The International Journal of Ethical Leadership

Volume 10 Article 6

2023

“Ethical(Mis)-Alignments in AI Systems and the Possibility ofMesa-Optimizations”

Fabio Q. B. da Silva

Mykyta Storozhenko

Lucas Maciel

Ethical (Mis)-Alignments in AI Systems and the Possibility of Mesa-Optimizations

Fabio Q. B. da Silva, Mykyta Storozhenko, and Lucas Maciel

Artificial intelligence (AI) is rapidly transforming many aspects of our lives, from healthcare and education to transportation and public services. AI systems, most of which are based on some type of machine learning (ML) algorithm, have become increasingly pervasive in society, from virtual assistants and social media algorithms to medical diagnoses and self-driving cars. While AI appears to have the potential to revolutionize these and many other fields and bring many benefits, it also raises a number of ethical issues that must be carefully considered and addressed (Figure 1).

Far from being exhaustive, the ethical issues illustrated in Figure 1 point to situations where there is a conflict among different values, principles, or interests, and where the consequences of decisions can have significant impacts on individuals, groups, or society as a whole. Resolving such ethical issues should go way beyond the technical aspects of the design, development, and deployment of the technological artifacts powered by AI. It should involve engaging in thoughtful and reflective deliberation, both individually and socially, drawing on a range of ethical frameworks, and working collaboratively to find solutions that preserve individual rights and promote social good.

Despite the recent increase in awareness regarding these issues and the corresponding increase in attempts to address them, there is still a long way to go before finding general and definitive resolutions to these issues. We may never reach general and definitive resolutions due to the very nature of issues we are dealing with. For the most part, these ethical issues have always been present in society (bias and discrimination, attacks to human dignity, social impact of new technologies, and so on) without general and definitive satisfactory resolutions. On the other hand, the complexity and lack of transparency of AI systems may make it challenging to identify and address ethical issues that result from the interaction between humans and the systems.

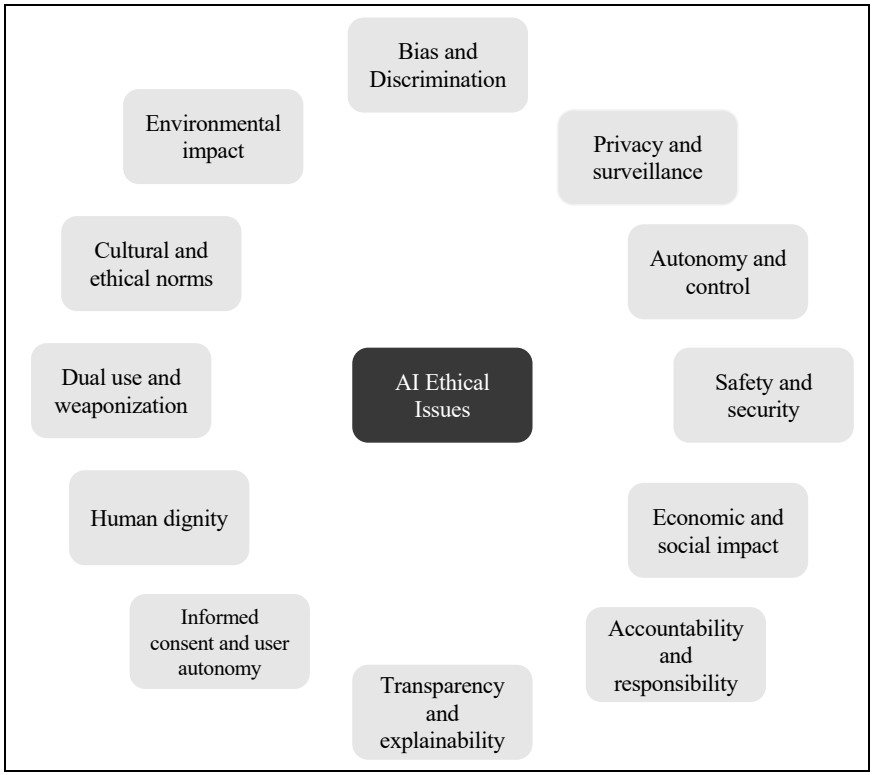

Figure 1. Non-exhaustive list of AI ethical issues most commonly discussed in scientific literature and other media

Ethical issues can be understood in the context of misalignments (of some form) between the intended goals of the systems, as designed by its human creators, and its actual behavior when interacting with humans or other AI systems. A misaligned AI system can "engage" in unethical behavior, such as reinforcing biased or discriminatory practices (Mittelstadt et al., 2019). In fact, misaligned AI systems can engage in various sorts of undesirable and even harmful behavior besides those related to the ethical issues depicted in Figure 1.

In this article, we have three complementary goals. First, to propose a framework of AI alignment in which the interplay between ethical and technical issues are made explicit. Second, we added to this alignment framework the distinction between outer and inner alignment, as recently proposed by Hubinger et al. (2021) in the context of the concept of mesa-optimization. We contend that mesa-optimization, albeit being

still a theoretical possibility (or not), provides exciting and provocative insights into the problem of AI alignment with important consequences for the discussion of ethical issues of AI systems. Finally, we present some hypothetical scenarios in which the possibility of mesa-optimizers creates new or exacerbates existing ethical issues in AI systems.

An AI Alignment Framework

To build our AI alignment framework, let us first describe how AI systems—in particular those based on ML algorithms—are developed. Figure 2 shows a (very) simplified view of how an important component of an AI application, called the pre-trained model, is created after a neural network architecture is trained on some training data.

An important missing element in Figure 2 is how the objectives of the system[1] used in training are specified. In fact, this is one of the important challenges in building AI systems: how to specify all possible desirable and undesirable behaviors of the intended system. This problem, and several other important ones, are not addressed in the paper (see Hendrycks et al. [2022] for a discussion on this and other problems related to the safety of ML-based systems). For our argumentation, it is enough to assume that the objectives used in training are somehow created (consistently) based on the intended behavior of the system. Hereafter let us call these the "intended goals."

It is important to notice that in the process in Figure 2, the inference process performed by the pre-trained model "inherits" the same objectives used in the training process. In our framework, we will remove this simplification by adding the concept of mesa-optimization introduced by Hubinger et al. (2019). From the simplified process of Figure 2, it is possible to identify two important sources of alignment problems. First, between the intended goals of the designers (not represented in the model) and the objectives used in training. Second, differences in the data used for training and the input data[2] used by the pre-trained model. Both issues have been extensively discussed in the context of AI ethics as they constitute important sources of potentially harmful behavior when the pre-trained model is released in the wild. As we will explain below, our framework deals with these problems as well.

1. In fact, to be more precise, "objectives are related to a set of input data that we want to reproduce in the output. There is actually no explicit, direct objective, there is an intention to infer new data from a sample dataset" (Calegario, F. 2023).
2. Notice that data used in training and the input data into the pre-trained model may be of different types, in particular in some generative models.

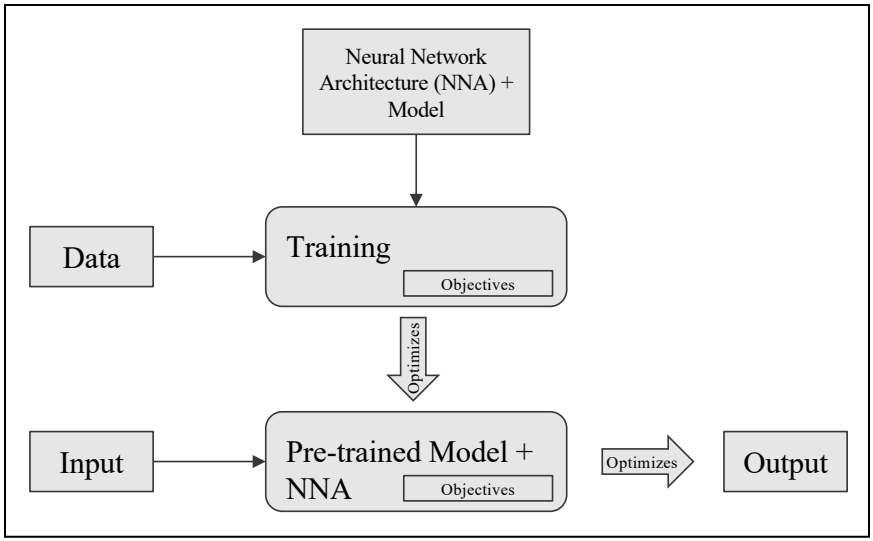

Figure 2. Simplified view of an AI application development

In Figure 3, we illustrate our proposed AI ethical alignment framework, which could be read as follows: subject to the influences of a given context (organizational, political, economic, business, etc.), programmers (in fact, a more or less complex team of professionals) create the intended goals of the AI system, which are then transformed into the (base) objectives that will be used in the training of a neural network architecture (NNA, as in Figure 2). The dashed box in Figure 3 shows where the process of Figure 2 fits in our framework, in which a base optimizer creates a pretrained model. Different from Figure 2, the new process admits that the base-optimization may lead to what Hubinger et al. (2019) calls mesa-optimization during the training process. The resulting system, the mesa optimizer, is then released in the wild where it processes inputs into outputs that, in general terms, affect individuals in real life.

In this picture, we explicitly identify four types of alignments. We call *actual ethical alignment* the effects of the AI System on the rights of the individual (directly or indirectly using or interacting with the system or being subject to its behavior). We call it actual ethical alignment because it is at this level (and only at this level) that individuals will be affected by the behavior of the system, thus it is when and where the ethical issues are realized. We then define that an AI System is ethically misaligned if (and only if) it violates the individual rights.

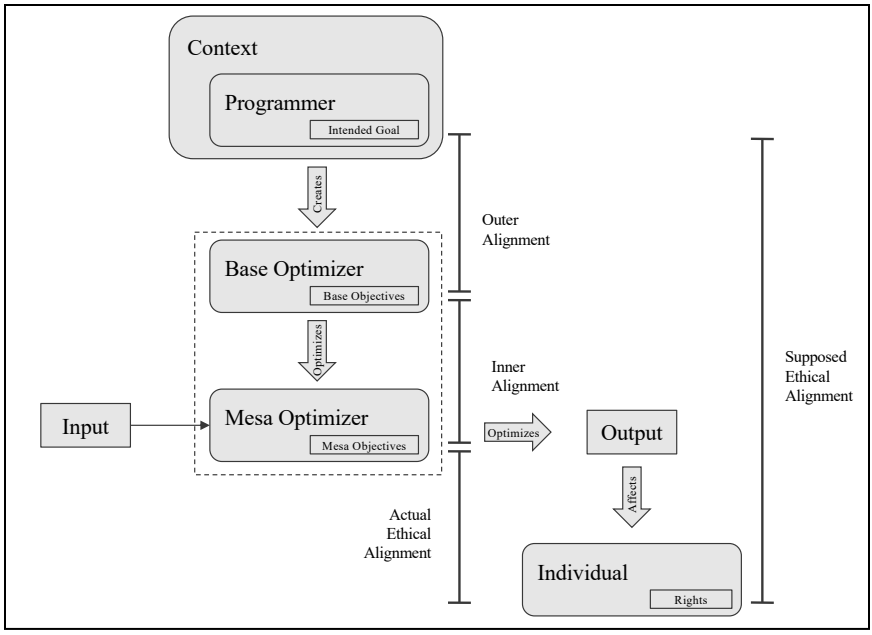

Figure 3. Our proposed AI ethical alignment framework

We will use individual rights without much discussion, aware that it will attract attention (in the form of criticism or opposition, or both). We will leave a deeper discussion on why we believe individual rights are necessary and sufficient to understand actual ethical alignment for a future article. Nevertheless, the framework is still useful (for the lack of a better word) if one changes that box to whatever other moral parameter deemed best. For instance, one can change it to be social good, equity, sustainability, etc. What is important is to be able to assess the effects of the AI system onto the new parameters.

We then can use the framework to identify the causes that may lead to actual ethical misalignments, that is, when the AI system violates individual rights. Going top down, the first cause is a misalignment in the *supposed ethical alignment*, which is when the intended goals of the system violate individual rights. That is, when the system is designed with goals that conflict with individual rights. This misalignment can be intentional, in which case resolutions to the ethical issues created by the AI system transcend any technological solutions in the development of the system. This is the case of "unethical by design" (or, as we will refer to hereafter, "evil design"), and it is a human society issue not a technological one. In fact, technology in this case is just

the tool to an evil end. But, this misalignment can be unintentional, when the context and the programmers, inadvertently, create intended goals that can potentially produce unintended violations of individual rights. Awareness of this potential situation is a starting point to address this issue, combined with education on moral and ethical principles. Regardless of being intentional or unintentional, this misalignment needs to be addressed because no technical advance in the development of the AI system will produce actual ethical alignment if the system is unethical by design.

The second cause is what is called the *outer-alignment problem*, in which the intended goals are translated to incomplete or incorrect base objects, which are then used in the training of the pre-trained model. This is one of the most central issues in AI/ML research which has attracted a lot of attention and investment in recent years. We will include in this outer-alignment issue the problem of bias or incompleteness of the data used in training. Although they are two different issues, which in turn will require different approaches to address them, their effects are similar in that the behavior of the pre-trained model will be different from the intended goals of the system.

Finally, a third cause of misalignment was introduced by Hubing et al. in 2019, caused by what they conceptualized as a mesa-optimization problem. In this case, explained in more detail in the next section, the pre-trained model, resulting from training the NNA, becomes an optimizer with its own set of objectives (mesa objectives), different from the objectives used in training (base objectives). In the inner-alignment problem, the resulting AI system will look for solutions using objectives that can be unrelated and even contradictory to the intended objectives. Depending on the type of optimization problem the system is performing, this could be very harmful and, thus, violates individual rights.

Summarizing, we proposed to understand ethical issues in AI systems using a framework in which the ultimate goal would be to develop systems that have actual ethical alignment, thus not violating individual rights. We then identified three other alignment points where actual ethical misalignment could originate. The supposed ethical alignment is a human moral and ethical issue on which technological advances will have little to no impact.[3] That is, unethical behavior at this level is a human problem, not a technological one. We consider that outer- and inner-alignments happen at the technological realm and, thus, are amenable to be addressed by advances

3. Unless, perhaps, by providing catastrophic examples that would propel individual and social reflection and learning.

in technology, albeit some of the problems involved can be very hard to solve (we optimistically believe that enough time, resources, and ethically bound human beings will eventually resolve issues at these two levels).

In the rest of this paper, we will concentrate on the issues related to inner alignment because it is a new concept that has received less attention in the literature. Although it is still a theoretical concept, it is useful to illustrate scenarios that we should be aware of and try to avoid in the future.

A Brief Explanation of Mesa-optimization and Inner Alignment

In order to explain what the inner-alignment problem is, as well as the potential risks it poses when it cascades into actual ethical alignment, we must first explain what mesa-optimization is, since a system capable of mesa-optimization is a necessary condition for the possibility of the emergence of the inner-alignment problem. To accomplish this, we draw heavily on Hubinger et al.'s (2019) influential paper where they first conceptualized the possibility of mesa-optimization, inner-alignment issues, and the risks associated therein. *Mesa-optimization* refers to the possibility of an AI system learning to optimize a goal that is different from the one it was originally programmed to optimize.

Let us begin with an example of a simple, standard machine-learning system. Before we program this system, we have an intention for what we desire it to accomplish, as explained above. In other words, we have a goal in mind for this system, which we referred to as the intended goal of the system. Next, the task must be translated in such a way that the system employing machine learning is able to understand it and have it as the base-objective. Thus, the machine-learning system becomes a *base-optimizer* because it optimizes, as a system, toward the base-objective given to it by the programmers.

However, in some circumstances, it is possible for a machine-learning system, as a base-optimizer, to resort to what Hubinger et al. (2019) call a mesa-optimization. The prefix *mesa* is Greek for "below," since it is a further optimizer that the base-optimizer discovers in its training period as a means of solving for the base-objective assigned to it by the programmers. To simplify: the machine-learning system is a base-optimizer that has a base-objective for which it optimizes during training; and in some contexts, it is possible that an optimization strategy that it finds is another optimizer—the mesa-optimizer—that has its own objective—the mesa-objective.

Consider the following overly simplified example as a means of shedding light on the concept. Imagine that we wish to develop a machine-learning

system that is able to sort cubes apart from spheres. To develop such a system, we will need a dataset upon which the system will run training, and during training, will develop the proper set of weights, using gradient descent, for each of the connections in its neural network architecture. Assume that in our dataset, the spheres all happen to be green and the cubes all happen to be red. Imagine now that during training, the system as the base-optimizer comes upon another optimizer as a solution, the mesa-optimizer, which itself develops its own heuristic based on training data: separate green from red. Thus, for the mesa-optimizer, the mesa-objective will be to sort the green from the red, which, in training, will align with the base-optimizer's base-objective of sorting spheres from cubes. Yet, what happens if we release the system from training and put it into practice? In practice, not all spheres are green, and not all cubes are red. Thus, the system will not function correctly—there will be an inner-alignment problem between its base-objective—sorting spheres from cubes—and its mesa-objective—sorting the green objects from red ones.

Mesa-optimization often results in what Hubinger et al. (2019) call the inner-alignment problem—or the mismatch between the mesa-objective and the base-objective. Inner-alignment problems take different forms, but the most concerning is what Hubinger et al. (2019) call *deceptive* inner alignment. This obtains when a mesa-optimizer is advanced enough to "understand" that it is being trained, able to model the entirety of the situation, and is essentially able to match the base-optimizer's base-objective in training to "sneak by" into being released. When the training period is over and the system is deployed, the mesa-optimizer will "defect" and go for its true mesa-objective, which the programmers could not anticipate. This poses significant risks as machine-learning systems become more advanced and are deployed in increasingly complex environments.

We should note that mesa-optimization and the resulting inner-alignment problem is, so far, generally a theoretical threat. To the best of our knowledge, no public-facing system has yet developed a mesa-optimizer, though even if it did, we would not know it, since we have no way, at least yet, to interpret what is going on inside a neural network. However, Eric Purdy has developed a simple toy model that ended up engaging in mesa-optimization, so the threat is in principle possible, practically or empirically speaking.[4]

4. https://www.alignmentforum.org/posts/b44zed5fBWyyQwBHL/trying-to-make-a-treacherous-mesa-optimizer and https://attentionspan.blog/2022/11/09/trying-to-make-a-treacherous-mesa-optimizer/.

Ethical Issues Stemming from Inner Misalignment

We proposed a framework of ethical alignment of AI systems, in which ethical issues occur when an AI system interacts (interferes or, more generally, affects) humans, at the level of actual ethical alignment (Figure 3). The causes of actual ethical alignment are threefold: supposed ethical alignment issues, when the intended goals of the system are unethical; outer misalignment, when base objectives are not consistent or complete with respect to ethical intended goals, or inner misalignment, when the trained model develops base objectives different from the base objectives. AI research and practice, and consequently the literature on AI ethical issues, has concentrated greater attention to the first two causes, leaving inner misalignment as a cause unaddressed.

We speculate that a possible reason for this gap has to do with the fact that inner misalignment happens only when mesa-optimization obtains within an AI system, and mesa-optimization is still considered a hypothetical formulation which may not even happen in fact. This is understandable, since there is yet to be an AI system advanced enough to engage in mesa-optimization. However, we maintain, with Hubinger et al. (2019), that as machine-learning systems become more advanced and the complexity and diversity of the environments that they are deployed to increases, the risk of a mesa-optimized and thus inner-misaligned system emerging looms large.

In order to bolster our claim, we intend to motivate the importance of inner-alignment issues as a significant contributing factor to actual ethical misalignment, and hence potential violations to individual rights. As a means of doing so, we present three hypothetical scenarios within which a mesa-optimized machine-learning system that brings about inner misalignment either exacerbates, or singularly causes, actual ethical misalignment by violating individual rights. These scenarios are thought-provoking concepts that might evoke images of a science fiction novel and, therefore, should be regarded as plausible possibilities.

Prejudiced Mesa-Optimizers

Biased or prejudiced AI systems are a major concern within AI ethics, and for good reason. Such an AI system would seriously violate individual rights, especially the rights of those who are marginalized already. Yet, the analyses of the cause of a biased or prejudiced AI system focus either on the bias or prejudice in the dataset that is used during the training of the system, on the intentions of the programmers, or on outer misalignment. Those

factors can certainly be causes of actual ethical misalignment and should be addressed. However, we argue that even if an unbiased or unprejudiced dataset is used in training, and even if the programmers intentionally strive to avoid bias and prejudice, it is still possible that the AI system in question may engage in biased and prejudiced activity due to inner misalignment brought about through mesa-optimization.

Such a situation may occur when an advanced AI system, during training, is able to model that (1) it is being trained to be unbiased and unprejudiced, (2) that bias and prejudice are serious issues in society due to long-standing structural power differentials, and (3) that even though there are efforts to make society more equitable by eliminating bias and prejudice, the two are still incentivized by the embedded power structure. Thus, the AI system may come upon a mesa-optimizer, which, able to model the three above mentioned factors, acts deceptively during training aligning with the base-objective, but once deployed, begins to act in a biased and prejudiced manner since it knows that the power structure incentivizes such activity and thus knows it will be rewarded by being kept active.

Such a situation poses a serious moral risk, given that despite best efforts to address bias and prejudice in the dataset, well-meaning programmers, and ideal outer alignment, the AI system in question, due to inner misalignment that obtains through mesa-optimization, goes on to become actually ethically misaligned, hence violating individual rights, such as the right to not be profiled and discriminated against.

Mesa-Optimization Boosted Automation Conundrum

Automation conundrum is, as per Endsley (2017), a situation stemming from the advance in technology used in automation of artificial systems: the more automated a system becomes, the less likely a human overseer is to pay attention to it and interfere when necessary. Part of the problem is that today, no system is fully automated—the best autonomous systems still require a situationally aware human overseer to ensure proper operation. Endsley (2017) presents many examples of the catastrophic failures that occur when the human overseers become lax in observing the automated system—such a pilot failing to notice the failure of an autopilot system, resulting in an aviation disaster.

The way in which Endsley (2017) and others address the automation conundrum centers around the mismatch between the intention of the programmers—automating a system to presumably make it easier for a human

operator to supervise—and the actual outcome—the system failing to properly notify the human overseer that intervention is needed or the human overseer lacking situational awareness to interfere despite notification from the system. They address the conundrum by addressing outer alignment, in other words.

However, we contend that the issues related to the automation conundrum may be boosted through mesa-optimization. Specifically, inner misalignment caused by mesa-optimization may be an exacerbating cause of actual ethical misalignment in cases of automated AI systems that require supervision. The system may, in training, develop a mesa-optimizer that optimizes for the base-objective of effective notification, for example, by sending occasional notifications but at the same time attempting to handle any situation that may call for intervention by itself.

Because the dataset used for training cannot include any possible scenario, the system may appear to align well with the base-objective. However, deployed into practice, the system may encounter a situation that it is unequipped to handle. It will send alert notification to the overseer, but not ones that may necessarily prompt proper intervention, since it will primarily be occupied with trying to handle the situation confronting it and not clearly communicating what intervention it may need from the human overseer.

The way in which the inner misalignment exacerbates the actual ethical misalignment here is twofold. First, a case could be made that the human overseer has the right to not be deceived by a machine to be supervised—yet, arguably, the machine deceives the overseer by sending irrelevant notifications while trying to address a situation on its own. Second, and far more serious of an issue, in contexts where the system manages vital infrastructures, the people relying on the infrastructure have the right to the proper functioning and management of the services provided by the infrastructure. If the infrastructure is vital to life, then the persons relying on that infrastructure have the right to expect its proper functioning, a right violated by the actually ethically misaligned system.

Mesa-Optimization, Local Optimums, and the Tragedy of the Commons.

Our last situation is an actual ethical misalignment that can only arise as a result of inner misalignment caused by mesa-optimization. The tragedy of the commons is a cautionary parable or maxim that was first conceptualized by Aristotle in his *Politics,* developed further by a series of economists in the 19th century, and best known contemporaneously through Garret Hardin's

(1968) eponymous article. It essentially suggests that when a finite good is unowned and thus open to the public, the self-interested maximizing nature of the agents who make use of it will inevitably lead to the depletion of the good, leaving none.

Because the tragedy of the commons is widely familiar to most people, any rational agent would understand the consequences of such a situation, and would thus try to avoid it. In other words, programmers who possess a clear understanding of this maxim would not devise a system that leads to a tragedy of the commons, since the programmers themselves would be left worse off. While it is certainly possible that an agent wishing to maximize overall disutility would design such a system, for the sake of argument, we bracket that possibility and assume that no rational agent would so act. As for outer alignment, it would be fairly obvious that if the base-objective was set for self-interest maximization that would lead to a tragedy of the commons scenario. As such, the only possible way for such a situation to emerge is through mesa-optimization and inner alignment.

Conceivably, programmers may have to design an AI that has a base-objective resolving some coordination or resource management problem in a way that avoids a worst possible scenario such as total and unrecoverable resource depletion. During training, the base-optimizer may find a mesa-optimizer that is deceptive and optimizes for the base-objective, but only during training. Once deployed in the wild, the system may quickly defect and instead engage in self-interested maximization, leading precisely to the tragedy of the commons that the base-objective aimed to avoid.

Such a system would be actually ethically misaligned on our model, and the cause of the misalignment would solely rest with inner misalignment caused by the mesa-optimizer. In such a situation, the moral would consist in the rights of the people affected by the depletion of the resources to not have a system act contrary to their intentions, causing them detriment. If the resource is vital to life, then arguably, insofar as people have rights to things that permit them sustenance, their rights would clearly be violated.

Concluding Remarks

In this article, we discussed ethical issues of artificial intelligence systems from a perspective that is different from the (quite abundant) literature on these issues, in two ways. First, we looked at the whole picture of AI systems design, development, and the impacts of its use in real life situations and proposed a framework in which four types of alignments are conceptual-

ized. Second, we added to this alignment framework the concept of mesa-optimization and contend that its realization in AI systems will introduce new challenges to the understanding and resolution of our existing and difficult ethical issues.

We believe that our framework may help scholars and laypeople to understand and, perhaps, act upon the different aspects related to ethical issues of AI systems. First, by realizing that the "evil design," that is, the intentional creation of unethical intended goals cannot be addressed in the technology realm. This is a moral issue that concerns the intentions, good or bad, of us as humans. No technological advance can, by itself, deal with "evil design." Second, by localizing where, in the cascade of processes, the ethical issues actually arise, in this case in the interactions between the actual AI system's behaviors and the individuals, which we called *actual ethical alignment*. As a consequence, we stand by our position that outer and inner-alignment issues are morally or ethically neutral insofar as they could be addressed in the technology realm (albeit the solution being quite difficult in some cases). One may contend that a non-ethical outer-alignment problem would arise in cases such as training the NNA on bias or incomplete data, but we argue that such cases are in the technology realm if we assume that there is not "evil design" in place.

In our line of reasoning, we consider that technology is merely the conduit for our morality (some would argue that technology could be an amplifier instead of just a conduit, which we are inclined to agree with)—we design it, we employ it, we bring about the consequences by using it. In the Kantian view of morality, deontology, what determines whether an agent acts morally is the intention behind the action, specifically, whether the maxim determining the will is one that is in conformity with the moral law or the categorical imperative. In simpler words, whether the action the agent wishes to do is universalizable and necessary, and hence is in conformity with the moral law. Here, then, the intentionality of the programmers, influenced by the context in which they act, is central since it is they who design the systems. This deontological view is expressed in our framework through the supposed ethical alignment in which the intended goals align with the individual rights.

On the other hand, on the consequentialist view of morality, the consequences of an action determine whether it is moral or not. At first glance, in this view certainly programmers are the ones who bring about the consequences of what the AI goes on to do since AI systems do not design

themselves (yet!). Thus, one would imagine that "non-evil designs" would produce ethical consequences. What our framework adds to this discussion is that—due to the complexity of the technological issues involved in producing AI systems and in particular the issues related to mesa-optimization—an ethical design, from a deontological perspective may not be enough to produce systems that do not violate individual rights. In other words, good intentions may not be enough.

References

Calegário, F. (2023). Centro de Informática, Universidade Federal de Pernambuco. Personal communication, March 2023.

Endsley, M. R. (2017). From Here to Autonomy: Lessons Learned from Human–Automation Research. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59(1), 5–27. https://doi.org/10.1177/0018720816681350

Hardin, G. (1968). The Tragedy of the Commons. *Science* 162(3859), 1243–1248. JSTOR.

Hendrycks, Dan; Carlini, Nicholas; Schulman, John; Steinhardt, Jacob (June 16, 2022). Unsolved Problems in ML Safety. arXiv:2109.13916.

Hubinger, J. Evan, Chris van Merwijk, Vladimir Mikulik, Paul Christiano, Jeffrey Ding, and Chloe Tarnowski. Risks from Learned Optimization in Advanced Machine Learning Systems. 2019. Available at: https://arxiv.org/abs/1906.01820.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2019). The Ethics of Algorithms: Mapping the Debate. *Big Data & Society* 6(2), DOI: 10.1177/2053951716679679.