

AI-AUGMENTED CONSEQUENTIALIST ETHICS (AI-ACE): A NEW FRONTIER FOR THE CONCEPT OF RESPONSIBILITY

Ética Consequencialista Aumentada por IA (AI-ACE): uma nova fronteira para o conceito de responsabilidade

Marcus Gansauskas Bruzzo

University of Tartu

Tartu, Tartumaa, Estonia

marcus.bruzzo@gmail.com

<https://orcid.org/0000-0002-3689-5008> 

A lista completa com informações dos autores está no final do artigo 

Abstract: This paper explores how consequentialism in ethics, which bases the 'morality' of an action on the predictability of its results, is being redefined by Artificial Intelligence. Traditionally, human responsibility is limited by the cognitive capacity to foresee consequences (non-culpable ignorance). The central thesis is that AI, by processing vast "Data Lakes," generates a "hyper-vision" that makes previously unattainable patterns visible, dramatically expanding the domain of what is "reasonably predictable." This creates AI-Augmented Consequentialism, where the agent's moral duty comes to include consideration of the predictive potential of available technology. Consequently, ethical responsibility expands, eroding the claim of ignorance as a moral excuse (erosion of culpable ignorance), and the failure to use this capacity can be seen as negligence. However, the supposed hyper-vision faces issues of system opacity (black box) and 'moral deskilling', where human autonomy can be sacrificed in favor of algorithmic optimization. Although AI should enhance human ethical maturity, not replace it, human dependence on AI for prediction becomes inescapable.

Keywords: ethics; consequentialism; AI; artificial intelligence; deskilling.

Resumo: Este artigo explora como a teoria ética do consequencialismo - que determina a "moralidade" de uma ação com base na previsibilidade de seus resultados - está sendo redefinida pela Inteligência Artificial. Tradicionalmente, a responsabilidade ética humana é restringida pela capacidade cognitiva de prever consequências (ignorância não culpável). A tese central é que a IA, ao processar vastos "Data Lakes," gera uma "hipervisão" que revela padrões previamente não vistos, expandindo significativamente o escopo do que é considerado "razoavelmente previsível." Este desenvolvimento estabelece o Consequencialismo Aumentado por IA (AI-Augmented Consequentialism), onde o dever moral de um agente agora incorpora a capacidade preditiva da

tecnologia disponível. Consequentemente, a responsabilidade ética aumenta, minando a defesa da ignorância como desculpa moral (erosão da ignorância culpável); a falha em utilizar esta capacidade pode, portanto, ser vista como negligência. No entanto, esta suposta hipervisão é desafiada por questões de opacidade do sistema (o problema da caixa preta) e o "deskilling moral" (perda de habilidades morais), onde a autonomia humana pode ser sacrificada pela otimização algorítmica. Embora a IA deva idealmente aprimorar a maturidade ética humana em vez de substituí-la, a dependência humana da IA para uma previsão precisa torna-se inevitável.

Palavras-chave: ética; consequencialismo; IA; inteligência artificial; desqualificação.

1. Consequentialism and the limits of culpability imputability

Imputation of culpability by implementing "due diligence" to AI consulting of scenarios for decision-making, ties human ethics to non-human stances for the first time in history. We do not refer here to the seat belt conundrum proposed by Bruno Latour, in which human society is deeply entangled with a myriad of technically materialized moral decisions, often prior to our conscious awareness of them. In his example of the seat belt, Latour shows how a mechanical device compels safe behavior by requiring the driver to fasten it before driving. In doing so, it transfers substantial ethical responsibility to technical systems: "I cannot be bad anymore. I, plus the car, plus the dozens of patented engineers, plus the police are making me be moral" (Latour in Bjørk, 1992, p. 126). Today, however, we are confronted with far deeper layers of decision-making machines and with new forms of interaction between technical systems and human intentions, in which agency, responsibility, and moral judgment are increasingly redistributed.

The mere fact that an augmented perception by use of AI extracting new information from immense data volumes exists, configures culpability to humans once they deny the use of it in ethical decision-making processes, especially in consequentialism. This is referred to as "erosion of culpable ignorance". But beforehand, we should clarify "consequentialism" in metaethics. Classical consequentialism defines the morality of an action based on the nature of its results (Peterson, 2013) in such a way that the predictability of these same results, at the moment of choice, determines the choice itself (Craig, 20025, p. 143). In other words, I anticipate the scenarios resulting from a given decision made now, and I define my current decision based on the predictability of its consequences. This means that the ability to predict directly determines the choice I must make at this moment in a back-and-forth movement where:

1. I project future scenarios;
2. Choose the best one, and;
3. Return to the present moment making a decision that should lead us towards it.



We acknowledge variations within consequentialist theories and therefore limit the present scope to intentional action, specifically act-consequentialism. It is important to note that consequentialism – historically linked to the utilitarianism of Jeremy Bentham and John Stuart Mill – also adopts a quantitative approach to moral evaluation, seeking a measurable reduction of suffering. However, it is distinguished by allowing present decisions to be revised in light of anticipated future outcomes. Consequentialism constitutes one of the main families of contemporary normative ethical theories, maintaining that the moral correctness of an act is determined exclusively by its consequences and by the consequences of its alternatives (Peterson, 2013; Pettit, 1997). It must be emphasized that such predictability is necessarily limited to what can be foreseen. Indeed, this constitutes one of the most common critiques of consequentialist ethics: the moral agent may “lack time and knowledge,” which constrains “what consequentialists can do by way of calculation” (Craig, 2005, p. 144). Furthermore, consequentialism rejects the need to assign intrinsic value to actions themselves. It prioritizes the concept of the “Good” over that of the “Right” in its justificatory structure (Brink, 2020, p. 381; Portmore, 2020), thereby discarding duty and virtue as intrinsic moral foundations. Moral judgment, in this framework, derives exclusively from effective outcomes and does not necessarily function as the starting point of action. In other words, adherence to beliefs in duty or virtue is not required, even though a consequentialist action may retrospectively come to embody such values.

The structure of guilt also varies according to ethical stance. In deontological morality (“*deon*” meaning duty), commonly associated with the Christian perspective, there exists an obligation to act in a prescribed manner (Craig, 20025, p. 143). While deontology emphasizes the intrinsic moral nature of actions, consequentialism adopts a logic of “promoting” values rather than merely “honoring” them (Brink, 2020, p. 382). A classic example of this approach is the Categorical Imperative formulated by Immanuel Kant in *The Metaphysics of Morals* (1797). For Kant, the Categorical Imperative refers to a category of action that “states an obligation regarding certain actions” (Kant, 2013, p. 33), a condition compatible with the Abrahamic worldview, which conceives the world as a determined system governed by divinely grounded laws. In this framework, guilt receives a distinctive treatment, as it concerns the respect or violation of moral obligation. Kant designates this condition as *culpa*, defining it as transgression when he writes: “An act contrary to duty is called a transgression (*reatus*)” (Kant, 2013, p. 34). Transgression, therefore, consists in disrespecting a supposed obligation. Accordingly, Kant distinguishes categories of



culpability within deontology, arguing that “an unintentional transgression that, nevertheless, can be imputed, is called a mere fault (*culpa*). Intentional transgression (that is, transgression linked to the awareness of it being a transgression) is called a crime (*dolus*)” (Kant, 2013, p. 34).

Guilt merely acknowledges the link between decision and consequence, while intent (*dolus*) registers the perception of the intention that the consequences would be bad for someone. Since consequentialism does not impose a duty to act in a certain way, the actor's decision can be nothing more than a choice. The orchestration of the choice, however, determines their culpability or intent, in the sense that, if it is proven that they made every effort to foresee all possible consequences of their action, there is no way to consider it intent (*dolus*) that the result would be bad.

Here we are faced with the limit of ethical judgment, in the effort to foresee results. Therefore, in consequentialism, the decision-maker can be exonerated from blame even in the face of negative outcomes. This happens because the evaluation criteria are not limited to the final result (positive or negative) and its quantitative measurement, as in the case of Bentham's utilitarianism. The underlying logic is that if the agent used all available resources to foresee the consequences, the act acquires a sense of ‘good faith’, we might say. This condition shifts the judgment of guilt from the final results of an action to the evaluation of the procedure followed during the decision-making process, which will ultimately determine whether it was an act of good faith or bad faith. In these lines, one might say, that the if captain of the Titanic made use of all available technological resources for iceberg forecasting, to the point of being absolutely certain that accelerating the ship would be acceptable, an iceberg encountered in its path would constitute an accident, as an unforeseen incident, and exempt him from culpability as it had not applicable negligence.

Evidently, considering limitations in interpreting possibilities due to a lack of tools enabling this anticipatory vision could lead to negligence, but we must note that without restraint, this logic can be paralyzing because it assumes that, up to the present point in history, no human being has had access to omniscience to be fully responsible for any decision. In a way, every decision in history has been made partially in the dark. Thus, consequentialism, in a sense, makes predictability a limiting factor on agency. It does not require the agent to know all the causal ramifications of their actions, which can be massive and simply inscrutable (Railton, 2003, p. 166).

Hence, given the proven effort of predictability, any negative outcome is understood to be beyond human capacity to predict. This landscape has procedurally changed



throughout history with the increase in the number of tools, methods, and technological instruments available for predictability, and we point to the basic premise that, as more technical resources become available, the responsibility of the decision-maker increases, in the sense of taking into account all the instrumentation available that allow decision-making processes. We can cite as an illustrative case the decision of a pilot to take off his plane considering the weather forecast regarding possible storms far beyond his mere field of vision. At some point in history, a take-off decision could be made only based on visual judgment of the clouds. The philosophical premise of the decision in this case implies that the instrumental scope of the devices capable of reading future states must be taken into account in the act of decision-making, taking to a technological augmentation. The mere individual's knowledge of the existence of these instruments, therefore, should influence his responsibility upon the use of them.

In the hybrid world between humans and machines, the theme of 'extension' has been quite common. We find examples from the 20th century in Marshall McLuhan's work in "Understanding Media: The Extensions of Man," from 1964, where we read that even electricity could be taken "as an extension of our own central nervous system, including the brain" (McLuhan, 2012, p. 200), or in the creator of cybernetics, Norbert Wiener, who in 1950, in an interview with Time magazine, stated that this "cybernetic revolution" would be the replacement of the human brain, just as the Industrial Revolution was the replacement of the human arm.¹

We chose to leave aside positivist interpretations of hybridism between humans and machines, at this moment. Human modifications through the use of machines are the subject of study in various fields, but here, our scope is limited only to the effects of expansion of the technically predictable scenarios. In the end, we will return to the question of the impact on human abilities due to the condition of ethical deskilling. Thus, we can focus on increasing predictive capacity made possible by accumulating data in technologies capable of processing large volumes of records. It is in this scenario that Artificial Intelligence stands out as a historic innovation functioning as an enhancer of predictability. AI-powered technologies achieve this through their ability to track and identify patterns in vast areas and quantities of data, volumes that would be unattainable for human analysis.

Consequentialism, as presented, starts from an evaluation that is intrinsically linked

¹ "Science: The Thinking Machine", Time, 23 jan. 1950.

to the agent's ability to anticipate the consequences of the decision. Within the framework of normative ethics, foreseeability is a crucial factor for moral responsibility: as stated, an agent is often exempt from guilt if harm resulting from their action was not a possibility that a reasonable person should have foreseen (Chakraborti, 2023 p. 268). This limit to human foresight serves as a safeguard, distinguishing moral error from mere misfortunes or biological capability, something that with augmentation through AI, will tend to be reconsidered.

Responsibility is an old question. Since within the Aristotelian framework, for instance, the epistemic condition is a fundamental requirement for the attribution of responsibility: the agent must know what he is doing and be able to foresee the probable consequences of his conduct (Chakraborti, 2023, p. 265). When an individual reaches the limit of what is "reasonably foreseeable," he enters a state commonly known as "non-culpable ignorance" (Bykvist, 2020, p. 310). In these circumstances, consequentialism recognizes that, although the final result may be negative, the original action cannot be classified as morally wrong if the agent did not possess the cognitive means to anticipate the disaster (Chakraborti, 2023, p.265). In this manner, a profound ignorance (a scenario where the subject has no idea that their action has a dangerous characteristic) is cited as the strongest condition for exempting an agent from responsibility (Chakraborti, 2023, p. 270). Here, one must consider the lack of awareness about one's own ignorance and cognitive biases such as the popular Dunning-Kruger effect, which we have chosen to summarize as "ignorance about one's own ignorance" (Floridi, 2014, p. 83), and the assertive behavior derived from this condition.

Regarding the limits of predictability, the gray area containing the greatest complexity of definition, Chakraborti (2023, p. 265) asserts that the ethics of this modality should focus on the objectively probable consequences that any reasonable person, in such a position, would be able to foresee. In this definition, it is noted that the complexities of phenomenological, subjective and psychological aspects are often not fully encompassed by the yardstick of "capacity to see" for the purposes of moral judgment. However, upon reaching the limit of what human reason can project, the agent ceases to be the "master" of the outcome, and the event begins to be treated from the perspective of risk or fatality, detaching itself from the moral character of the individual (Chakraborti, 2023, p. 265; Peterson, 2013, p. 102). Brought to historical materiality, linked to the allowance for prediction, the presence or lack of technological apparatuses such as AI tend to engender judgments on consequentialist decisions. One might determine someone 'not guilty' simply because a given technology was not 'at hand' at the time his decision was taken, for



purposes that overpasses individual aspects.

Other accounts of the idea must be pointed out. Although the term 'consequentialism' is considerably modern, the seeds of consequentialist thought date back to antiquity. In ancient India, the Charvaka (or Lokayata) school already promoted an ethics based on the pursuit of pleasure and happiness in this life, grounded in the empirical consequences of decisions (Chakraborti, 2023, p. 435). In ancient Greece, the hedonism of Aristippus and Epicurus established pleasure as a central value, although Epicurus specifically emphasized the absence of pain and tranquility of the soul (Chakraborti, 2023, p. 438) and the goal was to achieve Ataraxia through rational decisions that altered the consequences of living. Epicureanism, by grounding the conditions of life in the direct consequences of previous decisions and actions, assumes an ethical logic where learning from the past allows for projecting data and adjusting present conduct in light of future projections (Chakraborti, 2023, p. 438). Furthermore, during the Renaissance, thinkers like Pierre Gassendi revived forms of hedonism, adding a universalist trait: the idea that the happiness sought should be for everyone, not just the individual (Chakraborti, 2023, p. 440), and also containing consequentialist traits due to the predictability of actions in the impact on the lives of others. Later, Francis Hutcheson (Idem) is credited with enunciating the first English-language version of the "principle of utility," arguing that actions are morally better to the extent that they promote the greatest happiness (Craig, 2005, p. 427), something that fits within the utilitarian and consequentialist lineage through anticipation.

Modern consequentialism derives directly from the Classical Utilitarianism of the 18th and 19th centuries, systematically formulated by Jeremy Bentham and John Stuart Mill (Chakraborti, 2023 p. 441; Wood, 2020). Jeremy Bentham: Introduced the concept of social utility, proposing the "Greatest Happiness Principle" as a criterion for evaluating laws and public policies. Here, "the concept of social utility stands for social usefulness" (Chakraborti, 2023 p. 443), whereas for Bentham, pleasure was evaluated quantitatively through a "hedonic calculus". That is, it was supposed to aim on the mathematically verifiable reduction of suffering. John Stuart Mill refined utilitarianism by introducing the distinction between higher and lower pleasures, linked respectively to intellectual and sensory faculties, while maintaining the principle of proportionality, according to which actions are right to the extent that they tend to promote the greatest possible happiness, and not merely the greatest quantity of pleasure (Chakraborti, 2023, p. 448; Brink in Portmore, 2020, p. 392). During this period, utilitarianism functioned as a tool for social reform, challenging institutions based on privilege and defending rights such as universal suffrage and education, like in the case in



which John Stuart Mill (James Mill's son) wrote in favor of women's suffrage" (Portmore, 2020, p. 17).

In the 20th century, the theory expanded beyond classical utilitarianism, becoming what we now call the genre of "Consequentialism". The current status is marked by a diversity of subspecies that attempt to respond to traditional criticisms, such as Rule Consequentialism, which evaluates acts not in isolation, but by their conformity to rules whose general acceptance would produce the best results (Portmore, 2020, p. 4). Multilevel and Sophisticated Consequentialism also proposes a distinction between the criterion of rightness (what makes an act right) and the decision-making procedure (how the agent should deliberate), allowing agents to act out of friendship or loyalty without constantly calculating overall utility (Portmore, 2020, p. 407; Railton, 2003, p. 156). Multidimensional Consequentialism, introduced to address the complexity of values, argues that the status of an act depends on several irreducible dimensions, such as individual well-being, equality, and risk (Peterson, 2013, p. 3). Currently, consequentialism enjoys a preeminent position due to its ability to be "consequentialized": the idea that any plausible moral theory (including those based on rights or virtues) can have its structure translated into a consequentialist form through the ranking of outcomes (Portmore, 2020, p. 1; Hurley, 2020, p. 24). In theory, by taking the characteristics that a theory considers relevant and building them within the consequences, any consensus on the hierarchization of outcomes can fit into a consequentialist "mechanic" (Portmore, 2020, p. 23).

However, we return to the point that the theory still faces the seemingly insurmountable challenge of uncertainty and actor ignorance, where the focus has shifted from the "actual" consequence to the expected or "epistemic" value (Bykvist, 2020, p. 310; Jackson, 2020, p. 326), given that human agents do not possess what we might call total foresight. In the contemporary era, consequentialism is vigorously applied in new areas, such as the ethics of Artificial Intelligence (Floridi, 2023, p. 57; Tsamados *et al.*, 2022, p. 94), where large volumes of data are used to increase the predictability of future harm and optimize social outcomes (Floridi, 2023, p. 5), giving rise to what we discussed earlier as AI-Augmented Consequentialist Ethics.

2. Data Lakes and the Generation of Enhanced Predictability

There is a growing interest around Artificial Moral Advisors (AMAs), that is "systems powered by artificial intelligence that are explicitly designed to assist humans in making ethical decisions" (Meyers, 2025, p. 1). These are made possible by the inherent ability of

AI to navigate huge amounts of data in search of previously invisible patterns for humans. The data revolution, characterized by what Floridi (2014, p. 13) calls “the age of the zettabyte” and “tsunami of bytes”, allows humanity to store and process quantities of information that exceed by orders of magnitude everything that was recorded until the end of the 20th century. The use of data lakes (vast repositories of raw and unstructured data) serves as fuel for Artificial Intelligence algorithms that seek not only to understand causalities but also to identify correlations in “small patterns” invisible to the human eye (IBID). It should be emphasized, however, that the concept of ‘predictability’ does not refer to ethical or moral decisions, but to the expansion of the visibility of scenarios, once, according to several researchers, a ‘moral prediction of AI’ would be incompatible with human psychology (Liu *et al.*, p. 438). Artificial Intelligence augments or extends the scope of available information from which human ethical decisions will be made, therefore, the hypothesis of decision-making entirely performed by machines is not being argued here.

So, it is fundamental to highlight that the virtualization of information and data extraction depend on preliminary processing of documents, files, sorts of data. Even in this initial phase, such as the “Extract, Transform, Load” (ETL) stage, reading and interpreting errors can occur. For example, commonly, legal files originally in “.pdf” format are converted to the “.ls.json” format, which, according to Lazovic (2022, p. 163), “enables detection of all the aspects of the legal text contained in the data set”. This very process might undermine all predictability of patterns once the ‘detection’ fails to address factual information. Machines do not predict, but find patterns that lead to prediction, as the predictability we are referring to here is that of the human sphere. Technology, thus, encompasses the scope of the predictable as a basis for human action, in the circumstances that configure consequentialist ethics. Predictability is, therefore, the possible human anticipation of future scenarios. This note is important because the terminology may overlap with “predictability” within Machine Learning processes, which is “the extent to which a system’s outputs or effects can be anticipated”. In other words, it is the degree to which one can answer the question, “What will this system do?” (Holland, 2020, p. 5). The (ETL) stage turns raw data into potential information meaning that a simple mistake in this single step, must be considered as something that undermines the culpability of the human decision at the end.

Still, we proceed to consider an undeniable potential augmentation of the ‘knowable’ through current data technologies. It is mandatory to assume that, while the human mind operates with perceptual and computational limitations (Chakraborti, 2023, p. 642), AI is a system that can process statistical evidence in real time within an expanded data scope



called by Floridi (2023, p. 14) as zettaflood. This global infrastructure of data allows for predictions about several areas such as climate trends, market behaviors and social risks with unprecedented accuracy, transforming into identifiable patterns what was previously incalculable (Floridi, 2023, p. 181-182; Tsamados et al., 2022, p. 94). Potential identifiability means potential responsibility. This capability generates superior predictability, transforming what was previously “uncertain” into “calculable risk”, which constitutes the technical basis for increasing the agent's responsibility (Floridi, 2023, p. 183). The availability of potential information, offered by Artificial Intelligence in a worldwide network of systems (including commercially available, off-the-shelf AI solutions such as Software as a Service SAAS (SaaS)) has a direct impact on human decision-making dynamics.

How does this impact the epistemic condition of consequentialism? As informational friction decreases and information becomes ubiquitous, the ability to predict harm increases, eroding the “right to ignore” (Floridi, 2014, p. 42) and making agents (human and organizational) increasingly responsible for the consequences of their actions in the world.

3. Difficulties for Factual Application of AI-Augmented Consequentialism

The advent of AI for prediction introduces what we hereby call AI-Augmented Consequentialist Ethics (AI-ACE). In this new framework, the criterion for the rightness of a choice is no longer limited to what the agent “can” individually predict, but to what the technical system at its disposal “is capable” of projecting. Hence, the mixing of biological to technical structures of manipulation of the ‘knowable’ poses new challenges to define the scope of ethics and the ‘ethical’ acting. Although we do not adhere to transhumanist or posthumanist positivist perspectives, the growing intermingling of human and machine factors in decision-making presents itself as a historically inescapable reality. Here, the Actor-Network paradigm proposed by Bruno Latour may suffice, as it allows us to account for the deep imbrications of conscious decision-making within extensive networks of interference among heterogeneous actors, without isolating any of them from their distinct natures. This approach avoids conceiving humans as cyborgs and instead understands them as beings engendered within networks composed of ideologies, documents, practices, instruments, tools, political events, and other mediating agencies. The bibliography on Actor-Network Theory (ANT) is rather extensive and exceeds the present scope.

The technical system that enhances knowability must be taken as a component in the performance we call ‘decision’. There is a causal chain of responsibility that intermingles machines and humans according to a given consequence. A network of influences, as



Latour puts it, that “is not only a technological arrangement” (Latour, 2013, p. 32) and should not be understood as a rigid or fixed network such as a road network, but considering that “behind the concept of network, there is always that movement, and that surprise” (Idem, p. 33). But how do algorithms link to human decisions? The order, as presented by Floridi, is that algorithms can be used:

- 1) to turn data into evidence for a given outcome;
- 2) which is used to trigger and motivate an action that may have ethical consequences.

According to him, steps 1 and 2 can be achieved semi-autonomously by algorithms, which complicates the 3rd action, namely: “to attribute responsibility for the effects of actions that an algorithm may trigger” (Floridi, 2023, p. 95). Nevertheless, despite the fact that there is not a clear cut between what is machine and what is human responsibility, it efficiently proves a form of causal connection, even if an emergent one.

This technological enhancement reshapes moral responsibility. In classical consequentialism, the agent is often exempt from responsibility for unforeseen outcomes on the grounds of non-culpable ignorance (Railton, 2003, p. 166). However, the hybrid machine–human causal chain characteristic of AI-Augmented Consequentialist Ethics (AI-ACE) expands what counts as “predictable”. If an Artificial Moral Advisor (AMA) can process environmental data to recommend optimal moral actions according to the user’s standards, then failing to consult such a system, or failing to anticipate harms it would have detected, may reasonably be interpreted as negligence (Liu *et al.*, 2022, p. 439). As a result, the limits of prediction are continually pushed forward, and the principle “ought implies can” becomes more demanding: when technology enables reliable foresight, agents acquire a moral obligation to take this information into account. Technological augmentation thus becomes integrated into the notion of “reasonably foreseeable consequences”, understood as outcomes that any reasonable person could anticipate (Chakraborti, 2023, p. 265). In this context, reasonable action increasingly presupposes consultation with AI systems used as reference points for decision-making. Moreover, as Kirsten Martin observes, “while technology and algorithms are constructed by humans, technology’s scripts endure to influence the behavior, acts, and beliefs of individuals in return” (Martin, 2019, p. 840). This feedback loop reinforces a cultural tendency to assign superior epistemic authority to machine-generated predictions, further transforming how responsibility is distributed in moral practice.

This potential of diligence proper to the machine does not, however, mean the direct



application of its capabilities to human life. For that to happen, a clear governance process is necessary regarding data extraction, processing, and the preparation of informational material that are ethically important to human beings. The risks increase as ethics, seen within the fields of engineering, is diffuse, since it is “predominantly driven by self-assembled communities and associations, who develop their own standards” (Kazim, 2021, p. 5). Luciano Floridi (2023, p. 57) summarizes this problem with the concept of “Explainability”, which is understood as “incorporating both the epistemological sense of intelligibility as an answer to the question ‘how does it work?’ and the ethical sense of accountability as an answer to the question ‘who is responsible for how it works?’” (ibid.). To ensure that AI does not become *de facto* a Black Box, we must guarantee the intelligibility of the system by governance principles of “explainability”.

As Castells indicates (1996, p. 260), there will always be a division of power determined by access to the system's rules, between those he calls the decision-making “Networkers” on the one side, and the “Networked” on the other side, as those operated by a given system. AI operates with the use of symbolic systems as a code-language removed from the dimension of social reality, in a purely conventional structuring. This makes its codes and deciphering dependent on the power to access specific information, which are factors determined by political relations, for example, in the case where companies keep access-keys to algorithmical rules as trade secret and business strategies (such as how the algorithm for social media content-delivery works). Power-dependent access to the black box require spheres external to that of a digital system only, such as the Government determination of the breaking of secrecy to company databases in case of criminal investigation. Either way, these are never accessible to the general public apart from political power clashes. We understand that this configures the weakest point in the pipeline for implementing AI-enhanced ethics, and are especially liable to the processes called by Floridi as “ethics shopping”, “ethics bluwashing”, “ethics lobbying”, “ethics dumping”, and “ethics shirking” (Floridi, 2023, p. 69). All these are processes that relegate ethical discussions to the mere public relations departments of companies, accommodating ethical discourse through one of these means regarding practices that do not provide genuine transparency.

4. The Visibility of Invisible Patterns and the Shifting Norm of Accountability

We therefore start from the premise that the technological advent of AI transforms both the status and the application of consequentialist ethics by introducing techniques capable of revealing patterns that are cognitively inaccessible to direct human interpretation

(Tsamados *et al.*, 2022, p. 226). This new technological landscape, we argue, opens the possibility for a profound revision of legal epistemology. For instance, a judge could impute criminal liability to an individual who failed to consider data available through widely recognized and accessible AI systems before making a given decision. In such cases, refusal to consult this information may be interpreted as “technical negligence” or even as “administrative misconduct by omission”. Human–machine hybridity represents a genuine novelty in the fields of axiology and metaethics. It marks a new frontier for ethical reflection, one in which moral decisions emerge from hybrid processes that combine logical–biological reasoning with computational–virtual predictions.

Considering that the decision, in consequentialist ethics, has its ultimate validity through the elimination of all foreseeable consequences, AI may come to be understood as a perceptual augmentation, an increase in observational amplitude, or an epistemic extension whose negligence in considering may constitute blatant omission. The problem, therefore, arises from the imputation of guilt through accountability as a “due diligence”, which comes to encompass AI as an augmentation of human cognition for decisions that assume consequentialist ethics as an alternative. The hypothesis for “culpability” then derives from the same basis, on the premise that, once the potential for increased predictive power with AI is possibly known by the agent, any decision that does not consider it becomes negligence and automatically imputes culpability.

Examples of practical application span virtually all areas, such as cases where machine learning algorithms can identify associations in health, safety, and economic data that allow for preventive interventions before damage materializes (Chakraborti, 2023). This “hyper-vision” means that the justification for moral action ceases to be a narrative based on macroscopic evidence and becomes grounded in highly complex statistical inferences about “small patterns” (Floridi, 2014, p. 15) that extend, even at the hypothetical level, the human obligation to consider the machine's perspective before making decisions. These “small patterns” are “patterns with real added-value [that] lie in their immense databases” (ibid.) that only machines can extract, making humans, therefore, dependent on it. All areas incessantly produce tons of data, containing immense amounts of potentially useful information, such as curing cancer by cross-referencing all digitized diagnoses in the world, but a human being would never be able to access these “small patterns” within the data set by itself. Imputation of culpability by implementing “due diligence” to AI consulting of scenarios for decision-making, ties human ethics to non-human stances for the first time in history.

The logical conclusion we draw is that AI not only assists in consequentialist ethics, but necessarily redefines it from now on: moral responsibility now encompasses the management of risks that were previously considered “acts of God” or mere unforeseen fatalities (Floridi, 2014, p. 42) or any other imagined reasons that supported the dark areas of decision-making. Hence, we are faced with the fact that historically we simply considered it impossible to navigate such large volumes of data that we ourselves create. We can refer to this very process as “erosion of culpable ignorance”, meaning, a gradual shrinking of dark areas of knowledge that might make culpability less rigid. The less ‘assumably knowable’ the scope of data is, the smaller the area of potential culpability.

5. Risks: Dependence and Ethical Deskilling

Acting on the hypothesis of AI-Augmented Consequentialist Ethics (AI-ACE), which we understand as being inescapable in face of its clear advantages, we foresee the emergence of other accommodative demands, such as the generation of dependence on this hyper-vision created by AI, and we highlight this issue as a one that poses severe existential risk regarding “the psychological, social, and political impact of AI” (Kazim, 2021, p. 4). The most prominent of these risks is the ethical deskilling, as indicated by researcher Shannon Vallor (2015) in her article entitled “Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character”. Moral and ethical are equivalent, although we prefer the term “ethics” to distance from its religious “moral” counterpart. In short, her argument centers on the condition that, as ethical deliberation is outsourced to algorithms, human beings lose the ability to practice independent ethical judgment (Vallor, 2015; Liu *et al.*, 2022 p. 438).

The human agent becomes a passive agent under these circumstances, reduced to only choosing whether or not to follow the algorithmic recommendation of the machine. This has consequences both behaviorally and in neuroplasticity, where humans gradually reduce exercising the reflection necessary for the development of an action considered culturally virtuous. The reduction goes as follows: “As the human agent does not need to understand the link between their moral values and the AMA’s moral advice, they might be reduced to having only one choice of whether or not to conform to the Artificial Moral Agents” (AMA) (Liu *et al.*, 2022 p. 439). Some of us might come to a point of not having to exercise choice in ethical dilemmas at all, as Vallor puts it: “the systematic replacement of reflective discernment with mindless prediction; the efficient sacrifice of shared flourishing to expected utility; the exchange of humane creativity and open-ended progress for local optimization of

content delivery” (Vallor, 2024, p. 11). This boundary of responsibility becomes even more complex in high-tech data systems.

In Artificial Intelligence scenarios, for example, the opacity of the system (the black box problem) can make the results of an algorithm impossible to predict, even for its creators. We read that “According to the recent literature, factors contributing to the overall lack of algorithmic transparency include the cognitive impossibility for humans to interpret massive algorithmic models and datasets.” (Tsamados *et al.*, 2022, p. 218). For example, if a human-machine team encounters conditions that differ significantly from those they have previously encountered, the operator’s mental model may not be able to anticipate how exactly the system will respond.” (Michel, 2020, p. 10), reflecting the inherent problems of “understandability”, due to the different nature of reasoning between humans and their machines. Even considerably transparent systems, once working together, may lose transparency and understandability. Consider that “multiple individual AI agents (for instance, a swarm) operate as a collective network to achieve a goal, the process by which that goal is achieved may be highly uninterpretable” (Idem).

If a future harm becomes genuinely unpredictable due to computational complexity, the traditional framework of consequentialism tends to protect the developer from direct moral guilt, since moral obligation (the “duty”) is limited by the ability to act based on available information - the principle that “ought implies can” (Floridi, 2023, p. 183; Chakraborti, 2023, p. 265). Furthermore, reliance on opaque systems (black boxes) can lead to “agency laundering”, where individuals and organizations avoid responsibility for morally questionable decisions by blaming the algorithm (Floridi, 2023, p. 109). This leads us to conclude that if the agent does not understand the reason for the decision (lack of explainability), it loses its autonomy and becomes a mere interface of the technology.

The dispute over decisions between humans and non-human agents (AI) introduces a dilemma of existential responsibility and importance. AI is an agency without intelligence, capable of solving problems without needing to understand the meaning of its actions (Floridi, 2023, p. 46). However, moral responsibility demands intentionality and consciousness, qualities that AI does not possess (Chakraborti, 2023, p. 648). From an existentialist perspective, the choice of an advisor (whether a priest or an AI) is, in itself, a commitment. But a choice made by the human agent cannot have responsibilities over their consequences delegated (Liu *et al.*, 2022 p. 439). The dispute arises when the “best” technical decision (provided by the AI) conflicts with human judgment. We see that “algorithms are implemented with the hope of being more neutral”, that is, “thereby



suggesting that the decisions are better than those performed solely by individuals” (Martin, 2019, p. 837). If the human yields to the machine to maximize results, they may, at best, be preserving material well-being but sacrificing their dignity as an autonomous agent (Floridi, 2023, p. 106). The “many-hands problem” in AI systems makes the attribution of blame diffuse, creating a vacuum of responsibility (Tsamados *et al.*, 2022). where no one is held accountable for future harm, despite increased predictability.

Conclusions

Artificial Intelligence permanently alters the status of consequentialism by raising the available predictability beyond the biological level. AI-Augmented Consequentialism requires that human agents assume responsibility for a significantly broader range of future consequences once they are aware of the predictive expansion made possible by data analysis. In this way, the status of culpability is altered, in what has been called the “erosion of culpable ignorance,” occupying previously blind areas, the dark spots, with accountability through technical amplification. AI-Augmented Consequentialist Ethics (AI-ACE) makes appeals to ignorance increasingly illegitimate as moral excuses, since the sphere of uncertainty is substantially reduced by the predictive potential of AI applied to large volumes of data. What was previously an inability to analyze vast datasets in search of relevant micro-patterns is now being systematically overcome. This new investigative capacity is not only feasible but increasingly constitutes a normative requirement within consequentialist decision-making, which is grounded in the anticipation of outcomes. Consequently, ignoring such predictive resources may be interpreted as administrative misconduct, negligence, or a failure to exercise due diligence.

The price of this expanded ‘vision’ is the risk of moral atrophy and the loss of decisional autonomy. The adequate fulfillment of expectations regarding AI systems as powerful elucidators of future scenarios through the capture of patterns invisible to humans places humanity in an oracular relationship with machines. The constant recourse to their verifications, considered more reliable, coupled with the obligations now established by the certainty of increased predictability, establishes a relationship of human dependence on machines as the source of determinations that enable a decision, even if it is a purely ethical, moral, human decision. Here, we envision a worrying scenario where the occlusion of black boxes, the technological hermeticism of AI systems, can prevent the full verification and revalidation of their results and internal processes, feeding back into human action based on errors in readings and algorithmic manipulations, which, in an advanced scenario



contemplating human dependence on machines for decisions, constitutes a potentially catastrophic condition. The lack of speed caused by a lack of transparency can severely delay identifying the source of errors in misreadings that lead to poor predictions and flawed judgments when using AI. This is the most problematic scenario considered here.

In short, while technology should be seen as an agency resource to be governed and not as a substitute for practical human wisdom, the ability to read vast fields of data creates a conundrum for humanity: it makes AI a technology that must be incorporated into the decision-making process, while simultaneously engendering in AI-enhanced decisions the risk of misreading and manipulating deep data that could lead to poor judgment or tragedies. For humanity not to become a passive "bug" in the system it created, it is vital that existential responsibility remains firmly anchored in human judgment, ensuring that AI's hyper-vision serves to augment, not replace, our ethical maturity. However, governance over the weight of directions provided by AI systems or purely human decisions is still far from being developed. Meanwhile, the AI-Augmented Consequentialist Ethics (AI-ACE) tends to expand discreetly and unofficially in its scope and application. This transformation is occurring as ordinary citizens increasingly turn to AI for advice, business owners rely on it for management consulting, and legal professionals and legislators routinely integrate it into complex decision-making processes. Within these networks, AI acquires the status of an actor in the sense proposed by Bruno Latour, or of an agent endowed with agency in the terminology of Luciano Floridi. This widespread adoption therefore intensifies the need for coordinated communication and systematic governance, as societies confront the new normative conditions introduced by AI-Augmented Consequentialist Ethics (AI-ACE).

References

- BIJKER, Wiebe E.; LAW, John. **Shaping Technology/Building Society**: studies in sociotechnical change. Cambridge, MA: MIT Press, 1992.
- BRINK, David O. Consequentialism, the separateness of persons, and aggregation. *In*: PORTMORE, Douglas W. (ed.). **The Oxford handbook of consequentialism**. New York: Oxford University Press, 2020. p. 378–398.
- BYKVIST, Krister. Consequentialism, ignorance, and uncertainty. *In*: PORTMORE, Douglas W. (ed.). **The Oxford handbook of consequentialism**. New York: Oxford University Press, 2020. p. 310–330.
- CASTELLS, Manuel. **The rise of the network society**. Oxford: Blackwell, 1996.
- CHAKRABORTI, Chhanda. **Introduction to ethics**: concepts, theories, and contemporary issues. Cham: Springer, 2023.
- CRAIG, Edward. **The shorter Routledge encyclopedia of philosophy**. London/New York: Routledge, 2005.
- FLORIDI, Luciano. **The ethics of artificial intelligence**: principles, challenges, and opportunities. Oxford: Oxford University Press, 2023.
- FLORIDI, Luciano. **The fourth revolution**: how the infosphere is reshaping human reality. Oxford: Oxford University Press, 2014.
- HOLLAND MICHEL, Arthur. **The black box, unlocked**: predictability and understandability in military AI. Geneva: United Nations Institute for Disarmament Research, 2020.
- HOOKER, Brad. The role(s) of rules in consequentialist ethics. *In*: PORTMORE, Douglas W. (ed.). **The Oxford handbook of consequentialism**. New York: Oxford University Press, 2020. p. 441–460.
- HURLEY, Paul. Consequentializing. *In*: PORTMORE, Douglas W. (ed.). **The Oxford handbook of consequentialism**. New York: Oxford University Press, 2020. cap. 2.
- JACKSON, Frank. Consequentialism and action guidingness. *In*: PORTMORE, Douglas W. (ed.). **The Oxford handbook of consequentialism**. New York: Oxford University Press, 2020. cap. 17.
- KANT, Immanuel. **Metaphysics of morals**. Petrópolis: Vozes, 2013.
- KAZIM, Emre; KOSHIYAMA, Adriano S. A high-level overview of AI ethics. **Patterns**, v. 2, n. 100314, 2021.
- LATOUR, Bruno. **An inquiry into modes of existence**: an anthropology of the moderns. Cambridge, MA: Harvard University Press, 2013.
- LAZOVIC, S. **Perspectives on AI, machine learning and predictability in the Italian legal system**. Tese (Doutorado) – Università degli Studi Niccolò Cusano, Roma, 2022.



LIBERMAN, Alida. Consequentialism and promises. *In*: PORTMORE, Douglas W. (ed.). **The Oxford handbook of consequentialism**. New York: Oxford University Press, 2020. cap. 15.

LIU, Yinqi; MOORE, Adam; WEBB, J.; VALLOR, Shannon. Artificial moral advisors: a new perspective from moral psychology. *In*: **Proceedings of the ACM/AAAI conference on AI, ethics, and society (AIES '22)**. New York: ACM, 2022. p. 1–10.

MARTIN, Kelly. Ethical implications and accountability of algorithms. **Journal of Business Ethics**, v. 160, p. 835–850, 2019.

MCLUHAN, Marshall. **Understanding media: the extensions of man**. São Paulo: Cultrix, 2012.

PETERSON, Martin. **The dimensions of consequentialism: ethics, equality and risk**. Cambridge: Cambridge University Press, 2013.

RAILTON, Peter. **Facts, values, and norms: essays toward a morality of consequence**. Cambridge: Cambridge University Press, 2003.

TSAMADOS, Andreas *et al.* The ethics of algorithms: key problems and solutions. **AI & Society**, v. 37, p. 215–230, 2022.

VALLOR, Shannon. Moral deskilling and upskilling in a new machine age. **Philosophy & Technology**, v. 28, n. 1, p. 107–124, 2015.

WOOD, Nathan. **Virtue rediscovered: deontology, consequentialism, and virtue ethics in the contemporary moral landscape**. Lanham: Rowman & Littlefield, 2020.

DADOS DE PUBLICAÇÃO

CONTRIBUIÇÃO DE AUTORIA

Concepção e elaboração do manuscrito: M. G. Bruzzo.

USO DE INTELIGÊNCIA ARTIFICIAL

Os autores utilizaram o ChatGPT (Gemini 3) para revisão gramatical do texto em inglês. Todo o conteúdo argumentativo foi desenvolvido integralmente pelos autores.

FINANCIAMENTO

Não se aplica.

CONFLITO DE INTERESSES

As pessoas autoras declaram não haver interesses conflitantes.

DISPONIBILIDADE DE DADOS DE PESQUISA E OUTROS MATERIAIS

Os dados completos foram publicados no próprio artigo. Todo o conjunto de dados que dá suporte aos resultados deste estudo está incluído no corpo do artigo.

LICENÇA DE USO

As autorias cedem à revista ethic@ os direitos exclusivos de primeira publicação, com o trabalho simultaneamente licenciado sob a Licença [Creative Commons Attribution](#) (CC BY) 4.0 International. Essa licença permite que terceiros remixem, adaptem e criem a partir do trabalho publicado, atribuindo o devido crédito de autoria e publicação inicial neste periódico. As autorias têm autorização para assumir contratos adicionais separadamente, para distribuição não exclusiva da versão do trabalho publicada neste periódico (ex.: publicar em repositório institucional, em *site* pessoal, publicar uma tradução, ou como capítulo de livro), com reconhecimento de autoria e publicação inicial neste periódico.

PUBLISHER

Universidade Federal de Santa Catarina. As ideias expressadas neste artigo são de responsabilidade das pessoas autoras, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES

Darlei Dall'Agnol

HISTÓRICO

Recebido em: 06-04-2026

Aprovado em: 31-08-2026

Publicado em: 30-09-2026

