REINFORCEMENT LEARNING SYSTEM FOR A MORAL AI AGENT

BORIS D. GROZDANOFF[1]

Institute of Philosophy and Sociology, BAS

Defence and International Security Institute (DISI)

grozdanoff@gmail.com

Abstract

In 2015 Savulescu and Maslen argued, influentially, that a moral artificial intelligence (MAI) could be developed with the main function of moral advising human agents. The suggested MAI would monitor factors that affect moral decision making, would make moral agents aware of biases and would advise them on "right course of action". Effectively, since the MAI is an AI, it would "Gather, compute and update data to assist human agents with their moral decision-making." The data would encode information about agents and environment, moral principles and values. In this paper I critically analyze the suggestion of MAI and I argue that in the contemporary domain of digital technology and AI driven industry a version of MAI can and should be developed for practical purposes. My main argument is that even if, as Savulescu stresses, "a single account of right action cannot be agreed upon" this should not prevent us from harnessing the computational and reasoning power of AI in order to regulate the exploding AI industry. For this purpose, I suggest a modification of MAI – from a mere moral advisor to a full blown Artificial Moral Agent (AMA), capable besides MAI's gathering, computing and analyzing data for moral advice to real time moral analysis of complex real-life situations (like one involving autonomous vehicles, drone swarms AI led local warfare and AI led nuclear warfare) and choosing an optimal moral decision as well as instructing the AI system to execute it in real time. I also suggest the encoding framework of AMA, based on multidimensional graphs, that would be able to quantize the acquired moral relevant data and moral reasoning engine, operating rationally on models, rendered from the data.

Keywords: Ethical AI, Reinforcement learning, neural networks, moral artificial agent, systematic ethics.

In 2015 Savulescu and Maslen argued convincingly that a moral artificial intelligence (MAI) could be developed with the main function of morally advising human agents.[2] Their proposal is one of the first practically oriented solution suggestions in the broader context of the so-called ethical AI (EAI). The EAI, however, is most of the time framed in terms of an AI being ethical and much less often, if at all, grasped as an engine for ethical comprehension, advice, or actions based on such advice. The *status quo* both in the academic world and the social layers of government, including states' and industry management, is exclusively focused on defining a *set of norms*, widely accepted by policy makers and professional academics, as ethical. This set, as we may well see in the near future, being aware of non-trivial differences that regulate interests and goals of different states,[3] is to be employed later in programming the actions of AI-based and AI-run electronic systems at large. Foreseen and unforeseen consequences for the life and, according to some, even survival of humanity, could hardly be overestimated; they are going to be anything but trivial for the future of human society.

Thus, it is an important and time-sensitive task to speed up the formulation of both theoretical systems of ethical AI and their practical implementation. Savulescu's and Mahlen's suggestion of MAI follows this trend in a novel way. While most professional ethicists would perhaps agree that in theory an AI system could contribute to the centuries old philosophical discipline of ethics, few of them, if any, would accept that it would *outperform* humans in this task. I, having a tremendous intellectual respect for the achievements of history of ethics, do disagree, however, and perhaps along the same line of the proposed MAI. Where I differ is in the *modesty* of the MAI proposal, quite apt for the normal tempo of evolution of academic ideas. Such modesty, however, follows the natural speed of evolution of academic ideas and not at all the pace of modern technological development; thus, it is inadequate as a timely response to the pressing need of yesterday's AI, which is fully disinterested in the fact of the presence of an ethical system for AI, or the lack thereof. In what follows I will offer, firstly, a critical overview of the ethical AI theme, and secondly, a detailed proposal for an ethical AI agent.

Unified Ethics, Codification and Algorithms

The modern academic discussion on MAI is diverse and rich. I will focus on a particular branch of skepticism[4] towards MAI, which varies between technological impossibility and moral inferiority of MAI compared to the morality of humans.

We need to approach this argument along the length of its reach. It is beyond doubt that ethics is not merely not "completely codified" but there is also, uncodified and unrelated, a myriad of ethical views, principles, theories, values and other ethical components, that are in rivalry among each other. There are many ethical islands opposing each other in direct contradiction of main claims or principles, and this naturally prevents any sort of codification that would prefer to use logic as an instrument, let alone mathematics. The reasons for the lack of “complete” codification are largely three: no universally accepted ethics, logical tension between most ethical positions and theories (eventually candidates to be codified), and lastly, complete lack of perspective regarding the most apt candidate among the ethics rivals. However, we need to position this argument within the practical reality of *ethical human practice* given *actual AI and machine presence*, which engulfs many important aspects of our lives, the key one being, certainly, the ethical one.

The first fact that this argument must face is the fact that, universally accepted or not, codified or not, for centuries humanity has relied on certain principles when behaving and reasoning ethically. Thus, we must accept that humanity has never relied exclusively on *doctrines* in order to be ethical, with a layer of exceptions for the religious ethical doctrines, which both in Eastern and Western cultures have managed to influence the ethical beliefs and behaviors of enormous groups of people across history. The philosophical discipline of ethics, at least in the Western world, has chosen long ago to use human reason as an instrument in the ethical endeavour. It is due to this reason mainly that the segmented, both ideologically and historically, Western ethical positions, have never reached the statistical success of the religious ethical positions. We need not at all, of course, to investigate the depths of intra-ethical tensions in order to realize that they bear virtually no weight for the majority of humanity, or to realize that human behavior is seldom governed by *any* particular ethical theory, no matter how superior it may be to other ethical theories. That is why we need to approach the task before machine ethics in the real setting it takes place in, and not purely on a theoretical level.

We live in a world operated more and more by myriads of machines and the questions of the ethical aspects of these operations are certainly non-trivial for humanity. But we must not impose greater requirements on machines than we have tried and failed to impose on ourselves: the very humanity that worries more about machines being unethical to it, than about humans being unethical to themselves. This is a sort of a demonstration of an elevated self-image which positions humans as the natural leader in the determination of ethical values, principles and the like. Not, say, animals, and not machines. I will try to disagree. Let me illustrate. Consider a simple calculator machine, with a hardware and a software, that performs a very limited number of operations. Now let us add a simple principle of an agreed upon ethical nature to the firmware of the calculator. Once we have encoded the principle, no mather its content, and the operational system of the machine is compatible with its execution in its very limited ethical world of addition and square rooting, the calculator would meticulously execute its program and thus would *follow* the instructions of the principles.

Amazingly, there is no single ethical principle that a human being is able to follow in this way. We would always reason, doubt its content, its ethical success, the domain of application and what not; only to find an occasion when we would decide that it is more ethical *not to follow the principle* in a situation that by definition requires its execution. The natural reaction would be to find shelter in human naturalism, perhaps, and more scientifically, in evolution: we are humans and thus rational beings that reason and do not follow principles blindly, unless they pass the strictest of our rational tests. But is not the strive to codify ethics "completely" exactly this? To arrive at a set of principles, theses, values and the like that are universally accepted as ethical, and to follow them collectively? And only then to request from the machines we have created to follow them as well? It feels a bit improper.

But even this side note is not particularly important for elucidating the problem of artificial ethical agents. We need to distinguish between *autonomous machine ethical agents*, which by definition are able to arrive at *novel* ethical decisions not "codified" within them and to implement them in real life scenarios (e.g. autonomous vehicles, presented with one of the potentially infinite line of the notorious trolley problem varieties), and *mere artificial ethical agents*, which, much like the calculator from our illustration, blindly follow the rules encoded in them by human programmers, informed by professional ethicists. Novel ethical decisions will benefit from the enormous computational power of AI, that, given a suitable system architecture, would be able to find the optimal ethical algorithm of solutions both in a narrow and in a general ethical sense. For this to happen, a few conditions must be satisfied. First, we need to distill the normative problems into a quantitative formulation; a mechanism for this I suggest below. Second, we need a modular proof of concept AI system that can demonstrate following a rule and establishing normative priority between rival ethical scenarios; the suggestion below demonstrates this as well. And thirdly, we need to integrate the modules into a general normative AI system that would be able to find ethical priorities in real life complex situations, and *prove this priority* mathematically. This is a task that, with sufficient work on the first two, can be expected to be solved successfully. An artificial ethical intelligence would be able, given its initial embedded ethical structure and large amounts of data to train, to find an ethical solution for a given action, based on a no-nonsense rational procedure, and executed on computational machine like modern computers.

Artificial intelligence cannot wait for the emergence of academic consensus on what is ethical and what not. The industry embeds AI technologies much faster than ethical consensus can emerge in contemporary international ethics. While academic discussions on duties, obligations, discrimination and many other points of interest for today's digital ethics continue to take place, China has been deploying (for years now[5]) a truly global AI system of "Social Credits", which most of humanity, and not at all just liberals, sees as an Orwellian prophecy come true. Or, to give another example, as discussions on what is discrimination and what is not slowly protract their more or less green horns in the darkness of the unchartered space before the reason of the would-be ethical agent, military headquarters around the world have already deployed AI-embedded systems of warfare, whether in the ICBM domain or in the 6th generation unmanned autonomous jet fighter swarm domain. Their systems are not at all interested in questions like whether it is permissible, let alone bad, to kill the entire population of a city, of, say, a few million residents, completely *indiscriminately* of any feature of theirs, be they social status, race, color of skin, gender, wealth and the like. Such systems, which would actually gradually replace human-led warfare decision in favor of AI-made warfare decisions, would merely *calculate* the *quantitative* effects of adversarial scenarios. And if the figures go one way and not another, missiles and drones would rage war that is much more real than the abstract academic worries of otherwise completely legitimate ethical problems.

Sometimes science needs to conform to non-scientific reality, if only because the alternative would be undoubtedly and enormously worse. The AI explosion in technologies, both narrow (like image and pattern recognition[6]) and general (like Deep Minds' latest impressive attempts to lead fully adequate conversations with humans[7]), demonstrate that modern ethics no longer has the luxury of acting slowly. Instead, if it is to participate in the chain of events, it would need to develop a *practically implementable ethical system for AI*, such that no matter its theoretical bias, if any, it would *measurably improve* the chance of well-being and survival of individuals, groups and humanity as a whole. In the fully conceivable case where such a human-created ethical system were to lead to the demise of humanity, we can at least find a bitter consolation in McMurphy's line from *One flew over a cuckoo's nest*[8] where he attempts an obviously impossible physical task and fails: "At least I tried".

If humanity reaches its demise because it never even attempted to control the technologies it develops, or because that attempt failed, it would not matter much. But in the attempted scenario humanity would have at least tried to use its rationality for its own evolution and survival. And this should count for something, even if it is cosmic alien archeologists who discover the ruins of our wise hi-tech civilization after millennia.

This difference in pace between the evolution of AI technology and socially accepted ethical consensus is anything but trivial for human society; both local, in the shape of state-defined society, like the ones in the UK, USA, Germany, Russia and China, or within state-segmented societies, like certain state academic, art, juridical and the like societies (or layers of state society, as it could be), or global, like an international academic society or international industrial society. Historically, ethics has evolved much like the rest of the philosophical disciplines driven by reason, intuition, culture and other contributing factors. But rarely, if ever, has a philosophical discipline been driven by an industry or the effects of an industry like it is by AI technologies today. Ethics for AI needs to catch up with AI, and fast.

A common skepticism regarding AI debate that sooner or later touches upon matters of machine doomsday for humanity, whether in a “mild” version, like the one presented in the original Matrix movies trilogy,[9] or in a “harder” version, like the one presented in the Terminator movie franchise,[10] can be summed up as follows: there is no rationally demonstrated reason to neither accept nor even anticipate an AI doomsday scenario and therefore all worries about such possible paths of machine evolution do not deserve the hype they receive today. Some of that hype, however, does not come from science fiction writers like Isaac Asimov or Robert Sheckley, or even from multimillion-dollar movie directors like David Cameron and the Washovsky sisters. Some of the worries come from successful industrial managers like Elon Musk or physicists like Stephen Hawking. But none of these remarks could form a good enough reason for an academic to *deduce* the potential dark prospects of an AI doomsday.

Another one, albeit pretty formal in nature, can. Here is my version of it. AI is a fact of everyday life, and in its evolution it looks much more like a revolution and an explosion than like the familiar gradual century-by-century adaptation. AI and the AI-based industry have already become factors of the environment of humanity, therefore we must inquire about our adaptation to them.

As a rational inquiry into this combined system, AI and humanity need only ask a couple of questions in order to establish whether the worry about human survival in preconceived doomsday AI scenarios deserves academic and practical attention. Here is a logical argument, modal as it is and one of many possible, which seems to me to be valid and also sound:

1. AI is a fact

2. AI evolution from narrow types to more general (and eventually full generality) is a fact

3. AI evolution is impressively fast as of 2020

4. There is no known reason that AI evolution would stop or even slow down significantly

5. AI has the potential to drive and manage virtually all digital technologies which employ significant amounts of data

6. All digital technologies progress towards the acquisition and management of huge amounts of data (the so called Big Data realm of the digital industry)

7. Management of digital systems by AI technology is growing steadily

8. Digital systems manage virtually all aspects of modern human life, more directly in the developed world and less directly, but still rather effectively, in the developing world (AI-driven warfare would undoubtedly affect the developing world; AI-developed drugs or foods and AI-denied communications or resources, or AI's poor management of those would certainly affect the developing world as well)

9. *Therefore*, it is conceivable that AI, in its revolution, could take on a path where it would separate its evolution from the control of humanity.

10. If (7), we have no reason to accept that the paths of AI evolution and human evolution would coincide

11. They *might* deviate (11a) or even intersect (11b) and collide head-on (11c).

12. Because of the possibility of 11's a, b and c, humanity needs to act proactively and not *ex post* for its own interest, evolution, well-being and survival.

It is this line of reasoning that sheds light on the contrast between the standard pace of evolution of socially accepted ethical norms, values, duties and rules, and the revolution of AI. Savulescu's and Mahlen's proposal is directed towards overcoming this contrast. For it is all too clear that academic practice would not change its pace, and this is a no-go for the pressing AI ethical vacuum problem. The MAI suggestion is presented, again, completely in line with common academic pace. In reality, society needs the programming of the MAI, as an initial stage, to be conducted under the expert supervision of professional ethicists, but *following the industry line of production pace rather than the academic one*. Effectively, this amounts to the interdisciplinary work of a team of coders, project managers and ethicists in a full-day, IT-style working environment.

In what follows I would suggest a different and hopefully more practically feasible proposal. I will suggest that what society needs is not the modest and otherwise potentially useful (if successful) MAI. Instead, *what society needs today* is an Artificial Moral Agent (AMA). AMA is a full blown AI system, developed, trained and deployed for ethical data acquisition, ethical analysis, ethical decision-making and ethical decision-execution (EDE), where the EDE is physically being performed by an actual physical AI system governed by AMA on a structural level. AMA is not to be a side application to the operational AI system. Instead, in order to be truly ethical, the AMA should be developed as an integrated structural part of the AI System and in such a manner that if AMA does not operate the AIS does not operate. This is the only way we can have a theoretical guarantee that an AI, hopefully an AGI as well, would be ethical in all its actions.

Sample Deep Neural Network with Reinforcement Learning

From the available multitude of artificial neural network architectures, I chose as a most promising approach the one which uses deep neural networks (DNNs) and the highly successful methods of reinforcement learning (RL). Compared to traditional machine learning algorithms like k-means, random forests and decision trees, the DNNs have a much greater learning capacity and a much higher number of learnable parameters. Both are essential in solving complex AI tasks, like building a functional ethical AI system, harnessed by the power of DNNs. The task here, however, is not to suggest a conceptual model for Savulescu’s MAI, but to suggest a conceptual model for an AMA. The system, when built, trained and optimized, would be able to act and not merely to suggest, given its embedding into an actual hardware, e.g. an autonomous car.

The input to a MAIA can be the ordered set of a *normative pattern* (*np*). I introduce the *np* as an *n*D analogue of raw 2D visual images, comprised of pixels with defined properties like color, brightness, hue, saturation, sharpness and others. The reason for the introduction of the *np* is methodological: while images are the exemplary test bed for RL deep neural network systems, typically built for image recognition, few people within the AI community, especially among the AI architects who build, train and test the code, are familiar with the normative environment in ML and DL technologies. The introduction of *np* as analogy shapes the normative environment in a familiar quantitative form that can be immediately approached by the AI architects. Where ethicists talk about good and bad, the normative pattern would be the mediating component that extracts the quantitative features of normative agents, states, actions, values, properties and rules in order to form a quantitative representation of the normatively relevant environment (NRE).

The ordered set of an image would be mirrored by the *np*. The differences, no matter how non-trivial, would be differences in principle and mostly, richness. For example, while in an image every pixel has a two-value unique address {x:n, y:m}, and certain properties {r:, g:, b:; h:, c:, b:, s:, etc. for color specification, brightness specification, hue, contrast, etc.}, in the *np* we will have an n-dimensional hyper shape that would provide an n-number address for each normative "pixel", that is, an element of the *np*.

What remains as a task before the AMA architect is to encode all relevant normative properties that would comprise the elements of *np*. One difference, obvious at the very outset, is that while all pixels in an image eventually differ in values for all its properties, all pixels are *of the same kind* and thus are one and the same element of the image. Patterns learnt by, say, modern convolutional neural networks, are patterns built by pixels. In the *np* setting, however, we do not have the luxury of a single normative element. Instead, we have a variety of normative elements that can possess properties, which are the carriers of the values to be reflected in the indices for them in the *np* ordered set, or sets, as it were.

Thus, an element for the normative environment is a normative action, normative effect, normative agent, normative consequence, normative history, normative behavior, normative rule (like "do x", "~do y"), and others. That is why every *np* developed for an AMA should fall into one of two categories: a narrow one, with a custom selection of normative elements of its *np*, or a general one, with a rich selection of normative elements. In the simple concept case presented here, I will show that a DNN can be built to learn a very simple normative behavior, which is comprised of the general normative elements “kill” and “do not kill”, say, in the specific action of an autonomous car where kill would be taken as equal to “run over a pedestrian”. Both elements are at the end of concrete actions of the AMA, varying in the normatively relevant environment for the purpose. Thus, for example, if the kill or not kill actions are to be learnt in an ideal simplified environment, which only has a single driver, a road, sidewalks and pedestrians, we will have 4 kinds of normative elements to build our *np* around and thus we will have a 4 dimensional hyperspace for our elements. In it, however, we can, besides giving unique 4D addresses to each one of them, express all encountered elements of this kind in the wild: this includes all roads, all sidewalks, all pedestrians, all drivers. In a richer, more sophisticated system, we need to delve deeper into the properties of those, and also into their kinds.

For example, the properties of pedestrians are their position, speed, direction, size, visibility, etc. Their kinds are the likes of kids, elderly, sportsmen, joggers, etc. The AMA would be taught which element is a road or pedestrian, which kind it is, what is its characteristic behavior and only then what its actual real time behavior before the radar-sensoric “perceptual” system of the car is. Then the AMA would be trained to learn numerous sets of actions of the moral agent, i.e. the car’s real life autopilot. Those are limited to the hardware-related actions of the car that include brakes, turns, stops, acceleration, horn, and others. The sets can be permutated pre-training into properly segmented data sets, to be fed into the input layers of the AMA upon training. Eventually, the AMA, after hundreds of thousands, perhaps millions or even billions of phases of training, would learn which sets of actions lead to a killing within a particular setting, and which do not. After having achieved a success rate of above 98%, widely considered a good one, the AMA would most probably be a more ethical driver than the vast majority of human drivers; not due to malevolence on their behalf, but due to its sheer technological and informational superiority. A well-trained AMA would be a much safer driver than even the best of human drivers. As an illustration, we can take the very recent AI jet flight simulator, which managed to beat an expert F-16 pilot[11]. If we focus on the ethical aspect of this “game”, we can see that well devised AI agents can surpass humans in many ethically relevant domains.

The remarkable success of RL in projects like AI sector leader Deep Mind’s more recent ones[12] does not need advertising. What is more important for our purposes here is the suitability and potential of RL for our task of building an ethical AI agent. I would like to stress a couple of reasons for choosing RL for the task. The success of RL methods in games like Go and Atari is nothing short of remarkable and counts as a good enough reason. But the structure of RL, which approaches the tasks through the perspective of its key components, an “agent” and an “environment”, is extremely apt for our purposes. Seeing as AMA needs to be developed as an ethical “agent”, the extraction of the “advice” functionality, needed for a MAI advisor, is nothing but a trivial downgrade of the functionality of the AMA. In RL the environment is the “world” within which the agent figures and functions. The functioning of the agent is expressed as a formal interaction with the environment, which can take the form of either passive observation or an action that *aims to change said environment*. Interaction is discrete and structured in a distinct finite number of steps. At any given step the agent “observes” the current state of the environment and “decides” to act (or not to act). After an action has been performed by the agent, the environment changes from a state S_n to a state S_{n+1}, but it can also change without any input from the agent as well.

At the heart of the RL methods is a reward signal. This signal travels from the environment towards the agent and essentially carries information about a predefined hierarchy of the world, typically cashed in terms like “good” or “bad”. It is important to note for the skeptics of MAIs, that it is this difference, highly non-trivial from the point of view of virtually any modern theory of ethics, that is *best captured via quantification,* i.e. expressed by numbers. If one were to point at the surface-level where the allegedly unquantifiable normativity becomes quantifiable, it is this difference in states of the agent’s world that is best captured in numbers. Thus, the agent receives the “reward” signal and can orient itself towards both the ordered state of the world (say, 87% bad) and its own action (immediate, or algorithmically structured like a set of actions, effectively a strategy or a *policy,* as it is termed in RL) that would maximize the return, which is formalized in RL methods as the “cumulative reward”.

In a nutshell, via a RL system of DNNs which has been carefully structured, formalized, fed with sufficient amounts of data, trained and optimized, an agent (including an AMA) can learn to “behave” so that it can achieve its goal, in this case the goal of being ethical by justifiably choosing all actions that lead to “not killing a pedestrian by running over them” instead of “killing a pedestrian by running over them”. The concept uses a very simple set of actions, candidates for being ethical actions. Such candidates can be listed for other AMAs, like AMAs for non-discrimination, legality of action, universal healthcare and the like. All those AMAs would be narrow in the above sense. The task of integrating them into a more complex AMA system is a different task, that would necessitate the functional compatibility of AMA modules.

Another note belongs here. Since skeptics attack MAI from more than one side, we need to point that at its prototype version the AMA is only established as a *narrow* or a very specific (to a certain well-defined task) AI system. It is a proof of concept and has neither the pretentions nor the code to address more complex normative tasks, let alone *general* normative tasks. The idea here is to demonstrate the technological possibility and potential of RL methods for a narrow AMA technology – one that can be thought of as a natural beginning, and second, one that is much needed in industry and state. For example, the autopilot of an autonomous car is one such technology, albeit not extremely narrow but in fact quite a complex “narrow” AMA, that needs, however, to solve a short list of predefined tasks in real time.

The expectation for the growth of this technology is to either combine separate narrow AMA as modules, where the system would be integrated but still modularly functional, or, based on that, to build an integrated general MAIA that would perform two distinct tasks:

1. To find the optimal normative behavior within a real-life complex normative space
2. To solve real-life complex normative tasks in real time (and prove its being optimal via explainability and traceability)

In our system the state S of the world is represented as a “complete” (only complete within this world) description of the agent’s environment. An agent’s “observation” O is represented as a partial description of S. The information is carried by sets of numbers, whether real-valued vectors, matrices or n-dimensional tensors. In simple cases, like image recognition, the set of numbers can be the set of the values of the color three-space of the RGB schema. If we want richer information and thus better informed “decisions” of RL agents, such as the immensely successful implementation of RL in systems that managed to beat human champions in many Atari games merely learning from values and patterns of pixels, we can add space of contrast, hue, saturation and brightness. In any case, those sets of numbers would provide the values for S and O.

Possible actions and inactions of the RL agent are formalized in the so called “action space” or S_a. It contains the set of all actions that are available to the agent. In Go DNN RL systems such actions can be to put a stone down. In Atari DNN RL systems such actions are typically *discrete* and can be “move” {L, R, U, D} with a value for each parameter of the motion (say, move 10 pixels to the left); “shoot” {laser beam, rocket, etc.}, “evade” {lasers, bullets, etc.} “collect rewards” {treasures, coins, etc.}, and many others. In a narrow AMA RL agent, we can make only a very short list of actions available to it. For example, in an autopilot AMA, the agent is physically limited by the physical actions of the car itself. Thus, the list of available actions would be identical to the list of possible actions of the car {break, accelerate, stop, turn, blink, flash lights, press horn, etc.} along with the respective values for each action, just like the RGB values in image recognition RL systems {break with full force (100%), accelerate 30%, blink (left), flash lights (twice), etc.}.

The RL agent of AMA will decide on a policy (deterministic or stochastic) in order to choose its actions. The policy is parameterized and formalized as an ordered set of computable functions, where sets of parameters are taken as input. In the case of DNN, these parameters are its weights and biases. It is exactly those parameters that play a crucial role for the agent’s “learning” process during its quest to find the optimal policy to, say, not run over a pedestrian on the sidewalk. The learning is delivered via intense and well-tuned training on large amounts of data that, in our case, effectively presents myriads of scenarios for the AMA agent’s environment and its actions. Eventually, when trained on sufficient amounts of data, and when loss is minimized by skillful employment of the optimizer (e.g. gradient descent function), the agent becomes extremely skillful, much more so than even an expert driver, at avoiding pedestrians on sidewalks.

The major limitation of RL, as Deep Mind team puts it, is that standard RL systems need huge amounts of data in order to learn well enough. While attempts are being made to diminish the required amount of data, I would like to suggest an *ad hoc* solution for our AMA. We can develop, in parallel, a simple DNN of the generator type found in GANs, or the generative adversarial networks. The generator DNN would have the sole purpose of generating myriads of possible combinations between the components in the *np*, and those would be fed into the

RL DNN for learning purposes. This can help solve the technical problem with the requirement of sufficient amounts of relevant data.

This proposal frames a proof of a concept AMA system that can be developed, trained and tested, if the appropriate resources in AI efforts and investment are secured. Since it is an AMA and not a MAI, it is worth noting that the MAI functionality can be trivially deduced from it as a mere recommendation that falls short of hardware action, e.g. the physical action of an autonomous car.

Practical Reality Vs Theoretical Reality

There is a big divide in approaching AI ethics problems. Ethicists, stepping on the centuries-old shoulders of philosopher-giants, would prefer to dwell in the realm of *abstract reality,* where they can find all the comfort of working calmly and effectively, with no intellectual rush. The AI experts, however, are typically not ethicists by training (if at all). They know how to make an AI, whether it be a “simple” face recognition LSTM neural network system that can read the plates of your car’s number at 100 km’s/h, or the Deep Mind’s Alfa Go which managed to beat the best humans on Earth in a game where even an Oxford graduate in math would struggle to reach the national semi-finals for humans. They dwell in the physical, practical, non-abstract, no-prisoners-world. Both worlds could not be more different. The AI architects have no time to read and contemplate Plato’s philosophy; they need to code the neural network yesterday, otherwise all “real” hell would break loose on them and a lot of other people, companies, or even states. The “practical” people would love to be informed by the professional ethicists, but in reality this never happens: Snow’s *Two cultures* rarely talk to each other and even more rarely understand each other. Bridges between the two are precious little gems that occur only a couple of times in a nation’s century. The ethical AI, however, does not care about social and cultural divisions. Ironically, it would not discriminate. Unless you train it to, that is. A driverless car would run over everyone all the same, young and old, black and white, man and women, heterosexual, bisexual, transgender and in general all humans who happen to be at the wrong side of a badly calculated equation. It is therefore that what we need to understand and work on in order to establish an effective bridge between the historical achievements of ethics (that would take a lifetime for a philosopher to grasp only a fraction of) and the exploding AI technology of the day. But good management might facilitate the task and help bridge at least some of the gaps. Teamwork sounds like a good prospect: a team of ethicists and a team of open minded AI architects who can talk to each other, and eventually understand what the other side is saying.

Practical problems do not allow for the luxury of centuries-long contemplation. Far from it. Good or bad, right or wrong, this is the way modern society functions. Meanwhile, if we try to «repair it», we would need to solve a wide range of practical problems. Not to prove that the theory on which we base our solution is great or even better, but rather to at least try to enter the game of practical life-and-death scenarios and attempt to make a difference. Hopefully for the best, whatever this might turn out to actually be for both ethicists and AI architects.

Let me illustrate with a provocation. The Trolley Problem, in all its multivaried counterfactual glory, emerged as a theoretical problem. It focused the rational power of many a truly powerful ethicists across many university departments, countries and cultures. It is probably safe to say that there is no single preferred solution, predominantly accepted by the academic community. It is also safe to accept that this problem, like multiple others, does a pretty good job of shaping the limits of human rationality when it comes to questions of good and evil. Now, we can stay content with these views, and feed them to the AI architect responsible for coding the AI which would run the driverless trains of near future, only to find herself late. For the AI architect, who has already coded Tesla's and Daimler-Benz' autonomous auto-pilots, has passed their beta version long ago and now codes a version that would not only operate their cars in the real world, but also become legalized in, say, a conservative part of the world like the European Union.

Reality is running faster than theory, again. Theory can either rest content with itself or try to catch up. Trying to catch up is, obviously and rationally, preferable. Even if you get it wrong, many times there is at least a slim chance that one day your team would get onto something that society (and not elite philosophy departments) would judge as better than what we have seen in the streets so far. Or even as «good». The trolley problem, in all its notoriety, needs to be «solved» today. Just not in the theoretical realm of abstract intellectual comfort, where your manager is not shooting at you because you did not deliver. But in the practical realm. Would the practical good be the same as the ideal theoretical good? Surely not, even a priori. But this is not its adequate measure. For in the real world good and bad are already operating as we speak, and the practical solution to the TP only needs to be better than what happened yesterday in a town nearby, when we all covered our eyes whilst watching the news.

This slight ethical «edge» would, however, be much more valuable that even the most promising candidate for an ideal good and bad. It would have one non-trivial difference which society would probably see as an advantage and not as a deficiency: it would have done *real good*, no mater how small. Or at least it would have tried. While us clever philosophers sit in armchairs, contemplating the concepts of "bad" and "good", real people suffer and die out there in the practical world, with no one to offer them a helping hand. This is real ethics, and its real problems need real solutions, even if two centuries from now advanced theory may show them to be flawed.

When answering the usual question about the best gear, a photographer would say it did not matter all that much, what mattered was that *I was there*. The position of the AI architect is similar to that of the photographer. Yes, we prefer to have hundreds of millions of dollars, a state of the art photo studio, the best models in the world and a Hasselblad at hand at all times, while an army of assistants recreates our imaginary reality as we see it in our preferred vision. After taking a hundred shots, one of them goes on a cover, and the audience is in awe for it sees a dream world that they have never dreamed of. This, comfortable as it is for a handfull of chosen professionals, is as far from the real world as it gets. The real world drives fast and reckless, the pedestrians have no names and faces, everyone is rushing in the focus of her narrowing life, and the weight of the real value of life and death becomes more and more abstract. Not unlike a video game. A penny of good in a real world situation is worth millions in abstract armchair intellectual success. That is, until the latter becomes truly valuable and manages to *transfer to real world*. This is the task before the AI architect and it is more practical than theoretical. Actually, a theorist might draw inspiration from the riches of real world rational mistakes and come up with a better version of its ethical system. For we all wait for it and it would better be true ethics, and not a fake one.

Another argument in favour of ethically flawed solutions of the trolley problem goes like this. It just does not seem fair to ask of the machinist more than we ask of ourselves. It translates quite seamlessly to machines: is it too much to ask of a machine to be better than us, while we are all too forgiving to ourselves about anything from the slightest of mischiefs to much bigger things? Is our bar as high for us as it is for machines? It better be, right? Well, it does not seem to be. In reality, the non-armchair world, ethical solutions need to be and *are* executed myriads of times, in real time. Good, bad, flawed, benevolent but fruitless, pure evil, good by pure chance, all kinds are there. The numbers for today's 8 billion people global society are practically incalculable. We cannot even begin to think how to form an equation in order to arrive at only a minute fraction of the ethical depth of today's human reality, the one delivered by numbers.

Do we solve the task of everyday deeds with the same rational fervour as we do when we try to razor-cut the precise limits of evil in the likes of the trolley problem? Do we acknowledge the real normative weight and value of a physical penny dropped out of sheer sympathy in the hat of a hungry street musician? How bad could the Tesla's code for autonomous driving have been if it crashed into a pedestrian due to miscalculation or a solution vacuum in version 4.2? In order to even begin to evaluate the «objective» bad value we need to factor in many *other* values, like lives saved due to correctly functioning AI autopilot, lives saved due to prevented reckless driving by otherwise «good» humans. To illustrate the point of the bar once again – in the Balkans, where I come from, no autonomous machine can do as much harm to humans as humans ourselves. We are guilty of all sorts of bad driving, from pure teenage per horsepower reckless driving, through an innocent glass of wine behind the wheel till record pro-mill drunk and intended crash due to suboptimal psychological state and a history of, say, family abuse in the past. The TP solution architect should, if she is worth her salt, be able to quantify, qualify and rationalize these factors *as well* before she finds a rational, universal and later - accepted solution of the TP. For while we are engaged in the search for ideals, a lot of real, non-abstract pedestrians lose their lives, due to real reasons and not statistical anomalies.

The AI architect, for narrow purposes, and not for an AGI style doomsday scenarios, needs only do incrementally better than current statistics in order to justify implementation of its ML system into the car that we would all go and buy tomorrow from the showroom. The modest formula *less bad in real world is better than more good in abstract world* ($-B_{RW} > +G_{AW}$) seems to have a good set of solutions for a wide range of values for B and G. These need to be captured by an apt and practically successful ML system that operates simple machines like cars, trains and IoT toasters. Rocket launchers, gene editing, social engineering techniques and the like would naturally require more complex real world data and core complex ML algorithms. But let us not forget that we are facing an era of technological evolution: if a group of scientists does not like the ML solution of another group, for, say, social benefits distribution, they can come up with their own, hopefully better, solution. It does not seem much different from our human-led reality, only without the corruption, bias and inherent or acquired malevolence that only humans — and not animals, let alone machines — are capable of.

NOTES

[1] As an analytic philosopher, I am indebted to my mentor and friend Nenad Miscevic, as well as to Prof. Janos Kis, for introducing me to the fascinating field of analytic ethics. I have also benefitted enormously from the weekly seminars of Julian Savulescu and Nick Bostrom in Littlegate house in Oxford during the period between 2008 – 2010.

[2] "… moral artificial intelligence (moral AI) could be developed to help agents overcome their natural psychological limitations. The moral AI would monitor physical and environmental factors that affect moral decision-making, would identify and make agents aware of their biases, and would advise agents on the right course of action, based on the agent's moral values. In being tailored to the agent, the moral AI would not only preserve pluralism of moral values but would also enhance the agent's autonomy by prompting reflection and by helping him overcome his natural psychological limitations." (Savulescu and Maslen, 2015, in *Beyond AI*) and "We therefore argue that the contender for serious consideration is a type of "weak" moral AI that does not involve creating new agents nor undermining human freedom but, instead, gathers, computes and updates data to assist human agents with their moral decision-making. This data will comprise information about the individual agent and his environment, about his moral principles and values and about the common cognitive biases that affect moral decision-making. The moral AI will use this data to alert the agent to potential influences and biases, will suggest strategies for ameliorating these influences and biases, and will advise the agent of particular courses of action at his request." (ibid. p. 84)

[3] China's active system of social credit being perhaps one of the best illustrations, as a contrast to EU's continuing efforts for an ethical AI that would, among other things, respect and preserve human rights such as privacy.

[4] Very recent skeptical position if presented by Serafimova: "… even if one assumes the existence of well-working ethical intelligent agents in epistemic terms, it does not necessarily mean that they meet the requirements of autonomous moral agents, such as human beings." (Serafimova 2020, p.1); she follows up the general direction of Mittelstadt end others: "… that decision-making based upon algorithms remains "a standout question in machine ethics" (Mittelstadt et al., 2016, p. 11)."; see also " … whether or *not algorithmic decision-making can achieve a level of moral autonomy similar to that of human decision-making*. " (Serafimova 2020, p.2) and most of all: "This is due to the fact that "ethics has not been completely codified" (Serafimova 2020, p. 2).

[5] See China's system for social credits.

[6] AI systems for pattern recognition Bernstein, Alexander & Burnaev, Evgeny. (2018). Reinforcement learning in computer vision. 58. 10.1117/12.2309945.

[7] Deep Mind's Duplex AI demonstrated the Turing-test style success with a couple of real life conversations led by it with humans. For a discussion see Chowdhury, Sandeep. (2019). *How do ethics influence Google Duplex's acceptance and the incoming wave of Proactive Voice Assistants?*.

[8] The protagonist in Kesey's novel *One Flew Over the Cuckoo's Nest* (1962).

[9] For a discussion of the machine- AI doomsday scenario, depicted in Matrix movie trilogy see Haslam, Jason. "Coded Discourse: Romancing the (Electronic) Shadow in "The Matrix"." College Literature 32, no. 3 (2005): 92-115. Accessed November 10, 2020. http://www.jstor.org/stable/25115289.

[10] The Terminator movie Franchise doomsday machine scenario addresses well a niche, that became enormously popular in pop culture but also in philosophy discussions.

[11] The AI simulator that beat F-16 expert pilot (for details see the BI entry in the literature above) is a state of the art AI that can be recreated in a variety of tasks by using (mostly) open source NN architecture and sufficient data, which in the F-16 case might not be open source and understandably so.

[12] Deep Mind puts a lot of effort into developing RL methods and some of its greatest successes are based on RL. For details see Deep Mind's team papers on RL in the literature.

BIBLIOGRAPHY

Barreto, André et al. (2020) *Fast reinforcement learning with generalized policy updates* in «Proceedings of the National Academy of Sciences» Aug 2020, 201907370; DOI: 10.1073/pnas.1907370117

Bernstein, Alexander & Burnaev, Evgeny (2018) *Reinforcement learning in computer vision*. 58. In «Proceedings of Conference: Tenth International Conference on Machine Vision» (ICMV 2017) 10.1117/12.2309945.

Boyle, Alan (2019) *Terminator' is back! AI experts do a reality check on Hollywood's new robo-nightmare.* in Geek Wire, November 2, 2019; https://www.geekwire.com/2019/terminator-back-ai-experts-reality-check-hollywoods-new-robo-nightmare/ .

Hao, Karen (2019) *Is China's social credit system as Orwellian as it sounds?* In MIT Technology Review,

https://www.technologyreview.com/2019/02/26/137255/chinas-social-credit-system-isnt-as-orwellian-as-it-sounds/

Haslam, Jason (2005) *Coded Discourse: Romancing the (Electronic) Shadow in "The Matrix"*. In College Literature 32, no. 3: 92-115. Accessed November 10, 2020. http://www.jstor.org/stable/25115289.

Mittelstadt et al. (2018) *The ethics of algorithms: Mapping the debate*, in Big Data and Society, Volume: 3 issue: 2; https://doi.org/10.1177/2053951716679679.

Pickrell, Ryan (2020) *A US Air Force F-16 pilot just battled AI in 5 simulated dogfights, and the machine emerged victorious every time* Aug 21, 2020; https://www.businessinsider.com/ai-just-beat-a-human-pilot-in-a-simulated-dogfight-2020-8 .

Savulescu J., Maslen H. (2015) *Moral Enhancement and Artificial Intelligence: Moral AI?*. In: Romportl J., Zackova E., Kelemen J. (eds) «Beyond Artificial Intelligence. Topics in Intelligent Engineering and Informatics», vol 9. Springer, Cham. https://doi.org/10.1007/978-3-319-09668-1_6

Serafimova, S. (2020) *Whose morality? Which rationality? Challenging artificial intelligence as a remedy for the lack of moral enhancement*. Humanit Soc Sci Commun 7, 119 https://doi.org/10.1057/s41599-020-00614-8 .

Snow, Ch, P. (2001) [1959]. *The Two Cultures*. London: Cambridge University Press. p. 3. ISBN 978-0-521-45730-9.