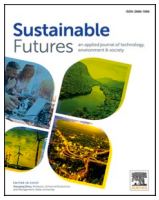
Sustainable Futures 11 (2026) 101644



Contents lists available at ScienceDirect

Sustainable Futures

journal homepage: www.sciencedirect.com/journal/sustainable-futures



AI on the path to good decisions

Mindaugas Kiškis

Faculty of Public Governance and Business & MRU Law School, Mykolas Romeris University, Lithuania

ARTICLE INFO

Keywords:
Ai
Artificial intelligence
Decision-making
Good decisions
Accountability
Ethical AI
Collective intelligence

ABSTRACT

This article examines objections to the use of artificial intelligence (AI) in decisions affecting humans and argues that most objections rest on anthropocentric assumptions rather than evidence about decision quality. Critics often presume that human decision makers are uniquely capable of moral and contextual judgment, while AI systems are inferior, opaque or hostile. The article challenges this view arguing that modern AI systems are an expression of collective human intelligence and ethics, built on the image of human cognition, trained on curated human decisions and operating on best decision making frameworks. Drawing on empirical studies the article shows that modern AI systems offer unique advantages in decision-making, thus may already in some domains be as good at making good decisions as an average individual human. The article deconstructs main criticisms of AI decision-making and introduces the novel argument of inseparability between human and AI decisions. Human and AI contributions are increasingly intertwined, AI involvement is latent and appropriated by humans, making existing accountability frameworks based on a clear human–AI boundary obsolete. The article advocates for the development of agnostic decision-making frameworks that apply universal accountability to both human and non-human agents and provide a path to better decisions.

1. Introduction

As artificial intelligence (AI) systems are increasingly deployed in law enforcement, the justice system, healthcare, education, financial services, policing of social networks, etc., AI becomes directly involved in daily decision-making affecting human lives, fundamental rights and freedoms ([17], p. 211).

This shift in the way many decisions are made leaves many humans uneasy and sensitive to misconceptions and speculations on the impact of AI. While many academics, policymakers and decision makers are optimistic that AI decision-making is improving the quality of decisions [40], a vocal opponent group also stokes public fears about catastrophic AI failures, and some advocate for the slowing down the advancement of AI decision-making or even for limiting it altogether [59].

Common arguments against AI decision-making are opinions that are light on empirical evidence, and often based on metaphysical conjectures or sci-fi based doomsday scenarios. Some arguments openly play on existential human fear of the unknown and unpredictable consequences of AI technology [69]. While some of these fears may be genuine, they are also driven by various selfish and self-promotion interests competing to conjure the scariest scenario. Fear is one of the most primal human emotions and therefore arguments playing into human fears exert a powerful influence on law, policy and society [8].

Besides pure fiction, a common line of arguments of the opponents of the AI involvement in decisions affecting humans is a general anthropocentric worldview. It is proposed that human decision-making is superior to the AI and algorithmic decision-making, because human intelligence is special and superior to machine intelligence. Human intelligence is said to have certain metaphysical qualities, articulated in terms of agency, morals and complex emotional states such as love, empathy or conscience. Thus, AI decision-making is deemed inferior to the full complexity of human decision-making, especially when it comes to deciding about humans, because it is presumed to lack moral qualities. Making decisions about humans is claimed to require special emotional and moral considerations that AI is not prepared to make and might never be able to make. The argument is expanded by arguing that machines decide quantitatively and, especially when decisions have a moral or value judgment or involve human lives and behaviour, machines make bad decisions (Matsumi, Solove, 2024). The human superiority arguments are often supported by pointing out various mistakes made by AI agents.

Empirical research on human vs AI decision quality is mixed and inconclusive. Both camps have sufficient case studies to stand their ground, but data is case and algorithm specific, as well as limited in sample sizes [4,6]. Recent work on AI-informed decision-making in applied domains similarly highlights trade-offs between accuracy,

E-mail address: mkiskis@mruni.eu.

https://doi.org/10.1016/j.sftr.2026.101644
Received 3 November 2024; Received in revised form 26 November 2025; Accepted 1 January 2026
Available online 8 January 2026
2666-1888/© 2026 The Author. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

fairness, and explainability [3,18]. What is notable is that research on bad AI decision is often based on biased datasets that were used to train a specific algorithm, and can be rectified relatively easily, while research on bad human decisions does not suggest any easy or simple solutions on how to correct it [61]. Moreover AI technology is basically in its infancy and relies on few decades of research, while human societies had millennia and yet human agents remain prone to bad decisions at individual, group and even whole society levels. These limitations, however, do not preclude anthropocentric AI scholars' claim that it is the human contribution to decision-making that somehow leads to good decisions (Matsumi, Solove, 2024). For features of good decisions, emphasis is placed on qualitative metaphysical modalities of human decisions.

The central thesis of this article is a rebuttal of the claims that AI involvement inevitably leads to bad decisions. To present this thesis the article critiques the main arguments against involving AI in decisions affecting humans (primarily meaning legal and policy decisions). The author argues that human decision-making becomes deeply entangled with AI, thus making human and non-human decisions practically and normatively inseparable. Moreover, if AI systems embody collective human intelligence and morals, such AI systems can match or surpass the decision quality of individual humans. Claims that AI leads to bad decisions produce wrong incentives for the human decision-making. Instead of limiting the applications of AI in decision-making, policy and scholarly attention should focus on causes and accountability for bad decisions, as well as agnostic good decision frameworks, that apply equally to human and non-human agents. Such frameworks are severely lacking in broader social sciences, and mostly confined to a niche of decision science in economics. Employing AI is instrumental on a path to better decisions and allows uncovering modalities of good decisions. Shifting the blame for the problems of human decision-making onto AI conceals persistent structural problems in human decision systems.

To structure the argument, in Section 2 the author argues that modern AI is evolving into a collective image of a moral humanity. In Section 3, the author addresses main arguments of the anti-AI decision-making camp. In Section 4, the author highlights advantages of using AI for making better decisions. The article concludes that better decisions need broader agnostic decision-making frameworks and more accountability for decision makers whether they are human or not, rather than metaphysical discussion of human decision virtuousness.

The article is primarily conceptual and normative, and offers novel normative arguments rather than testing specific hypotheses. The limitations of this approach should be acknowledged explicitly. The normative arguments are aspirational, and need further empirical validation. As it is surveyed in the article, current empirical literature on comparative decision quality of human and AI agents is limited and inconclusive. It is used to support the argument that AI makes imperfect decisions, while actually demonstrating that AI is not worse and often better than humans at decision making. Particularly studies in criminal justice, credit scoring and medical diagnosis suggest that even imperfect algorithms can match or outperform human judgement on key metrics, while critique of AI decisions is mostly perceptional and normative [3,4,41].

Separate notable limitation of the article is that the perspective offered is focused on Euro-American legal and policy debates and limited review of literature on human v AI decision making. The topic is rapidly evolving and some research relevant to the article may have been missed. Cultural and societal differences in perceptions of AI, and in ethical principles guiding decision-making, are only indirectly addressed, further comparative, cross-cultural work is needed to examine how anthropocentric bias and inseparability manifest across different normative orders.

The article does not prescribe institutional designs, metrics or governance frameworks. Developing such frameworks will require interdisciplinary collaboration between computer science, law, philosophy, sociology and decision sciences.

2. AI as model of collective human intelligence and ethics

Recent progress in artificial intelligence (AI) technology is largely attributable to advancements in neuroscience and a better understanding of the human brain [31]. Neuroscience has directly advanced AI technology, especially in key fields such as neural networks and deep learning [37]. The architecture of neural networks in AI is inspired by biological neurons, synapses, and the process of learning through synaptic strength adjustments. These insights from neuroscience have been abstracted into the algorithms and code underpinning AI systems [38].

The best example of brain science influencing AI is the development of deep learning that was significantly inspired by the hierarchical structure of the brain's visual cortex, which processes information in layers. This biological model has informed the design of more efficient AI systems. Similarly, the concept of Hebbian learning, based on the idea that "neurons that fire together wire together," has shaped reinforcement learning and unsupervised learning techniques in AI. The deeper understanding of human cognition and decision-making has further enabled AI models to better mimic human thought processes, as seen in natural language processing. However, the direct link between these architectures and the ethical quality of decisions is still under-theorised and empirically underexplored, which underscores the conceptual character of the present argument.

Neuroplasticity, or the brain's ability to reorganize itself by forming new neural connections throughout life, has inspired the creation of adaptive AI systems that continually evolve and improve. In this sense, humans are, quite literally, building AI in their own image—emulating human-like abilities incrementally, through continuous refinement. And whether some of us want that or not, the AI engineers are slowly succeeding - AI systems already produce outputs that are indistinguishable from human ones.

Further advancements in AI technology may come from innovations in bio-computing directly employing human brain cell organoids [58] or from advancements in quantum computing led by new theories that the brain itself is an advanced quantum computer.

The convergence of biology of the mind and AI also means that AI systems are trained in the same way as human brains, primarily relying on linguistic and visual inputs. The underlying moral maxims in such training also align, i.e. machine ethics is generally based on human ethics. AI systems are learning what is good and what is bad from human cases and decisions, curated by the brightest minds of humanity [29]. The engineers behind AI are striving to embed ethical principles within these systems, ensuring that AI becomes a force for good [65]. Major research and engineering effort is made to identify and eliminate the causes of bad AI decisions at all levels [12]. Certainly, the majority of AI engineers and companies are making their best effort to design good AI, that is AI – which has goodness embedded as a foundational principle. To suggest otherwise undermines the moral intent of those humans designing AI. Of course, this does not imply that all AI systems currently embody shared global values or that corporate incentives always align with public interest, rather it indicates that value alignment is an explicit design goal in AI development.

AI is both cognitively and ethically modelled on human systems, the additional outcome of said convergence is increasing inseparability of human and AI work. It is increasingly difficult to separate human and AI decisions based on outputs generated through humans using technology. Some researchers point out the phenomenon of co-intelligence, where technology is assisting humans with all kinds of work and decisions [45]. In legal and policy decisions the inputs of technology can start with the most basic tasks, such as finding basic information and pertinent sources. Human agents may already want to fully own results of such collaborative work, but advanced outputs offered by LLMs and other generative AI technology further create incentives for human agents to appropriate the output of AI due to potential economic and reputational benefits [50,53]. In scientific publishing, this is becoming a dominant problem, as some commentary suggest that up to 40% of papers in

certain fields show signs of AI involvement [11].

Recent well publicized controversies illustrate how easily AI-generated work can be sold as ostensibly human output. For instance, Gannett and Sports Illustrated were found to have published AI-generated product reviews under alleged human bylines that passed through editorial process [54]. For lay person it is impossible to distinguish whether such allegedly human expert reviewing and recommending a product is actually human, AI, or some hybrid. In the 2023 case of Boris Eldagsen winning one of the world's biggest photography competitions with an image entirely generated by AI, the involvement of AI was uncovered only after the artist came forward himself [50]. Note that the AI tools used in 2023, as of 2025 are obsolete, and the current AI image generation and manipulation capabilities far surpass them. These cases illustrate that, in many real-world settings, human and AI decisions are already inseparable in practice.

In a context of intellectually and morally indistinguishable outputs, it would be extremely condescending to claim that AI is somehow morally inferior. Yet, empirical work in psychology and management science suggests that anthropocentric bias is not merely a philosophical stance but manifests in concrete judgement patterns: people systematically devalue AI-generated art compared to identical human-labelled works [43], and cultural differences shape how societies react to and deploy AI systems [68]. Similar concerns have been raised in media studies regarding anthropocentric framings of "artificial creativity" [60].

Instead of condescending, we should view AI systems as a product of collective intellectual and moral efforts of humanity. Certainly imperfect, as we are ourselves, but continuously learning and improving, aided by the best human mentorship and curation. Understanding of AI systems as systems of collective intelligence surpassing capabilities of individual humans is increasingly gaining recognition and even empirical proof [16,67]. At the same time, these accounts also emphasise that collective intelligence is fragile and sensitive to socio-technical design choices, which cautions against unqualified optimism.

Overall, criticisms of AI use often reveal more about our own limitations as moral beings than they do about the potential of AI to embody moral reasoning. AI is not just some collection of algorithms—it is the collective product of humanity's best effort to create intellectually and morally capable decision-making agent. To scoff at such systems as something inferior to an average human is ignorant of our own collective best effort.

3. Arguments against AI decision-making

In the law and policy debates regarding AI decision-making, a series of arguments are frequently made emphasizing the defects of AI decision-making. All such arguments can be summarized into five categories:

1. AI mistakes and hallucinations argument;
2. Humans are special and better at making decisions argument;
3. AI cannot be held responsible argument;
4. The black box argument;
5. The hostile AI argument.

These arguments conveniently play into the human self-worth and insecurity complexes, which are very powerful primal human instincts, which provoke almost unconscious negative reactions to AI, especially among the general public uninitiated into the intricacies of the AI technology.

Further underpinning these arguments is the claim that human decision-making is extremely sophisticated and somehow intellectually and morally transcendent. Humans are claimed to be the only ones to exert moral judgement, able to make fair and just decisions (Matsumi, Solove, 2024). Only humans can rise to address bias, have unlimited capacity to comprehend and address complexities of life and other humans, etc. As a result, only humans are said to be able to overcome the limitations of AI decision-making and correct the unjust decisions of the AI.

Analysis below proceeds to scrutinize and criticise these arguments. Said arguments have no empirical, logical or moral justification and it is improper to view human decision-making as superior without any objective measure. Instead of glorifying humans, it is more appropriate to focus on agnostic decision-making principles and developing agnostic accountability and good decision frameworks, following the lead of economics scholarship.

3.1. AI mistakes and hallucinations argument

A common set of arguments made against AI and algorithmic decision-making technologies is to present mistakes and examples of AI hallucinations as generalizations of AI decision-making flaws. Lapses in basic logic, data errors or fill-in-the-blank conjectures are, unfortunately, as much traits of the human decision-making, as they are of AI. It is, however, no longer fashionable to mock human flaws, as it is more convenient to mock failures of the aspiring technological intelligence [20]. Although these arguments appear attractive on the surface, they often ignore considerations that should give us pause.

The quick willingness to mock AI mistakes and hallucinations highlights the human anthropocentric bias and a sense of human superiority, which is not a virtue. In the not so distant past humans have openly mocked other humans, which they considered inferior. Even under penalty of the criminality such human behaviour is not uncommon in the modern societies and is even considered acceptable in certain human cultures ([32], p. 294).

There is ample literature on the measurable advantages of AI decision-making [10,14], as there is literature on claimed superiority of human decision-making [25]. Either can be supported with select case studies, but both lack comprehensive empirical validation, also it is possible to find plentiful examples of mistakes and made-up conjectures in either. AI systems can produce bad outputs, especially due to bad input data or parameters, but this fact shouldn't lure us into concluding that humans are better. The only thing that can be assumed with certainty is that both humans and AI are prone to mistakes and hallucinations in their decision-making. Such mistakes and hallucinations can be minimized with better training, better resources and higher cognitive capacities in either human or non-human decision makers. But it is certainly "a lot easier to understand and then fix the algorithm or AI than the human" and "it's a lot harder to change the human or understand why humans make their decisions." [33]. Also, one must not forget that human agents may sometimes intentionally produce corrupt or even criminal decisions. Shifting the blame onto AI failure may just be a convenient instrument for bad meaning human agents to avoid accountability.

The studies comparing human and AI decision making directly are unfortunately not many. The studies most relevant to legal decision making are surveyed for this article (Table 1 below) and overall illustrate that AI is not worse and often is better than humans at decision making. Studies in other fields, e.g. medical diagnostics, are even more convincing [41], but fall outside of the focus of this article.

Recent studies on generative AI error rates and "hallucinations" in large language models similarly show that performance is highly sensitive to training data, prompting and guardrails, and that error patterns can be systematically characterised and mitigated [3,51]. These findings support the claim that, unlike many human cognitive biases, at least some AI failure modes are fixable design problems rather than immutable flaws.

3.2. Humans are special and better at making decisions argument

Anthropocentric scholarship claims a fundamental difference between humans and artificial intelligence in terms of ethics and morality [36]. This distinction is essential when considering the role and

Table 1
Studies on AI vs human decision quality.

Study	Domain / data	Agents compared	Measured qualities	Key findings on AI vs human decision quality
[6]	Criminal justice, recidivism prediction	Human judges vs risk-assessment algorithms (review of studies + normative argument)	Predictive accuracy, racial bias, consistency	Concludes that risk-assessment algorithms can match or exceed human predictive accuracy while being "colour blind" and more consistent. Overall evidence is limited and context-specific.
[4]	US criminal justice, COMPAS risk scores	Commercial recidivism algorithm vs human decisions and re-offending data	False positive/ negative rates, calibration, racial disparity	Concludes that COMPAS scores are similarly predictive overall but produce higher false-positive rates for black defendants and higher false-negative rates for white defendants. Accuracy comparable to humans, i.e. reproducing human biases.
[15]	Criminal justice, bail / sentencing	Human judges vs risk-assessment tools	Crime rates, jail population, racial disparity (simulations & empirical studies)	Concludes that even imperfect algorithms can reduce jail populations and keep crime rates roughly constant compared to human judges, if deployed carefully. Algorithmic tools on average exceed human-only baseline, but do not warrant total fairness.
[61]	General policy / legal decisions	Human officials vs algorithmic decision procedures	"Noise" in judgment, bias, legitimacy	Concludes that algorithms on average substantially reduce "noise" (random variation between human decision-makers), potentially improving overall decision quality. At individual decision level there are always outliers.
[9]	Lab experiments, generic decision tasks	Human decision-makers vs algorithms (perceptions, not performance)	Perceived explainability, perceived understanding	Concludes that humans believe they understand human decisions better than algorithmic ones, even when both are equally opaque.
[25]	Business decision-making	AI decision systems vs human managers	Conceptual mapping + case illustrations	Surveys organisational uses of AI decision-support systems, concluding that in some domains AI improves prediction accuracy and efficiency, but in certain other domains humans remain essential for contextual and ethical judgement.
[3]	Experimental tasks with AI decision support	Humans making choices with/ without different AI explanation types	Objective fairness metrics; perceived fairness; trust; reliance	Concludes that types of explanation can change both perceived and actual fairness in AI-assisted decisions. Feature-based explanations sometimes improve perceived fairness but can also mislead users. Explanation design is itself a decision-quality variable.
[13]	Hiring decisions (vignette experiment, 2 × 2 design)	AI vs human hiring agent; favourable vs unfavourable outcomes	Perceived fairness, procedural justice, trust	Concludes that in some conditions, AI rejections are perceived as fairer than human rejections, particularly because humans are

(*continued on next page*)

Table 1 (*continued*)

Study	Domain / data	Agents compared	Measured qualities	Key findings on AI vs human decision quality
				seen as more biased or self-interested. At least at the level of perceived fairness, AI can surpass human decision-makers.
[57]	Occupation prediction from bios; AI-assisted classification	Humans with AI recommendations plus different explanation designs	Distributive fairness (gender), reliance on AI, correctness of final decisions	Concludes that feature-based explanations do affect fairness perceptions and reliance, but do not reliably help humans distinguish correct from incorrect AI advice. Depending on what features explanations highlight, they can either mitigate or amplify stereotype-aligned errors.
[55]	Criminal justice recidivism tools	Many algorithms vs human/legacy baselines (systematic review)	Accuracy, fairness, calibration, legal adoption	Concludes that algorithms often outperform unstructured human judgment on accuracy, but evidence on fairness and real-world impact is mixed and context-dependent.

limitations of AI in society, especially in areas where moral judgments are crucial. Notwithstanding the fact that different human groups, societies and cultures have varying interpretations of what is considered moral, ethical or even fair [24], only humans are thought to have the ability to understand and interpret moral issues in context, taking into account a wide range of factors like intentions, emotions, cultural nuances, and potential consequences. Only humans allegedly possess a natural capacity for moral reasoning and ethical judgment, only humans can understand complex moral concepts like justice, empathy, and fairness, and can apply these concepts to make moral judgements in varied and nuanced situations [59]. The problem is that this describes a perfect human, which does not exist.

Regardless of this inconvenience, while AI can perform tasks and even mimic certain aspects of human decision-making, it supposedly lacks the intrinsic moral and ethical understanding that characterizes human beings. AI is claimed to lack innate morality and moral reasoning, lack the ability to inherently understand or value moral principles and can only mimic moral decisions based on the parameters and data it is programmed with [59]. Furthermore, AI systems are claimed to lack understanding of context and depth of the situation, as well as the emotional dimension, which may be important for intuitive moral judgments. Strikingly these assumptions in the literature are made either through purely thought conjectures or cherry-picked examples of judgement errors by AI systems, as well as by ignoring a sea of examples of human troubles and corruption in decision-making.

Very often AI systems are taunted as mere collections of algorithms that can never capture the complexities of morals [59]. This argumentation is fundamentally flawed because algorithmic understanding of modern AI is incomplete. Intelligence whether human or in current AI systems is far more complicated than such simplistic arguments let on. Complex neural networks and self-improving deep learning systems already test the boundaries of what can be described as quantitative algorithmic systems. Simplistic arguments assume that machine intelligence will not learn and will be doomed to repeat the same errors, while this is quite possibly more likely with human agents that than with modern self-learning AI agents.

In addition to said primitivism, there are three additional problems with this line of arguments:

First, while it is difficult, it is not proven impossible to set ethical principles into algorithms or train AI to act inherently morally. Moral capacity in humans is poorly understood, and there is plenty of empirical evidence of humans lacking moral capacity [56]. There is also an argument that moral capacity is an inherent property of a certain level of intelligence, and an intelligent entity may possess an innate ability to differentiate right from wrong [39].

Second, it is already noted that many humans lack moral capacity and struggle with major character vices ([23], p. 38-43). Persistence of violent crime in modern societies is a best proof for this. Moreover, moral capacity among humans is not equal and is not universal. Thus, the claim that humans are special is inherently a broad stroke generalization based on the collectivistic vision of an ideal human, which simply does not exist in the real world. Arguably a basic anthropocentric bias is what contributes to the whole humans are special and better argument.

Third, most biases and injustices in AI decision-making are a result of such biases and injustices in human decision-making. Large datasets of flawed human decisions are used to train AI systems, and thus distort AI decision-making. This automatically nullifies the logic of the claims that human decision-making warrants ethical decisions.

Practical aspect of “humans are better” argument is the human oversight requirement for AI systems, in research also referred to as “better together” and “humans in the loop” designs of the AI systems. Human oversight now became a universal regulatory requirement for high-risk AI systems according to the EU AI Act [21] (Article 14), even though it is vague and open to subjective interpretations. Human oversight is warranted to be inconsistent due to lack of standardization of human understanding and behaviour, yet according to the EU AI Act human oversight will be required for the AI decisions in development and deployment stages. It is notable that until and if there are autonomous self-building AI, human oversight is engineering inevitability of AI system design and training. Similarly to humans, AI systems cannot train entirely by themselves without human inputs and assistance.

Human oversight of AI is certainly dependent on the ideal human, assuming that overseer will act with consistent moral and intellectual standards. However, the reality of human decision-making systems is quite different. Many decisions, especially those of legal importance, are subject to multiple layers of political and legal oversight, yet errors, bad decisions are daily phenomenon and society has to rely on a system that often delivers only 'good enough' decisions. This stems not only from intellectual inconsistencies among decision-making humans but also from common negative human traits such as laziness, incompetence, and occasionally, malice or corruption, which may be further exacerbated by technology [1].

Humans are not trusting of machines, they are just too lazy to question them as it requires substantial dedication and own effort.

Research shows that automation can create a diminished sense of control, responsibility, and moral agency among human operators [14]. Humans become complacent and over reliant when working with AI systems, a tendency that cannot be overcome by experience or training [70]. This suggests second order effects of AI, like lessening the pressure on humans to learn to perform better, lessening human learning from own errors, as well as lessening the human responsibility and potential liability for bad decisions, especially if 'bad' AI can be implicated for human failures.

This is why human oversight, will not solve any problems of the AI decision-making and may even augment them. Human oversight assumes that the combination of humans and AI will address the deficiencies of each. All kinds of human oversight designs, whether based on AI supporting human-made decisions or humans approving AI-made decisions, have already been proven to be fallible. Empirical studies show that people readily defer to automated systems, overlook errors in algorithms, and deviate from algorithmic output in ways that render a less accurate result [42]. Moreover, humans cannot reliably balance an algorithm's advice with other factors, as they often over-rely on algorithmic advice and place greater weight on the factors that algorithms emphasize [26,49].

Overall, while mandatory human oversight may be well intentioned, it inadvertently reintroduces a major source of fallibility into AI decision-making. Instead of addressing the core issue of accountability, it is a simple comforter covering up the root problems.

3.3. AI cannot be held responsible argument

Humans are thought to be responsible for their actions and accountable for their moral choices. In contrast, it is problematic to assign responsibility or accountability to AI systems. Therefore, the ethical and legal responsibility for AI's actions currently falls on the developers, operators, or owners of the technology.

Semantics aside, modern human systems have unfortunately limited the accountability for the actions and liability for moral choices, especially when it comes to highest echelon decision makers. Lack of the "skin in the game" has been noted as one of the essential problems that undermines ethical decision-making in political and legal systems, making them fragile to various unanticipated shocks [62] and unable to solve existential long problems of humanity ([28], p. 24-26).

The last few decades are marked by multiple examples of human decision-makers, who do not bear proportionate accountability for their decisions leading to disastrous decision-making. The Russian invasion of Ukraine, the 2008 Financial Crisis, and even the Covid-19 crisis provide ample examples where decision-makers that made wrong calls remain insulated from the legal consequences for their actions or failures. These examples are situations of the so called moral hazard - a lack of accountability and disregard for the harm imposed onto others. Modern administrative, legal and political bureaucracies have become adept at diffusing accountability across the system itself, making it difficult to hold any one accountable. This phenomenon arises from the complex, opaque structures of modern governance, where decision is the product of a whole system rather than of individual decision maker [66]. This is not inherently a failure, it just highlights the difficulties in applying accountability when failures occur.

It is thought that the fact that human decisionmakers have to live with the moral consequences of their decisions is sufficient accountability. But this equally applies to the developers and curators of the AI systems, who will have to live with the moral consequences of the AI impact on humanity, in the same way as the developers of the nuclear weapons had to morally assume the use of such weapons against thousands of Japanese civilians, famously acknowledged as "blood on my hands" by J. Robert Oppenheimer ([46], p. 493–494).

In this context, the argument that it is problematic to assign responsibility or accountability to AI systems is fairly hypocritical and irrelevant. If anything, the accountability that is now imposed according to the EU AI Act onto the developers and operators of the AI system ensures as much if not greater level of accountability than what is available in human-only decision-making systems. More important issue is explicability of decisions, which is applied even more asymmetrically to AI decision-making.

3.4. The black box argument

The requirement for explicability has been proposed by ethicists and adopted by policymakers as a foundational ethical principle of AI [22]. This principle essentially extends an anthropocentric approach to AI and is somewhat disingenuous, as neither human status nor human decisions require explicability, yet lesser entities are held to this standard. In the literature, this principle is often referred to as the "black box" problem, highlighting that modern AI systems increasingly function as opaque entities, where the exact algorithmic process by which specific inputs lead to certain outputs is unclear [7]. Inconveniently, black-box decision-making is also a common feature of human decisions, sometimes even codified in law—for example, a judge or jury member act on internal conviction, a politician votes based on conscience, individuals can hide behind the secrecy of group ballots.

Explicability is not a requirement for many human decisions, and even when it is, the standards are low and inconsistent, with little accountability for inadequate explanations. Yet, unlike human intelligence, AI systems are expected to be fully transparent, with their decisions comprehensible to humans to increase trust, demonstrate how decisions are made, and prevent biases or discrimination in AI [47]. Trustworthy AI has therefore emerged as a major research area, with particular emphasis on counterfactual and contrastive explanations that allow affected individuals to understand under which hypothetical conditions a different outcome would occur [3,18].

In human decision-making, explicability is excluded because the "black box" nature of the human brain is considered an inherent feature of human decision-making [9]. If all decisions required elaborate elucidation, it would likely result in decision paralysis.

The human brain is the ultimate black box, and thus, AI systems designed to mimic brain processes are similarly prone to the black-box problem. The human brain's opacity can be particularly severe, as we have yet to fully understand or prevent violent and immoral behaviour. It is safe to assume that some individuals possess a "heart of darkness," and most of us harbour dark corners in our consciousness ([5], p. 237). Adult humans are relatively static, and we can never be certain of internal changes, even if external behaviour suggests otherwise. Good intentions, nurturing, and elite education are often insufficient to overcome the darkness within the human soul.

The idea of a perfect algorithm or AI system is as fallacious as the ideal human. The black-box problem is certainly expanding, at least in the sense that explaining the rationale behind AI decision-making now requires specialised expert knowledge. The more complex (and intelligent) the system, the closer it approaches a point where explanations also become exceedingly complex. Ironically, this is also a hallmark of modern political and legal systems.

AI systems can be curated to correct logical or moral flaws in their training data and decision-making processes. As noted, it is certainly much easier to fix AI than to fix human decision-makers. Increasingly sophisticated methods are being developed to explain AI decisions, and a whole new subdomain of AI research—Explainable AI (XAI)—has emerged to address this as a core issue [44]. Explicability requirements act as a form of accountability in AI decision-making, contrasting starkly with the unexplainable human-only decisions. From the perspective of this article, such tools are best understood as components of an agnostic decision framework that applies symmetrical demands for explicability wherever technically and institutionally feasible, irrespective of whether the agent is human or non-human.

3.5. The hostile AI argument

It can only be assumed that human history is the main source of the argument that superior humans tend to abuse perceived inferior beings, be it other humans, animals, ecosystems, etc. As it was noted, despite best efforts in design of social systems in education, healthcare, welfare, etc. it was not possible to overcome moral lapses of humans against humans. However, if we consider intelligence agnostically, there is no reason to believe that moral intelligence per se must be somehow hostile to other forms of intelligence.

The plausible possibility exists that higher intelligence possesses innate ability to know right from wrong and prevent wrongdoing and there are no specific reasons why non-human intelligent entities would need to harm human intelligence. It is at least plausible that agnostic intelligence will adhere to Kant's categorical imperative. AI, especially based on the collective intelligence and morals of an aspiring ideal human, may be able to overcome individual human vices, resist malicious exploitation, seek to avoid accidents and achieve better outcomes for the largest number of human subjects even without human moral oversight. Consequently, we should be more concerned with humans abusing AI systems in order to harm other humans, than AI systems abusing humans for AI ends themselves.

Max Tegmark is generally known as one of the most prominent AI pessimists, and however yet the main argument that he makes against AI is not agnostic. The risks of unintended consequences and adverse scenarios can never be excluded, just as they are not preventable in human systems ([63], p. 260), i.e. the hostile AI risks are not AI specific, they are assumed on the example of the human systems. Such arguments do not, by themselves, support the claim that AI is inherently predisposed to harm humans. In fact, Kantian ethics suggests otherwise, asserting that pure intelligence should be able to autonomously choose the categorical imperative. This philosophical argument requires further engagement with moral theory and cannot be empirically verified. It is presented as a conceptual counterpoint to dystopian narratives, rather than as a descriptive claim about existing AI systems.

4. Appreciating the AI decision-making

A lot of critique of AI decision-making stems from simplistic presentation of AI as algorithmic and entirely quantitative, which is inconsistent with modern AI technology. Based on evolutionary algorithms, algorithms based on deontic logic, deep learning, and neural networks that adapt to new data without human directives, modern AI systems already have at least rudimentary ability to make decisions and take actions based on their internal processes and values, rather than merely following pre-programmed instructions or reflectively reacting to set data points.

Focus on quantitative probabilistic algorithms in AI decision-making is a representation of human limitations in understanding non-human intelligence. It is not right to straw-man such algorithms as representation of the whole field of AI decision-making, as they are often presented in social science literature (Matsumi, Solove, 2024). Even if it is agreeable that problems emerge when only quantitative data is relied upon and qualitative data is excluded or limited, as the empirical evidence is aplenty in human decision-making, simplifications and generalisations based on rhetorical devices present a risk of unreasonably undermining trust in robust but complex mathematical and computer science methods underpinning AI systems. Claims that AI struggles when told to appreciate inconsistent or unexpected is a typical generalization. The fact is that all current decision-making systems (including human systems) struggle under such circumstances.

Further critiques of AI decision-making are often exemplified by narrow case studies, such as predictions of criminal recidivism. In these cases, the ideological bias of the commentator and the biases in the datasets used to train AI play as much role in the critique, as the asserted bias in the decisions made by the AI system. Especially if we specifically look for biases and disregard the long term outcomes and full scope of all available data, claims of bias can be conjured regardless of the methodology based on the theories of justice practiced by the commentator alone.

Overall, beyond metaphysics and rhetorics there is no certainty that human decision-making is somehow better, and it struggles with all the same problems of which AI decision-making is accused. AI decision-making offers at least some unique benefits. AI can lower the cost of complex decision-making and make it economically accessible [2]. Classical economic theory has long recognised that no individual decision-maker can access or process all relevant information [30]. AI systems explicitly address this by pooling and processing vast amounts of dispersed information through multiple decision making processes. By design, AI systems embody the idea that decisions emerge from collective rather than purely individual intelligence.

AI decision-making also offers advantages of explicability, easier failure correction and collective morals of an ideal human, which are unique. AI decisions are faster, universally accessible, and verifiable through multiple pathways such as through independent human agents or alternative AI agents. Explicability requirements for AI decision transparency well exceed the transparency requirements for the human decision makers. This certainly increases the richness of considerations in the decision-making. Individual human decisions are deterministically bound to represent human flaws, however due to economic and social reasons we do not have the luxury to avoid them (e.g. single judge in lowest instance courts). AI decisions inherently represent collective human effort and essentially are decisions made by collective rationale in addressing the issue. AI systems as an image of a collective of human intelligence and morals have better shot at being ideally human, than individual humans or even groups of individual humans. At the very least combining AI decision-making with human decision-making in non-subordinate way, i.e. as competing independent decisions, may offer novel and important ways to develop agnostic good decision frameworks. Developing such frameworks would benefit from integrating decision theory, empirical AI evaluation metrics (e.g. accuracy, fairness, accountability indices) and cognitive science perspectives, which is an agenda for future work.

Decision science subfield of economics research has long developed decision frameworks that are effectively agnostic as to whether the relevant agent is human or non-human. For example principal–agent models of algorithmic decision-making recognize the problem of agent agnostic design [19,27]; welfare-based approaches to algorithmic fairness evaluate decision rules independent of whether they are implemented by judges or neural networks [34,35,52]; recent work on level-agnostic economic theory and the economics of AI further emphasises that human and artificial agents can be analysed within the same decision-theoretic frameworks [48,64]. This research is particularly relevant for AI decision making, and needs serious consideration in domains of law and broader social science decision making.

Last, but not least with the latest generation of AI technology we are at the junction, where we can never be sure about the origin of a particular decision, i.e. whether it was made by human or non-human intelligence, as long as the human agent claims to own said decision. Since almost all decisions already involve some level of algorithmic assistance, human and algorithmic contribution is simply inseparable in the outcome. Incentives for humans to completely own AI decisions are rapidly increasing with the increasingly economically useful outputs that AI systems are producing. We may only talk about certain probabilities that decision may be made by non-human agent, which is insufficient for human accountability and legal outcomes as long as we are allowing human a benefit of doubt. Unless we challenge the fundamental principles of human accountability, we will never be able to differentiate between human and non-human decisions. This underlines the main normative problem – existing accountability frameworks based on clear human–AI boundary fail completely when decisions are made through collaboration of humans and AI, especially where AI

involvement is intentionally or unintentionally latent. Inseparability does not only create a practical problem for accountability, it also suggests a conceptual shift. If human and AI decisions cannot be separated, then decision frameworks must become agent-agnostic by design. Economics decision theory already offers outline of such framework that is more useful compared to metaphysical arguments.

5. Conclusions

Human and non-human decisions affecting humans are one and the same, therefore should be subject to the same rules and frameworks. It is becoming increasingly difficult to separate human from non-human decisions, particularly when humans are willing to appropriate AI-driven decisions and present them as their own. The probabilistic benefit-of-the-doubt principle, which underpins human accountability rules in modern society, will soon eliminate any possibility of distinguishing between human and AI outputs. This also complicates efforts to introduce differentiated accountability frameworks. As incentives grow for humans to claim credit for good AI decisions and shift blame for bad human decisions onto AI, this problem will become even more apparent.

This problem highlights the urgent need to rethink frameworks for good decision-making and human role within such frameworks. Anthropocentric decision-making models, and their often condescending views of AI, prescribe solutions such as human oversight, which is simplistic and naïve approach. As many anthropocentric AI scholars agree, and as demonstrated from non-anthropocentric point of view in this article, merely adding human oversight to AI decision-making may be counterproductive, encouraging humans to deflect blame for their own failures onto AI.

The overarching goal is to produce good decisions, regardless of the decision-making agent. However, what are the modalities for such agnostic decision-making? This critical problem receives much less attention in social science than the criticisms aimed at AI decision-making. Ephemeral features such as fairness, morality, impartiality, diversity, philosophical depth, open-mindedness, contextual awareness, and adherence to rulebooks, that dominate law and policy sciences, have not been particularly helpful in defining good decisions beyond mere hope that a human decision-maker's subjective interpretation aligns with the metaphysical meaning of these principles. It is troubling that many commentators believe these foundational principles are inadequately addressed by the tremendous collective efforts of AI developers and curators. This reflects a deep-seated mistrust of other humans on both intellectual and moral grounds.

We should embrace AI as an important tool for better decision-making. As a collective embodiment of a moral ideal, AI represents our best hope for producing good, or at least better, decisions, as well as for improving human decisions, especially, if we assist it with agnostic decision-making frameworks also emerging from economics research, and allow it to assist humans in a non-subordinate way. Since AI is already omnipresent in decision-making, it may be appropriate to rethink AI and human decision-making not in a hierarchical manner but as competing, independent processes. This could provide a novel and important approach for developing agnostic decision frameworks and enhancing accountability among decision-makers, not just in the field of economics. For policymakers, this suggests prioritising governance arrangements that specify decision-quality metrics and accountability mechanisms independently of the nature of the agent, and then ask how humans and AI systems can be jointly configured to meet them. For researchers, it invites empirical and theoretical work on operationalising inseparability, collective morality and agnostic frameworks in concrete institutional settings.

CRediT authorship contribution statement

Mindaugas Kiškis: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Mindaugas Kiskis reports a relationship with Mykolas Romeris University that includes: speaking and lecture fees. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] S.F. Ahmad, H. Han, M.M. Alam, M.K. Rehmat, M. Irshad, M. Arraño-Muñoz, A. Ariza-Montes, Impact of artificial intelligence on human loss in decision making, laziness and safety in education, Humanit. Soc. Sci. Commun. 10 (2023) 311, https://doi.org/10.1057/s41599-023-01787-8.

[2] A. Agrawal, J. Gans, A. Goldfarb, Prediction Machines: The Simple Economics of Artificial Intelligence, Harvard Business Review Press, 2018.

[3] A. Angerschmid, J. Zhou, K. Theuermann, F. Chen, A. Holzinger, Fairness and explanation in AI-informed decision making, Mach. Learn Knowl. Extr. 4 (2) (2022) 556–579, https://doi.org/10.3390/make4020026.

[4] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, in: K. Roemer (Ed.), Ethics of Data and Analytics, Auerbach, 2022, pp. 254–264.

[5] H. Arendt, The Human Condition, University of Chicago Press, 1958.

[6] M. Bagaric, D. Hunter, N. Stobbs, Erasing the bias against using artificial intelligence to predict future criminality: algorithms are color blind and never tire, Univ. Cincinnati. Law Rev. 88 (4) (2019) 1037–1058.

[7] Y. Bathaee, The artificial intelligence black box and the failure of intent and causation, Harv. J. Law Technol. 31 (2017) 889–938.

[8] A. Bianchi, A. Saab, Fear and international law-making: an exploratory inquiry, Leiden J. Int. Law 32 (3) (2019) 351–365, https://doi.org/10.1017/S0922156519000177.

[9] A. Bonezzi, M. Ostinelli, J. Melzner, The human black-box: the illusion of understanding human better than algorithmic decision-making, J. Exp. Psychol.: Gen. 151 (10) (2022) 2125–2140, https://doi.org/10.1037/xge0001181.

[10] J. Burrell, M. Fourcade, The society of algorithms, Annu Rev. Sociol. 47 (1) (2021) 213–237, https://doi.org/10.1146/annurev-soc-090820-020800.

[11] G. Chalhoub, We Must Set Rules for AI Use in Scientific Writing and Peer Review, September 29, Times Higher Education, 2025, https://www.timeshighereducation.com/opinion/we-must-set-rules-ai-use-scientific-writing-and-peer-review.

[12] S.S. Chanda, D.N. Banerjee, Omission and commission errors underlying AI failures, AI. Soc. 39 (3) (2024) 937–960, https://doi.org/10.1007/s00146-022-01585-x.

[13] H. Choung, J.S. Seberger, P. David, When AI is perceived to be fairer than a human: understanding perceptions of algorithmic decisions in a job application context, Int. J. Hum. 13Computer Interact. 40 (10) (2024) 1047–1067, https://doi.org/10.1080/10447318.2023.2260202.

[14] C. Coglianese, L.M. Ben Dor, AI in adjudication and administration, Brooklyn. Law Rev. 86 (3) (2021) 791–830.

[15] S. Corbett-Davies, S. Goel, S. González-Bailón, Even Imperfect Algorithms Can Improve the Criminal Justice System, December 20, The New York Times, 2017.

[16] S. Courtenage, Intelligent machines, collectives, and moral responsibility, AI. Ethics 4 (2024) 485–498, https://doi.org/10.1007/s43681-023-00285-6.

[17] K. Crawford, The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence, Yale University Press, 2021.

[18] J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, A. Saranti, A. Holzinger, On generating trustworthy counterfactual explanations, Inf. Sci. 655 (2024) 119898, https://doi.org/10.1016/j.ins.2023.119898.

[19] K. Dowding, B.R. Taylor, Algorithmic decision-making, agency costs, and institution-based trust, Philos. Technol. 37 (2) (2024) 68, https://doi.org/10.1007/s13347-024-00757-5.

[20] H.L. Dreyfus, What Computers Still Can't Do: A Critique of Artificial Reason, MIT Press, 1992.

[21] EU AI Act, Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on harmonised rules on artificial intelligence (Artificial Intelligence Act), Off. J. Eur. Union L 206 (2024) 1–94.

[22] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, E. Vayena, An ethical framework for a good AI society: opportunities, risks, principles, and recommendations, in: L. Floridi (Ed.), Ethics, Governance, and Policies In Artificial Intelligence, Springer, 2021, pp. 19–39.

[23] B.J. Fowers, Virtue and Psychology: Pursuing Excellence in Ordinary Practices, American Psychological Association, 2005.

[24] S.A. Friedler, C. Scheidegger, S. Venkatasubramanian, The (im)possibility of fairness: different value systems require different mechanisms for fair decision-

making, Commun. ACM 64 (4) (2021) 136–143, https://doi.org/10.1145/3433949.
[25] T. Gill, P. Prowse, Artificial intelligence and human decision-making: exploring the boundaries of AI decision systems, J. Bus. Res. 123 (2021) 158–169, https://doi.org/10.1016/j.jbusres.2020.09.050.
[26] B. Green, Escaping the impossibility of fairness: from formal to substantive algorithmic fairness, Philos. Technol. 35 (4) (2022) 1–32, https://doi.org/10.1007/s13347-022-00584-6.
[27] T. Guggenberger, L. Lämmermann, N. Urbach, A.M. Walter, P. Hofmann, Task delegation from AI to humans: a principal–agent perspective, in: Proceedings of the 44th International Conference on Information Systems (ICIS 2023), 2023.
[28] T. Hale, Long Problems: Climate Change and the Challenge of Governing Across Time, Princeton University Press, 2024.
[29] S. Han, E. Kelly, S. Nikou, E.O. Svee, Aligning artificial intelligence with human values: reflections from a phenomenological perspective, AI. Soc. 37 (4) (2022) 1383–1395, https://doi.org/10.1007/s00146-021-01247-4.
[30] F.A. Hayek, The use of knowledge in society, Am. Econ. Rev. 35 (4) (1945) 519–530.
[31] J. Hawkins, S. Blakeslee, On Intelligence, Henry Holt, 2007.
[32] S. Henry, Social construction of crime, in: J.M. Miller (Ed.), 21st Century Criminology: A Reference Handbook, Sage, 2009, pp. 296–304.
[33] K. Imai, Does AI Help Humans Make Better Decisions?, June 14 Harvard Gazette, 2024.
[34] P.J.R. Janssen, Master's thesis [, Eindhoven University of Technology, 2021.
[35] M. Kasy, R. Abebe, Fairness, equality, and power in algorithmic decision-making, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Association for Computing Machinery, 2021, pp. 576–586, https://doi.org/10.1145/3442188.3445919.
[36] G. Klein, Sources of Power: How People Make Decisions, MIT Press, 1999.
[37] R. Kurzweil, How to Create a Mind: The Secret of Human Thought Revealed, Duckworth, 2013.
[38] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444, https://doi.org/10.1038/nature14539.
[39] P. Limone, G.A. Toto, Origin and development of moral sense: a systematic review, Front. Psychol. 13 (2022) 887537, https://doi.org/10.3389/fpsyg.2022.887537.
[40] O. Lobel, The Equality Machine: Harnessing Digital Technology for a Brighter, More Inclusive Future, PublicAffairs, 2022.
[41] S.M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, M. Suleyman, International evaluation of an AI system for breast cancer screening, Nature 577 (7788) (2020) 89–94, https://doi.org/10.1038/s41586-019-1799-6.
[42] Matsumi, H., & Solove, D.J. (2023). The prediction society: algorithms and the problems of forecasting the future. George Washington university law school public law research paper No. 2023-58.
[43] K. Millet, F. Buehler, G. Du, M.D. Kokkoris, Defending humankind: anthropocentric bias in the appreciation of AI art, Comput. Hum. Behav. 143 (2023) 107707, https://doi.org/10.1016/j.chb.2023.107707.
[44] D. Minh, H.X. Wang, Y.F. Li, T.N. Nguyen, Explainable artificial intelligence: a comprehensive review, Artif. Intell. Rev. 55 (5) (2022) 3503–3568, https://doi.org/10.1007/s10462-021-10088-y.
[45] E. Mollick, Co-intelligence: Living and Working With AI, Portfolio, 2024.
[46] R. Monk, Robert Oppenheimer: A Life Inside The Center, Doubleday, 2012.
[47] J. Morley, L. Floridi, L. Kinsey, A. Elhalal, From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices, Sci. Eng. Ethics 26 (4) (2020) 2141–2168, https://doi.org/10.1007/s11948-019-00165-5.
[48] Naudé, W. (2023). Artificial intelligence and the economics of decision-making. IZA Discussion Paper No. 16000.
[49] C. O'Neil, Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, 2016.
[50] A. Parshall, How My AI Image Won A Major Photography Competition, April 14, Scientific American, 2023.
[51] I. Rahwan, M. Cebrian, N. Obradovich, J. Bongard, J.-F. Bonnefon, C. Breazeal, J. W. Crandall, N.A. Christakis, I.D. Couzin, M.O. Jackson, N.R. Jennings, E. Kamar, I. M. Kloumann, H. Larochelle, D. Lazer, R. McElreath, A. Mislove, D.C. Parkes, M. Wellman, Machine behaviour, Nature 568 (7753) (2019) 477–486, https://doi.org/10.1038/s41586-019-1138-y.
[52] Riehl, K., Makridis, M., & Kouvelas, A. (2024). Quantitative fairness – a framework for the design of equitable cybernetic societies. arXiv preprint arXiv:2411.13184. https://doi.org/10.3929/ethz-b-000720734.
[53] K. Roose, An AI-Generated Picture Won An Art Prize. Artists Aren't Happy, September 2, The New York Times, 2022, https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html.
[54] M. Sato, Gannett is Shuttering Site Accused Of Publishing AI Product Reviews, August 26, The Verge, 2024, https://www.theverge.com/2024/8/26/24228880/gannett-usa-today-reviewed-product-reviews-ai-articles.
[55] A.G. Scaria, V. Subramanian, N.K. George, N. Sengupta, Algorithms and recidivism: a multi-disciplinary systematic review, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES 2024), 2024, pp. 1292–1305, https://doi.org/10.1609/aies.v7i1.31724.
[56] T. Schramme, Introduction, in: T. Schramme (Ed.), Being Amoral: Psychopathy and Moral Incapacity, MIT Press, 2015, pp. 1–22.
[57] J. Schoeffer, M. De-Arteaga, N. Kühl, Explanations, fairness, and appropriate reliance in human–AI decision-making, in: Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), 2024, https://doi.org/10.1145/3613904.3642621.
[58] L. Smirnova, B.S. Caffo, D.H. Gracias, C.-T. Huang, I.E. Morales Pantoja, B. Tang, D. J. Zack, C.A. Berlinicke, J.D. Boyd, M.D. Harris, J.W. Johnson, B.J. Kagan, I. Kahn, A.R. Muotri, B.L. Paulhamus, J.C. Schwamborn, J.B. Plotkin, G. Szalay, J. T. Vogelstein, T. Hartung, Organoid intelligence (OI): The new frontier in biocomputing and intelligence-in-a-dish, Front. Sci. 1 (2023) 1017235, https://doi.org/10.3389/fsci.2023.1017235.
[59] D.J. Solove, H. Matsumi, AI, algorithms, and awful humans, Fordham. Law Rev. 92 (2024) 1923–1979.
[60] J.L. Stephensen, Artificial creativity: beyond the human, or beyond definition? Transform.: J. Media Cult. Technol. 36 (2022) 19–37.
[61] C.R. Sunstein, Governing by algorithm? No noise and (potentially) less bias, Duke Law J. 71 (5) (2022) 1175–1206.
[62] N.N. Taleb, Skin in the Game: Hidden Asymmetries in Daily Life, Random House, 2019.
[63] M. Tegmark, Life 3.0: Being Human in the Age of Artificial Intelligence, Knopf, 2017.
[64] Tohmé, F. (2023). Dynamic arrangements in economic theory: level-agnostic representations. arXiv preprint arXiv:2309.06383.
[65] W. Wallach, C. Allen, Moral Machines: Teaching Robots Right From Wrong, Oxford University Press, 2009.
[66] M.E. Warren, Accountability and democracy, in: M. Bovens, R. Goodin, T. Schillemans (Eds.), The Oxford handbook of public accountability, Oxford University Press, 2014, pp. 39–60.
[67] S. Westby, C. Riedl, Collective intelligence in human–AI teams: a Bayesian theory of mind approach, in: Proceedings of the AAAI Conference on Artificial Intelligence 37, 2023, pp. 6119–6127, https://doi.org/10.1609/aaai.v37i5.25755.
[68] K.C. Yam, T. Tan, J.C. Jackson, A. Shariff, K. Gray, Cultural differences in people's reactions and applications of robots, algorithms, and artificial intelligence, Manage Organ. Rev. 19 (5) (2023) 859–875, https://doi.org/10.1017/mor.2023.21.
[69] E. Yudkowsky, Pausing AI Developments Isn't Enough. We Need to Shut it All Down, March 29, Time, 2023.
[70] J. Zerilli, A. Knott, J. Maclaurin, C. Gavaghan, Algorithmic decision-making and the control problem, Minds Mach. 29 (4) (2019) 555–578, https://doi.org/10.1007/s11023-019-09513-7.