this section, a second claim is advanced: ethics is incapable of governing technology, and that is a consequence of that shift.

It is clear, at this point, that no degree or quality of forms of ethics can reverse this trend. The very nature of action has changed into will of power, and action refuses to be constrained by the limits of reason. Thus, an *ethos* (action) of technological power without a *logos* that precedes it and gives it meaning is free to produce an action that is not simply morally wrong, but leaves no room for understanding or redemption (i.e., evil). In other words, in the technological culture, technology can be evil because the culture itself is tainted by will of dominion. The same logic can be extended to ethics:

forms of ethics that are developed to limit the effects of technological progress (i.e., AI self-empowerment) are ineffective if not subordinated to a rationale (*logos*), which, in turn, restrains the use of technological power and guides the appropriate exercise of such power. But in a technological culture, Guardini argues, practice is not subordinated to action. Thus, one can't accept the rationale of technological progress and believe that one can restrain it with ethics.

In conclusion, in this section I showed the inadequacy of forms of ethics in limiting the self-empowering tendency of AI. In fact, the nominalist emphasis on superiority of will (voluntarism) in the age of AI has already established the primacy of *ethos* (action) over *logos* (rationale). No longer constrained in a rational framework, action is no longer morally answerable. As a result, an *ethos* of power divorced from responsibility is driven by a thirst for self-empowerment which can be boldly described as 'demonic.'

Conclusion

I discussed the culture transformation – from scientific to technological culture – by adopting the categories of Guardini's thought. A definition of culture is provided in terms of *ethos* (the embodiment of ideas in social practice) and *logos* (the rationale behind practice). Then I considered evil in the context of culture transformation, through a complex re-articulation of the *ethos-logos* relationship. The endgame of the reflection pursued throughout this article is that in a technological culture, evil emerges naturally as an inherent trait of power associated with technology.

Acknowledges

I am indebted to two anonymous reviewers of an earlier version of this paper for providing insightful comments and directions for additional work, which has resulted in this improved version.

References

Arendt, Hannah. 1963. *Eichmann in Jerusalem: A Report on the Banality of Evil*. New York: Viking Press.

Arkin, Ronald C. 2009. *Governing lethal behavior in autonomous robots*. Boca Raton: CRC Press.

Boddington, Paula. 2017. *Towards a Code of Ethics for Artificial Intelligence.* Berlin: Springer.

Bostrom, Nick. 2002. "Existential Risks – Analyzing Human Extinction Scenarios and Related Hazards." *Journal of Evolution and Technology* 9: 31–33.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies.* Oxford: Oxford University Press.

Bowles, Cennydd. 2018. *Future Ethics.* N.d.: NowNext Press.

Calder, Todd. 2015. "Evil Persons." *Criminal Justice Ethics* 34(3): 350–360.

Card, Claudia. 2002. *The Atrocity Paradigm: A Theory of Evil.* Oxford: Oxford University Press.

Cassirer, Ernst. 1985. *Symbol, Technik, Sprache: Aufsätze aus den Jahren 1927–1933,* edited by Ernst Wolfgang Orth & John Michael Krois. Hamburg: Meiner.

Coeckelbergh, Mark. 2020. *AI Ethics.* Cambridge, Mass: MIT Press.

Formosa, Paul. 2008. "A Conception of Evil." *Journal of Value Inquiry* 42(2): 217–239.

Guardini, Romano. 2001. *The End of the Modern World*, Trans. Joseph Theman and Herbert Burke. Wilmington, DE: ISI Books.

Guardini, Guardini. 1994. *Letters from Lake Como: Explorations in Technology and the Human Race*, Trans. Geoffrey W. Bromiley. Grand Rapids, MI: Eerdmans.

Guardini, Romano 1993. *Gläubiges Dasein: Die Annahme seiner selbst.* Mainz: Matthias-Grünewald-Verlag.

Gunkel, David J. 2012. *The Machine Question: Critical Perspectives on AI, Robots, and Ethics.* Cambridge, Mass: MIT Press.

Habermas, Jürgen. 1968. *Technik und Wissenschaft als "Ideologie".* Frankfurt am Main: Suhrkamp.

Heidegger, Martin. 1962. *Die Technik und die Kehre.* Pfullingen: Neske.

Haybron, Daniel. 2002. "Moral Monsters and Saints." *The Monist* 85(2): 260–284.

Jaspers, Karl. 1931. *Die geistige Situation der Zeit.* Berlin & Leipzig: Walter de Gruyter & Co.

Kekes, John. 2005. *The Roots of Evil.* Ithaca: Cornell University Press.

Lin, Patrick, Jenkins, Ryan, and Abney Keith. 2017. *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence.* Oxford: Oxford University Press.

Lin, Patrick and Abney, Keith. 2011. *Robot Ethics and Social Implications of Robotics.* Cambridge, Mass: MIT Press.

Martin, Brian. 2019. "Technology and Evil." *Social Epistemology Review and Reply Collective* 8(2): 1–14.

Millare, Roland. 2016. "The Primacy of Logos Over Ethos: The Influence of Romano Guardini on Post-Conciliar Theology." *The Heythrop Journal* 57(6): 974–983.

Pope Francis. 2015. *Encyclical Letter Laudato Si*, given in Rome at Saint Peter's on May 24.

Russell, Luke. 2014. *Evil: A Philosophical Investigation.* Oxford: Oxford University Press.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control.* New York: Vikings.

Severino, Emanuele. 2016a. *The Essence of Nihilism.* London and New York: Verso.

Severino, Emanuele. 2016b. *Nihilism and Destiny.* Milan: Mimesis International.

Rimbaud, Arthur. 1879. *Une saison en enfer.* Paris: Gallimard, Bibliothèque de la Pléiade. The English translation is taken from Arthur Rimbaud, Complete Works, trans. Paul Schmidt (New York: Harper and Row, 1976), http://www.mag4.net/Rimbaud/poesies/Farewell.html, accessed February 14, 2018.

Singer, Marcus G. 2004. "The Concept of Evil." *Philosophy* 79(2): 185–21.

Sparrow, Robert. 2007. "Killer Robots." *Journal of Applied Philosophy* 24(1): 62–77.

Sparrow, Robert. 2016. Robots as "Evil Means"? A Rejoinder to Jenkins and Purves. *Ethics & International Affairs* 30(3): 401–403.

Spengler, Oswald. 1931. *Der Mensch und die Technik: Beitrag zu einer Philosophie des Lebens.* München: C.H. Beck.

Szollosy, Michael. 2017. "Freud, Frankenstein and our Fear of Robots: Projection in our Cultural Perception of Technology." *AI and Society* 33(3): 433–43.

Steiner, Hillel. 2002. "Calibrating Evil." *The Monist* 85(2): 183–193.

Tegmark, Max. 2017. *Life 3.0. Being Human in the Age of Artificial Intelligence.* New York: Knopt.

von Balthasar, Hans Urs. 2010. *Romano Guardini: Reform from the Source.* San Francisco: Ignatius Press.

Vinge, Vernor. 1993. *The Coming Technological Singularity. How to Survive in the Post-Human Era.* San Diego State Univ., San Diego, CA, United States. Published on December 1, 1993. Available online at https://ntrs.nasa.gov/search.jsp?R=19940022856.

Wallach, Wendell and Allen, Colin. 2008. *Moral Machines: Teaching Robots Right from Wrong.* Oxford: Oxford University Press.

Winner, Langdon. 1986. *The Whale and the Reactor: A Search for Limits in an Age of High Technology.* Chicago, IL: University of Chicago Press.