International Journal of Swarm Intelligence and Evolutionary Computation

Hurtado, Int J Swarm Intel Evol Comput 2016, 5:3
DOI: 10.4172/2090-4908.1000137

Review Article Open Access

The Ethics of Super Intelligence

Manuel Hurtado*

Stanford High School Summer College, Spain

Abstract

The future arrival of super intelligence and its impact in society raises numerous concerns. Grounded in the research hitherto elaborated by the field of machine ethics, this paper contemplates the challenge of formulating a code of ethics that regulates super intelligence's behavior. The first section discusses the need for this code of conduct and contends why it should be centered on ethics. The second section examines the various complexities of this endeavor, and puts forth a theoretically plausible approach. Lastly, the final section further raises the question of whether human beings should truly have the final say given a disagreement between the ethics of man and machine.

Keywords: Super intelligence; Elvex's robotic cranium; Artificial Intelligence

The Ethics of Super Intelligence

As Dr. Susan Calvin cold-bloodedly drew her electron gun, pulled the trigger, and shot a lethal burst of electrons into Elvex's robotic cranium, an odd conglomeration of fear and relief churned my stomach. As Elvex ceased to be any more, a feeling deep inside me uttered that, fantastic as it might appear, Isaac Asimov's brief narrative entitled *Robot Dreams* (1986) was denoting an often overlooked truth: the concept of super intelligence, defined as "an intellect that is much smarter than the best human brains in practically every field, including scientific creativity, general wisdom and social skills" [1], portends a potentiality worthy no longer of mere science fiction's fancifulness, but rather exhaustive scientific study.

Indeed, the prospect of super intelligence, once alien to the realm of academia, has over the past decades become an increasingly popular focal point of study. The pace at which technology has evolved in the past and is currently evolving has led us to maintain that it will not be long until fully automated, super intelligent, non-human entities environ their creators. As a matter of fact, I.J. Good's renowned "Intelligence Explosion", whereby the fabrication of the first advanced, artificially intelligent entity will catalyze an indefinite progressive evolution of machine intelligence - otherwise considered to be the onset of what Ray Kurzweil regards as the period of 'Singularity' - is believed to be the root of this seemingly fictitious technology. As Anderson [2] explains it in his paper entitled Ethical Issues in Advanced Artificial Intelligence,

"several authors have argued that there is a substantial chance that super intelligence may be created within a few decades, perhaps as a result of growing hardware performance and increased ability to implement algorithms and architectures similar to those used by human brains."

And yet, for all their inherent mysticism, the means through which humanity will eventually procure super intelligence are naught but a minute, trivial bit of the full picture. Rather, we had better draw our attention towards the more impending matter: the impact that super intelligence will have in our society.

Disregarding the widespread notion of machines' suitability for the "three Ds" - that is, dull, dangerous, and dirty jobs - super intelligence's elevated computational power and ensuing proficiency in any task known to man will bring about a proliferation of machines whose roles in society will be infinitely more complex than they are now, exhibiting both a degree of dexterity that eclipses human capabilities and a will to reengineer themselves ad infinitum. It is indisputable; therefore, that super intelligence will surpass its creator.

An effective method to guarantee super intelligence's harmless behavior is, hence, in place. For this reason, academia's increasingly popular field of machine ethics has taken to the investigation, discussion, and reflection of the moral dimension of artificial intelligence and machines. Throughout this paper, I will use machine ethics' underpinning notions to explore the plausibility of developing a code of ethics for the future's super intelligent machines.

To do so, I will first attend to the debate of whether machine ethics is, in fact, the most suitable approach to the development of behavioral guidelines that mitigate any potential risks that super intelligence might bring by having it ethically evaluate its possible courses of action. Secondly, I will examine the viability of different practical approaches to controlling super intelligence, highlighting the complications of the attainment of less advanced, present artificial intelligence ethics, and subsequently outlining what I consider to be a theoretically feasible modus operandi. Finally, I will contemplate the differences and similarities between human and artificial ethical actors in order to further raise the question of who would ultimately have the last word given a contradiction between human ethics and super intelligence's ethics.

The Need for Machine Ethics in our Pursuit of Super Intelligence

The prospect of super intelligence is unquestionably attractive. However, the mere thought of coexisting with a lifeless entity infinitely more intelligent than any biological creature known to man is enough to spark distress in even the most fervent of its advocates. Capable of surpassing human achievement in practically any field or activity, the power of super intelligence must not only be regarded as an ideal source of widespread benefit to man but as a potential root of uncertainty and harm as well. Therefore, it comes as no surprise that the short answer to the seminal question "Why do we need machine ethics?" is, simply put, because it is in the ethical or unethical behavior of super intelligence

***Corresponding author:** Manuel Hurtado, Stanford High School Summer College, Spain, Tel: 333881283; E-mail: manuhram@gmail.com

Received August 26, 2016; **Accepted** August 29, 2016; **Published** September 03, 2016

Citation: Hurtado M (2016) The Ethics of Super Intelligence. Int J Swarm Intel Evol Comput 5: 137. doi: 10.4172/2090-4908.1000137

that the prosperity or demise of our existence lies. Notwithstanding, let us explore a more thorough and convincing response.

From a historical standpoint, the development of super intelligence might be looked upon as a marked parallelism to that of computers [3]. As noted by Asimov [4] albeit the exponential expansion of the computer industry throughout the second half of the 20th Century and, more prominently, the first two decades of the 21st Century has been accompanied by a myriad of societal benefits that have facilitated man's survival, the computerization of our culture has also been the root of numerous unpropitious trends such as, but not limited to, cyber-terrorism, child pornography, and the black market.

Hence, in the discussion of the need for machine ethics, understandably call for the consideration of the negative impacts that futuristic developments entail, asserting that, without foresight, emerging technologies have come at a cost - a remark that becomes all the more critical when discussing super intelligence. Still, their research goes on to claim that rather than labeling these fears as sufficient motives for the termination of our pursuit of non-human intelligence, these concerns underline the necessity of contemplating the risks that the materialization of such technologies supposes and, subsequently, the need for our collective effort to ensure their mitigation. As a matter of fact, it is these very preoccupations which form the bedrock for machine ethics as the field seeks to develop a sense of action that might allow autonomous beings to not only refrain from acting unethically but also have the inherent will to consistently act ethically. Therefore, the field must progress on par with, if not lie at the crux of, technological advancement.

And yet, opposition to the development of machine ethics still remains passionately adamant. Arguably, it doesn't take a profound instruction on the inner workings of machines to understand how electrical systems work. In the most fundamental sense, therefore, an antagonist to the field of machine ethics might claim that since any electrical machine has an absolute dependence on the flow of electricity through its circuits, rather than trouble ourselves with the philosophical quandaries that obscure the attribution of moral sense to these non-human beings, the technological development of super intelligence should be the sole focus of our attention, for, in the event of its misdemeanor, turning off a switch will suffice to prevent any potential injury from being inflicted on human beings. But this is a rather shortsighted claim, considering the latent ramifications of an antagonistic form of super intelligence: albeit it is true that these beings could be turned off with a single switch, the extent of their future involvement in society will make them so imperative to the adequate functioning of our societal structure that simply "turning them off" would practically amount to suicide [5].

In other instances, by resorting to naught but a vague apprehension of its very definition, dissentients might assert that, given the insurmountable brainpower of super intelligence, scientific furtherance should not pay much heed to the ethical virtues of the technology, but rather its actual creation, for not only will it inherently strive to do good (assuming that the definition of good is clear - herein lies another problem, which will be explained in more depth later on) but it is also through the delegation of important decisions to this entity that social benefit can be maximized. Notwithstanding, as Bergman [5] elegantly points out:

"The option to defer many decisions to the super intelligence does not mean that we can afford to be complacent in how we construct the super intelligence. On the contrary, the setting up of initial conditions, and in particular the selection of a top-level goal for the super intelligence, is of the utmost importance. Our entire future may hinge on how we solve these problems."

And yet, it is machine ethics' extensive contemplation of the difficulties entailed by choosing an ethical theory which both suits our society's needs and is commensurate with our expectations of moral machine behavior that most adversaries of the field disregard. As a result, aware of the potential harm that artificial intelligence and super intelligence might lead to, machine ethics' opponents like Bostrom [6] have devised what is referred to as "Safety Engineering". As it is subtly implied by its name, this emerging field seeks to formulate pathways leading to safe artificial intelligence, autonomous machines, and ensuing super intelligence through the incorporation of recognition of the need for "safe machines" in the field of engineering. Approaching the problem of autonomous systems' correct behavior from a more empirical standpoint, safety engineering discards the deliberation on the ethical dimension of non-human intelligence and favors instead practical experiments in environments that permit the adequate control of these forms of advanced technology, allowing for the study of their behavior. A set of guidelines governing the means to ensure proper machine behavior would therefore ensue.

There are two rebuttals to this perspective: first and foremost, considering the extensive amount of variables that pertain to a single action in the real world, we cannot possibly expect that the study of machine behavior in a controlled environment will suffice to adequately understand the resulting consequences of such demeanor when confronted with the outside world. Furthermore, provided this limited study could actually manage to fully grasp all the consequences of a single action, it seems rather implausible that a team of human programmers would be capable of taking them all into account when programming the machine's response to a given situation. Noting with concern that interaction with the outside world is filled with these decision making processes, we can conclude that safety engineering's approach to correct machine behavior seems non-viable.

Secondly, safety engineering falls short of understanding the full extent of machine ethics' purpose. While the former perceives artificial intelligence and autonomous machines as mere tools, the latter bears in mind that it is in our best interest to cooperate with them. Therein lies the bright line distinguishing safety engineering's pursuit of preventing unethical behavior in autonomous, intelligent systems from machine ethics' campaign to motivate these systems to act ethically. Unlike safety engineering, the cornerstone of machine ethics is not to refrain super intelligence from having unethical thoughts, but rather to make super intelligence think ethically so that all of its mental processes, be they as they might, will be permeated by a will to help, respect, and value humanity. Hence, when their evolution reaches the point at which they edit and engineer themselves, we will know not that their ethical dimension remains unedited, but rather that this ongoing evolution is, in and of itself, ethical.

On the other hand, machine ethics provides some convincing arguments for its pursuit. There are strong reasons to believe that machines - and thus super intelligence - would amount to a better ethical actor than man himself. According to research by Bostrom [7] this is due to three fundamental reasons: First, machines have a greater computational power, which facilitates the prevision of consequences of actions and therefore makes ethical decisions more accurate; second, human beings display a tendency towards bias when making ethical decisions, usually favoring those close to them, while machines do not; and third, whether or not due to their remarkably inferior processing

speed, human beings are likely to fail to consider all the possible actions that might be taken in a given situation. Other advantages of incorporating ethics into super intelligent systems include their capacity to carry out an action repeatedly and competently at high speeds, as well as their ability to share information between them at an equally efficient rate. Perhaps most importantly, unlike human beings, machines are adept at making decisions unemotionally, which "means that they can strictly follow rules, whereas humans tend to favor themselves and let emotions get in the way of clear thinking. Thus, machines might even be better suited to ethical decision-making than human beings [8]". Furthermore, as Gips [9] points out, the inherent detachment entailed by the consideration of human virtues, ethics, and morals in machine ethics will enable us to understand ourselves more profoundly. In other words, by attempting to formalize our ethical behavior and make our own morality the subject of this field's study, not only will we plow the seeds of a brighter, super intelligence-encompassing future, but we will also reap the fundamental benefit of further comprehending what it means to act, think, and exist ethically.

In conclusion, therefore, machine ethics should not only be regarded as the preferable means to approach the challenge of ensuring advanced artificial intelligence's adequate, safe, and beneficial behavior, but it should also be seen as a theoretical-practical venture to break down and formalize the ethical dimension of human beings. Ergo, the complete answer to the question "Why do we need machine ethics?" may very well be: because it is the field which uses the knowledge of the ethical self-hitherto developed by our species in order to analyze what correctness constitutes in our present society so that it can help ensure the safe, fruitful propitiousness of our seemingly unrealistic, technologically-dependent future [10].

The Approach towards the Ethics of super intelligence

The endeavor to control super intelligence by instilling in it a sense of morality that will dictate its behavior is not an exclusive competence of machine ethics. As a matter of fact, there is substantial literature on the topic that adopts an approach unlike that of this emerging scientific-philosophical field. An example of an alternative approach is that of Bostrom [7]. In his book, entitled super intelligence: Paths, Dangers, Strategies puts forth the concept of a "box" in which super intelligence can be contained, which, he notes, would render the powerful entity inside the enclosure harmless by isolating it from any contact with the outside world save a single, controlled communication channel with scientists. Furthermore, Bostrom argues, the controlled environment would allow scientists to determine the super intelligence's knowledge of our real world. Notwithstanding its undoubted effectiveness, however, this method of control is, to my mind, rather futile, for albeit it mitigates the dangers of super intelligence, it does so at the cost of exploiting its potential to aid human beings in the search for a solution to global issues such as hunger, poverty, and inequality, inter alia.

In light of this unsuitableness, Hall [11] proposes yet another method for control which happens to be slightly more akin to that sought by machine ethics. Coining the term "motivational control", the author suggests giving this advanced form of artificial intelligence a sound, beneficial, ultimate goal whose achievement should be the supreme objective of each and every one of super intelligence's actions. As he explains it himself, "Its top goal should be Friendliness. How exactly friendliness should be understood and how it should be implemented, and how the amity should be apportioned between different people and nonhuman creatures is a matter that merits further consideration" [12]. According to the author, because a sound, rational person would whose ultimate goal is X would not turn into Y if, in doing so, it would contradict its pursuit of X, super intelligence would refrain from acting in such a way that contradicts its friendliness towards humanity. Despite appearing detached from any ethical considerations, this method for control's alarming lack of clarity and clear need for further consideration - to which Bostrom himself alludes - is, in fact, all but a desperate call for machine ethics.

The ethical dimension of non-human intelligence, therefore, is central to the complex consideration of super intelligence's reliability when interacting with the real world. Even so, some of machine ethics' approaches to the control of super intelligence - strongly resembling sci-fi science - seem disproportionately implausible at this point in time. Namely, it has been put forth that, taking into consideration the power that super intelligence will provide us and the developments that the field of neuroscience will achieve in the future, it should not be ludicrous to contemplate the possibility of mentally scanning a human brain and incorporating that scan to an artificial neural network. The artificial intelligence would thus possess the ethical thoughts of the human being. On the other hand, one could propose that, given its superior intellect, super intelligence could be taught ethical virtues, as it is done with young children - a Turing Child approach of sorts. Notwithstanding, not only are both of these propositions currently inviable but they also entail super intelligence's arrival prior to the development of its ethics. Their pursuit would therefore result in the potential risk of creating an unsafe entity that may either trick us into believing that it is learning to act ethically when in truth it is not, or it might just blatantly refuse to adopt the ethical behavior we seek to impose on it - in which case its potential will cease to be exploitable, lest we are willing to risk the consequences.

Hence, in its pursuit of super ethics, machine ethics must first address the more tangible issue of artificial intelligence's ethics, for it is in the hands of this upcoming human-level intelligence that the creation of a safe, rational, and benevolent super intelligence largely lies. To address this concern, it must first be noted that these forms of intelligence need to undergo a pivotal transition from being implicit ethical agents who are programmed to act ethically (or at least avoid acting unethically) to explicit ethical agents - autonomous entities capable of reasoning appropriately in the face of an ethical dilemma and make a justified decision [13].

In venturing into the exploration of the means by which artificial intelligence's ethics - and the ensuing super ethics - can be attained, it is critical to first undertake the fundamental question "can machines think?". According to Kurzweil [14], the answer is, simply put, no: in his paper, Minds, Brains and Programs, the author uses the famous example of the Chinese Room to disprove the claim that machines can understand what they are being told, maintaining instead that their computational processes are naught but a set of rules being followed but not comprehended. Therefore, he concludes, while the machine appears to understand, in truth, it does not.

While the point of this paper is not to discuss this aspect of machine behavior, I will try to refute Searle's argument as succinctly as possible. In essence, the answer to the question of whether machines are capable of thought boils down to the definition that "thought" is given. From my point of view, thinking is the process through which human beings process information by using knowledge that has been acquired previously. Human beings understand that eating food entails chewing because they have learnt this based on experience.

Through methods like deep learning, Artificial Intelligence is

capable of processing data, altering its algorithms based on a trial and error basis, and process new data using these new algorithms, only to repeat these steps indefinitely and continuously hone its performance. Could a machine then not relate eating with chewing? Would this, then, not be considered thinking? As a matter of fact, do human beings not learn through rules? Does a child not learn to speak, read, and write, amongst countless other things, through rules? Furthermore, there is vagueness in how we define thinking as an actual state. How can we prove somebody is actually thinking? The best proof we can ascribe to thought is a behavior that demonstrates it. Would a machine which behaves as if it is thinking not be considered to be thinking then?

The underlying notion on which Searle's argument rests is that the different parts that make up the 'Chinese Room' - the human in charge of translating the input, the book containing the rules of translation, etc. - do not individually understand Chinese. Instead, they are merely gears that work in unison to give this impression.

And yet, is this not akin to how our brain digests sensory input? Let us briefly examine, for instance, the act of listening. While our ears are capable of picking up and processing sound waves, it would be misguided to ascertain that they understand speech. In a similar fashion, it would be erroneous to ascribe this ability to the neurons within our brains that process the sensory input from our ears by transmitting electrochemical signals.

The list could go on indefinitely, but the bottom line is this: When listening or undertaking almost any mental process, human beings display understanding, or thought. Yet because a single ear or a piece of our brain would not suffice to mentally digest speech, it is fair to claim that it is our system, and not each one of its individual components, which is exhibiting this behavior. Equally so, it is not the duty of human in charge of translating the Chinese input or that of the book containing the translation rules to understand or think about the Chinese symbols being interpreted. Rather, the comprehension of the content is the product of their collaboration: the system - or what is the same, the Chinese Room as a whole. Evidently, a more thorough discussion on this matter is required, but due to the space available on this paper this will have to suffice.

Another concern relevant to the consideration of the practical approach to machine ethics is artificial intelligence and automated systems' limited capacity to take their surroundings into account.

This awareness must transcend beyond mere hardware-based recognition of real-world elements around them and incorporate a deeper, more profound understanding of the consequences of their actions in a real-world scenario. The complication underlying this aspect of machine ethics is that it is not easy to clearly formalize and compute the fundamental effects of actions in these scenarios. Put differently, it is not short of difficult to clearly define key words in ethical considerations such as "good", "bad", "beneficial", "detrimental", etc. and further make a machine comprehend their meaning. In addition, even if we did manage to achieve this latter goal, we cannot be sure that we possess an adequate conception of what these words constitute. Not only is this meaning obscured by basic considerations such as conflicts between ethical theories, but the multiplicity of cultures around the globe and the subsequent variations in what they individually consider to be right and wrong further complicates the quest of granting these terms a computationally-applicable meaning. The dimension of this computational problem is made clearer by the contemplation of how specific "ethical laws" designed to ensure that machines apply their understanding of such key words needs to be. It is clearly not the same to tell a self-driving car to "stop at red lights" than "do not cause harm" [15].

As a matter of fact, the laws hitherto formulated by our legal systems lack a clarity that is essential to dictating an automated computer's behavior [16]. Albeit one might argue that super intelligence entails an inherent comprehension of the world as a whole, we must understand that artificial intelligence, as its precursor and co-creator, does not excel at this to such a high degree. Therefore, how to guarantee that intelligent machines understand key words that characterize actions as "positive" or "negative" and thus act in such manner that maximizes the wellbeing of humans remains a subject in need of study.

A last point of contention in the exploration of applicable machine ethics is the question of whether or not emotions should be an integral part of ethical machines. Despite classical literature's opposition to the presence of emotions in the process of rational decision making, more recent research on the topic labels emotions as an element necessary to making these choices.

Personally, I side with the classical perspective on the matter. While it might be true that emotions play an important role in our decisions, it is precisely the absence of emotions and the bias they lead to which characterizes humans as the inferior ethical actors. Nevertheless, this does not mean that forms of non-human intelligence should not understand emotions, for this is a critical aspect of the evaluation of an action's consequences.

That is, in spite of the fact that it is necessary that machines are capable of understanding the emotions that a human being might feel as a result of a certain action being performed, it would be detrimental to the attainment of an ethical artificial intelligence if emotions actually affected how the decision-making process is carried out. Put differently, it is imperative that an understanding of emotions is taken into account in the ethical calculations carried out by ethical machines, not that a machines' emotions - if at all existent - determine if or how the ethical calculations will take place.

And yet, the greatest problem faced by machine ethicists continues to be the determination of the best ethical theory to incorporate in the automated systems. Much of this field's literature concurs that there is no single one which can be considered to be absolutely correct.

This is arguably due to the fact that all ethical theories and their discussions are subject to the controversial issues described above, and it is therefore no easy task to choose or formulate a single theory that satisfies them all. Notwithstanding, the most practical approaches towards the creation of an ethical artificial intelligence have been governed by utilitarianism and action-based ethics, namely. The former's appeal lies in that it provides a simple method to compute and determine the correctness of an action. By subtracting the pain caused to a person from the pleasure that person receives, the machine could easily make a choice when faced with an ethical dilemma. Furthermore, because the information a machine would require making its calculations is virtually the same as that required by a human being, the formalization of this ethical theory is a relatively straightforward task. According to Anderson et al. [2], however, Utilitarianism cannot be considered an ethical theory appropriate for the challenge faced by machine ethics given that it cannot only violate people's rights, since it is capable of justifying blatantly immoral actions (enslaving the few, for instance, for the benefit of the many) but it also fails to take our notion of justice into account, for it judges actions based on their consequences as opposed to what is just - what people deserve. Action-based ethics, on the other hand, evaluates the action's morality in itself.

As is the case with W.D. Ross' prima facie duties [17,18] - essentially a set of variables that must be taken into account when considering an ethical action - this method of calculation allows the actor to extend his ethical scope beyond the consideration of the pain or pleasure caused by an action and evaluate instead the justifiability of the action itself. Because there is no absolute duty, it is the ethical actor's responsibility to give each duty a specific weight depending on the situation. This makes the ethical theory infinitely more applicable, since it is malleable enough to be used by automated systems in different environments.

A demonstration of the application of action-based ethics carried out by Anderson [3] involved the programming of a system that would require the user to assign the different weights to each duty for a single action. Subsequently, through a series of computations, the program would determine whether the action should be taken or not. According to the researchers, this program could be further enhanced by taking into account the effects of these duties on the different individuals impacted by the action. Furthermore, it was pointed out that the software could potentially be allowed to attempt to make the ethical decisions on its own by assigning weights to the duties autonomously [19].

The researchers could then compare the computer's results with what they considered to be ethical, and "teach" the machine what the correct weights should be. The machine would then relate the correct weights to the specific characteristics of that particular case. As a result, through this process of trial and error, the system would learn to assign the weights in a way that is considered to be ethical for a specific situation, and progressively become better at it. Although this method is an effective and controlled means of formalizing a "human" approach to ethical decision-making, it restricts the "correct" assignment of the weights to the judgment of the researchers. Machines operating in a real-world context, however, would be faced by a whole host of situations where the assignment of the weights requires knowledge that transcends beyond the scope of the scientists' knowledge. In these cases, the development of the ethical program would greatly benefit from the input of experts in the different fields of ethical machines' application. As it is explained by La Chat [15]. "For example, one computer program seeks to capture the medical diagnostic ability of a certain physician who has the reputation as one of the best diagnosticians in the world. The computer programmer working with him tries to break this procedure down into a series of logical steps of what to the physician was an irreducible intuition of how to go about doing it. With a lot of prodding, however, the diagnostician was soon able to break these intuitions down into their logical steps [20,21]. Perhaps this is true with all "intuitive" thinking, or is it? If we assume that ethics is a reasonable, cognitive undertaking, we are prone to formalize it in a series of rules, not exceptionless rules but something like W. D. Ross's list of prima facie obligations: a list of rules, any one of which might be binding in a particular situation."

In order to further facilitate the formalization of human beings' approach to ethical decision-making so as to make it computational, then, it would also be ideal to merge this approach to action-based ethics with the concept of casuistry. Based on the idea of comparison between cases, casuistry proposes that ethical decision-making be addressed by contrasting different situations and their characteristics in order to relatively decide what the best course of action is for a specific case. By drawing an exhaustive analogy between 16th-Century Jesuit Matteo Ricci's Memory Palace, where the storage of memory is facilitated through the mental simulation of a palace with numerous rooms and the attachment of that which one wishes to remember to those rooms and the items contained within them, Searle describes casuistry as the modus operandi of approaching an ethical decision by juxtaposing the case at hand with other ethical situations of the same nature and subsequently comparing their individual characteristics, or circumstances that define them. Put differently, in relation to Ricci's mental edifice, casuistry would amount to walking around the palace's rooms, referring to the ethical decisions or situations, and contrasting their interiors, or particular features/characteristics. As a result, the cardinal perquisite of implementing a casuist approach, Jonsen explains, is that "the ultimate view of the case and its appropriate resolution comes, not from a single principle, nor from a dominant theory, but from the converging impression made by all of the relevant facts and arguments that appear in each of those spaces" [22].

Hence, by adopting a casuist procedure, the machine could potentially be exposed to millions of situations where a human being makes a decision regarding an ethical dilemma that is believed to be morally correct by ethicists. This information could then be processed through refined methods at which machines are progressively excelling such as deep learning. This would facilitate the evaluation of the factors involved in a situation immensely, for instead of having a programmer manually compute all the possible variables that are involved in a single case, the machine could learn to draw patterns between the situations and thus learn to recognize these variables or features in previously unseen scenarios. Anderson [3] and Wallach et al. [23] system, for instance, could learn to form patterns relating the appearance of certain factors in different ethical situations and the weights assigned to each one of Ross' duties for those situations. This way, the presence or absence of one of these variables could translate into a more accurate assignment of weights. Such a pattern recognition sprouting from casuistry would also highly simplify machines' understanding of emotions. By having human beings label the emotions present in different situations and having the machine compare multiple scenarios, the system could be able to better grasp the causes that sparked those emotions and therefore act in a way that maximizes wellbeing. Through the fundamental methodology of comparison that casuistry proposes, therefore, not only would the scope of ethical machines' learning be widened significantly, drawing conclusions from a myriad of real-life cases as opposed to a narrowed research database, but it would also facilitate machines' grasp of a situational factors that human beings subconsciously account for, if not potentially overlook, when making ethical decisions.

I concur with Anderson et al. [2] insofar as the integration of ethical machines in society is concerned. As it is proposed in their paper, entitled Towards Machine Ethics [23,24]:

"We suggest, first, designing machines to serve as ethical advisors, machines well- versed in ethical theory and its application to dilemmas specific to a given domain that offer advice concerning the ethical dimensions of these dilemmas as they arise. The next step might be adding an ethical dimension to machines that already serve in areas that have ethical ramifications, such as medicine and business, by providing them with a means to warn when some ethical transgression appears imminent. These steps could lead to fully autonomous machines with an ethical dimension that consider the ethical impact of their decisions before taking action."

In essence, in suggesting that machines first advise human beings by processing data pertaining to an ethical circumstance and then coming up with a plausible course of action, Yampolskiy [24] are essentially alluding to an augmented cognition of sorts. This approach bears a strong resemblance to the decision support systems discussed by David

Martinez in his paper entitled Architecture for Machine Learning Techniques to Enable Augmented Cognition in the Context of Decision Support Systems. As Treatise and Martinez [19] explains, "The field of augmented cognition facilitates reaching insight after a significant amount of processing is done in the front-end of the decision support system," whose "objective is to drive, via a human-machine interaction, to the shortest decision time with the right amount of data volume."

In other words, the main objectives of decision support systems are collecting and processing data in order to facilitate its understanding, developing models of human cognition that can be extrapolated to machine learning, and providing assistance in decision-making. To do so, the author points out, these artificial advisors first acquire data from the external world through multiple sensors or machine-to-machine communication. The data is then grouped into the appropriate categories, and analyzed through various computational processes in order to transform information into knowledge. Finally, a probabilistic measurement offers possible courses of action to the user and provides numerical estimates of their consequences. If at this point the user feels the decision support system is lacking information, she may ask for more data. This is the underlying basis for the ethical advisor to which Anderson [3] allude.

And yet, the utility of modeling the first ethical machines as decision support systems capable of augmenting and learning from human cognition lies in the Human-Machine Interaction (HMI) that these systems involve. As highlighted by Treatise and Martinez [19], the corrective feedback the user provides to the machine is critical in order to make probabilistically-reached decisions more accurate and minimize false positives or false negatives. This supervised learning would play a key role in the improvement of machines' ethical decision-making. Furthermore, as ethical machines become more autonomous, their understanding of human cognition and behavior should also be increased. Therefore, it would be ideal to integrate a degree of collaboration between the user and the ethical advisor in Anderson et al.'s gradual approach. As noted by Miller and Ju [20], there are many benefits to reap from the cooperation between human beings and automated systems. While the former excel at handling novel situations, the latter are superior when it comes to executing preset actions given a determined set of inputs. To do so, it would be imperative that the user be predisposed to act ethically and abide by the predefined moral standards that we seek to make machines understand. It would also be necessary that an effective communication between the machine and the user be established, whereby the automated system can understand human beings' mental approach to ethical decision-making. Not only would this allow the automated system to gain a "powerful extra dimension of capability". According to Miller and Ju [20] but it would further allow the machine to learn what the user takes into account when facing ethical decisions.

Moreover, as Miller and Ju [20] explain, both the user and the machine must possess a clear notion of each other's roles in this cooperation:

"The necessity for the computer to hold a model of the [user], and for the [user] to hold a model of the computer presents a design challenge—designing understandable systems and feedback mechanisms so that the two entities can truly share control. With sensors and machine intelligence enhancing the capabilities of the [user], and backstopping human failings, and with human intelligence expanding the capabilities of the automated systems, the two can be considered to extend or expand each other's capabilities."

Were these prerequisites to be met satisfactorily, such collaboration could potentially improve human beings' ethical decision-making capabilities in the short run thanks to the provision of relevant data and, in the long run, enhance machines' understanding of human beings' ethical notions, facilitating their supervised learning of ethics.

By adopting Anderson gradual approach in the integration of ethical machines that abide by action-based, casuist machine ethics, and first structuring these automated systems as decision support systems intended to enhance human cognition, not only would humans be exempt from having to judge machines for their actions, since it would ultimately be humans who would be making the decisions, but this gradual process of integration would also grant us more control with regards to the real-world scenarios that machines are exposed to and translate into short and long term benefits. This controlled exposure, then, would further enable us to collect data on ethical machines' behavior in real-life contexts, allowing us to hone the ethics of artificial intelligence and, in the future, its infinitely more powerful successor.

The Implications of an Ethical super intelligence

The attainment of an ethical super intelligence capable of perceiving the world in a manner akin, if not superior, to that of its creator for the very purpose of giving him counsel is, in truth, a rather disturbing thought. And yet, from a historical standpoint, such a pivotal cataclysm seems all but predictable: throughout its existence, humanity has not been obedient to a single entity, or held the word of a single entity to be true, but has instead progressively transitioned from the worship of one entity to another, gradually detaching itself from the realm of the ethereal and moving on to that of the physical - while the ancient Romans worshiped their Gods and granted them responsibility for the occurrences of the world around them, and the Renaissance bequeathed this accountability to man himself, it now appears as if it is the oncoming technological era of the Singularity which will bestow this power to man's creation: technology. Now, as if the Roman Gods had created man for the sole purpose of yielding them their will, man is at the brink of a revolutionary epoch in which it will be advised by the product of his intelligence. Super intelligence, however, is going to evolve. Whatever entity it is we manage to contrive will enhance itself exponentially. Merely thinking that, at some point or another, we will be advised by an entity too complex for even us - its creator - to understand is unquestionably frightening. Will it actually behave ethically, then? We have hitherto addressed the query of how to make super intelligence as ethically right as possible. And yet, what would happen if this human ethics-bred superethics turns out to be 'righter' than its creator's ethics? Put differently, what if superethics and human ethics turn out to disagree? Who would have the last word? Man, or machine?

In order to properly address this question, we must first scrutinize the similarities between human beings and super intelligence. To do this, I will try to refute different claims aimed at differentiating them. In doing so, it is not my purpose to claim that machine is man's equal, but rather to further raise the question of whether the creator and its creature truly are blatantly distinguishable.

Referred to as the "Bright line argument" by Moor [21] in his paper entitled The Nature, Importance, and Difficulty of Machine Ethics, this claim states that only full ethical agents can be regarded as ethical agents - agents capable of making reasonable, justified ethical decisions. However, as the author himself goes on to explain, this assertion is misguided for two fundamental reasons: First and foremost, it implies a disregard for other lesser types of ethical agents, such as implicit (an autopilot system on a plane that has been programmed to take its

passengers to the correct destination) and explicit (a machine capable of making a choice when faced with a controversial ethical dilemma) agents. Albeit not as proficient in resembling a human's ethical decision-making process, these agents nonetheless clearly display a form of ethical behavior that must not be undermined. Secondly, in response to the allegation that, since consciousness, intentionality, and free will are the key characteristics of full ethical agents, or human beings, Moor contends that, even though non-human intelligence might fall short of exhibiting these traits, there is no empirical evidence to dispel the claim that the reality might not be otherwise at some point in the future. Furthermore, I would dare affirm that super intelligence will, in fact, possess these features. From machine ethics' theoretical standpoint, the foundations for this seemingly illusory ethical accomplishment are presently being laid: consciousness will be given to machines because, albeit computational, the incorporation of ethical programs in these systems will grant them an awareness of the consequences of their actions; intentionality is the very groundwork of machine ethics' theory, for at the very crux of the field's research lies the objective of having machines intend to minimize harm and maximize wellbeing; and lastly, these automated systems will also be furnished with free will, for they will choose how to act in every situation. Admittedly, their choice will be limited to a set of ethically-sound alternatives, but they are being given a choice nonetheless. Moreover, as LaChat [15] puts it, "If free will is real in some sense, there is again no reason to believe that it might not be an emergent property of a sophisticated level of technical organization, just as it might be asserted to arise through a slow maturation process in humans. I should also add that not all AI experts are convinced an AI could not attain free will."

Another argument intended to highlight the differences between human beings and machines is that of their supposedly different learning processes. Specifically, the argument states that human beings and machines cannot be regarded as equivalent ethical actors given the dissimilarity in the way in which they grasp ethics. While the former largely learns "moral rules by osmosis, internalizing them not unlike the rules of grammar of their native language, structuring every act as unconsciously as our inbuilt grammar structures our sentences" the latter would just require a chip containing an ethical program in order to operate ethically. Therefore, the argument goes, machines do not possess the profound understanding of the world around them that is imperative for adequate ethical decision-making. This latter claim - notwithstanding the truthfulness of Hall's previous assertion - can be refuted with the following observations: firstly, modern machine-learning algorithms, such as deep learning, literally enable machines to learn from the analysis of previous experience. Through a cyclical procedure of trial and error not unlike that proposed in the previous section, involving the combination of action-based theories, casuistry, and corrective feedback, machines theoretically could learn and be taught to act as an ethical human being would. Therefore, maintaining that ethical machines would not possess a gradually honed perception of their surroundings is erroneous. The fact that a machine's learning process would incorporate the in-depth scrutiny of millions of diverse scenarios is clear proof of the contrary, and could further support the claim that these systems' perception would be superior to that of human beings.

As a rebuttal to this reflection, it would be tempting to assert that humans, unlike machines, are aware of contextual factors that transcend beyond mere evaluations of their physical surroundings and englobe traditional and cultural beliefs that have a potential impact on ethical reasoning. In other words, as it is put by psychologist Lawrence Kohlberg, "situational factors are extremely important in moral action," for in many cases peer group and institutional shared norms may be moral or nonmoral in their content." Hence, one might contend, machines will never attain the moral reasoning that is characteristic of human beings. My response to this claim is simple, and not unlike that of Moor, which was presented earlier: there is no way to prove that this will not be plausible at some point in the future. As a matter of fact, alluding to the action-based, casuistry-guided, HMI-driven ethical approach outlined earlier, it seems conceivable that artificial intelligence could eventually learn to distinguish these cultural trends and take them into consideration when choosing an ethical course of action. Furthermore, bearing in mind the computational power super intelligence is deemed to possess, it is all the more believable to asseverate that it will excel at doing so.

And yet, it remains an insurmountable truth that a machine will never truly be man's equal. Although the ethical behavior of the former might bear a strong resemblance to the latter's - as I have tried to point out in the previous paragraphs - I do remain an adamant proponent of Luzac's publication entitled Man More than a Machine (1752), which stresses the differences between both creatures by dispelling any claims that might assert otherwise [18].

Indeed, there are in fact notable dissimilarities between human beings and artificially- intelligent entities, as it was explained in the first section: machines, unlike their counterpart, are exempted from being misguidedly swayed by emotions when making ethical decisions. Whilst a program comparable to that which was proposed previously would grant machines a comprehension of the emotions relevant to the evaluation of an action's impact, this fundamental understanding is central only to the computational process carried out by the machine, not the structure of the process itself. For these same reasons, machines are exclusively capable of overcoming the forces of self-interest and common sense. Furthermore, machines are not subject to the Law of Conscious Realization, whereby moral action precedes and catalyzes moral thought. This translates both into man's arguably innate tendency towards moral, ethically-correct action versus machines' increased reliability as far as ethical behavior is concerned, for the implausibility of the latter to set action before thought ensures that ethically-adequate thought will be followed by equally suitable behavior. Lastly, a stark difference between human beings and an ethically- correct super intelligence lies in the degree of awareness and therefore accuracy that the machine would manage to attain as a result of its computational power. Hence, coupled with an adequate action-governing, ethical program, the awareness of this elevated number of variables when carrying out the decision-making process implies a pronounced superiority of super ethics over more rudimentary human ethics.

It would appear, then, that not only are human beings and machines utterly clashing, but the latter's ethical dimension appears to be superior to that of the former. In other words, albeit super ethics is unquestionably different from human ethics, it does, at least mildly, come across as the better form of ethical reasoning. And yet, does this then mean that man is inexorably bound to listen to super ethics, holding its mathematically-wrought counsel to an unparalleled regard? Put differently, if the ancient Romans were to be told by our infinitely different, more evolved and arguably more knowledgeable, modern society that slavery, given its unethical justification, should be abolished in its entirety, ought the Romans to pay heed to our advice, or turn a blind eye to our counsel, adamantly convinced of the ascendancy of their knowledge?

The point I seek to draw with this analysis is not that machines are

superior to man, nor is it that, in consequence, human ethics should be subordinate to super ethics. Rather, my intention is to underline the question of who would ultimately be right by pointing to the flaws inherent in the seemingly obvious yet misleading answer machines are created by man, and thus it is man who determines what is ultimately right. In lieu of this evasive retort, this examination proposes the further consideration of the question through further studies of the parallelisms and dissimilarities between human beings' and machines' ethical reasoning. In any case, supplementary analysis is urgently in place, for although the correct answer to the unsettling question "Should machines have the last word?" remains unclear, the potential reverberations of the wrong one augur nothing but the onset of a somber, apocalyptic calamity.

Conclusion

The seemingly fantastical thought of attaining a viable, adequate code of ethics for a futuristic super intelligence is, in conclusion, not entirely surreal. Rather, by discussing the plausibility of materializing these ethics in an artificial form of lesser intelligence, it seems like machine ethics has laid the groundwork that may enable us to deliberate on this subject. While the exactly correct practical approach is yet to be determined, future research must fail not to be mindful of the various complex requirements that have been outlined in this paper, for the adequacy of such systems depends on whether or not they are met satisfactorily. Notwithstanding, in our pursuit of the correct ethical program that will govern the actions of a super intelligent entity, we must not wander away from the equally impending considerations of what implications such a machine, or its possible disagreement with man, might entail.

The path towards the attainment of a feasible, safe machine ethics is an obscurely long and winding one, and albeit academia's efforts have granted us the knowledge to steadily tread it, there is still much to be done. The furtherance of research governing the practical application of ethical theories in machine ethics is in place. In spite of the fact that some authors claim that materializing such ethical software without first concurring on a single, correct ethical theory is unbecoming, I disagree. I sincerely believe that the development of digital software designed to make a machine "think ethically" will not only enable us to expand our understanding of advanced artificial intelligence's computational interpretation of the real world - hence contributing to the arduous development of appropriate computational structures - but it will also allow us to carry out an assay of the suitability of different ethical theories or even variated combinations of them. In advocating the progression of such a hands-on approach to machine ethics, however, by no means is it my purpose to discredit its less practical counterpart. Quite on the contrary, I hold the philosophical deliberations of machine ethics in the highest regard, for they lie at the core of the field's purpose. At the same time, however, I am of the opinion that, given the fact that both the practical and theoretical dimensions of machine ethics are intrinsically intertwined, a reciprocal collaboration between both would be all but greatly beneficial to the field as a whole.

To conclude, in order for the benefits of super intelligence to be reaped by society, it is imperative that its code of ethics be developed in parallel to super intelligence itself. And yet, what if super intelligence is never materialized? Little of the effort put into this scientific-philosophical endeavor will be lost. To quote LaChat [15] "To the contrary, the failure (…) might eventually bring us to the brink of a mysticism that has, at least, been partially 'tested.' Would it be more mysterious to find intelligent life elsewhere in the universe or to find after unimaginable aeons that we are unique and alone?" The more menacing question, hereafter, is: what if super intelligence is materialized before we manage to formulate its code of ethics? The answer, I am afraid, is deserving not of an elaborated academic discussion, but rather the eeriest science fiction novel, and while I am no prolific creative writer, my best guess is that it ends with a metallic, mathematically-palpitating and inert automaton, as cold-blooded as Isaac Asimov's Dr. Susan Calvin, prosaically reciting, "Veni, vidi, vici" while the lifeless remnant of its creator's existence silently cries "et tu, Brutus.

References

1. Allen C, Wallach W, Smit I (2011) Why machine ethics? Machine ethics pp: 51-61.
2. Anderson M, Anderson SL, Armen C (2004) Towards machine ethics. American association for artificial intelligence.
3. Anderson SL (2011) Machine metaethics pp: 21-27.
4. Asimov I (1986) Robot dreams. Byron press visual publications.
5. Bergman R (2002) Why be moral? A conceptual model from developmental psychology. Human development pp: 104-124.
6. Bostrom N (2002) Ethical issues in advanced artificial intelligence. Oxford: Oxford University.
7. Bostrom N (1998) How long before super intelligence?
8. Bostrom N, Yudkowsky E (1998) The ethics of artificial intelligence. Cambridge handbook of artificial intelligence. New York, United States of America: Cambridge University Press.
9. Gips J (1991) Towards the ethical robot. The second international workshop on human and machine cognition: Android epistemology. Boston: MIT Press.
10. Goodall NJ (2014) Machine ethics and automated vehicles. road vehicle automation in switzerland: Springer international publishing pp: 93-102.
11. Hall JS (2011) Ethics for machines. New York, United States of America: Cambridge University Press pp: 28-44.
12. Hibbard B (2015) Ethical artificial intelligence. Bill Hibbard p: 3.
13. Kohlberg L (1987) The development of moral judgment and moral action, a cognitive-developmental view. New York: Longman.
14. Kurzweil R (2016) Super intelligence and singularity, Schneider, science fiction and philosophy: From time travel to super intelligence pp: 146-170.
15. LaChat MR (1986) Artificial Intelligence and ethics: An exercise in the moral imagination. The ai magazine 7: 70-74.
16. Legg S (2008) Machine super intelligence.
17. Lin P, Abney K, Bekey G (2011) Robot ethics: Mapping the issues for a mechanized world. Artificial intelligence pp: 942-949.
18. Luzac E, de La Mettrie JO (1752) Man more than a machine.
19. Treatise W, Martinez OD (2014) Architecture for machine learning techniques to enable augmented cognition in the context of decision support systems. Foundations of augmented cognition: Advancing human performance and decision-making through adaptive systems: 8th international conference, AC 2014, held as part of HCI international 2014, Heraklion, Crete, Greece, Proceedings pp: 148-156.
20. Miller D, Ju W (2015) Joint cognition in automated driving: combining human and machine intelligence to address novel problems.
21. Moor JH (2011) The nature, importance and difficulty of machine ethics. New York, United States of America: Cambridge University Press pp: 13-20.
22. Searle JR (1980) Minds, brains and programs. The behavioural and brain sciences pp: 417-457.
23. Wallach W, Allen C (2009) Moral machines. Oxford University Press.
24. Yampolskiy RV (2013) Artificial intelligence safety engineering: Why machine ethics is a wrong approach. Philosophy and theory of artificial intelligence pp: 389-396.