AI & SOCIETY (2025) 40:3163–3185
https://doi.org/10.1007/s00146-024-02128-2

MAIN PAPER

# Competing narratives in AI ethics: a defense of sociotechnical pragmatism

David S. Watson[1] · Jakob Mökander[2,3,4] · Luciano Floridi[4,5]

Received: 8 July 2024 / Accepted: 4 November 2024 / Published online: 27 December 2024


## Abstract

Several competing narratives drive the contemporary AI ethics discourse. At the two extremes are *sociotechnical dogmatism*, which holds that society is full of inefficiencies and imperfections that can only be solved by better technology; and *sociotechnical skepticism*, which highlights the unacceptable risks AI systems pose. While both narratives have their merits, they are ultimately reductive and limiting. As a constructive synthesis, we introduce and defend *sociotechnical pragmatism*—a narrative that emphasizes the central role of context and human agency in designing and evaluating emerging technologies. In doing so, we offer two novel contributions. First, we demonstrate how ethical and epistemological considerations are intertwined in the AI ethics discourse by tracing the dialectical interplay between dogmatic and skeptical narratives across disciplines. Second, we show through examples how sociotechnical pragmatism does more to promote fair and transparent AI than dogmatic or skeptical alternatives. By spelling out the assumptions that underpin sociotechnical pragmatism, we articulate a robust stance for policymakers and scholars who seek to enable societies to reap the benefits of AI while managing the associated risks through feasible, effective, and proportionate governance.

**Keywords** Artificial intelligence · Ethics · Epistemology · Explainability · Fairness · Governance · Machine learning

## 1 Introduction

Rapid advances in machine learning (ML) research and the accelerated adoption of artificial intelligence (AI) systems across societies have sparked intense debate among policymakers, scholars and the broader public about AI's societal implications. While harboring many diverse perspectives (Gilardi et al. 2024), the discourse is largely dominated by two competing narratives. Proponents of increased automation emphasize gains in efficiency and scientific progress, while critics counter that AI poses intolerable risks to individuals and society, disproportionately harming disadvantaged communities.

In a sense, there is nothing new about this dichotomy, which echoes political debates going back at least as far as the Industrial Revolution (Hobsbawm 1952; Johnson and Acemoglu 2023). However, with the growing power and ubiquity of AI systems, the struggle between these competing narratives—which we call *sociotechnical dogmatism* and *sociotechnical skepticism*, respectively—has assumed a new urgency. In this article, we explore the philosophical assumptions underpinning the two narratives and establish the theoretical and practical advantages of a pragmatic synthesis between them.

Before proceeding, two clarifications are in order. First, the sociotechnical dogmatism and sociotechnical skepticism we describe are competing *narratives,* not distinct groups of people. Narratives are selective depictions of reality that

Davis S. Watson and Jakob Mökander are joint first authors, having contributed equally to this article.

✉ Jakob Mökander
jakob.mokander@keble.ox.ac.uk

David S. Watson
david.s.watson11@gmail.com

Luciano Floridi
luciano.floridi@oii.ox.ac.uk

[1] Department of Informatics, King's College London, London, UK

[2] Oxford Internet Institute, University of Oxford, Oxford, UK

[3] Center for Information Technology Policy, Princeton University, Princeton, USA

[4] Digital Ethics Centre, Yale University, New Haven, USA

[5] Department of Legal Studies, University of Bologna, Bologna, Italy

humans use to make sense of the world (Bruner 1991). The dogmatic and skeptical narratives surrounding AI run across different research, business, and policy communities, and permeate their knowledge production to varying extents. Importantly, however, few people subscribe wholesale to strong versions of either narrative. Our aim in this article is thus to scrutinize claims and assumptions, not people or organizations.

Second, sociotechnical dogmatism and sociotechnical skepticism are intended as *ideal types* in the Weberian sense.[1] They are umbrella terms that cover a variety of similar, though not synonymous, labels. The former shares traits with techno-optimism (Danaher 2022), -solutionism (Morozov 2013), and -chauvinism (Broussard 2018); the latter may be caricatured as techno-pessimism (Königs 2022) or identified with the field of critical data studies more broadly. We have chosen to use the terms *dogmatism* and *skepticism* to highlight important links between the present-day AI ethics discourse and classical debates in ethics and epistemology, as well as to avoid unwelcome associations with overlapping terminologies.

There is undeniable merit to both narratives. The dogmatic narrative is right in emphasizing that the benefits of AI systems—ML algorithms included—are both economic and social. Consider the case of drug discovery. In the last decade, ML-based models trained on biomolecular data have become essential to drug discovery pipelines (Keshavarzi Arshadi et al. 2020). As the COVID-19 pandemic led to worldwide lockdowns, ML played an important part in accelerating the development of vaccines that saved millions of lives (Sharma et al. 2022). Similarly, concerning road safety, predictive analytics based on past accident data feed into decisions on highway redesigns that reduce the risk of future incidents (Mannering et al. 2020). These examples illustrate a more general point: abstaining from using ML systems altogether may incur significant social opportunity costs (Floridi et al. 2018).

At the same time, the skeptical narrative is correct in drawing attention to cases where ML systems cause harm. There is substantial evidence—from domains as diverse as predictive policing (Browning and Arrigo 2021), recruitment (Köchling and Wehner 2020), and credit scoring (Mendes and Mattiuzzo 2022)—that ML systems threaten to codify and amplify injustices already present in society. In one dramatic example, Obermeyer et al. (2019) found racial bias in a health screening algorithm that affects millions of Americans. Subsequent simulations suggest that rectifying the disparity would nearly triple the number of Black patients receiving medical attention for chronic illnesses. In short, there is no question that ML systems can replicate existing social inequalities (McGregor 2021) and often fail to perform as advertised (Narayanan and Kapoor 2024). To deny such evidence would be intellectually dishonest and morally inept.

As the above analysis suggests, the problem with the dogmatic and skeptical narratives is not that they are *wrong* but that they are *incomplete*. The dogmatic narrative is so focused on the ends of technological progress that it can be blind to the unjust means that attend such advances. Worse, it is often deployed by people with a vested interest in opposing democratic efforts to regulate the tech industry.[2] The skeptical narrative, meanwhile, has a tendency of attributing evils to technology that exist not only in algorithmically mediated systems but in most human organizations of sufficient size and complexity. Such puritanism is self-defeating since it leaves no room for evaluating the relative merits and constraints of specific policy options or for improving on an imperfect status quo with incomplete information. Strong versions of sociotechnical dogmatism and skepticism both erode the space for human autonomy and deliberation regarding political ends, and are ill-equipped to conceptualize the opportunities and challenges related to fairness, accountability, and transparency in ML.

A more nuanced approach is possible. Such an approach would have to be principled but flexible, acknowledging the irreducible context-dependence of value judgments. It would have to be relational, not relativist, preserving the autonomy of stakeholders to determine their tolerance for error and prioritize between different normative ends. It would have to experiment with both technology design and structural reform, combine quantitative and qualitative modes of reasoning, and incorporate mechanisms for feedback, learning and redress into empirically grounded efforts to develop policies that target specific social challenges. We submit that *sociotechnical pragmatism* meets all these desiderata.

In this article, we conduct a critical review (Grant and Booth 2009) of the AI ethics literature, centering the narratives driving the contemporary discourse.[3] In doing so, we offer two main contributions. First, we show how ethical and epistemological considerations are intertwined in the AI ethics discourse by tracing the dialectical interplay between dogmatism and skepticism across disciplines. By

[1] According to Weber (1904), an ideal type is formed by the one-sided accentuation of particular points of view, according to which individual phenomena are arranged into a unified analytical construct.

[2] Not all attempts to regulate or control technology are democratic. State control of digital platforms has led to increased surveillance and repression in authoritarian regimes (Feldstein 2021).

[3] A critical review is a narrative synthesis of a body of literature. It involves (i) a non-comprehensive search to identify dominant themes and (ii) an interpretative process that combines the reviewer's theoretical premise with existing theories in ways that allow for synthesis and interpretation of diverse studies (Sukhera 2022).

mapping how different assumptions are related, logically and genealogically, we help policymakers and scholars make sense of the growing cacophony of voices weighing in on this debate. Second, we demonstrate through examples how pragmatism does more than dogmatic or skeptical alternatives to (i) promote the design and use of AI systems that are legally compliant, ethically sound, and technically robust, and (ii) channel the power of technological innovation to serve socially beneficial ends.

A final remark. For this article, we have a broad audience in mind—including AI researchers and developers, the informed public, and policymakers. Recent developments in the European Union[4] and the United States[5] show that governments around the world are exploring different approaches to AI regulation. It is of critical importance to get the framing of those regulations right. The self-regulation called for by AI developers has proven ineffective so far (Floridi 2021). At the same time, overly restrictive regulations may lower economic growth and reduce living standards. Further, some ideas for how to reform policy and practice are infeasible to implement or based on flawed assumptions about the role that technology plays in driving societal change. These will at best miss the mark and at worst lead to significant social costs. For all these reasons, we hope that the sociotechnical pragmatism we espouse will inform the larger discourse on how to reap the benefits of AI while mitigating the associated risks.

The article proceeds as follows. In Sect. 2, we survey a wide range of scholarship to trace the interplay between sociotechnical dogmatism and sociotechnical skepticism from the Industrial Revolution to the present day. In Sect. 3, we introduce sociotechnical pragmatism as a theoretical stance that provides a constructive basis for designing and regulating AI systems. In Sect. 4, we consider two case studies—fairness and explainability in ML—and highlight pragmatism's real-world implications in these domains. In Sect. 5, we conclude by stressing the essential role of agency and context in understanding and driving social change.

[4] The European AI Act was finally approved on 13 June 2024 as Reg. (EU) 2024/1689; 2. The original draft of the AI Act was published by the European Commission in April 2021.

[5] In Oct. 2023, President Biden issued an Executive Order on Safe, Secure, and Trustworthy AI. In Feb. 2022, the Algorithmic Accountability Act of 2022 was introduced to the US Senate.

## 2 Framing the debate

In this section, we illustrate what we mean by sociotechnical dogmatism and sociotechnical skepticism, focusing on the assumptions underpinning these narratives. But first, some remarks about scope and methodology are in place.

To narrow down the scope, we focus our review on the AI *ethics* discourse. To begin, questions that are primarily *legal* or *technical* in nature fall outside the scope of this study.[6] That said, we still refer to contributions made by legal scholars or technical AI experts where these have direct implications on debates surrounding fairness, accountability and transparency in ML.

Further, the AI ethics discourse harbors many diverse voices and perspectives. While some researchers focus on near-term issues related to privacy or bias (O'Neil and Gunn 2020), others examine long-term issues related to artificial general intelligence (AGI) (Bengio 2024). Researchers also ground their analysis in different ethical frameworks, be they consequentialist (Vamplew et al. 2018), rights-based (Yeung et al. 2020), or virtue-ethical (Hagendorff 2022). Not all of these views can be neatly mapped onto a one-dimensional spectrum.[7] For example, researchers focusing on the large-scale societal risks that AGI may pose in the future have argued for a slow-down of technological innovation—not because AI systems fail to work as intended but because they are bound to become too powerful.[8] This position shares some assumptions with dogmatism and some conclusions with skepticism. Still, though no doubt a simplification, our purported dichotomy is analytically useful for two reasons. First, the struggle between sociotechnical dogmatism and skepticism is a persistent and instructive feature in the history of technological development (Frey 2019). Second, the dialectic is mirrored in the age-old tension between realism and constructivism in epistemology (more on this in Sect. 4).

Finally, our critical literature review is not limited to academic articles but includes references to policy drafts, journalistic works, and statements made by politicians and business leaders. The reasons for this are simple: the AI ethics

[6] Readers interested in the legal and safety challenges AI systems pose are referred to the *Research Handbook on the Law of Artificial Intelligence* (Barfield and Pagallo 2024) and *Open Problems in Technical AI Governance* (Reuel et al. 2024) for excellent overviews of the legal and technical literature, respectively.

[7] Focusing on the discourse surrounding generative AI, Gilardi et al. (2024) identify four main types of narratives: (1) the existential risk narrative; (2) the effective accelerationist narrative, (3) The real, immediate societal risks narrative, and (4) the balancing risks narrative – while acknowledging significant diversity of views within each.

[8] One example is the open letter that called on companies to *Pause Giant AI Experiments* (Future of Life Institute 2023), which has been signed by over 33,000 people, including leading AI researchers.

discourse supersedes academic silos, and many communities influence policymaking and shape the design of technologies.[9] Moreover, the strengths of competing narratives are unevenly distributed across different carrier strata. While the skeptical narrative is promoted mainly by journalists, social advocacy groups, and academic researchers (Wilson 2017), the dogmatic narrative is often endorsed by investors and politicians encouraging a laissez-faire approach to regulation (Johnston 2020). Because the AI ethics discourse is intrinsically part of larger conversations around innovation, technology, and social change (Hilbert 2020; Heilinger 2022), we make a conscious effort throughout to situate it in its proper historical context.

### 2.1 Sociotechnical dogmatism

Sociotechnical dogmatism, as defined in this article, is an umbrella term for narratives that one-sidedly emphasize the power of technological advancement to fuel economic growth and social progress. These are powerful narratives, with deep roots in enlightenment ideals about objective knowledge and continuous material and moral progress. Sociotechnical dogmatism does not deny that individual technologies may have limitations and adverse effects. But in the long run, it holds, efforts to restrict or slow down technological advancements are misguided—and potentially even harmful. Silicon Valley investor Marc Andreessen recently summarized the essence of the dogmatic narrative in his *Techno-Optimist Manifesto*:

> Technology—new knowledge, new tools, what the Greeks called *techne*—has always been the main source of growth, and perhaps the only cause of growth, as technology made both population growth and natural resource utilization possible […] there is no material problem—whether created by nature or by technology—that cannot be solved with more technology. (Andreessen 2023)

Andreessen is blunt in spelling out the radical materialism that underpins dogmatism. Yet there is little new here: the essence of his statement was already recognized as a dominant post-enlightenment narrative in critiques formulated by scholars like Weber (1910), Horkheimer and Adorno (1944), and Lyotard (1984). In what follows, we identify the ethical and epistemological assumptions that underpin the dogmatic narrative.

At the core of sociotechnical dogmatism is an emphasis on the unique affordances of AI and big data—and a tendency to draw far-reaching conclusions on this basis. Consider the 2008 *Wired* cover story, in which then-Editor-in-Chief Chris Anderson notoriously embraced the "end of theory" perspective on big data:

> This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behavior, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves. (Anderson 2008)

Anderson's instrumentalism elevates predictive accuracy over explanatory insight as the goal of inquiry. While the scientific method requires structural reasoning to generate hypotheses and experiments to isolate cause and effect, proponents of this new behaviorism hold that big data has inaugurated an era of automated discovery requiring little or no human input (Hey et al. 2009; Desai et al. 2022). As the focus has shifted from datasets to the algorithms required to make sense of them, the dogmatic narrative has highlighted AI systems' potential to solve complex problems, unlock economic growth, and contribute to human flourishing.

In *The Creativity Code*, Du Sautoy (2019) argues that AI systems can do anything humans can—only better. Through examples from domains as diverse as music and finance, he attempts to show that machines can be not only efficient but also genuinely creative. Thanks to powerful AI systems, the argument goes, it will be possible to solve complex optimization problems like food production, disease prevention, and reductions in carbon emissions. Recent advances in ML research, such as transformers (Vaswani et al. 2017) and foundation models (Bommasani et al. 2021), have been followed by dogmatic claims about AI's seemingly godlike capabilities to perform cognitive tasks from researchers (Future of Life Institute 2023) and private sector actors (Goldman Sachs 2023) alike. Such sentiments, which border on science fiction (Leaver and Srdarov 2023), are best understood against a larger backdrop.

In *Enlightenment Now*, Pinker (2018) argues that, thanks to science and technology, life has improved for most of the earth's inhabitants over the past 300 years along a wide range of value criteria from child mortality to material well-being. Mayer-Schönberger and Ramge (2018) argue along similar lines that AI's capacity to improve efficiency and enable new solutions to complex problems is *Reinventing Capitalism* by creating data-driven markets that will result in more stable and productive societies. To be clear, there is nothing dogmatic about observing that technology has contributed to improved living standards. However,

[9] Ideas, as sociologists like Weber (1922), Wuthnow (1989), and Collins (2000) have noted, need carrier strata that systematize and promote them.

sociotechnical dogmatism does not stop there. It views such progress as the logical culmination of a scientific-rational worldview, projecting its consequences into the future with limited regard for collateral damage. Hence, there is a strong link between dogmatism and what Dafoe (2015) calls *Technological Determinism*. For example, Kurzweil (2005) argues that technology is necessary and sufficient to ensure positive social outcomes. On this view, the future is bright—provided that technological innovation is neither restricted nor slowed down.

Of course, this depiction is greatly simplified. As Danaher (2022) observes, dogmatic claims differ in the degree and temporal orientation of their optimism as well as the role they ascribe to technology. That said, sociotechnical dogmatism rests on the belief that, in the long run, good prevails over evil and technology plays an essential role in ensuring that outcome. What assumptions underpin this belief? Building on the work of Boden (1966), Danaher argues that dogmatism is underpinned by four premises:[10]

1) *The fact premise*, which states that it is possible to assemble relevant facts;
2) *The value premise*, which states that there are value criteria according to which the facts can be evaluated;
3) *The evaluation premise*, which states that the evaluation of the facts in light of the value criteria will be positive on balance; and
4) *The technological premise,* which states that technology plays a crucial role in ensuring that positive evaluation of the facts holds up in light of the values.

Sociotechnical dogmatism is characterized by an uncompromising acceptance of all four premises. However, each is open to legitimate critique, with attacks on the earlier premises signaling more radical forms of sociotechnical skepticism. We will employ this typology in later sections to categorize the most salient objections to dogmatic narratives.

As mentioned, sociotechnical dogmatism is often promoted by Silicon Valley entrepreneurs and other business executives (Tutton 2020). At times, policymakers and researchers also implicitly or explicitly propagate it, and for good reason. The living standards *are* higher today than ever before, and human ingenuity—including technological innovation—has played no small part in this. If properly designed and deployed, AI and other emerging technologies promise to accelerate this trajectory, e.g., by allowing governments to develop a more productive and equitable public sector (Margetts et al. 2024).

While highly influential, the dogmatic narrative is not without its critics. As a powerful counternarrative, sociotechnical skepticism paints a picture of human history that is very different from the one presented thus far.

## 2.2 Sociotechnical skepticism

In the context of AI ethics, sociotechnical skepticism is a collection of related narratives that consistently foreground the capacity of technology to cause harm or exacerbate preexisting socioeconomic injustices. These center on the fundamental inability of technology to address social and political issues, and the demonstrable harms that specific technologies have on individuals, groups, and the environment. Sociotechnical skepticism is associated with calls for more regulation, from greater oversight of the design and use of AI systems to bans on specific use cases. Perhaps the most radical attacks on dogmatism, however, arise from epistemological objections that challenge the very possibility of objective knowledge or moral progress.

In recent years, many scholars have lent voice to different facets of sociotechnical skepticism. In *To Save Everything, Click Here!*, Morozov (2013) critiques the "solutionist" impulse to regard all human affairs as ripe for optimization. On this view, efficiency is a false idol that distracts us from more pressing social goals. In *Weapons of Math Destruction*, O'Neil (2016) extends the analysis to advertising and criminal justice, demonstrating how algorithms implement pernicious feedback loops that disproportionately impact vulnerable communities. Along the same lines, Eubanks (2018) examines the effects of AI on poor Americans in *Automating Inequality*; Noble (2018) provides an intersectional critique of Google search results in *Algorithms of Oppression*; and Benjamin (2019) argues that the racism that has historically been encoded into the US legal system is now being built into technology.

Focusing on different—but equally subversive—effects of AI and other data-driven technologies, Zuboff (2019) argues that tech giants have inaugurated *The Age of Surveillance Capitalism*, in which human experience is systematically processed into behavioral data and used to develop prediction products that undermine autonomy and democracy. Similarly, in her book *Privacy is Power*, Véliz (2021) exposes how individual (data) privacy is being eroded by big tech and governments before outlining how to design and adopt privacy-friendly alternatives to Google, Facebook, and other online platforms. Rich in detail, all of these scholarly contributions demonstrate through example the limitations of present day AI systems and the political economy that shape their design and use.

The broad canvas painted by these works is clear: sociotechnical dogmatism is at best misguided and at worst a disingenuous cover for the self-interested agendas of profit-seeking companies or authoritarian regimes. Yet a careful analysis displays how different skeptical counternarratives

[10] Danaher (2022) uses *techno-optimism* when referring to what we in this article call sociotechnical dogmatism.

focus their critique on the different premises that underpin dogmatism. For example, several recent works have challenged the *evaluation premise*, i.e., that the benefits brought by ML systems outweigh the associated harms.

In *The Atlas of AI,* Kate Crawford (2021) traces the "oppressive logic" that underpins ML systems beyond the digital realm, from the exploitation of data workers in developing countries to the environmentally damaging extraction of rare earth minerals needed to produce machine hardware. Crawford does not deny that ML systems are powerful tools that can solve specific computational problems. Instead, she questions the evaluation premise by highlighting the social and environmental costs of such systems. Similarly, Bender et al. (2021) argue that the narrow utility of large language models must be balanced against a variety of costs—typically born by those who do not benefit from the resulting technology—including financial burdens resulting in high barriers to market entry, as well as substantial harms from misinformation and unlawful discrimination.

Other works have challenged the *technology premise*, i.e., that technology plays an important role in determining whether the future will be better or worse on balance. For instance, Broussard (2018) critiques sociotechnical dogmatism's irrational reliance on technological solutions for human problems. Because human problems are not exclusively material, we should not expect social problems to retreat before a digitally enabled utopia.

Challenges levied against the evaluation and technology premises license comparatively mild versions of skepticism. These critiques share with sociotechnical dogmatism the *fact premise*, i.e., that it is possible to assemble relevant facts about the natural world; and the *value premise*, i.e., that there are value criteria against which those facts can and should be evaluated. What is left for debate is substantial but not insurmountable—specifically, what the facts are, according to which value criteria they should be evaluated, and the role technology plays in that evaluation. However, other critics go further.

Strong versions of sociotechnical skepticism challenge not only the technology and evaluation premises but also the fact and value premises. For instance, adherents of the sociology of scientific knowledge (SSK)—a movement rooted in the post-structuralism of Foucault (1976) and the constructivism of Latour and Woolgar (1986), which arguably reached its apex in the "strong programme" of Barnes et al. (1996)—hold that all so-called facts are irreducibly subjective, and that all social and institutional relationships should be viewed in terms of asymmetrical power relations. This denies the possibility of assembling facts about the world and erodes any basis for identifying value criteria against which any state of affairs can be deemed better or worse.

Strong skepticism—i.e., narratives that challenge not only the *technology* and *evaluation premises* but also the *fact* and *value premises*—clearly undermine dogmatism. But, as we shall see in Sect. 4, they also make it hard to justify policy interventions on normative grounds or articulate any constructive vision for future societies. Of course, this does not mean that skeptical narratives are without merit. On the contrary, they serve important functions in surfacing social problems that need to be addressed (Martínez 2024) and expanding the Overton window of plausible policy options (Johnson et al. 2024). At their best, skeptical narratives push technology providers and governments to reduce AI harms, improve transparency and accountability throughout AI supply chains, and distribute AI's benefits more broadly. To succeed in this, however, they must find resonance in progressive political programs that have constructive as well as critical components, offering concrete, feasible policy options. This is where sociotechnical pragmatism comes in.

## 3 Sociotechnical pragmatism

In this section, we introduce and defend sociotechnical pragmatism as a constructive synthesis of the dogmatic and skeptical narratives. Our argument proceeds in two steps. First, we outline sociotechnical pragmatism as a theoretical stance. Second, we highlight its implications for contemporary debates in AI ethics.

### 3.1 Theoretical stance

With origins in nineteenth century American thought—particularly the works of Charles S. Peirce, William James and John Dewey—pragmatist philosophy has a rich history (Scheffler 1974). Central to all varieties of pragmatism is the primacy of agents and contexts over ideas and abstractions (James 1907). Conceptual advances are only valuable insomuch as they are useful. A theory with no practical implications is little more than a formal exercise.

How can sociotechnical pragmatism be situated in relation to dogmatism and skepticism? To start with, pragmatism is a philosophical tradition that does not separate knowing the world from acting within it (Peirce 1878). This is part of the "maker's knowledge" tradition (Floridi 2018). The emphasis on *agency*—the ability to make representations of the world and intervene in it (Hacking 1983)—is fundamentally incompatible with determinism in general and sociotechnical dogmatism in particular. On this account, the future is not determined by technological innovation alone but depends on the actions we take to shape it.

In contrast to strong versions of sociotechnical skepticism, however, pragmatism offers a robust foundation upon which to build constructive, progressive research and policy programs (Dewey 1948). The pragmatists hold that theories should be judged by their success when applied

to real-world situations (Legg and Hookway 2008). This maxim has immediate consequences. Opposing the use of AI "on principle" is nonsensical in pragmatist terms. AI systems may have undesirable consequences that call for them to be redesigned, and AI systems employed for unethical purposes should rightly be opposed. But such evaluations will necessarily be context-specific and reflect the values and goals of the communities that design or are subject to decisions produced by such systems. This brings us to a larger point.

Pragmatism points us to the kind of freedom that consists of humans taking full responsibility for our claims and actions (Brandom 1979; Rorty 2021). The denial of divinely ordained truths (and the privilege of powerful elites to interpret them) makes pragmatism a radical movement. However, the view that what counts as morally acceptable is a revisable cultural inheritance also makes pragmatists hesitant to accept utopian imperatives—whether dogmatic or skeptical. Instead, pragmatists tend to embrace progressive yet practical solutions. This involves democratic deliberation regarding normative ends (Dryzek 2004), combined with evidence-based policymaking (Sanderson 2009) and institutional innovation (Frega 2019) to empirically establish the most feasible and effective ways of achieving those ends.

Further, sociotechnical pragmatism stresses the reciprocal relationship between social and technical systems.[11] In applied contexts, algorithms inevitably form part of larger *sociotechnical systems* that encapsulate other artifacts, people, and organizations (Lazar and Nelson 2023). Individual components of sociotechnical systems cannot be analyzed in isolation (Leveson 2016). Many AI harms stem not from failures of individual technical components but from the dynamic ways different parts of the system interact and, of course, how they are used, by whom, and for what purpose (Lauer 2021). Technological fixes—while useful and sometimes necessary (Baxter and Sommerville 2011)—are therefore insufficient. Pragmatists hold that, to ensure good governance, technological solutions must be complemented by legal and cultural interventions (Maas 2022).

The pragmatist stance is deeply historical. As a driver of social change, technology is best understood as the complex web of knowledge, institutions, tools, and behaviors that enable us to solve real-world problems (Ede 2019). Both dogmatic narratives hailing technology as an inevitable force that drives progress and skeptical narratives discarding it as inherently oppressive stem from a view of technology as something separate from human society. But this view does not withstand empirical scrutiny. Successful invention requires not only scientific breakthroughs but also social utility and acceptance (Bronowski 1965). Thus, there are constant tensions between the need to utilize technology and adhere to established rules governing its use (van Dijk 2024).

Finally, pragmatism embraces multi-model thinking (Kaushik and Walsh 2019). Abstract frameworks like utilitarianism, Marxism, and human rights are indispensable for analyzing social dynamics. But their power lies in filtering the complexity of the world through particular lenses (e.g., utility, class struggle, inalienable rights) that obscure as much as they explain. Engaging with many different models is key to making sense of social phenomena (Page 2018). Critically, however, no single model provides conclusive guidance for how to evaluate technologies or design policies. As demonstrated in Sect. 2, sociotechnical dogmatism is underpinned by utilitarian thinking. Despite its conceptual merits, utility is hard to measure (Williams 1973), and interpersonal comparison of utility is complicated by assumptions about timescales and identity (Parfit 1987). The dogmatic narrative around AI's societal impact is irresponsible in glossing over these limitations.

In contrast, skeptical narratives tend to build on concepts rooted in the human rights literature, Marxism, or poststructuralism. These are powerful frameworks for evaluating social change–but there are limits. Rights are often inconclusive, as they stand in tension with other rights (Biggar 2020); AI systems may reconfigure power relations in ways that benefit one group over another (Lazar 2022) while still (in Rawlsian spirit) improving the lives of the worst off; and technology can be both empowering and alienating at the same time (Schroeder 2019). Many evils that skeptical narratives ascribe to AI (like social stratification) are deep-rooted social problems that have existed in all societies of sufficient size and complexity (Schumpeter 1942). By failing to diagnose socioeconomic ills with sufficient precision, such totalizing critiques struggle to guide social change and can unintentionally obstruct actionable avenues for improvement (Murdoch 1994). Rooted in multi-model thinking, pragmatism provides a robust stance for assembling evidence from different sources and interrogating it through complementary theoretical lenses.

Of course, adopting a pragmatist stance does not resolve all (or even most) ethical tensions associated with the design and use of technology. As Isaiah Berlin (1997) argues in *The Pursuit of an Ideal*, different values that are desirable in and of themselves can clash and require tradeoffs. It is important to stress that this "incommensurability"—which gives rise to moral dilemmas—is not exclusively (or even primarily) a conflict between social groups with different values. Rather, it is the result of the many conflicting impulses experienced by individual human beings, and thus a conflict between alternative yet incompatible modes of self-realization (Rorty

[11] The structures, values, and behaviors constituting human societies both shape (Flanagan et al. 2008) and are shaped by technology (Schroeder 2018).

2021). For pragmatists, these lessons from the theory of value pluralism do not constitute a conclusion but a starting point. Accepting that emerging technologies can bring both social benefits and potential harms only tells us that it is reasonable to subject these to proportional governance and oversight. It does not tell us what the nature or purpose of that governance ought to be.

This is where critical data studies as a discipline has an important role to play. Sociotechnical pragmatism can be conceived as exploratory problem solving, i.e., a practice in which a community examines the empirical world to improve a situation the community has decided needs change (Prasad 2021). By highlighting AI systems' shortcomings, critically oriented researchers and social advocacy groups help surface and frame such situations (Abebe et al. 2020). The point here is that all pragmatic research has a critical component—but not all critical research has a pragmatic component. Unwavering sociotechnical skepticism runs the risk of prioritizing abstract values over real-world consequences, effectively making the perfect the enemy of the good (Mulgan 2023).

To summarize, pragmatists are willing to experiment with both structural reform and technological solutions to empirically determine what does and does not work, preferably according to public and predetermined criteria. However, as opposed to sociotechnical dogmatism, pragmatism views emancipation as neither inevitable nor final (Gross et al. 2022). While unable to guarantee pro-social outcomes, good governance facilitates well-intentioned actions, deters malicious actors, and provides a foundation for communities to engage in an informed dialogue around what normative goals to prioritize and at what cost.

### 3.2 Practical implications

The theoretical stance outlined above has several direct implications for policymakers, technology providers, and researchers. Here, we highlight the five most important ones.

First, sociotechnical pragmatism holds that the design and use of AI systems can only be beneficial or harmful insofar as they advance or hinder specific normative goals. This requires policymakers and researchers to be clear about the problems they seek to address and the normative ends they want to achieve. In the contemporary ethics discourse, the term "AI" refers not to a specific technology but to a wide range of computational techniques, from logic-based automated decision systems to large language models based on deep neural networks (Narayanan and Kapoor 2024). Each computational technique comes with affordances and constraints, and gives rise to different ethical, technical, and legal risks depending on the use case for which it is employed (Maragno et al. 2023). Hence, totalizing claims about AI (or technology) as either emancipatory (as sociotechnical dogmatism claims) or oppressive (as sociotechnical skepticism implies) lack resolution and undermine human agency.

AI's societal impact is largely a matter of design (Floridi et al. 2021).[12] But pragmatic problem-solving demands specific problem formulations. A good example is provided by Wang et al. (2023). The authors demonstrate that ML systems employed to make predictions about individuals face fundamental limitations that design changes cannot address, and that such systems' performance often fails to satisfy technology providers' own claims. It is precisely because their study focuses on a specific type of AI system and limits the scope of its criticism to empirically evaluating these systems on their own terms that the authors can present verifiable results and draw actionable conclusions. This is the kind of normative clarity, conceptual precision, and methodological rigor sociotechnical pragmatism demands.

Second, sociotechnical pragmatism implies that a normative evaluation of the merits and limitations of AI systems cannot be conducted in isolation but only in relation to the available alternatives. Consider environmental impact as an example. While ML systems require vast amounts of energy to train (Cowls et al. 2021), they can reduce carbon emissions by substituting or enhancing wasteful processes (Tomlinson et al. 2024). Determining the net environmental impact is thus nontrivial and requires careful empirical analysis of specific use cases. Human decision-makers and ML systems also have different strengths and weaknesses from a fairness perspective (Hullman et al. 2022). Human judgment, for instance, can be influenced by prejudices and fatigue (Kahneman 2011). Using ML systems can therefore lead to fairer decisions in some circumstances (Lepri et al. 2017). The studies have repeatedly shown that even well-intentioned people are prone to biases against historically disadvantaged groups (Greenwald and Krieger 2006). However, a comparison between individual human decision-makers and algorithms is somewhat misleading.

Typically, the alternative to an algorithm is bureaucratic systems that arose precisely in response to the subjectivity and inconsistency of human decision-making (Barocas et al. 2023). While bureaucracies incorporate mechanisms to ensure procedural regularity and operational transparency (Strandburg 2019), they too have flaws and often lack mechanisms of redress. As Weber (1922) and Wiener (1950) noted, even well-functioning bureaucracies can appear dehumanizing to individual decision subjects.[13]

[12] In the words of Kranzberg (1986), technology is neither good nor bad; nor is it neutral.

[13] A potent critic of the bureaucratic state was Kafka. His 1925 novel *The Trial* tells the story of Josef K., a man arrested for unspecified crimes, who struggles to mount a defense without any charges. The tale has been used to support the "right to explanation" (Vredenburgh 2021).

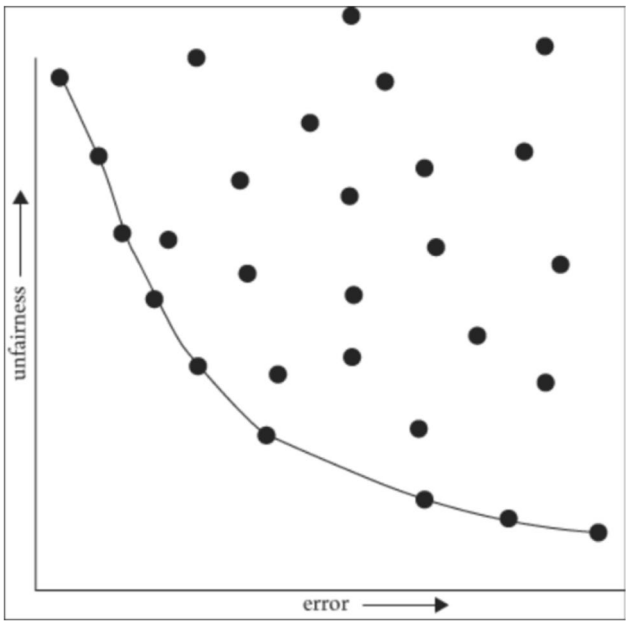


**Fig. 1** A schematic example of the Pareto frontier (Kearns and Roth 2019)

Moreover, bureaucracies too are sociotechnical systems in which human- and machine-centric information processes overlap (Di Maio 2014). That is why bureaucracies, despite procedural protections, are not immune to the imperfections and biases of the people who fill their ranks. It is also why identifying and mitigating such biases with computational methods is not just a speculative ideal but a real opportunity. The proposed US Algorithmic Accountability Act of 2022 accounts for this dynamic by requiring organizations deploying new algorithms to "describe the existing decision-making process" and "explain the intended benefits of augmenting it" (Mökander et al. 2022). The policymakers in other jurisdictions should take note.

Johnson and Zhang (2022) provide a good example of work along these lines. In *What is the Bureaucratic Counterfactual?*, they show that the alternative to algorithmic classification for social policy purposes is often categorical prioritization—a method with its own constraints. Most pertinently, it demands that continuous attributes be simplified into discrete categories that homogenize different levels of need. Surveying the real-world impact of housing vouchers and school financing, Johnson and Zhang do not claim that algorithmic or bureaucratic systems are superior in any universal sense. Instead, they show empirically that categorical prioritization in social policy has opportunity costs, and that there is an understudied potential for predictive algorithms to narrow inequalities.

Third, sociotechnical pragmatism accepts that different normative ends conflict and require tradeoffs. We consider two such cases in Sect. 4, where fairness and explainability are each purported to come at some cost to algorithmic performance. Though details vary, we may describe the general problem setting in terms of a *Pareto frontier*.[14] Imagine a two-dimensional space with axes for, say, accuracy and fairness. Given formal definitions of each, we may score decisions along both axes and locate them within this coordinate system (see Fig. 1). If there is no tradeoff, then we should be able to design arbitrarily fair and accurate models. But if these ideals are in tension, we find an empty space near the origin, indicating that maximally fair models incur some performance penalty and vice versa. We say that one system *Pareto dominates* another if and only if it is strictly better along at least one axis and no worse along any other. The Pareto frontier is constituted by the set of points that cannot be made more accurate without becoming less fair or vice versa. Note that there is no context-independent way to decide which point along the frontier we consider optimal, for this judgment depends upon our valuations of different desiderata for particular problems.

In *The Ethical Algorithm*, theoretical computer scientists Kearns and Roth (2019) argue that delicate tradeoffs like this cannot be navigated without confronting them head-on:

> Once we pick a decision-making model […] there are only two possibilities. Either that model is *not* on the Pareto frontier, in which case it's a "bad" model […] or it *is* on the frontier, in which case it implicitly commits to a numerical weighting of the relative importance of error and unfairness. Thinking about fairness in less quantitative ways does nothing to change these realities—it only obscures them (Kearns and Roth 2019).

Of course, the quantitative approaches to ensuring social justice have their own limitations and should be complemented by qualitative engagement (Narayanan 2022). However, the notion that putting numbers to problems somehow does violence to our underlying humanity or commits us to naïve dogmatism is itself reductive and limiting. To pragmatists, the quantitative methods are merely one tool among many for diagnosing and combatting social injustice—and a powerful one at that. Used correctly, they bring to light the complexity of the issues under discussion. Ignoring that option on philosophical grounds incurs devastating opportunity costs that society can ill afford.

Fourth, sociotechnical pragmatism seeks to foster the use of ML systems for socially beneficial purposes, while taking concrete measures to mitigate technological risks (Lappin 2025; Suleyman 2023). Consider healthcare as an example. Society has an obligation to put patients' safety first. Often,

[14] The concept now referred to as a Pareto frontier (and the associated notion of Pareto efficiency) is attributed to the Italian economist and sociologist Vilfredo Pareto. For details, see Lockwood (2008).

that means using available technologies to develop new drugs or diagnose patients early in the course of a disease. To ensure that new drugs are safe, pharmaceutical companies train ML algorithms to detect treatment response patterns (Nadler et al. 2020). AI systems can also improve the cost-efficiency of the healthcare system, freeing up resources to provide better care (Mainz et al. 2024). The fact that red tape related to such models could restrict the development of potentially lifesaving procedures shows that it is often impossible to "err on the side of caution." The real tension is therefore not between innovation and regulation (a false dichotomy) but between good regulation that promotes both growth and pro-social outcomes and bad regulation that fails in doing so.

This insight permeates the framing of emerging AI regulations. Take European AI Act as an example. While the use of "high-risk" AI systems (like those used in medical diagnostics) is encouraged, the AI Act mandates that such systems undergo "conformity assessments" before they are put on the market and that their outputs are monitored over time (Mökander et al. 2021). The regulatory provisions of the European AI Act also go hand-in-hand with incentives for innovation and investment in enabling digital infrastructure (Novelli et al. 2024). Similarly, The UK Government (2023) has pursued *A pro-innovation approach to AI regulation*, acknowledging AI's potential to improve the quality of public services while relying on sector-specific regulators to oversee the use of AI in their respective domains. Of course, it is hard to strike the right balance. Yet the rhetoric used by policymakers indicates that there is an appetite for a pragmatic stance that combines rigorous and proportional AI governance mechanisms with investments in research and innovation to solve real-world problems.

Fifth, sociotechnical pragmatism asserts that procedural regularity and transparency contribute to good governance (Morley et al. 2021). Consider the functions of datasheets (Gebru et al. 2021) and model cards (Mitchell 2019). By providing information on how ML models are trained and tested, such transparency enhancing tools enable downstream developers to design applications that account for the known limitations of the model in question (Mökander et al. 2023). Awareness of ML models' limitations is also a prerequisite for designing procedural guardrails. For example, human-curated validation of ML systems can aid the design and use of bias-aware and efficient data-driven reasoning in healthcare settings (Boman 2023).

Other mechanisms to improve technical robustness and procedural transparency include algorithmic impact assessments (Selbst et al. 2021; Thomas et al. 2024), disclosure requirements (Kamalnath and Varottil 2022), red teaming (Longpre et al. 2024), and AI audits (Metaxa et al. 2021; Mökander 2023). Different mechanisms fill different functions. For example, internal audits help check that the engineering processes involved in designing ML systems meet specific expectations or standards (Raji et al. 2020). Their goal is to identify and mitigate risks before harm occurs. In contrast, external audits help technology providers verify claims about the systems they design and deploy (Brundage et al. 2020). Independent third-party audits thereby provide a basis for holding technology providers accountable. The fact that these governance mechanisms help address both near-term and long-term AI risks (Christian 2020) means that pragmatists should focus on building broad coalitions for change (Arnold and Toner 2024).[15]

A final point. The practice of transparently communicating normative tradeoffs not only sparks ethical deliberation among developers of ML systems but also informs the public discourse concerning what society we want to live in, and what compromises we are willing to strike in bringing that society to fruition. The purpose of technically informed evaluation tools is not to guarantee ethical outcomes (an impossible goal). Rather, it is to make implicit technology design choices visible, foreground tensions between competing values, give voice to different stakeholders, and arrive at resolutions that—even when imperfect—are at least publicly defensible (Whittlestone et al. 2019).

The significance of this position does not lie in the normative conclusions it justifies, but in the way it reorients our thinking about legitimacy toward political practice (Fossen 2017). Sociotechnical pragmatism puts a premium on justification through informed, rational, and honest discourse. However, decades of research in philosophy, as well as in social and behavioral sciences, have shown that notions like justification and explanation are significantly more complex than intuition suggests (Miller 2017; Danks 2022).

## 4 Contemporary debates in AI ethics

In this section, we use the conceptual tools introduced in the previous sections to analyze the contemporary discourse on AI fairness and explainability. These topics are essential for AI practitioners, policymakers, and end users, as they lie at the heart of debates regarding whether and how to deploy ML systems in high-risk settings like finance or healthcare. Throughout, we leverage real-world examples to show how sociotechnical pragmatism does more to promote fair and transparent AI than either dogmatic or skeptical alternatives.

[15] For details, see Baum (2018), Prunkl and Whittlestone (2020), or Sætra and Danaher (2023).

### 4.1 Fairness

In 2021, Mark Rutte, the prime minister of the Netherlands, offered to resign following controversy surrounding a data-driven welfare fraud detection system referred to as "SyRI." Simplified, SyRI was an algorithm that used statistical patterns to flag potential benefit fraud based on data aggregated from government agencies (Meuwese 2020). While its stated purpose was to make state administration more efficient, SyRI was found to systematically discriminate against minorities, wrongly accusing over 26,000 families of benefit fraud (van Bekkum and Borgesius 2021). Private companies have faced similar controversies. In 2018, Amazon's automated recruitment tool was found to discriminate against female candidates. Trained on résumés from past hires, the system had learned to prefer male candidates and to downgrade applications containing words associated with women (Langenkamp et al. 2020). These and other incidents demonstrate how AI systems trained on unrepresentative or incomplete or data can perpetuate existing societal biases and create new ones.

Algorithmic bias is a serious social problem, as it may result in allocational or representational harms (Mitchell et al. 2021). Promisingly, researchers have made great progress in advancing the theory and practice algorithmic fairness. This includes the assembly of more representative datasets (Ding et al. 2021; Le Quy et al. 2022; Fabris et al. 2022; Kirk et al. 2024), the development of new algorithms and training techniques to reduce the bias of ML models (Friedler et al. 2019; Mandal et al. 2020; Wan et al. 2023), and the proliferation of fairness tools and metrics to test and evaluate ML systems both prior to and after their deployment (Bellamy et al. 2019; Bird et al. 2020; Pagano et al. 2023). These are significant engineering and design achievements, and have led to the emergence of a new industry focused on providing *ethical assurance*. Today, countless startups and consulting companies offer AI auditing services designed to help clients *ensure* that AI systems are "fair" or "ethical" (Shneiderman 2020; Mökander 2023).

However, it is important to remain realistic about the limits of technical fixes in this domain. To begin with, thick ethical concepts are difficult to operationalize, as they have both descriptive and evaluative content (Williams 1985). Moreover, the focus to date has overwhelmingly been on prediction tasks like those found in supervised learning, with much less attention paid to the ethical impact of other ML applications such as unsupervised or reinforcement learning (Jabbari et al. 2017; Polonioli et al. 2023; Watson 2023).

With these challenges in mind, how do we define fairness in the context of AI ethics? A substantial subgenre of the ML literature is devoted to formalizing criteria in an explicit effort to answer this question (Pessach and Shmueli 2022; Caton and Haas 2024). Prominent examples of fairness definitions include:

- *Fairness through unawareness*. A model is fair if sensitive attributes are not included in the training data.
- *Demographic parity*. A model is fair if predictions are independent of sensitive attributes.
- *Equality of opportunity*. A model is fair if predictions are independent of sensitive attributes after conditioning on the true outcome.

This list is hardly exhaustive. A tutorial by Narayanan (2018) surveyed no fewer than 21 competing definitions of algorithmic fairness. Others have emerged since (e.g., Kusner et al. 2017; Kim et al. 2018; Romano et al. 2020). A thorough analysis of these and other fairness criteria is beyond the scope of this article.[16] It suffices to observe that while each captures some intuitive notion of fairness, impossibility theorems have shown that many of the most popular formal criteria are mutually incompatible except in trivial cases (Chouldechova 2017; Friedler et al. 2021; Kleinberg et al. 2017). This suggests that while mathematical formulae can help clarify the tradeoffs inherent in any socially sensitive decision-making context, they cannot in principle "solve" the problems posed by algorithmic fairness.

The impulse to automate our way out of fundamental social problems like systematic discrimination and structural inequality is a clear example of sociotechnical dogmatism in action. The computer scientists and tech companies share some responsibility for promoting the notion that problematic models and datasets can be rectified through technical solutions that make them safe for deployment. Microsoft,[17] Google,[18] Amazon,[19] and IBM[20] all offer fairness toolkits and dashboards that operate on their respective cloud platforms, putting model auditing and bias mitigation tools behind user-friendly graphical interfaces. Though the accompanying whitepapers typically document the limitations of these approaches—"it is not possible to fully 'debias' a system" reads the abstract of Microsoft's Fairlearn paper (Bird et al. 2020)—it is safe to assume that most users do not engage with this technical material, opting instead to outsource the work of AI fairness to established brands.

The dogmatic approach to algorithmic fairness is not only theoretically limited but may also produce real-world harm when uncritically employed in applied contexts (Selbst et al.

[16] For more comprehensive discussions, see Saxena et al. (2019) or Barocas et al. (2023).

[17] https://fairlearn.org/

[18] https://www.tensorflow.org/responsible_ai

[19] https://aws.amazon.com/sagemaker/clarify/

[20] https://aif360.res.ibm.com/

2019). Optimizing an ML systems' output to strictly satisfy any fairness definition implies privileging some individuals and groups and at the expense of others, as different fairness definitions conflict and require tradeoffs (Corbett-Davies et al. 2024). Optimizing an ML systems' performance for fairness can also come at the expense of other values like predictive accuracy. Focusing on the much-discussed examples of predictive policing and criminal recidivism, Corbett-Davies et al. (2017) demonstrate that there is often a real tension between improving public safety and satisfying prevailing notions of algorithmic fairness.

Other failure modes are subtler. The techno-solutionist approach to algorithmic fairness risks ignoring the possibility that in some cases better outcomes may be attained without technology (Selbst et al. 2019); giving AI deployers and users a false sense of security that undermines the need for continuous ethical reflection (Leslie 2019); and eroding humans' sense of their non-instrumental obligations to each other (Mökander and Schroeder 2024). In addition, there are legitimate concerns about skewed incentives in the AI assurance industry. Several studies have highlighted the risk of "fairwashing", whereby AI systems are made to appear more ethical than they are through post-hoc rationalization with respect to some fairness metric(s) (Aïvodji et al. 2019; Burr and Leslie 2023). Others have warned that AI assurance companies with an interest in keeping client relationships friendly risk being too lenient when conducting fairness audits (Costanza-Chock et al. 2022; Munn 2023).

These are well-documented risks (OECD 2024), and both pragmatic and skeptical narratives in AI ethics stress the limitations of mathematical fairness definitions and caution against algorithmically oriented approaches to improving public policy. However, strong versions of sociotechnical skepticism go further. For example, Hoffmann (2019) argues that the very logic of computation is fundamentally unfit to address problems of social injustice, as both are founded on the same rationalist mode of hierarchical labeling and sorting. The problem formulation itself—in terms of variables and averages, optimizing metrics for prespecified groups—fails to question how the categories that algorithmic fairness seeks to protect are themselves contested, socially constructed, and reductive (Hanna et al. 2020; Cantwell Smith 2019). On this view, data science privileges mathematical order over lived experience (McQuillan 2018). Building on the works of Bowker and Star (2000) and Gonen and Goldberg (2019), Birhane (2021) writes:

> The mathematization and formalization of social issues brings with it a veneer of objectivity and positions its operations as value-free, neutral, and amoral. The intrinsically political tasks of categorizing things such as ''acceptable'' behavior, ''ill'' health, and ''normal'' body type then pass as apolitical technical sorting and categorizing tasks. Unjust and harmful outcomes, as a result, are treated as side effects that can be treated with technical solutions such as ''debiasing'' datasets rather than problems that have deep roots in the mathematization of ambiguous and contingent issues, historical inequalities, and asymmetrical power hierarchies. (Birhane 2021)

Goals like "greater transparency," "better algorithms," or "more representative datasets" do little to resolve these issues and may even exacerbate them by normalizing neoliberal modes of agency in which users must navigate a complex marketplace of algorithmic alternatives with limited information (Ananny and Crawford 2016). On this view, the language of power is more appropriate than that of facts and principles when assessing the ethical and social implications of ML (Waelen 2022). Efforts to advance technical solutions to algorithmic bias will only add fuel to the fire.

These objections are provocative and perspicacious. They amount to a full-throated challenge to the *value premise*—perhaps even the *fact premise,* as discussed in Sect. 2. And the challenge has bite: the pervasiveness of metrics and quantification in modern societies does affect individual behavior, social interactions, and institutional practices (Mau 2019). For example, ranking systems often lead to increased self-monitoring and competition. However, abstaining from quantifying human abilities and social relationships is a risky strategy for anyone who hopes to wield policy as an instrument of social change.

When difficult decisions must be made, the current state of the art is to appeal to expert judgment, aided by some procedural guardrails (Gasser and Schönberger 2024). For all its merits, this strategy is vulnerable to human bias, caprice, and outright fraud (Kahneman et al. 2021). That existing datasets are tainted with human error (both individual and social) is a central premise in the skeptical arguments against automation. Historical decisions on credit lending, job hiring, and criminal justice are commonly cited examples. However, for the same reasons, the skeptical impulse to oppose technological advances that could improve this state of affairs may be counterproductive, effectively reverting responsibility back to the very institutions that created the problem in the first place.

The pragmatist synthesis is to observe that whether AI systems can be fruitfully deployed in high-risk domains is an empirical question that must be addressed on a case-by-case basis (Binns et al. 2018). The procedural solution is to map out the tradeoffs and ask difficult questions. Which notion of fairness is most relevant to the task at hand? What will be the social impact of prioritizing one notion of fairness over others in different contexts? Are we willing to sacrifice model performance (e.g., in terms of predictive accuracy)

to obtain fair outcomes, however we choose to operationalize the concept? If so, then how much of a drop in accuracy is tolerable? Perhaps most importantly, how can relevant stakeholders—including users and data subjects—influence these choices?

The answers to these and other questions can and should vary across domains. A one-size-fits-all solution threatens to anoint an "algorithmic leviathan" with the power to arbitrarily exclude individuals from a wide range of socially significant opportunities (Creel and Hellman 2022). It is vital not just to ask hard questions but also to be transparent about the process and the challenge of building fair models. The dogmatic narrative tends to downplay the risks and overestimate the effectiveness of current tools to mitigate algorithmic bias. Such bluster may help short-term sales but harm long-term user trust (Hildago et al. 2021). The skeptical narrative, meanwhile, is reflexively in opposition. It demands superior outcomes that either lie beyond the limits of the Pareto frontier or defy quantification altogether. This is not a recipe for progress. Perhaps both narratives are polarized by the perceived excesses of their counterparts, strategically staking out extreme positions in an effort to rebalance a discourse they feel has become lopsided. If so, then both would be better off starting from a shared commitment to sociotechnical pragmatism, rather than converging on it through meandering dialectical cycles that delay or annul the benefits they claim to seek.

### 4.2 Explainability

ML models are increasingly common in scientific research, where algorithms can help predict the outcomes of complex physical phenomena at scales ranging from the subatomic to the celestial. In a notable recent example, researchers introduced a probabilistic deep learning model for forecasting seasonal Arctic Sea ice concentration (Andersson et al. 2021). Not only did their model vastly improve upon the prior state of the art—it also suggested a previously overlooked feedback loop between Arctic Sea ice and atmospheric circulation that has since been incorporated into many large-scale climate models (Eyring et al. 2024). This relationship was revealed in model post-processing with variable importance techniques, a popular form of explainable AI (XAI) that is widely used in scientific exploration and hypothesis generation (Zednik and Boelsen 2022).

XAI is a vast and growing field of scholarly inquiry aimed at helping humans understand the potentially opaque behavior of complex statistical models such as deep neural networks (Saeed et al. 2023; Räuker et al. 2023). Methodological approaches to computing explanations vary (Linardatos et al. 2021; Rudin et al. 2021) and serve different purposes. The primary aims of XAI are threefold (Watson and Floridi 2021): (i) *to audit*, e.g., ensuring that a credit risk scoring algorithm does not unfairly rely on protected attributes; (ii) *to validate*, e.g., testing whether an image classifier exploits a watermark in the data; and (iii) *to discover*, e.g., generating novel hypotheses for disease mechanisms using explanations from a diagnostic model. In this section, we focus on the epistemological aspects of XAI, as characterized by goals (ii) and (iii). This avoids redundancies, since goal (i) largely overlaps with the fair ML agenda, and highlights connections between the XAI discourse and classic debates from the philosophy of science.

Whether the opacity of ML models is due primarily to technical complexity or corporate policy is a matter of some debate (Burrell 2016; Kroll 2018). In either case, the demand for XAI is undeniable. To highlight the dialectic between dogmatism and skepticism in this subfield of AI ethics, we will focus on post-hoc, model-agnostic XAI tools. For concreteness, consider the SHAP algorithm (Lundberg and Lee 2017), which attempts to explain individual model outputs by computing weights for all input features. This is an example of a local linear approximator, intended to aid interpretation by indicating the magnitude and direction of feature contributions in terms of simple scores that can be visualized and compared across variables. SHAP is *post-hoc* in the sense that it is applied to a target model $f$ after training, as opposed to intrinsic alternatives that aim to make $f$ interpretable in the first place; and *model-agnostic* in the sense that SHAP treats $f$ as a black box and makes no assumptions about its underlying architecture.[21] Such XAI tools are popular with practitioners due to their flexibility and modularity. However, critics object that these traits are precisely what make methods like SHAP so unreliable in practice. After all, a local linear approximation to a highly nonlinear function is guaranteed to introduce errors that could mislead users about the true behavior of the target model $f$.

The dogmatist narrative stakes out a familiar position on XAI, arguing that new and improved methods will eventually resolve whatever perceived issues arise from humans failing to grasp the inner workings of complex ML models. This optimism is on display in many technical works, where authors promote their solution by demonstrating that it meets some favored criteria for interpretability (Lipton 2018). Shapley values are motivated by foundational results from cooperative game theory (Shapley 1953), where they are known to uniquely satisfy a set of reasonable axioms (Sundararajan and Najmi 2020). This mathematical basis may seem reassuring, but several commentators have questioned the relevance of these axioms to XAI (Huang and Marques-Silva 2024; Bilodeau et al. 2024).

[21] Model-specific variants such as DeepSHAP (for neural networks) and TreeSHAP (for tree-based ensembles) exist too (see Lundberg et al. 2020). Yet we restrict our focus to the more generic algorithm.

Tech companies tend to amplify the dogmatic message and, undisturbed by scholarly reservations, offer explainability toolkits intended to make ML algorithms more palatable for users who may lack the required expertise to design, troubleshoot, and modify models prior to deployment (see Microsoft[22] and Google's[23] cloud offerings, for example). These well-designed and actively maintained software libraries tempt stakeholders into believing that we can always understand the behavior of ML models, no matter how complex. Indeed, many XAI projects are explicitly motivated by the goal of promoting user trust (Bhatt et al. 2020), which may be low when the target model is opaque to humans.

The skeptical narrative takes a slightly different guise in XAI, as the risks in this case are primarily epistemic rather than ethical. The worry is that XAI may provide a misleading explanation that gives users a false impression of understanding, perhaps contributing to a faulty decision with harmful consequences. For instance, Lipton (2018) argues that the very notion of "interpretability" is underspecified, with various tools addressing unique problems raised by distinct notions of a fundamentally ambiguous term. No wonder XAI methods often give inconsistent explanations of the same model prediction (Krishna et al. 2022). This issue is echoed by Doshi-Velez and Kim (2017), who conclude that "there is little consensus on what interpretability in ML is and how to evaluate it for benchmarking." Perhaps the strongest critic of post-hoc XAI methods is Rudin (2019), who argues that for high-risk decisions, we should restrict ourselves to globally interpretable models:

> Explanations must be wrong. […] If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation (Rudin 2019).

Rudin is right to point out that post-hoc explanations cannot guarantee 100% fidelity to the target model, for the same reason that maps are usually smaller than the territory they describe.[24] But just as we rely on maps to navigate unfamiliar terrain, so we can use XAI tools to learn about the behavior of a model, at least within some bounded subregion.

Scientists rely on idealizations and abstractions to get a handle on complex phenomena in many domains (Floridi 2008; Potochnik 2017). There are no frictionless planes or infinite populations, but physicists and geneticists freely make use of such assumptions to strip away irrelevant details and focus on mechanisms of interest. Similarly, definitions of key scientific terms are seldom clear-cut. Biologists have yet to settle on a precise definition of the word "gene" (Hopkin 2009; Portin and Wilkins 2017), while competing interpretations of "entropy" persist in thermodynamics (Brissaud 2005; Swendsen 2011). It is worth reminding proponents of strong sociotechnical skepticism—who demand full consensus on contested terms and complete fidelity from post-hoc XAI tools—that such luxuries are often lacking even in mature natural sciences.

This illustrates how the debate between dogmatism and skepticism with respect to XAI reflects the longstanding tension between realism and constructivism in epistemology. Simplified, naïve realists believe in the objectivity and singularity of truth, and view humans as passive receptors of information about the world (Comte 1865). The propositions are true, false, or meaningless (Schlick 1985). In contrast, radical constructivists (whether romantic or postmodern) deny the existence or ontological independence of an external world (Berkeley 1734) and hold that all so-called facts are irreducibly subjective (Barnes et al. 1996). True explanations, they argue, cannot exist and would be unknowable if they did.

According to pragmatism, however, realism and constructivism present a false dichotomy. Because pragmatism does not separate knowing the world from acting within it (Hacking 1983), it replaces the passive and declarative knowledge that something is the case with the interactive and practical knowledge of something being the case (Floridi 2011). It follows that one explanation can be preferable to another given a specific purpose, context, and level of abstraction (Floridi 2008). All models may be wrong, but that does not mean that anything goes. Facts are not subjective but *relational* (Floridi 2011). While individual propositions can be neither proven nor disproven in isolation, they stand and fall with their implications for the broader *systems of knowledge* to which they are connected (Quine and Ullian 1970). Insofar as they can be tested or reproduced, explanations can also be deemed more or less reliable (Mayo 2018).

A pragmatic approach to XAI begins with the agent and the context. Who is seeking an explanation and why? Doctors and patients may prefer different levels of detail, for instance, in their respective inquiries regarding an algorithmic diagnosis (Watson 2022). Perhaps the doctor's explanation is more faithful to the target model and underlying biology than the patient's, but this does not mean the latter explanation is false or useless. On the contrary, if patients lack the requisite expertise in medicine or ML, they may be ill-served by a technical account that describes nonlinear interactions between biomolecular pathways. Insofar as the goal is to secure trust among stakeholders (including patients and healthcare providers), then the accuracy of a

[22] https://interpret.ml/

[23] https://cloud.google.com/explainable-ai

[24] Notwithstanding the best efforts of imperial cartographers in Borges's (1946) classic *On Exactitude in Science*.

**Table 1** Summary of dogmatic, skeptical, and pragmatic positions in ethics and epistemology

| | Ethics | Epistemology |
|---|---|---|
| Dogmatism | Technology fuels economic growth and social progress. Efforts to restrict or slow down technological advancement are misguided and ultimately harmful | Propositions are true, false, or meaningless. What can be known can be stated precisely. A true explanation is complete and objective (absolutism, positivism) |
| Skepticism | Technology exacerbates preexisting inequalities. There are no purely algorithmic solutions for social, economic, and/or political issues, which must be confronted with qualitative engagement rather than quantitative measurement | All so-called facts are irreducibly subjective. Knowledge is socially constructed, not naturally discovered. True explanations (if any such things exist) are fundamentally unknowable (relativism) |
| Pragmatism | Tradeoffs between competing values are inevitable. We must confront these tradeoffs and deliberate between them in a transparent and inclusive manner, striking provisional compromises to advance shared goals | What is true is not as important as what is useful. Theories only gain meaning through their impact on practice. Knowledge is a social project, and explanations are radically context-dependent (pluralism, relationalism) |

system relative to viable alternatives must be a central concern (London 2019).

Once again, we find ourselves at a familiar fork in the road. While the dogmatic impulse is to conclusively "solve" the "problem" of algorithmic explanation, the skeptical impulse is to declare that such solutions are impossible a priori. The pragmatic view is willing to accept that XAI tools can inform and improve decision-making under appropriate conditions. Doctors and patients may benefit from interactive XAI tools (SHAP included) that provide opportunities for follow-up questions and to elaborate on unexpected or unclear aspects of an initial explanation. Users may rationally choose different tradeoffs between accuracy and simplicity, effectively locating themselves at different points on a Pareto frontier between these goals. Of course, the pragmatic view also entails an acknowledgement that XAI tools are only part of the toolkit, alongside socio-structural explanations that offer complementary insights (Smart and Kasirzadeh 2024). Ultimately, it is an empirical matter whether and to what extent any given method succeeds in helping people better understand ML models.

## 5 Conclusion

The societal and ethical implications of AI have sparked much debate. At the extremes of this discourse are two competing narratives: *sociotechnical dogmatism* and *sociotechnical skepticism*. In this article, we discussed the ethical and epistemological assumptions underpinning those two narratives and the pragmatic synthesis we espouse. For brevity and clarity, these narratives are summarized in Table 1.

It is worth re-emphasizing that we use the terms sociotechnical dogmatism, skepticism, and pragmatism as ideal types, to simplify and accentuate their underlying assumptions. Reality is more nuanced.[25] Few researchers or policymakers subscribe wholesale to any pre-packaged narrative and individual articles typically draw on and combine different traditions in the history of ideas. In practice, pragmatic insights permeate both the dogmatic and skeptical narratives—and this is a good thing. Still, contrasting the extreme positions is analytically useful to highlight the fault lines in the contemporary AI ethics discourse.

As we have shown, sociotechnical pragmatism constitutes a constructive and coherent stance that allows researchers and policymakers who seek to identify and mitigate the risks of emerging technologies to navigate between the Scylla of dogmatism and the Charybdis of skepticism. This stance has both epistemological and ethical aspects. In epistemology, pragmatism replaces objectivity with intersubjectivity: for practical purposes, there is no "point of view of the universe" (Singer and de Lazari-Radek 2014) or "view from nowhere" (Nagel 1986). However, pragmatism does not succumb to relativism either. Questions are asked within a given context, for a specific purpose, and at a particular level of abstraction (Floridi 2008). Once these parameters are fixed, the relative merits of competing answers can be distinguished by rational agents.

In ethics, pragmatism is the attitude that what counts as morally acceptable is not an insight produced by something non-human but a revisable cultural inheritance. Since we have no reason to believe that this inheritance should offer theoretical closure, we should not expect normative tensions to be overcome by technological innovation (as dogmatism claims) or solved by more restrictive regulation for specific

[25] The AI ethics discourse is rich and multifaceted. Yet nuance is not always a virtue of theoretical work. At times, it can even obstruct the development of concepts that are intellectually interesting, empirically generative, and practically useful (Healy 2017).

technological systems (as skepticism implies). Instead, policymakers should focus on developing an *ethical infrastructure* (Floridi 2017) that (i) emphasizes the role of agency and context when evaluating the advantages and limitations of specific AI systems and (ii) promotes trust in digital technologies through procedural transparency and regularity.

Of course, there are limits to what can be achieved through good governance and ethically aligned design. Hardin's (1968) thought experiment regarding *The Tragedy of the Commons* illustrates how dynamics on one level of abstraction (e.g., the individual) can lead to undesirable consequences on other levels (e.g., the environment). Moreover, Hayek's (1973) distinction between *cosmos* and *taxis* reminds us that sociotechnical systems have emergent properties, i.e., qualities that cannot be deduced from the system's parts. The dogmatic and skeptical narratives fail to account for these complexities. To address sociotechnical problems, pragmatism demands that we are specific in our problem formulations, explicit about our assumptions, open to combining quantitative and qualitative modes of discovery, and prepared to continuously redesign technologies and policies based on evidence and learning (Berman and Fox 2023).

In Sect. 4, we showed how sociotechnical pragmatism does more to promote fair and explainable AI than dogmatic or skeptical alternatives. Importantly, however, the two case studies discussed are only illustrative. The tension between dogmatic and skeptical narratives also permeates the discourse with respect to other desiderata of ML models, like "accuracy" and "performance." While a review of those domains lies outside the scope of this article, the pragmatist stance we have outlined can be applied to them as well. For example, while ChatGPT may be a "bullshit generator" (Narayanan and Kapoor 2022), it can still be a valuable tool for specific applications, especially when users are aware of and account for its inherent limitations (Floridi 2023). Similarly, ML-based predictions (framed as a process of evaluating a model's ability to approximate an outcome of interest) can be useful for researchers in the social sciences—without committing them to accept any dogmatic claims about deterministically forecasting the future (Verhagen 2022).

AI poses enormous opportunities and challenges across a range of important sectors. As a result, policymakers face legitimate and challenging questions about whether and how to integrate black box ML models into areas like finance, healthcare, and defense. We have argued against the polar extremes of sociotechnical dogmatism, on the one hand, which urges an accelerated rate of technological deployment; and sociotechnical skepticism, on the other, which cautions against automated approaches, especially in domains with significant social, economic, and political consequences. There is nothing simple about our proposed alternative. Sociotechnical pragmatism is inherently messy, deliberative, and reformist. It commits us to specifying priorities and contexts, confronting (possibly unpleasant) tradeoffs, constructing mechanisms to reflexively adapt and improve our own decision-making, and building broad coalitions for change.

As AI grows ever more powerful and prevalent in the years to come (Suleyman 2023), societies will face tough choices about how to regulate emerging technologies. Though there are valuable lessons to be learned from the dogmatic and skeptical narratives, we argue that neither provides an actionable approach to achieving a just and progressive information society. With hard work and a little luck, sociotechnical pragmatism might stand a chance.

**Acknowledgements** The authors wish to thank Arvind Narayanan, Ralph Schroeder, Magnus Boman, Hannah Rose Kirk, Andrew Strait, and David Hagan—as well as two anonymous reviewers—for helpful comments on earlier versions of this manuscript. Their feedback has greatly improved the article. Any opinions expressed or remaining mistakes belong solely to the authors.

We hereby declare that this article is our original work and is not under consideration for publication by any other journal. Further, we have acknowledged all sources used and cited these in the reference section.

**Funding** Not applicable.

**Data availability** Not applicable.

## Declarations

**Conflict of interest** No conflicts of interest to report.



## References

Abebe R, Barocas S, Kleinberg J, Levy K, Raghavan M, Robinson DG (2020) Roles for computing in social change. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3372871

Aïvodji U, Arai H, Fortineau O, Gambs S, Hara S, Tapp A (2019) Fairwashing: the risk of rationalization. In: International conference on machine learning. pp 161–170. PMLR. https://proceedings.mlr.press/v97/aivodji19a.html

Ananny M, Crawford K (2016) Seeing without knowing: limitations of the transparency ideal and its application to algorithmic

accountability. New Media Soc 20(3). https://doi.org/10.1177/1461444816676645
Anderson C (2008) The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. WIRED. https://www.wired.com/2008/06/pb-theory/
Andersson TR, Hosking JS, Pérez-Ortiz M et al (2021) Seasonal Arctic sea ice forecasting with probabilistic deep learning. Nat Commun 12:5124. https://doi.org/10.1038/s41467-021-25257-4
Andreessen M (2023). The Techno-Optimist Manifesto. Andreessen Horowitz. https://a16z.com/the-techno-optimist-manifesto/
Arnold Z, Toner H (2024) AI regulation's champions can seize common ground—or be swept aside. Lawfare. https://www.lawfaremedia.org/article/ai-regulation-s-champions-can-seize-common-ground-or-be-swept-aside
Barfield W, Pagallo U (eds) (2024) Research handbook on the law of artificial intelligence, 2nd edn. Edward Elgar Publishing
Barnes B, Bloor D, Henry J (1996) Scientific knowledge: a sociological analysis. University of Chicago Press
Barocas S, Hardt M, Narayanan A (2023) Fairness and machine learning. Massachusetts Institute of Technology (MIT), Cambridge
Baum SD (2018) Reconciliation between factions focused on near-term and long-term artificial intelligence. AI Soc 33:565–572. https://doi.org/10.1007/s00146-017-0734-3
Baxter G, Sommerville I (2011) Socio-technical systems: from design methods to systems engineering. Interact Comput 23(1):4–17. https://doi.org/10.1016/j.intcom.2010.07.003
Bellamy RK, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Zhang Y (2019) AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. IBM J Res Develop 63(4/5):4–1. https://ieeexplore.ieee.org/document/8843908?denied=
Bellanova R, Irion K, Lindskov Jacobsen K, Ragazzi F, Saugmann R, Suchman L (2021) Toward a critique of algorithmic violence. Int Political Sociol 15(1):121–150. https://doi.org/10.1093/ips/olab003
Bender E, McMillan-Major A, Shmitchell S, Gebru T (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. https://doi.org/10.1145/3442188.3445922
Benjamin R (2019) Race after technology: abolitionist tools for the new jim code, 1st edn. Polity, Cambridge, UK
Bengio Y (2024) International Scientific Report on the Safety of Advanced AI (Doctoral dissertation, Department for Science, Innovation and Technology). https://www.gov.uk/government/publications/international-scientific-report-on-the-safety-of-advanced-ai
Berkeley G (1734) A treatise concerning the principles of human knowledge. Hackett Publishing Company, Inc., Indianapolis: 1982
Berlin I (1988) On the pursuit of the ideal: sir Isaiah Berlin's address at the award ceremony of the senator giovanni agnelli international prize. Turin
Berman G, Fox A (2023) Gradual: The Case for Incremental Change in a Radical Age. Oxford University Press
Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, Ghosh J, Puri R, Moura JMF, Eckersley P (2020) Explainable machine learning in deployment. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3375624
Biggar N (2020) What's wrong with rights? Oxford University Press
Bilodeau B, Jaques N, Pang Wei Koh, Kim B (2024) Impossibility theorems for feature attribution. Proceedings of the National Academy of Sciences of the United States of America, 121(2). https://doi.org/10.1073/pnas.2304406120
Binns R, Max Van Kleek Veale M, Ulrik Lyngs Zhao J, Shadbolt N (2018) It's reducing a human being to a percentage"; perceptions of justice in algorithmic decisions In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI'18). https://doi.org/10.31235/osf.io/9wqxr
Bird S, Dudík M, Edgar R, Horn B, Lutz R, Milan V, Walker K (2020) Fairlearn: a toolkit for assessing and improving fairness in AI. Microsoft Tech Rep. MSR-TR-2020-32. https://www.microsoft.com/en-us/research/uploads/prod/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf
Birhane A (2021) Algorithmic injustice: a relational ethics approach. Patterns 2(2):100205. https://doi.org/10.1016/j.patter.2021.100205
Boden MA (1966) Optimism Philosophy 41(158):291–303. https://doi.org/10.1017/s0031819100058848
Boman M (2023) Human-curated validation of machine learning algorithms for health data. Deleted Journal, 2(3). https://doi.org/10.1007/s44206-023-00076-w
Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Liang P (2021) On the opportunities and risks of foundation models. arXiv:2108.07258
Borges JL (1946) On exactitude in science. In: Collected Fictions. Penguin, New York
Bowker GC, Star SL (2000) Sorting things out: classification and its consequences. MIT press, Boston
Boyd D, Crawford K (2012) Critical questions for big data. Inf Commun Soc 15(5):662–679. https://doi.org/10.1080/1369118X.2012.678878
Brandom R (1979) Freedom and constraint by norms. Am Philos Q 16(3):187–196. https://www.jstor.org/stable/20009758
Briggs J, Devesh K (2023) The Potentially Large Effects of Artificial Intelligence on Economic Growth. https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html
Brissaud J-B (2005) The meanings of entropy. Entropy 7(1):68–96. https://doi.org/10.3390/e7010068
Bronowski J (1965) Science and human values. Harper & Row
Broussard M (2018) Artificial unintelligence: how computers misunderstand the world. MIT Press
Browning M, Arrigo B (2021) Stop and risk: Policing, data, and the digital age of discrimination. Am J Crim Justice 46(2):298–316. https://doi.org/10.1007/s12103-020-09557-x
Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H, Yang J, Toner H, Fong R, Maharaj T, Koh PW, Hooker S, Leung J, Trask A, Bluemke E, Lebensold J, O'Keefe C, Koren M, Ryffel T (2020) Toward trustworthy AI development: Mechanisms for supporting verifiable claims. ArXiv.org. https://doi.org/10.48550/arXiv.2004.07213
Bruner J (1991) The narrative construction of reality. Crit Inq 18(1):1–21. https://www.sas.upenn.edu/~cavitch/pdf-library/Bruner_Narrative.pdf
Burr C, Leslie D (2023) Ethical assurance: a practical approach to the responsible design, development, and deployment of data-driven technologies. AI and Ethics 3(1):73–98. https://doi.org/10.1007/s43681-022-00178-0
Burrell J (2016) How the machine "thinks": Understanding opacity in machine learning algorithms. Big Data Soc 3(1):1–12. https://doi.org/10.1177/2053951715622512
Cantwell Smith B (2019) The promise of artificial intelligence: reckoning and judgment. Massachusetts Institute of Technology (MIT)
Caton S, Haas C (2024) Fairness in machine learning: a survey. ACM Comput Surv 56(7):1–38. https://doi.org/10.1145/3616865
Chouldechova A (2017) Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. Big Data 5(2):153–163. https://doi.org/10.1089/big.2016.0047

Christian B (2020) The alignment problem: machine learning and human values. W. W. Norton & Company
Collins R (2000) The sociology of philosophies: a global theory of intellectual change. Belknap Press of Harvard University Press, Cambridge
Comte A (1865) A general view of positivism. Cambridge University Press, Cambridge: 2009
Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. In: Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining, pp 797–806). https://doi.org/10.1145/3097983.3098095
Corbett-Davies S, Gaebler JD, Nilforoshan H, Shroff R, Goel S (2024) The measure and mismeasure of fairness. J Mach Learn Res 24(1):14730–14846. https://doi.org/10.5555/3648699.3649011
Costanza-Chock S, Raji ID, Buolamwini J (2022) Who audits the auditors? recommendations from a field scan of the algorithmic auditing ecosystem. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency. pp 1571–1583. https://doi.org/10.1145/3531146.3533213
Cowls J, Tsamados A, Taddeo M, Floridi L (2021) The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. AI Soc 38(1)
Crawford K (2021) Atlas of AI: power, politics, and the planetary costs of artificial intelligence. Yale University Press, New Haven, Connecticut
Creel K, Hellman D (2022) The algorithmic leviathan: arbitrariness, fairness, and opportunity in algorithmic decision-making systems. Can J Philos 52(1):26–43. https://doi.org/10.1145/3442188.3445942
Dafoe A (2015) On technological determinism: a typology, scope conditions, and a mechanism. Sci Technol Human Values 40(6):1047–1076
Danaher J (2022) Techno-optimism: an Analysis, an Evaluation and a Modest Defence. Philosophy & Technology, 35(2). https://doi.org/10.1007/s13347-022-00550-2
Danks, D. (2022). Governance via explainability. In: The Oxford Handbook of AI Governance. Oxford University Press, Oxford. https://academic.oup.com/edited-volume/41989
Desai J, Watson D, Wang V, Taddeo M, Floridi L (2022) The epistemological foundations of data science: a critical review. Synthese, 200(6). https://doi.org/10.1007/s11229-022-03933-2
Dewey J (1948) Reconstruction in philosophy. Beacon Press, Boston
Di Maio P (2014) Towards a Metamodel to Support the Joint Optimization of Socio Technical Systems. Systems 2(3):273–296. https://doi.org/10.3390/systems2030273
Ding F, Hardt M, Miller J, Schmidt L (2021) Retiring adult: new datasets for fair machine learning. Adv Neural Inf Process Syst 34:6478–6490. https://doi.org/10.5555/3540261.3540757
Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1702.08608
Dryzek, J. S. (2004). Pragmatism and democracy: in search of deliberative publics. The Journal of Speculative Philosophy, 18(1): 72–79. https://doi.org/10.1353/jsp.2004.0003
Ede A (2019) Technology and society: a world history. Cambridge University Press
Eubanks V (2018) Automating Inequality: how high-tech tools profile, police, and punish the poor. St. Martin's Press, New York
European Commission (2024) The artificial intelligence act. https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf
Eyring V, Collins WD, Gentine P et al (2024) Pushing the frontiers in climate modelling and analysis with machine learning. Nat Clim Chang 14:916–928. https://doi.org/10.1038/s41558-024-02095-y
Fabris A, Messina S, Silvello G et al (2022) Algorithmic fairness datasets: the story so far. Data Min Knowl Disc 36:2074–2152. https://doi.org/10.1007/s10618-022-00854-z
Feldstein S (2021) The rise of digital repression: How technology is reshaping power, politics, and resistance. Oxford University Press
Flanagan M, Howe DC, Nissenbaum H (2008) Embodying values in technology: theory and practice. Information technology and moral philosophy, 322–353. https://doi.org/10.1017/cbo9780511498725.017
Floridi L (2008) The method of levels of abstraction. Mind Mach 18(3):303–329. https://doi.org/10.1007/s11023-008-9113-7
Floridi L (2011) A defence of constructionism: philosophy as conceptual engineering. Metaphilosophy 42(3):282–304. https://doi.org/10.1111/j.1467-9973.2011.01693.x
Floridi L (2017) Infraethics–on the conditions of possibility of morality. Philos Technol 30(4):391–394. https://doi.org/10.1007/s13347-017-0291-1
Floridi L (2018) What a maker's knowledge could be. Synthese 195(1):465–481. https://doi.org/10.1007/s11229-016-1232-8
Floridi L (2021) The end of an era: from self-regulation to hard law for the digital industry. Philosophy & Technology 34(4):619–622. https://doi.org/10.1007/s13347-021-00493-0
Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5
Floridi L, Cowls J, King TC, Taddeo M (2021) How to design AI for social good: Seven essential factors. Ethics Gov Policies Artif Intell pp 125–151. https://doi.org/10.1007/s11948-020-00213-5
Floridi L (2023) AI as agency without intelligence: On chatgpt, large language models, and other generative models. Philosophy & Technology, 36(1). https://doi.org/10.1007/s13347-023-00621-y
Fossen T (2017) Language and legitimacy: is pragmatist political theory fallacious? Eur J Polit Theo 18(2):293–305. https://doi.org/10.1177/1474885117699977
Foucault M (1976) The archaeology of knowledge. Harper, New York
Frega R (2019) Pragmatism and the wide view of democracy. Palgrave Macmillan, Cham, Switzerland
Frey CB (2019) The technology trap: capital, labor, and power in the age of automation. Princeton University Press, New Jersey
Friedler SA, Scheidegger C, Venkatasubramanian S (2021) The (im)possibility of fairness. Commun ACM 64(4):136–143. https://doi.org/10.1145/3433949
Friedler SA, Scheidegger C, Venkatasubramanian S, Choudhary S, Hamilton EP, Roth D (2019) A comparative study of fairness-enhancing interventions in machine learning. In: Proceedings of the conference on fairness, accountability, and transparency, pp 329–338. https://doi.org/10.1145/3287560.3287589
Future of Life Institute (2023) Pause giant AI experiments: an open letter. Future of Life Institute. https://futureoflife.org/open-letter/pause-giant-ai-experiments/
Gasser U, Mayer-Schönberger V (2024) Guardrails: guiding human decisions in the age of AI. Princeton University Press, New Jersey
Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, IIIHD, Crawford K (2021) Datasheets for datasets. Communications of the ACM, 64(12), 86–92. https://doi.org/10.1145/3458723
Gilardi F, Kasirzadeh A, Bernstein A et al (2024) We need to understand the effect of narratives about generative AI. Nat Hum Behav. https://doi.org/10.1038/s41562-024-02026-z
Goldman Sachs (2023) The potentially large. https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

Gonen H, Goldberg Y (2019) Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1903.03862

Grant MJ, Booth A (2009) A typology of reviews: an analysis of 14 review types and associated methodologies. Health Info Libr J 26(2):91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

Greenwald AG, Krieger LH (2006) Implicit bias: scientific foundations. Calif Law Rev 94(4):945–967. https://doi.org/10.2307/20439056

Gross N, Reed I, Winship C (2022) The new pragmatist sociology: inquiry, agency, and democracy. Columbia University Press, New York

Hacking I (1983) Representing and intervening: introductory topics in the philosophy of natural science. Cambridge University Press, Cambridge

Hagendorff T (2022) A virtue-based framework to support putting AI ethics into practice. Philosophy & Technology, 35(3). https://doi.org/10.1007/s13347-022-00553-z

Hanna A, Denton E, Smart A, Smith-Loud J (2020) Towards a critical race methodology in algorithmic fairness. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3351095.3372826

Hardin G (1968) The tragedy of the commons. Science 162(3859):1243–1248. https://doi.org/10.1126/science.162.3859.1243

Hayek F (1973) Law, legislation and liberty: rules and order. Routledge, New York

Healy K (2017) Fuck nuance. Sociol Theory 35(2):118–127. https://doi.org/10.1177/0735275117709046

Hey T, Tansley S, Tolle KM (2009) The fourth paradigm: data-intensive scientific discovery, vol 1. Microsoft research, Redmond, WA. http://microsoft.com/en-us/research/uploads/prod/2009/10/Fourth_Paradigm.pdf

Heilinger, J. C. (2022). The ethics of AI ethics. A constructive critique. Philosophy & Technology, 35(3), 61. https://doi.org/10.1007/s13347-022-00557-9.pdf

Hidalgo CA, Orghian D, Canals JA, De Almeida F, Martin N (2021) How humans judge machines. MIT Press, Chicago

Hilbert M (2020) Digital technology and social change: the digital transformation of society from a historical perspective. Dialogues in Clinical Neuroscience, 22(2), 189–194. https://doi.org/10.31887/dcns.2020.22.2/mhilbert

Hobsbawm EJ (1952) The machine breakers. Past and Present 1(1):57–70. https://doi.org/10.1093/past/1.1.57

Hoffmann AL (2019) Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. Inf Commun Soc 22(7):900–915. https://doi.org/10.1080/1369118x.2019.1573912

Hopkin K (2009) The evolving definition of a gene. Bioscience 59(11):928–931. https://doi.org/10.1525/bio.2009.59.11.3

Horkheimer M, Adorno T (1944) Dialectic of enlightenment: philosophical fragments. Stanford University Press, Stanford: 2002

Huang X, Joao Marques-Silva (2024) On the failings of shapley values for explainability. International Journal of Approximate Reasoning, 109112–109112. https://doi.org/10.1016/j.ijar.2023.109112

Hullman J, Kapoor S, Nanayakkara P, Gelman A, Narayanan A (2022) The worst of both worlds: a comparative analysis of errors in learning from data in psychology and machine learning. In: Proceedings of the 2022 AAAI/ACM conference on AI, Ethics, and Society. pp 335–348. https://doi.org/10.1145/3514094.3534196

Jabbari S, Joseph M, Kearns M, Morgenstern J, Roth A (2017) Fairness in Reinforcement Learning. In: Proceedings of the 34th International Conference on Machine Learning. https://proceedings.mlr.press/v70/jabbari17a.html.

Jackson MC (2019) Critical systems thinking and the management of complexity. John Wiley & Sons

James W (1907) Pragmatism: a new name for some old ways of thinking. Longmans, Green and Co, New York. https://doi.org/10.1037/10851-000

Johnson S, Acemoglu D (2023) Power and progress: our thousand-year struggle over technology and prosperity. Hachette UK

Johnson RA, Zhang S (2022) What is the bureaucratic counterfactual? Categorical versus algorithmic prioritization in U.S. social policy. 2022 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3533223

Johnson EA, Hardill I, Johnson MT, Nettle D (2024) Breaking the Overton Window: on the need for adversarial co-production. Evidence & Policy 20(3): 393–405. https://cnrs.hal.science/hal-04287638/document

Johnston S (2020) Techno-fixers: origins and implications of technological faith. Mcgill-Queen's University Press, Montreal & Kingston

Kahneman D, Sibony O, Sunstein CR (2021) Noise: a flaw in human judgment. Little, Brown Spark, New York

Kahneman D (2011) Thinking, fast and slow. Farrar, Straus and Giroux, New York.

Kamalnath A, Varottil U (2022) A disclosure-based approach to regulating AI in corporate governance. SSRN Electron J. https://doi.org/10.2139/ssrn.4002876

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 100804–100804. https://doi.org/10.1016/j.patter.2023.100804

Kaushik V, Walsh CA (2019) Pragmatism as a research paradigm and its implications for social work research. Soc Sci 8(9):255. https://doi.org/10.3390/socsci8090255

Kearns M, Roth A (2019) The ethical algorithm: the science of socially aware algorithm design. Oxford University Press

Keshavarzi Arshadi A, Webb J, Salem M, Cruz E, Calad-Thomson S, Ghadirian N, Yuan JS (2020) Artificial intelligence for COVID-19 drug discovery and vaccine development. Front Artifi Intell 3:65

Kim MP, Reingold O, Rothblum GN (2018) Fairness Through Computationally-Bounded Awareness. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1803.03239

Kirk HR, Whitefield A, Röttger P, Bean A, Margatina K, Ciro J, Hale SA. (2024) The PRISM alignment project: what participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *arXiv preprint* arXiv:2404.16019.

Kleinberg J, Mullainathan S, Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.1609.05807

Köchling A, Wehner MC (2020) Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Bus Res 13(3):795–848. https://doi.org/10.1007/s40685-020-00134-w

Königs, P. (2022). What is techno-optimism? Philosophy & Technology, 35(3). https://doi.org/10.1007/s13347-022-00555-x

Kranzberg M (1986) Technology and History: "Kranzberg's Laws." Technol Cult 27(3):544–560. https://doi.org/10.2307/3105385

Krishna S, Han T, Gu A, Pombra J (2022) The disagreement problem in explainable machine learning: a practitioner's perspective. ArXiv.org. https://doi.org/10.48550/arXiv.2202.01602

Krishnan M (2019) Against interpretability: a critical examination of the interpretability problem in machine learning. Philosophy & Technology. https://doi.org/10.1007/s13347-019-00372-9

Kroll JA (2018) The fallacy of inscrutability. Phil Trans R Soc A 376(2133):20180084. https://doi.org/10.1098/rsta.2018.0084

Kurzweil R (2005) The singularity is near: when humans transcend biology. Viking, New York

Kusner M, Loftus J, Russell C, Silva R (2017) Counterfactual Fairness ArXiv (Cornell University). https://doi.org/10.48550/arxiv.1703.06856

Langenkamp M, Costa A, Cheung C (2020) Hiring Fairly in the Age of Algorithms. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2004.07132

Lappin S (2025) Understanding artificial intelligence: neither catastrophe nor redemption. Polity Books, Cambridge

Latour B, Woolgar S (1986) Laboratory life: the construction of scientific facts. Princeton University Press, Princeton, N.J.

Lauer D (2021) You cannot have AI ethics without ethics. AI Ethics 1(1):21–25. https://doi.org/10.1007/s43681-020-00013-4

Lazar S, Nelson A (2023) AI safety on whose terms? Science 381(6654):138–138. https://doi.org/10.1126/science.adi8982

Lazar S (2022) Power and AI: Nature and Justification, In: Justin B Bullock, and others (eds) The Oxford Handbook of AI Governance, Oxford Handbooks, https://doi.org/10.1093/oxfordhb/9780197579329.013.12,

Le Quy T, Roy A, Iosifidis V, Zhang W, Ntoutsi E (2022) A survey on datasets for fairness-aware machine learning. Wiley Interdisciplinary Rev 12(3):e1452. https://doi.org/10.1002/widm.1452

Leaver T, Srdarov S (2023) ChatGPT isn't magic: The hype and hypocrisy of generative artificial intelligence (AI) rhetoric. M/c Journal, 26(5). https://doi.org/10.5204/mcj.3004

Legg C, Hookway C (2008) Pragmatism (Stanford Encyclopedia of Philosophy). Stanford.edu. https://plato.stanford.edu/entries/pragmatism

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2017) Fair, transparent, and accountable algorithmic decision-making processes. Philosophy & Technology 31(4):611–627. https://doi.org/10.1007/s13347-017-0279-x

Leslie D (2019) Understanding artificial intelligence ethics and safety: a guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529

Leveson NG (2016) Engineering a safer world: systems thinking applied to safety. The MIT Press, Boston, p 560

Linardatos P, Papastefanopoulos V, Kotsiantis S (2021) Explainable AI: a review of machine learning interpretability methods. Entropy 23(1):18. https://doi.org/10.3390/e23010018

Lipton ZC (2018) The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Lockwood B (2008) Pareto efficiency. Palgrave Macmillan UK EBooks, 1–5. https://doi.org/10.1057/978-1-349-95121-5_1823-2

London AJ (2019) Artificial intelligence and black-box medical decisions: accuracy versus explainability. Hastings Cent Rep 49(1):15–21. https://pubmed.ncbi.nlm.nih.gov/30790315/

Longpre S, Kapoor S, Klyman K (2024) A Safe Harbor for AI Evaluation and Red Teaming. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2403.04893

Lundberg S, Lee S-I (2017) A unified approach to interpreting model predictions. ArXiv.org. https://doi.org/10.48550/arXiv.1705.07874

Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, Lee SI (2020) From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2(1):56–67

Lyotard J-F (1984) The postmodern condition: a report on knowledge. University of Minnesota Press, Minneapolis

Maas MM (2022) Aligning AI regulation to sociotechnical change. In: The oxford handbook of AI governance. Oxford University Press, Oxford

Mainz J, Munch L, Bjerring JC (2024) Cost-effectiveness and algorithmic decision-making. AI Ethics pp 1–13. https://doi.org/10.1007/s43681-024-00528-0

Mandal D, Deng S, Jana S, Wing J, Hsu DJ (2020) Ensuring fairness beyond the training data. Adv Neural Inf Process Syst 33:18445–18456. https://doi.org/10.5555/3495724.3497273

Mannering F, Bhat CR, Shankar V, Abdel-Aty M (2020) Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic Methods in Accident Research 25:100113. https://doi.org/10.1016/j.amar.2020.100113

Maragno G, Tangi L, Gastaldi L, Benedetti M (2023) Exploring the factors, affordances and constraints outlining the implementation of Artificial Intelligence in public sector organizations. Int J Inf Manage 73:102686. https://doi.org/10.1016/j.ijinfomgt.2023.102686

Margetts H, Dorobantu C, Bright J (2024) How to build progressive public services with data science and artificial intelligence. Political Quart. https://doi.org/10.1111/1467-923X.13448

Martínez MA (2024) Activist research as a methodological toolbox to advance public sociology. Sociology 58(4):832–850. https://doi.org/10.1177/00380385231219207

Mau S (2019) The metric society: On the quantification of the social. John Wiley & Sons

Mayer-Schönberger V, Ramge T (2018) Reinventing capitalism in the age of big data. Basic Books, New York

Mayer-Schönberger V, Ramge T (2022) Access rules freeing data from big tech for a better future. California University Of California Press, Oakland

Mayo DG (2018) Statistical inference as severe testing: how to get beyond the statistics wars. Cambridge University Press, New York, Ny

McGregor S (2021) Preventing repeated real world AI failures by cataloging incidents: the AI incident database. Proc AAAI Conf Artificial Intell 35(17):15458–15463. https://doi.org/10.1609/aaai.v35i17.17817

McQuillan D (2018) Data science as machinic neoplatonism. Philos Technol 31:253–272. https://doi.org/10.1007/s13347-017-0273-3

Mendes L S, Mattiuzzo M (2022) Algorithms and discrimination: The case of credit scoring in brazil. Ius Gentium, 407–443. https://doi.org/10.1007/978-3-030-90331-2_17

Metaxa D, Park JS, Robertson RE, Karahalios K, Wilson C, Hancock J, Sandvig C (2021) Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. Foundations and Trends® in Human–Computer Interaction, 14(4), 272–344. https://doi.org/10.1561/1100000083

Meuwese A (2020) Regulating algorithmic decision-making one case at the time: a note on the dutch "syri" judgment. 1(1):209–211. https://pure.uvt.nl/ws/portalfiles/portal/43647493/syri_case_note.pdf

Miller T (2017) Explanation in artificial intelligence: Insights from the social sciences. https://doi.org/10.48550/arxiv.1706.07269

Mitchell M, Wu S, Zaldivar A (2019) Model Cards for Model Reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19. https://doi.org/10.1145/3287560.3287596

Mitchell S et al (2021) Algorithmic fairness: choices, assumptions, and definitions. Annual review of statistics and its application 141–163. https://doi.org/10.1146/annurev-statistics-042720-125902

Mökander J, Floridi L (2022) From algorithmic accountability to digital governance. Nat Mach Intell. https://doi.org/10.1038/s42256-022-00504-5

Mökander J, Juneja P, Watson DS, Floridi L (2022) The US Algorithmic Accountability Act of 2022 vs. The EU Artificial Intelligence Act: what can they learn from each other? Mind Mach 32(4), 751–758. https://doi.org/10.1007/s11023-022-09612-y

Mökander J, Schuett J, Kirk HR et al (2023) Auditing large language models: a three-layered approach. AI Ethics 4:1085–1115. https://doi.org/10.1007/s43681-023-00289-2

Mökander J, Schroeder R (2024) Artificial intelligence, rationalization, and the limits of control in the public sector: the case of tax policy optimization. Soc Sci Comput Rev. https://doi.org/10.1177/08944393241235175

Mökander J, Axente M, Casolari F, Floridi L (2021) Conformity assessments and post-market monitoring: a guide to the role of auditing in the proposed European AI regulation. Mind Mach. https://doi.org/10.1007/s11023-021-09577-4

Mökander, J. (2023) Auditing of AI: Legal, ethical and technical approaches. Digital Society, 2(3). https://doi.org/10.1007/s44206-023-00074-y

Morley J, Elhalal A, Garcia F et al (2021) Ethics as a service: a pragmatic operationalisation of AI ethics. Mind Mach 31:239–256. https://doi.org/10.1007/s11023-021-09563-w

Morozov E (2013) To save everything, click here: the folly of technological solutionism. Publicaffairs, New York

Mulgan G (2023) When science meets power. John Wiley & Sons, New York

Murdoch I (1994) Metaphysics as a guide to morals. Penguin, New York

Nadler E, Arondekar B, Zhou J (2020) Treatment patterns and clinical outcomes in patients with advanced non-small cell lung cancer initiating first-line treatment in the US community oncology setting: a real-world retrospective observational study. J Cancer Res Clin Oncol 147(3):671–690. https://doi.org/10.1007/s00432-020-03414-4

Nagel T (1986) The view from nowhere. Oxford University Press, New York

Narayanan A, Kapoor S (2024) AI snake oil: what artificial intelligence can do, what it can't, and how to tell the difference. Princeton University Press

Narayanan A (2018) Translation tutorial: 21 fairness definitions and their politics. https://facctconference.org/static/tutorials/narayanan-21defs18.pdf

Narayanan A (2022) ChatGPT is a bullshit generator. But it can still be amazingly useful. AI Snake Oil, 6. https://www.aisnakeoil.com/p/chatgpt-is-a-bullshit-generator-but

Noble SU (2018) Algorithms of oppression: how search engines reinforce racism. New York University Press, New York

Novelli C, Casolari F, Rotolo A, Taddeo M, Floridi L (2024) AI risk assessment: a scenario-based, proportional methodology for the AI act. Digital Society 3(1). https://doi.org/10.1007/s44206-024-00095-1

O'Neil C (2016) Weapons of math destruction: how big data increases inequality and threatens democracy. Crown, New York

O'Neil C, Gunn H (2020) Near-term artificial intelligence and the ethical matrix. Ethics Artif Intell. pp 235–269. https://doi.org/10.1093/oso/9780190905033.003.0009

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. Science 366(6464):447–453. https://doi.org/10.1126/science.aax2342

OECD (2024) Assessing potential future artificial intelligence risks, benefits and policy imperatives. In: OECD artificial intelligence papers, no 27. OECD Publishing, Paris. https://doi.org/10.1787/3f4e3dfb-en

Pagano TP, Loureiro RB, Lisboa FV, Peixoto RM, Guimarães GA, Cruz GO, Nascimento EG (2023) Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big Data Cognitive Comput 7(1):15. https://www.mdpi.com/2504-2289/7/1/15

Page SE (2018) The model thinker: what you need to know to make data work for you. Basic Books

Parfit D (1987) Reasons and persons. Oxford University Press

Peirce CS (1878) How to make our ideas clear. Popular Sci Mon. 12:286–302. https://philpapers.org/rec/PEIHTM

Pessach D, Shmueli E (2022) A review on fairness in machine learning. ACM Computing Surveys (CSUR) 55(3):1–44. https://doi.org/10.1145/3494672

Pinker S (2018) Enlightenment now: the case for reason, science, humanism, and progress. Penguin, UK

Polonioli A, Ghioni R, Greco C et al (2023) The ethics of online controlled experiments (A/B Testing). Mind Mach 33:667–693. https://doi.org/10.1007/s11023-023-09644-y

Portin P, Wilkins A (2017) The evolving definition of the term "gene". Genetics 205(4):1353–1364. https://doi.org/10.1534/genetics.116.196956

Potochnik A (2017) Idealization and the aims of science. University of Chicago Press

Prasad M (2021) Pragmatism as problem solving. Socius: Sociological Research for a Dynamic World, 7, 237802312199399. https://doi.org/10.1177/2378023121993991

Prunkl C, Whittlestone J (2020) Beyond near-and long-term: towards a clearer account of research priorities in AI ethics and society. In: Proceedings of the AAAI/ACM conference on AI, Ethics, and Society, pp 138–143. https://doi.org/10.1007/s11023-022-09612-y

Quine WVO, Ullian JS (1970) The web of belief. Random House, New York: 2009

Raji ID, Smart A, White RN, Mitchell M, Gebru T (2020) Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2001.00973

Räuker T, Ho A, Casper S, Hadfield-Menell D (2023) Toward transparent ai: a survey on interpreting the inner structures of deep neural networks. In: 2023 ieee conference on secure and trustworthy machine learning (satml). IEEE, pp 464–483. https://ieeexplore.ieee.org/document/10136140

Reuel A, Bucknall B, Casper S, Fist T, Soder L, Aarne O, Trager R (2024) Open problems in technical ai governance. arXiv preprint. https://arxiv.org/pdf/2407.14981

Romano Y, Barber RF, Sabatti C, Candès E (2020) With malice toward none: assessing uncertainty via equalized coverage. Harvard Data Science Review. https://doi.org/10.1162/99608f92.03f00592

Rorty R (2021) Pragmatism as anti-authoritarianism. The Belknap Press of Harvard University Press

Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215. https://doi.org/10.1038/s42256-019-0048-x

Rudin C, Chen C, Chen Z, Huang H (2021) Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. ArXiv.org. https://doi.org/10.48550/arXiv.2103.11251

Saeed W, Omlin C (2023) Explainable AI (XAI): a systematic meta-survey of current challenges and future opportunities. Knowl-Based Syst 263:110273. https://doi.org/10.1016/j.knosys.2023.110273

Sætra HS, Danaher J (2023) Resolving the battle of short- vs. long-term AI risks. AI Ethics. https://doi.org/10.1007/s43681-023-00336-y

Sanderson I (2009) Intelligent policy making for a complex world: pragmatism, evidence and learning. Political Studies 57(4):699–719. https://doi.org/10.1111/j.1467-9248.2009.00791.x

Sautoy MD (2019) The creativity code: how AI is learning to write, paint and think. 4th Estate, London

Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y (2019) How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In: Proceedings

of the 2019 AAAI/ACM conference on AI, ethics, and society, pp 99–106. https://doi.org/10.1016/j.artint.2020.103238
Scheffler I (1974) Four pragmatists: a critical introduction to peirce, james, mead, and dewey. Routledge & Kegan Paul
Schlick M (1985) General theory of knowledge. Open Court Pub, Lasalle, Il (Originally published: 1918)
Schroeder R (2018) Social theory after the internet : media, technology and globalization (pp. 28–59). UCL Press, London. https://discovery.ucl.ac.uk/id/eprint/10040801/1/Social-Theory-after-the-Internet.pdf
Schroeder R (2019) 'Big Data: Marx, Hayek, and Weber in a Data-Driven World', in Mark Graham, and William H. Dutton (eds), Society and the Internet: How Networks of Information and Communication are Changing Our Lives, 2nd edn, https://doi.org/10.1093/oso/9780198843498.003.0011,
Schumpeter JA (1942) Capitalism, socialism and democracy. Harper, New York
Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J (2019) Fairness and abstraction in sociotechnical systems. In: Proceedings of the conference on fairness, accountability, and transparency, pp 59–68. https://doi.org/10.1145/3287560.3287598
Selbst A, Anthony D, Bambauer J (2021) An institutional view of algorithmic impact assessments. Harvard J Law Technol 35:117–192. https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf
Shapley L (1953) A value for n-person games. In: Contributions to the theory of games, pp 307–317. Princeton University Press, Princeton
Sharma A, Virmani T, Pathak V, Sharma A (2022) Artificial intelligence-based data-driven strategy to accelerate research, development, and clinical trials of COVID vaccine. Biomed Res Int 2022:e7205241. https://doi.org/10.1155/2022/7205241
Shneiderman B (2020) Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Trans Interactive Intell Syst (TiiS) 10(4):1–31. https://doi.org/10.1145/3419764
Singer P, de Lazari-Radek K (2014) The point of view of the universe: sidgwick and contemporary ethics. Oxford University Press
Smart A, Kasirzadeh A (2024) Beyond model interpretability: socio-structural explanations in machine learning. AI & Soc. https://doi.org/10.1007/s00146-024-02056-1
Strandburg K (2019) Rulemaking and inscrutable automated decision tools, Columbia Law Review 119(7). https://www.columbialawreview.org/wp-content/uploads/2019/11/-Strandburg-Rulemaking_and_Inscrutable_Automatic_Decision_Tools.pdf
Sukhera J (2022) Narrative reviews: flexible, rigorous, and practical. J Grad Med Educ 14(4):414–417. https://doi.org/10.4300/JGME-D-22-00480.1.PMID:35991099;PMCID:PMC9380636
Suleyman M (2023) The coming wave: technology, power, and the twenty-first century's greatest dilemma. Crown
Sundararajan M, Najmi A (2020) The many Shapley values for model explanation. In Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 859, 9269–9278. https://dl.acm.org/doi/abs/10.5555/3524938.3525797
Swendsen RH (2011) How physicists disagree on the meaning of entropy. Am J Phys 79(4):342–348. https://doi.org/10.1119/1.3536633
Thomas C, Roberts H et al (2024) The case for a broader approach to AI assurance: addressing "hidden" harms in the development of artificial intelligence. AI Soc. https://doi.org/10.1007/s00146-024-01950-y
Tomlinson B, Black RW, Patterson DJ, Torrance AW (2024) The carbon emissions of writing and illustrating are lower for AI than for humans. Sci Rep 14(1):3732. https://doi.org/10.1038/s41598-024-54271-x
Tutton R (2020) Sociotechnical imaginaries and techno-optimism: examining outer space utopias of silicon valley. Science as Culture 30(3):416–439. https://doi.org/10.1080/09505431.2020.1841151
UK Government. (2023) A pro-innovation approach to AI regulation. [White Paper]. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper
Vamplew P, Dazeley R, Foale C, Firmin S, Mummery J (2018) Human-aligned artificial intelligence is a multiobjective problem. Ethics Inf Technol 20(1):27–40. https://doi.org/10.1007/s10676-017-9440-6
van Dijk J (2024) Power and technology: a theory of social. In: Technical and Natural Power, John Wiley & Sons
van Bekkum M, Borgesius FZ (2021) Digital welfare fraud detection and the Dutch SyRI judgment. Eur J Soc Secur 23(4):323–340. https://doi.org/10.1177/13882627211031257
Vaswani A, Shazeer N, Parmar N, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 6000–6010. https://dl.acm.org/doi/https://doi.org/10.5555/3295222.3295349
Véliz C (2021) *Privacy is Power*. Penguin (Bantam Press) London, UK
Verhagen MD (2022) A pragmatist's guide to using prediction in the social sciences. Socius: Sociological Research for a Dynamic World, 8, 237802312210817. https://doi.org/10.1177/23780231221081702
Vredenburgh K (2021) The right to explanation. J Polit Philos. https://doi.org/10.1111/jopp.12262
Waelen R (2022) Why AI ethics is a critical theory. Philosophy & Technology 35(1). https://doi.org/10.1007/s13347-022-00507-5
Wan M, Zha D, Liu N, Zou N (2023) In-processing modeling techniques for machine learning fairness: a survey. ACM Trans Knowl Discov Data 17(3):1–27. https://doi.org/10.1145/3551390
Wang A, Kapoor S, Barocas S, Narayanan A (2023) Against predictive optimization: On the legitimacy of decision-making algorithms that optimize predictive accuracy. ACM J Respons Comput. https://doi.org/10.1145/3636509
Watson DS (2022) Conceptual challenges for interpretable machine learning. Synthese 200:65. https://doi.org/10.1007/s11229-022-03485-5
Watson DS (2023) On the philosophy of unsupervised learning. Philos Technol 36:28. https://doi.org/10.1007/s13347-023-00635-6
Watson DS, Floridi L (2020) The explanation game: a formal framework for interpretable machine learning. Synthese. https://doi.org/10.1007/s11229-020-02629-9
Weber M (1922) Economy and society: an outline of interpretive sociology. University of California Press, Berkeley
Weber M (1904) Objectivity in social science and social policy. In: The methodology of the social sciences, Free Press, New York: 1949
Weber M (1910) Remarks on technology and culture. Theory Cult Soc 22(4):23–38. Reprinted: 2005. https://doi.org/10.1177/0263276405054989
Weerts H, Dudík M (2023) Fairlearn: Assessing and improving fairness of AI systems. *ArXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2303.16626
Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The role and limits of principles in AI ethics. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3306618.3314289
Wiener N (1950) The human use of human beings: cybernetics and society. Houghton Mifflin, Boston US

Williams B (1973) A critique of utilitarianism. In: Smart JJC, Williams B (eds) Utilitarianism: for and against, Cambridge University Press

Williams B (1985) Ethics and the limits of philosophy. Harvard University Press, Boston

Wilson A (2017) Techno-Optimism and rational superstition. Techné: Research in Philosophy and Technology, 21(2), 342–362. https://doi.org/10.5840/techne201711977

Wuthnow R (1989) Communities of discourse: ideology and social structure in the reformation, the enlightenment, and European socialism. Harvard University Press, Cambridge, Mass, London

Yeung K, Howes A, Pogrebna G (2020) AI governance by human rights–centered design, deliberation, and oversight: an end to ethics washing. In: The oxford handbook of ethics of AI, Oxford University Press, Oxford

Zarsky T (2015) The trouble with algorithmic decisions. Sci Technol Human Values 41(1):118–132. https://doi.org/10.1177/0162243915605575

Zednik C, Boelsen H (2022) Scientific exploration and explainable artificial intelligence. Mind Mach 32:219–239. https://doi.org/10.1007/s11023-021-09583-6

Zuboff S (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power. PublicAffairs