9

AI and Political Theory

Linda Eggert, Jeffrey Howard, Ting-an Lin, Lorenzo Manuali, and Rob Reich

Abstract: In mapping the emerging landscape of political theory on artificial intelligence, this chapter identifies significant contributions to date as well as areas for future research, concerning (1) theories of justice, (2) democracy, and (3) rights. It addresses issues of algorithmic injustice, the digital public sphere, digital-age human rights, as well as the contested possibility of AI rights. Along the way, the chapter assesses what existing theories imply for new policy questions in a world with AI and what basic theoretical commitments or assumptions may need to be rethought.

Introduction

AI and Political Theory

Political theory is a rich and varied discipline, extending across philosophy, political science, international relations, law, and economics. Our focus in this chapter is on political theory as a predominantly normative – rather than empirical or historical – enterprise, concerned with assessing and defending values, principles, and reasons for action (Estlund, 2012; Miller, 2022).

The aim of this chapter is not to defend substantive positions but to set out some of the issues related to artificial intelligence that preoccupy political theorists. The task is to map emerging scholarship, identifying some of the most significant contributions to date, and to highlight propitious lines of inquiry for future research. In some cases, the task involves assessing what

existing theories imply for new policy questions raised by AI. In other cases, it involves rethinking basic theoretical commitments or assumptions. What makes AI a fertile and important topic for political theorists is its potential to contribute to or prompt us to rethink our understanding of substantive problems that political theory has traditionally sought to address.

The chapter is divided into three main parts: (I) Justice, (II) Democracy, and (III) Rights. It does not discuss many other important issues of interest to political theorists, some of which are taken up in other chapters in this volume. These include issues concerning national security and international relations (Chapter 8), economic inequality and the labor force (Chapter 7), policy making and the administrative state (Chapter 4), and research methods (Chapter 10).

Within the confines of this chapter, we cannot hope to adequately characterize the rich array of different subdisciplinary traditions and approaches in political thought and theory. But this may be just as well. AI's claim on the attention of political theorists of all stripes has begun to blur traditional boundaries between different methods and approaches (Risse 2023). While thinkers outside of analytic liberal-egalitarian political theory have long been concerned with the ways in which technology affects people's lives, much of contemporary analytic political theory is comparatively late to the party (Bernholz et al., 2021; Risse, 2023; Lazar, 2025). As political theory evolves as a field, AI may become a powerful catalyst for the convergence of different approaches within political thought.

I. Justice

Algorithmic Injustice

This section provides an overview of scholarly discussions of *algorithmic injustice*, broadly construed as situations in which AI systems introduce new forms of injustice or exacerbate existing ones.

Regarding the unjust impact of algorithms, considerable attention has been drawn to the issue of *algorithmic bias*: Algorithms produce biased outcomes that perpetuate or exacerbate existing forms of discrimination and oppression (O'Neil, 2017; Noble, 2018; Benjamin, 2019; Buolamwini and Gebru, 2018). Algorithmic biases occur, for example, when algorithms systemically assign higher health needs to White populations compared to people of color (Obermeyer et al., 2019) or when algorithms associate denigrating messages with women and people of color (Noble, 2018).

In response to concerns about algorithmic bias, the term *algorithmic fairness* was introduced to describe the vision of ensuring that algorithms operate fairly. (For an overview, see Hellman,

2025.) The initial scholarly debates differentiated various notions of fairness – primarily understood as ensuring statistical parity – and debated their suitability in assessing whether algorithms operate fairly (Kleinberg et al., 2016; Narayanan, 2018; Binns, 2020; Hedden, 2021; Hellman, 2025). Valuable critiques have been raised about the limitations of understanding algorithmic fairness as ensuring statistical parity. One prominent worry is that such an analysis problematically isolates AI systems from broader social contexts and overemphasizes technocentric responses (Hoffmann, 2019; Le Bui and Noble, 2020). These reflections have led to a "structural turn" that goes beyond exploring technical methods for debiasing AI systems and considers how AI systems shape broader power dynamics and contribute to injustices (Kalluri, 2020; Le Bui and Noble, 2020).

Major questions concerning algorithmic fairness include: What is algorithmic fairness? Why is it valuable? What does it require in practice? Different traditions of political theory have been drawn on to provide insights.

John Rawls' theory of justice (1971) and its implications for distributive justice have been widely engaged with and translated into practical guidelines by computer scientists. In political theory, Gabriel (2022) is a notable example of engagement with Rawlsian principles of justice that outlines their implications for AI. For example, what is known as Rawls's first principle of justice requires that the design and deployment of AI should protect the equal basic liberties of all. The principle of fair equality of opportunity implies moving away from a formal conception of fairness (or "de-biasing" AI systems) toward more substantive fair outcomes, such as how AI tools might help mitigate discriminatory impacts due to social membership. Furthermore, Gabriel suggests that the difference principle implies that, when AI systems are integrated into key social practices, they must work to the greatest advantage of the least advantaged members of society.

Scholarship in feminist and critical race theories offers a more multifaceted understanding of algorithmic (in)justice that extends beyond distributive justice. While the terminology varies, a common theme in this literature is its emphasis on power dynamics, social relations, and social status across different populations. This emphasis has led to suggestions to rethink the goals and practices of algorithmic fairness. For example, Iris M. Young's influential work on structural injustice (2010) has led scholars to reconceptualize the goal of algorithmic fairness as the pursuit of a more just social-technical structure in which unjustified power imbalances are mitigated (Lin and Chen, 2022; Kasirzadeh, 2022). Insights from intersectionality research (Crenshaw, 1991; Collins 2002) and relational egalitarianism (Anderson, 1999; Fraser, 1995) suggest that the goal of algorithmic fairness should encompass not only distributive equality but also people's ability to relate to one another as equals (Kong, 2026).

Concerning practices of pursuing algorithmic fairness, reflections on power dynamics have led to advocacy for participatory approaches to technological development (Costanza-Chock, 2018; D'Ignazio and Klein, 2020), emphasizing the importance of engaging a wide range of stakeholders (especially with marginalized perspectives) in the process. Though they offer a

promising direction, participatory approaches in AI design often face practical challenges and lack concrete guidelines. They have also raised concerns about "participation washing" (Birhane et al., 2022). How to best implement participatory visions remains an important question for future exploration.

Another body of literature concerns how AI models shape epistemic norms and practices, and produce unjust impacts. One major theme is how AI systems contribute to *epistemic injustice* – that is, injustice concerning the domain of knowledge. Studies have suggested that the algorithms behind social media exacerbate epistemic injustice through algorithmic targeting and sorting (Stewart et al., 2022) and constitute a novel form of hermeneutical injustice, as algorithmic profiling leads to epistemically fragmented environments (Milano and Prunkl, 2025).

The content-creation capacity of generative AI models has further impacted collective knowledge and epistemic practices, presenting new forms of testimonial and hermeneutical injustices through amplification of misinformation, intentional manipulation of wrongful content, and increased hermeneutical obstacles for marginalized groups (Kay et al., 2025; Mollema, 2025). Since algorithmic systems tend to encode and manifest the conventions of dominant groups, such as accents and communication styles, the constant cultural code-switching required of marginalized groups constitutes a new form of epistemic oppression: cultural smothering (Falbo and LaCroix, 2022). Similar concerns have been raised about AI models' tendency to generate content that reflects or prioritizes Western-centric perspectives, which may unjustly reinforce Western epistemic hegemony. (See the subsection on Global & Environmental Justice for further discussion.)

Finally, across many discussions about algorithmic injustice, the insights that non-Western traditions can provide are worth further exploration. Birhane (2021) has used Ubuntu's relational ethics to highlight the value of relationality and to rethink how AI practices shape the web of relations. Wong (2025) has drawn from Confucian philosophy to rethink trustworthiness and trustworthy AI. Green (2025) has compared the AI ethics principles developed by tech corporations in Silicon Valley and those proposed by the Roman Catholic Church to examine different approaches and facilitate potential dialogue between technology and religion. With AI systems' growing prevalence and global impact, the need to engage with diverse perspectives and epistemic traditions is as urgent as ever in theorizing the many faces of algorithmic (in)justice.

Rectificatory Justice

Rectificatory justice is broadly concerned with how to respond to injustice in its aftermath, in an effort to rectify or repair wrongful harm. While AI plausibly intersects with this topic in numerous ways, we restrict ourselves to two prominent aspects of the contemporary debate. The first focuses on issues of liability to compensation. Who is responsible for harms caused by AI?

Relatedly, who incurs the duty to remedy the harm caused? If political theory is to offer adequate normative guidance to policymakers designing regulations for AI, including by rethinking tort liability for AI-inflicted harms, it must wrestle with this question. The second aspect concerns the other main way in which the state responds to wrongdoing: criminal punishment. Here the most discussed issue concerns the use of AI technologies in the administration of criminal justice.

Consider, first, the general issue of responsibility for harm, on which compensation claims turn. In contexts in which people may suffer harms caused by AI, assigning responsibility is of special moral urgency. Think, for example, of crashes involving autonomous vehicles, harms and destruction caused by autonomous weapons, or harms caused by delayed, incorrect, or insufficient treatment as a result of errors by AI tools in healthcare. Consider, too, administrative contexts in which people may be unfairly burdened or denied resources as a result of mistaken automated decision making.

One prominent concern is that autonomous systems' autonomy – understood as a certain kind of independence from human agents – makes it difficult, if not impossible, to assign moral or legal responsibility to human persons for acts performed and harms inflicted by AI. Neither operators nor programmers nor manufacturers may be sufficiently directly morally or causally linked to AI-inflicted harm to be held meaningfully accountable. The influential notion of a 'responsibility gap' describes this presumptive 'gap' between machine action on the one hand and responsibility for those actions on the other (Matthias, 2004; Sparrow, 2007). More recently, the label has come to be used to describe several different concerns about ascribing responsibility for AI-inflicted harms (Santoni de Sio & Giulio Mecacci, 2021).

One question is, "Who is 'on the hook' for harms caused by AI?" More precisely, "Under what conditions do people become liable to punishment, and under what conditions do people incur obligations to compensate victims for harms caused by AI?" Another question is whether responsibility gaps really exist and, if they do, whether they can be closed. One complication here arises from the somewhat ambiguous nature of relevant "gaps." On the one hand, the question of whether a responsibility gap is bridgeable arises only if there really exists at least a *prima facie* gap. On the other hand, some views suggest that, if a *prima facie* gap can be closed, there is no relevant gap in the first place.

Perhaps harms caused by AI can be traced back to human agency to a sufficient degree to hold persons morally responsible for those harms (Simpson & Müller, 2016). Perhaps human agents have the normative power to deliberately take responsibility, to make themselves answerable, for harms caused by AI (Kiener, 2022). Or, perhaps, we should view AI systems as physically instantiated socially constructed institutions, such that any difficulty to attribute responsibility is no different from any other collective action problem (Robillard, 2018). Perhaps the gap can thus

be closed by treating institutions and group agents, rather than individual persons, as relevant responsibility-bearers (Taylor, 2021).

Not everyone assumes that responsibility gaps are necessarily objectionable. On some views, certain gaps might actually be desirable. They may be beneficial, at least in contexts of tragic moral choices, because they reduce the psychological costs of responsibility that would otherwise fall on human decision makers (Danaher, 2022b).

The notion of responsibility gaps raises the broader question of what, if any, bearing the possibility of holding someone responsible should have on the permissibility of delegating decisions to AI systems. The possibility of accountability – often used synonymously with responsibility in this context (Lechterman, 2022) – is sometimes treated as a necessary condition for the permissible use of AI. A common argument, for example, is that human agents must remain in control over autonomous systems precisely because, unlike human moral agents, AI cannot be held morally or legally accountable for its actions and decisions (Asaro, 2020).

If the use of AI ultimately reduces harm and leads to fewer lives lost, for example, in road traffic and armed conflict, but responsibility gaps cannot be closed, we may be left to choose between states of affairs in which less harm is caused but there is no possibility of accountability and states of affairs in which more harm is caused but responsible agents can be held to account (Eggert, 2023). One resulting question for future research in this area is how to weigh the harm-reducing benefits of AI against the importance of traditional accountability mechanisms. Further questions for future research include how accountability affects the permissibility of delegating certain potentially harmful actions and decisions to AI; how accountability should affect design choices; and, under what conditions and at what cost human agents must retain a certain degree of control over AI agents specifically for the purpose of exercising responsibility.

Now, consider the issue of criminal justice. Though traditionally the purview of legal theorists, criminal justice is now widely recognized as a fundamental concern of political philosophy. AI is deployed in the context of criminal justice in numerous ways. One way, not strictly concerned with rectification of crime but rather its prevention, concerns the role of AI in predictive policing (Jørgensen, 2025). AI is also increasingly deployed downstream, in decisions about the sentencing of offenders and about their suitability for release under parole, based on risk assessments for recidivism. The actual punishments that criminals receive could thus be determined by algorithmic tools (though some scholars would quibble about whether these predictive algorithms are sufficiently sophisticated to be called “AI”).

The use of predictive technology to determine the punishments of criminals is highly controversial (Lewis, 2013). Some of the controversy tracks general worries about predictive algorithms, especially if training data reflect racial biases (see the section on Algorithmic

Injustice). Insofar as there are duties to rectify the historical injustices of institutional racism, the use of criminal rectification measures that reinforce racist injustice seem particularly perverse, reinforcing the racist injustices that pervade the very institution of punishment (Shelby, 2022). A related concern is the phenomenon of statistical discrimination, whereby someone is treated not as an individual but rather on the basis of properties of the group of which they are a member. (For relevant background, without particular focus on AI, see Enoch et al., 2012.) Some theorists argue that predictive algorithms can be unjust, even if they do not produce any downstream harms, by virtue of involving impermissible epistemic claims (Stone and Lazar, 2022).

To the extent that punishment should be determined by backward-looking considerations of retribution or communicative condemnation, forward-looking risk assessments may seem irrelevant (Kelly, 2018). After all, crime prevention and punishment are two different domains. To the extent that one endorses an instrumental approach to the justification of punishment, based on its future effects, one will be less inclined to rule out forward-looking measures *ex ante*. Given the increasing popularity of instrumental approaches to criminal punishment (e.g. Tadros, 2011), it is no surprise that there is an active research agenda exploring when risk-based sentencing algorithms can and cannot be justified. (For useful overviews, see Ryberg, 2020 and 2024, and Ryberg and Roberts, 2022). Fruitfully, the question of AI has also prompted theorists to return to foundational theoretical issues, such as whether repeat offenders merit greater punishment than first-time offenders (Ewing, 2019).

Global and Environmental Justice

Certain practices surrounding AI systems raise issues of global justice, particularly concerning the unfair distributions of benefits and burdens between the Global North and the Global South, as well as such distributions' implications for global power dynamics. This subsection highlights specific dimensions of justice that gain heightened salience when examined through a global justice perspective, which critically accounts for the power asymmetries of the prevailing, and historically rooted, global order.

One aspect that has received extensive analysis is AI's impact on the environment. AI systems consume vast amounts of planetary resources and impose substantial environmental costs (Crawford and Joler, 2019). The manufacturing of AI infrastructure requires vast amounts of mineral resources. These resources are largely extracted from low- and middle-income countries and have been linked to geopolitical conflict and environmental harms in those countries (Crawford, 2021; Dauvergne, 2022). Training and using AI models further requires large amounts of energy, such as electricity and water, while energy consumption by data centers (Bender et al., 2021) exacerbates environmental degradation that disproportionately burdens already marginalized populations (Sultana, 2022; Ngcamu, 2023). While the extent of AI's energy consumption is an empirical issue, the information disclosed by tech companies remains

limited. The lack of transparency impedes proper evaluation and leads to polarized attitudes toward AI's environmental impact (Luccioni et al., 2025).

Besides potential benefits of AI and short-term employment impacts affecting tech workers (see Chapter 7 in this volume), another part of the story is the labor behind the development of AI systems, which marks another domain of AI's extractive process and unjust impacts. Currently, the AI supply chain relies on an *international division of digital labor* (Muldoon and Wu, 2023). The high-paid, well-regarded software engineering and AI research jobs are clustered in the Global North. But many other types of human labor behind the AI supply chain – from workers supporting the physical infrastructure of AI (e.g., miners and workers assembling electronic components) to data workers (e.g., those who undertake data preparation and annotation, verify algorithmic results, and review content for moderations) – tend to be low-paid. They also work under precarious and even harmful and exploitative conditions. And they are largely located in the rest of the world (Gray and Suri, 2019; Tubaro et al., 2020; Roberts, 2019; Crawford, 2021). Even though AI supply chain labor plays a crucial role in AI's development and deployment, corporate AI narratives and the design of microwork platforms have been found to obscure and devalue certain forms of human labor, thereby further enabling labor exploitation for the benefit of some large tech corporations (Gray and Suri, 2019; Crawford and Joler, 2019).

In addition to material injustices, concerns have been raised about AI's unjust impacts in reinforcing the Western hegemony of values and epistemologies. Currently, AI models are primarily trained on high-resource languages – those with large amounts of digitized data, such as English, French, and German (with English content being the most dominant one). While many other languages are spoken by huge populations, their countries have significantly lower quantities of digitized data. This has created a *digital language divide* in which English speakers are granted greater access to advanced AI tools than speakers of lower-resource languages (Ta et al., 2023). Furthermore, studies reveal that Western-centric, especially American-centric, perspectives are often encoded in AI models (Johnson et al., 2022). The skewed perspectives embedded in AI systems have given rise to concerns about Western hegemony and cultural imperialism (Birhane, 2023; Muldoon and Wu, 2023). This situation also echoes what Cave and Dihal (2020) call the "Whiteness of AI" – the fact that AI is predominantly portrayed and perceived as for and about White people, which perpetuates and exacerbates the racialized hierarchy of power.

Based on these observations, burgeoning scholarship has recently argued that the power dynamics shaped by AI systems and surrounding industrial practices mirror and further extend the extractive dynamics of colonialism (Kwet, 2019; Ricaurte, 2019; Couldry and Mejias, 2019; Coleman, 2019; Amrute, 2019; Oyedemi, 2019; Mohamed et al., 2020; R. Adams, 2021; Birhane, 2023; Muldoon and Wu, 2023; Nothias, 2025). While using slightly different terms – including *digital colonialism*, *data colonialism*, *algorithmic colonialization*, and *algorithmic coloniality* – the core concern is the concentration of economic and political power that big tech companies which are primarily based in the US hold over the rest of the world. These companies

conduct extractions in the name of progress, while imposing disproportionate harms and violence (including environmental degradation, labor exploitation, and cultural imperialism, as discussed above) on a broad range of populations, including populations of the Global South and other underserved populations, such as Black, Indigenous, and Latinx communities. Thus far, discussions have drawn on literature in decolonial studies (Maldonado-Torres, 2007; Quijano, 2007; Mignolo, 2012) to delineate the mechanisms that shape these global power dynamics (Muldoon and Wu 2023) and to explore ways to disrupt them (Mohamed et al., 2020; R. Adams, 2021).

Many questions about digital colonialism remain to be further theorized. Both the concept of digital colonialism and the adequacy of related terminology are issues of scholarly debate (Moosavi, 2020; Mouton and Burns, 2021; Táíwò, 2022). Further examination is also needed to historicize the idea of digital colonialism, analyze its connections to traditional forms of colonialism and coloniality, and explore potential responses to it. Last, but not least, while scholars largely converge on diagnostic aspects of digital colonialism, one area that warrants further scholarly exploration is the compatibility of their prescriptive approaches (Nothias, 2025).

Intergenerational Justice

According to some commentators, an already existing cluster of concerns about existential risks arising from climate change and other global threats now also includes potentially catastrophic risks posed by AI (cf. Bales et al., 2024). Rather than discuss the possibility of existential risk and the future of humanity, we will limit our discussion here to issues of intergenerational justice. Intergenerational justice concerns what duties, if any, present-day generations have to consider justice for past and future people (Meyer, 2021). Since AI creates new ways in which we in the present can exercise power over people who are no longer alive and people who do not yet exist, it is an appropriate subject of intergenerational justice.

One foundational question is whether valid justice claims exist only between people who can harm or benefit one another (Barry, 1989; Buchanan, 1990; Gauthier, 1987; Heyd, 2009; Hiskes, 2009). While AI has no bearing on the general question of between whom duties exist, and thus on whether duties exist only between people who can harm or benefit one another, AI does affect, in specific contexts, who is able to harm and benefit others. To the extent that AI affects whether and how we, in the present, might harm, wrong, or benefit people who are no longer alive or who do not yet exist, it may bear on questions of intergenerational justice. While political theorists have come to grapple with how AI mediates power relations between living people (Lazar, 2025), a nascent area of research thus concerns how AI mediates the power we, in the present, exercise over both people who are no longer alive and people who do not yet exist.

Consider, first, so-called afterlife technologies. Advances in generative AI have made it possible, in effect, to digitally revive the dead (Worth, 2025; Dickey, 2025; Hern, 2024; Yang, 2024). Users can now create interactive avatars and AI companions that are essentially deepfakes of deceased people. 'Griefbots' or 'deadbots' are mainly marketed as providing new, digital ways of grieving. But deepfakes and chatbots that mimic deceased people could, in principle, appear anywhere; including in schools or museums, allowing people to have conversations with historical figures. Anyone of whom there exists audio and video recordings that can be used to train models will be revivable as an avatar.

Setting aside reasons we might have not to digitally revive the dead, questions raised by afterlife technologies include the following. How might generative AI extend theories of privacy to the afterlife? What does valid consent require to digitally replicate someone and, relatedly, what should people be consulted concerning how their data may be used after their death? Finally, how should digital afterlife technologies be regulated to protect deceased people's rights and interests, and to minimize risks of psychological harm to living users of deadbots?

Now consider future generations. One characteristic of intergenerational relations is the asymmetry in power relations between those currently alive and those who are not. Whereas power relationships between contemporaries can change frequently, present generations necessarily have asymmetrical power over future generations (Meyer, 2021; Barry, 1977; Barry, 1989). Present generations can affect the circumstances of future generations. AI-related actions and policies that deplete the world's natural resources may well set back future generations' interests or reduce their options (cf Barry, 1999; Beckerman, 1999; Heyward, 2017). Besides the circumstances in which future generations will live, present-day generations also determine which future people will exist. That future people's very identities are contingent on the actions of those who come before them has given rise to a persistent debate about the very possibility of intergenerational justice claims (Parfit, 1984). Any applications of AI that might affect what specific people will come into existence may similarly become entangled in what, following Derek Parfit, has come to be known as the non-identity problem.

One central question in debates about intergenerational justice is whether we have reasons to discount future persons' interests (Sidgwick, 1907; Rawls, 1971; Parfit, 1984; Broome, 1994; Mintz-Woo, 2021). These concerns extend well beyond questions relating to technology. Nevertheless, decisions about the development of AI models today can account for or ignore the interests of future generations. AI's familiar promises and perils – from significant medical advances to jeopardizing the future of humanity – also raise questions about whether AI's benefits to existing or near future people may justify risks to those in the more distant future. If AI promises to be sufficiently beneficial to present-day generations, does that justify exposing future generations to certain risks? Might powerful AI models even help determine whether, and to what extent, future people's interests may permissibly be discounted?

II. Democracy

Since this volume contains a dedicated chapter on Democracy/Elections (Chapter 1), our focus here will be on democratic values, rather than institutions, practices, and processes. In a remarkable example of how social (in this case technological) developments can and should prompt us to rethink our theories (Dewey, 1927), the rise of AI has led democratic theorists to rethink values and paradigms core to the field (Bernholz et al., 2021).

Existing democratic governance practices involving AI have already spawned a considerable body of literature. This trend continues political theory's longstanding concern with not just governance by the state but also governance by private as well as other types of public institutions (N. P. Adams, 2018; Anderson, 2017; Kirby, 2022). While normative evaluation of governing institutions that rely on algorithms has been in the public consciousness for a while (Danaher 2016), contemporary AI has spurred theorists to consider whether and how AI should be used to mediate relations in a democratic society (Danaher, 2022a; Lazar, 2024, 2025) and in what circumstances the use of AI entails, for example, a democratic duty of explanation (Lazar, 2024).

Another body of literature concerns the democratization of AI. To "democratize AI" can, as a descriptive matter, mean many things – from merely increasing access to AI to ensuring democratic governance of AI. Seger et al. (2023) give a helpful categorization of various descriptive uses of the phrase "democratizing AI." Considering normative questions of whether and how we ought to democratize AI, many authors helpfully examine the semantic ambiguity the phrase licenses. Among the concepts most commonly discussed (with significant overlap) are the design, development, deployment, use, and governance of AI (Himmelreich, 2023; Lin, 2024; Noorman & Swierstra, 2023). (Note in particular Himmelreich's argument against the democratization of AI, conceived of as ensuring more democratic governance of AI. On AI's impact *on* democracy, see Coeckelbergh, 2024; Jungherr, 2023; Kreps and Kriner, 2023.)

Perhaps three of the most prominent areas on which AI has had substantive effects for political theory are (1) representation, (2) formal deliberation, and (3) the digital public sphere. It is these three areas which will be our focus. (For practical applications of AI to democratic processes, see Chapter 1 in this volume). We will begin with a note about how to categorize AI's various uses in democratic governance.

Conceptualizing AI Uses in Democratic Governance

AI is increasingly used to enhance, support, or implement democratic governance processes in both public and private institutions. These developments have sparked a growing literature in political theory concerned with when and how, if at all, AI should be incorporated into

democratic governance processes (Schneier and Sanders, 2025; Boehmer et al., 2025; Devine et al., 2023; Edelman, 2023; Fish et al., 2024; Konya et al., 2023; Lazar, 2025; Mendoza, 2023; Tessler et al., 2024).

Two different conceptual schemas for categorizing different uses of AI have emerged so far. The first is the Input-Process-Output (IPO) model (Han, 2025). This is a four-fold categorization that focuses on whether a given AI application in a democratic process (1) "originate[s] from the citizenry" (input) and (2) "whether AI decisions carry legal or authoritative weight" (authority of output). Han then assesses the four resulting categories of AI application (democratically binding, undemocratically binding, democratically unbinding, undemocratically unbinding) in terms of democratic values: "inclusive and equal participation, quality of decisions, deliberation, and the autonomy of citizens to set the political agenda" (Han, 2025).

The second schema is Lazar and Manuali's categorization of the various ways large language models (LLMs) are currently used in proposed democratic processes: *summarization*, *aggregation* of preferences, *facilitation* of deliberation, and *representation* (Lazar & Manuali, 2026). These uses are not mutually exclusive. For example, facilitation of deliberation may involve using LLMs to summarize conversation. Generalizing from Lazar and Manuali's focus on LLMs, we will examine how AI has been theorized and used in democratic governance most commonly: *representation* and *deliberation*.

AI in Representation

Perhaps the most prominent example of the influence of AI on democratic theory stems from efforts by computer scientists and democratic practitioners to use AI as representatives for people (Boehmer et al., 2025; Fish et al., 2024; Gudiño et al., 2024). One could apply existing theories of political representation to extract a normative analysis of such efforts.

The technological possibility that AI could predict and extrapolate our preferences over possible policy options with relative accuracy (Boehmer et al., 2025; Fish et al., 2024; Konya et al., 2022) raises the normative question of whether AI could *legitimately* represent us. Lechterman (2024) worries that the normative viability of such a proposal would reduce representation to the mere prediction of preferences. After summarizing some of the arguments in favor of digital representatives (what he terms "Avatar Democracy"), Himmelreich (2023) argues that digital representatives run into the same classic problem as human representatives. Namely, they might deliberate with other representatives and disagree with the preferences of their constituent(s) due to expertise that they have, at which point justifying and explaining their decision(s) to their constituent(s) (a requirement of legitimacy) will be difficult. Han (2025) sees promise in AI representatives to which duties by citizens can be delegated with those citizens' authorization, and which can be overseen by citizens to ensure that those AI representatives continue to act

within their remit. Taylor (2025) also gives a more tentative answer in the affirmative by outlining conditions under which current or near-term-future AI could act as a representative in a normatively robust sense – though he notes that many conditions (including ones involving AI representatives continuously consulting with citizens) would have to be met for that to be the case.

But beyond the question of whether AI can legitimately represent us, the technological possibility of AI representatives also allows us to challenge some of the tenets of representation theory. Taylor's (2025) approach is especially illuminating. It adjusts classic formulations of principles in representation theory in such a way that an AI representative could, assuming certain conditions and constraints, meet them. The challenge, then, is to figure out whether, and why, such revised principles are implausible, or to accept that the original formulations were mistaken. The technological possibility of AI representation thus allows us to try and refine representation theory, either by affirming or by adjusting the original formulations in light of AI-inspired challenges.

The relationship of AI to systems of representation will depend on the capabilities of AI models, of course. How accurate our predictions are with respect to the future capabilities of AI representatives (real or hypothetical) determine how we might be able to challenge current formulations in representation theory. For instance, Taylor's (2025) normative argument in favor of AI representatives depends on a suitably idealized AI representative, with a particular suite of abilities that no currently existing projects that attempt to instantiate AI representation seem to have. If theorists make assumptions that are implausible in light of present and future AI abilities, any normative proposal for an AI representative will be vulnerable to the critique that it is too remote a possibility to consider. It may also be subject to a 'problem of the second best.'

The rise of advanced AI has also led democratic theorists to consider, at least hypothetically, the possibility of a completely AI-run government. Lovett and Zuehl (2022) conclude that something would be missing from such a world: the non-instrumental value of collective self-determination, which they conceive of as a form of distinctly democratic autonomy. This is an example of how engaging with questions about AI may lead to a contribution in its own right to the literature relating to justifications for democratic governance. Contemporary AI has thus helped articulate a value that purely instrumentalist accounts of democracy have not been able to capture.

AI in Formal Democratic Deliberation

One prominent set of proposals for using AI in democratic governance involves employing AI in *democratic deliberation* (see Chapter 1 in this volume). Formal deliberative processes are costly, and it is especially difficult to train and pay for human moderators. Landemore (2024) asks if AI can be used, then, to bring "quality deliberation to the masses." She considers two different

proposals for scaled-up deliberation – Mass Online Deliberation and Multiple Rotating Mini-Publics – and several different roles for AI in scaling up these proposals to allow for more participation. She suggests that AI could facilitate, summarize, select participants, translate, fact-check, and aggregate data (Landemore, 2024). Still, whether AI can actually "scale up" deliberation remains to be seen. One recent study suggests that AI-based facilitation incurs a "penalty" in that people prefer human facilitators in deliberation and so participate less in AI-facilitated deliberation (Jungherr & Rauchfleisch, 2025).

Besides asking whether and how AI can "scale up" deliberation, one might ask what normative constraints should be placed on AI as we employ it in formal deliberative processes, such as in mini-publics. Mikhaylovskaya suggests that AI's guiding value *qua* facilitator ought to be (democratic) equality (Mikhaylovskaya, 2024). She proposes that AI (as opposed to humans) might be better at being an impartial facilitator and at making sure all voices are included. In her view, we ought to design AI facilitators in light of the non-instrumental democratic value of equality.

Considering the use of AI in citizen assemblies, McKinney (2024) offers both (1) a conceptual framework for the various functions of AI in citizen assemblies and (2) a normative framework that allows for the assessment of these functions with respect to democratic goods.

A common thread among those focusing on incorporating AI into formal deliberative processes emerges from the original deliberative-democratic tradition dating back to Cohen and Habermas (Cohen, 1989; Habermas, 1996). Deliberative democratic practitioners in the present context often focus on the ability of AI to *identify consensus* or *agreement* (Devine et al., 2023; Konya et al., 2023; Tessler et al., 2024). However, other deliberative democrats resist the claim that deliberation should (in some sense) aim at consensus or agreement (Fraser, 1990; Young, 2002). In a similar spirit, Palomo Hernández (2025) critiques one popular project that incorporates AI into a democratic decision-making process: the Habermas Machine (Tessler et al., 2024). In addition to critiquing the lack of engagement between participants, Palomo Hernández (2025, p. 1951) argues that the Habermas Machine "overemphasises the desirability of agreement in deliberative processes."

Other theorists have more recently taken an even less deliberation-focused approach. Instead of proposing to use AI in formal deliberative processes (such as in mini-publics), Farrell and Han (2025) suggest that we ask how we can use AI to create, maintain, and strengthen the power of *democratic publics*. Reminiscent of the agonistic democratic tradition (Mouffe, 1993), they remind us that a crucial aspect of politics, in general, and democracy, in particular, is the negotiation of power relations between groups, and that democracies require people to recognize themselves as part of publics to tackle shared problems (Dewey, 1927). Farrell and Han (2025)

thus call for both theorizing and empirical research related to the question of how AI can support the formation of publics and their negotiation of social problems.

The Digital Public Sphere

Consider now the place of much of the contemporary democratic action: the digital public sphere. Our public discourse is increasingly mediated by AI. We are, as Seth Lazar puts it, "connected by code" (2026). The most conspicuous way in which AI mediates our public discourse concerns its role on social media platforms. For better or worse, these platforms continue to serve as the most significant fora in which people argue with one another about important social and political issues, express their opinions, and learn information about current political events. Whether they want to be or not, Facebook, Instagram, X, TikTok, BlueSky, YouTube, and other spaces together constitute our new digital public sphere.

How does AI govern our communications in this space? First, AI systems enforce the rules on what speech is allowed. While humans formulate the general rules prohibiting harmful content (from hate speech to death threats to child abuse material), AI systems play a fundamental role in applying and enforcing them. Second, AI systems determine what content people see. Algorithms tuned to optimize our engagement control what posts appear in our feeds. Finally, machines loom large on social media because the content we encounter is itself often synthetically produced – the product of generative AI tools powered by large language models, such as ChatGPT, Claude, and Grok. Increasingly, chatbots are integrated into the platforms, themselves participants in our digital public conversation.

Each of these three areas raises important normative questions, including for political theorists – concerning the moral duties of both private corporations and policymakers tasked with regulating them. Consider the role of AI in content moderation: the process by which platforms govern the speech of their users. Insofar as such policies have a profound effect on users' communicative interests, they are arguably evaluable by normative standards of political theory; and when such policies are compelled by the state, they are inarguably so (see Lazar, 2026 and Howard, 2026a).

One question concerns permissible and impermissible uses of AI by platforms as part of their content moderation practices. Social media platforms deploy supervised machine learning (ML) models, trained on massive data sets of labelled training data. For each type of prohibited content, platforms build what is known as an ML *classifier* to adjudicate whether a fresh piece of content is allowed or disallowed. (This approach replaces older, cruder uses of technology to moderate content – such as simple lists of banned keywords.) Suppose a user on X or Facebook offers a robust critique of a politician – perhaps one that could arguably be interpreted as a threat

(see Fisher and Howard, 2025). Suppose the speech is then removed, thanks to an AI classifier. Why might the user object?

There are many potential objections – all applications of more general worries about deploying AI in decisions that impact our fundamental interests. Some have argued for the notion of a fundamental right to a human decision (for a full treatment of this view, see Huq, 2020). Others have raised worries about the inaccuracy of AI, though there are questions as to whether humans are really "better," not least, since they cannot do the relevant work at scale. In the content moderation context, AI classifiers struggle to interpret speaker intent or purpose, such as inferring sarcasm and nuance – leading to false positives (Douek, 2022). Likewise, efforts to evade rules by speakers outsmarting AI leads to false negatives. And there are familiar worries about bias, and that AI will end up producing unfairly worse results for members of historically marginalized groups (see the section on Algorithmic Injustice).

One particularly instructive objection concerns the *opacity* of AI models – the fact that they constitute a so-called "black box," such that even engineers often cannot explain fully why those models produce the verdicts that they do. Machine learning models can detect patterns in training data far more subtle and complex than humans ever could themselves. But this may make it difficult for users to understand and predict when, and why, their speech is likely to be affected – and to contest mistakes, holding power accountable. This impairs users' interest in what Kate Vredenburgh, writing about AI opacity in general, calls *informed self-advocacy* (2022). Note that, while this analysis is inspired by the case of AI, it has wider ramifications – helping us see, for example, what is objectionable about opaque bureaucratic rules that citizens cannot successfully navigate or contest. In this way, Vredenburgh's work fruitfully illuminates how attention to problems raised by AI prompts us to make progress in more fundamental normative questions. (For related analysis, see Lazar, 2024.)

These objections need not render the use of AI impermissible in the governance of online speech. Rather, they constrain what sort of AI should be used, and what requirements of transparency and oversight properly apply to it. Further, the costs of using AI must be weighed against the benefits. Content moderation is a defensive effort to prevent harm; and preventing harm requires speed. Faced with billions of user-generated posts per day, the large platforms cannot feasibly catch the surfeit of harmful content without the use of technology. Insofar as platforms have a duty to combat such content, they may have a duty to use AI (Howard, 2026a, though cf. Barnes, 2022).

What about the role of AI in *curating* public discourse, such as deciding what content users see in their recommended algorithm-powered newsfeeds? Political theorists are only beginning to address this question. As Lazar argues, settling the boundaries of permissible and impermissible online speech does not settle this problem (2026). Rather, we need a positive vision of what a

flourishing public sphere ought to involve. Despite historic attention to this question by luminaries, such as John Dewey (1926), Iris Marion Young (2002), Jürgen Habermas (1989 and 1996), and John Rawls (2005), there has been insufficient effort to update our conception of a public sphere for the digital age (for an early exception, albeit in legal theory, see Balkin 2007). But progress is afoot. Some work addresses the principles that should govern the amplification of content (Brown forthcoming; Lazar, 2026; Miller, 2021). Some work explores the ways that amplifying some (otherwise innocuous) speech can cause harm (Howard and Kira, 2026), as with forms of online public shaming (Billingham and Parr, 2020). Other work focuses on the discursive norms that citizens ought to follow in their online discourse (Cohen and Fung, 2021). And some work focuses on how to design online spaces to build community (Forestal, 2021). There is also a raft of work on such problems as polarization which, while not limited to the digital realm, may be exacerbated by it (Sunstein, 2017; Talisse, 2019), as well as on the problem of manipulation (Renzo, 2025).

A final set of concerns about AI involves the burgeoning use of LLM-powered chatbots. Insofar as these chatbots are used as new search engines, they bring back to the fore familiar questions about how citizens ought to receive their information (Sunstein, 2017), and what the normative principles are for governing search (Grimmelmann, 2014). Indeed, the question of the appropriate role of the news media in an age of algorithmic media, and how to rebuild shared media institutions that command widespread trust, remains seriously under-researched by political theorists relative to its importance (though see Heawood and Peter, 2023) – a problem only exacerbated by LLMs. Some existing work concerns the moral obligations of AI companies that produce public-facing chatbots (or other content-creation tools, such as image or video generators) – e.g., concerning how to align model outputs with important moral and political values (e.g., Kasirzadeh and Gabriel, 2023). Other work concerns the downstream obligations of platforms that host AI-generated content (Fisher et al., 2024).

Cutting across all these concerns is a research agenda on the political economy of Big Tech. Many argue that antitrust and anticompetition policy tools, or reforms to business models, are the appropriate remedy for the pathologies of our digital public sphere (Balkin 2021 and 2022; Hindman 2018; Reich *et al* 2021). In this way, concerns about AI are fruitfully reinvigorating broader debates about economic justice.

III. Rights and the Future

This section explores the future of rights in a world shared with AI. We will first consider the future of *human* rights in a world in which our lives increasingly depend on technology. We will then briefly consider the contested notion of *robot* rights.

Human Rights

Three 'generations' of existing human rights are commonly distinguished: (1) civil and political rights; (2) economic, social, and cultural rights; and (3) collective rights. AI's transformative impact on critical areas of human life and activity has led to calls to recognize a new, fourth 'generation' of human rights, fit for the digital age (von Schirach, 2021; Risse, 2023; Shany, 2023). These include a right to digital self-determination, the right that certain algorithms be "transparent, verifiable and fair," and that "major decisions" be "taken by a human being (von Schirach, 2021) as well as rights to protect our "epistemic actorhood" as knowers and knowns and, ultimately, a right to the exercise of *human* intelligence (Risse, 2023).

No consensus exists on the need for a new generation of human rights for the digital sphere. On the one hand, a longstanding worry is that positing too many human rights devalues their currency. If we treat everything important as a right, the worry goes, then rights become meaningless (Griffin, 2008). On the other hand, some commentators have argued, if human rights are to remain relevant in a changing world, the human rights framework must adapt as transformative technologies advance (Schulz & Raman, 2020). One foundational question AI brings to the fore in human rights theory, then, is how to adjust human rights to changing social norms and new technological advances while avoiding longstanding fears about conceptual pitfalls, including rights inflation, overreach, and expansionism. In short, as transformative technologies advance, should human rights evolve? If so, how?

Even if new technology requires a reassessment of human rights, the determination of *which* rights would make the list remains contested. As noted above, proponents of new human rights have invoked rights to protect people from new and distinct, AI-precipitated harms. Calls for new human rights raise the question whether non-human, artificial forms of agency create new and distinct ways of harming and/or wronging people. This possibility, in turn, raises the question of whether we need new legal protections for people – in particular, as subjects of algorithmic decision making. While legal rights depend on institutional recognition and enforcement, *moral* rights exist independently of political and legal institutions. What, if any, *moral* rights are available to humans, non-human animals, or non-human, artificial agents does not depend on what rights are *legally* recognized.

An enduring question is what 'grounds,' or what justifies, human rights. A prominent assumption is that all human beings have human rights simply by virtue of being human (Griffin, 2008; Simmons, 2015). But it is contested whether there is anything special about being human that can explain the existence of human rights. What feature of 'being human' could possibly explain – in a manner acceptable to people with different commitments, from different cultures, and different religious backgrounds– what makes humans 'special' and why we have certain rights simply by "virtue of our humanity"?

One popular response in human rights theory has been to divorce the defense of human rights from attempts to specify their foundations; to provide an account of their function in global political practice that does not rely on some contested view about the significance of humanity in virtue of which we have human rights (Rawls, 1999; Cohen, 2004; Beitz, 2009). While human rights theorists have been divided over the significance of the notion of "our humanity," the arrival of AI on the scene might sharpen views over time about whether there is, in the end, anything special about "being human." By providing a new foil to humans, non-human, artificial agents may help us advance our understanding of the notion of different entities' "moral status" (Risse, 2023; Liao, 2020 and 2010). They might bring into sharper relief what – if anything – makes humans relevantly unique. Or they may confront us with the sobering possibility that there is nothing special about human intelligence, or being human, after all, and that insisting otherwise is simply parochially "carbon chauvinist" (Sagan, 1973).

Robot Rights

Consider, finally, the possibility of robot rights (Coeckelbergh, 2010; Gunkel, 2018; Risse, 2023). If the features that ground rights are not unique to humans, we face the possibility that those same features ground rights of non-human entities, along with questions about whether those non-human entities can be subjects of justice. It is already a common view that we should recognize certain rights of non-human animals. (On the question of political membership for non-human animals, see Donaldson and Kymlicka, 2011.) A less common view is that we should recognize certain rights of inanimate objects, such as trees, rivers, or mountains. Instead of taking a position on the possibility of robot rights, we will note simply how much is at stake in getting the answer right. To say that AI systems are moral rights-holders would be to say that they are the kinds of entities that can be wronged. Generally speaking, the conceptual possibility of being wronged comes with the possibility of generating duties of compensation, and apology, and so on, as well as, possibly, permissions to enforce the protection of rights (Bryson et al., 2017; Basl & Bowen, 2020). Hence, to argue for AI systems' rights is, in effect, to argue for the possibility of duties to compensate AI systems, to apologize to AI systems, and perhaps even the possibility of permissions for AI systems to enforce the protection of their rights. Recognizing rights comes with serious normative commitments.

In sum, by providing a foil to humans, AI might help us make progress on the longstanding question of who has rights and why. AI also gives new urgency to questions about whether new technological possibilities generate the need for new human rights, and about nothing less than what, in the end, if anything at all, is 'special' about being human.

Closing Remarks

Different theoretical perspectives have always come to different conclusions. Transformative though it is, AI changes nothing about political theory at the foundational level. Foundational

questions at the heart of political theory concerning at least the nature of certain values are, in a sense, impervious to the contingencies of the 'real world.' Yet, longstanding concerns about, for example, structural racism now extend to algorithmic bias and discrimination. What one thinks about AI and the future of work (though we didn't discuss this here: see Chapter 7) will likely depend at least partially on one's views about meaning in life, distributive justice, and equality of opportunity. And one's views about robot rights will likely be informed by one's views about human rights and those of non-human animals.

This chapter presented a partial snapshot of the intellectual landscape during a moment of great upheaval. We cannot meaningfully address, not even in theory, the political questions of our time without grappling with the impact technology has on people's lives (Risse, 2023). What one should think about AI's appropriate place in our social and political lives may, in many cases, be determined by one's views in political theory. Political theory has an important role to play as we grapple with how to live with AI, and we hope to have provided a sample of starting points for future research.

References

Adams, N. P. (2018). Institutional Legitimacy. *Journal of Political Philosophy*, *26*(1), 84–102.

Adams, R. (2021). Can artificial intelligence be decolonized? *Interdisciplinary Science Reviews*, *46*(1–2), 176–197. https://doi.org/10.1080/03080188.2020.1840225

Amrute, S. (2019). Tech Colonialism Today. In *EPIC* [Keynote Address]. EPIC2019. https://www.epicpeople.org/amrute-tech-colonialism-today/

Anderson, E. (1999). What Is the Point of Equality? *Ethics*, *109*(2), 287–337. https://doi.org/10.1086/233897

Asaro, P. (2012). On banning autonomous weapon systems: Human rights, automation, and the dehumanization of lethal decision-making. *International Review of the Red Cross*, *94*(886), 687–709. https://doi.org/10.1017/S1816383112000768

Asaro, P. (2020). Autonomous Weapons and the Ethics of Artificial Intelligence. In S. M. Liao (Ed.), *Ethics of Artificial Intelligence* (pp. 212–236). Oxford University Press.

Aytac, U. (2024). Digital Domination: Social Media and Contestatory Democracy. *Political Studies*, *72*(1), 6–25. https://doi.org/10.1177/00323217221096564

Bales, A., D'Alessandro, W., & Kirk-Giannini, Cameron Domenico. (2024). Artificial intelligence: Arguments for catastrophic risk. *Philosophy Compass*, 19 (2). https://doi.org/10.1111/phc3.12964

Balkin, J. M. (2007). Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society. In P. S. Berman (Ed.), *Law and Society Approaches to Cyberspace*. Routledge. https://doi.org/10.4324/9781351154161

Balkin, J. M. (2021). How to Regulate (and Not Regulate) Social Media Symposium: Free Speech and Social Media Platform Regulation. *Journal of Free Speech Law*, *1*(1), 71–96. https://heinonline.org/HOL/P?h=hein.journals/jfspl1&i=71

Balkin, J. M. (2022). To Reform Social Media, Reform Informational Capitalism. In L. C. Bollinger & G. R. Stone (Eds.), *Social Media, Freedom of Speech, and the Future of our Democracy* (pp. 233–254). Oxford University Press. https://doi.org/10.1093/oso/9780197621080.003.0014

Barnes, M. R. (2022). Online Extremism, AI, and (Human) Content Moderation. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14295

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press. https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/

Barry, B. (1977). Justice Between Generations. In J. Raz & P. M. S. Hacker (Eds.), *Law, Morality, and Society: Essays in Honour of H. L. A. Hart* (pp. 268–284). Oxford : Clarendon Press. http://archive.org/details/lawmoralitysocie0000unse

Barry, B. (1989). *Theories of Justice: A Treatise on Social Justice, Vol. 1*. University of California Press.

Barry, B. (1999). Sustainability and Intergenerational Justice. In A. Dobson (Ed.), *Fairness and Futurity: Essays on Environmental Sustainability and Social Justice* (pp. 93–117). Oxford University Press. https://doi.org/10.1093/0198294891.003.0005

Basl, J., & Bowen, J. (2020). AI as a Moral Right-Holder. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 289–306). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.18

Beckerman, W. (1999). Sustainable Development and Our Obligations to Future Generations. In A. Dobson (Ed.), *Fairness and Futurity: Essays on Environmental Sustainability and Social Justice* (pp. 71–92). Oxford University Press. https://doi.org/10.1093/0198294891.003.0004

Beitz, C. R. (2009). *The Idea of Human Rights*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199572458.001.0001

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT'21), Association for Computing Machinery*, 610-623. https://doi.org/10.1145/3442188.3445922

Benjamin, R. (2019). *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.

Benn, C., & Lazar, S. (2022). What's Wrong with Automated Influence. *Canadian Journal of Philosophy*, *52*(1), 125–148. https://doi.org/10.1017/can.2021.23

Bernholz, L., Landemore, H., & Reich, R. (Eds.). (2021). *Digital Technology and Democratic Theory*. University of Chicago Press.

Billingham, P., & Parr, T. (2020). Online Public Shaming: Virtues and Vices. *Journal of Social Philosophy*, *51*(3), 371–390. https://doi.org/10.1111/josp.12308

Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524. https://doi.org/10.1145/3351095.3372864

Birhane, A. (2021). Algorithmic injustice: A relational ethics approach. *Patterns*, *2*(2), 100205. https://doi.org/10.1016/j.patter.2021.100205

Birhane, A. (2022). "Power to the people? Opportunities and challenges for participatory AI." *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. https://doi.org/10.1145/3551624.3555290

Birhane, A. (2023). Algorithmic Colonization of Africa. In K. Dihal & S. Cave (Eds.), *Imagining AI: How the World Sees Intelligent Machines* (pp. 247–261). Oxford University Press.

Boehmer, N., Fish, S., & Procaccia, A. D. (2025). *Generative Social Choice: The Next Generation* (No. arXiv:2505.22939). arXiv. https://doi.org/10.48550/arXiv.2505.22939

Broome, J. (1994). Discounting the Future. *Philosophy & Public Affairs*, *23*(2), 128–156. https://doi.org/10.1111/j.1088-4963.1994.tb00008.x

Brown, É. (forthcoming). Algorithmic Amplification and Democratic Equality. *Ethical Theory and Moral Practice*. https://philpapers.org/rec/BRODNR-7

Bryson, J. J., Diamantis, M. E., & Grant, T. D. (2017). Of, for, and by the people: The legal lacuna of synthetic persons. *Artificial Intelligence and Law*, *25*(3), 273–291. https://doi.org/10.1007/s10506-017-9214-9

Buchanan, A. (1990). Justice as Reciprocity versus Subject-Centered Justice. *Philosophy & Public Affairs*, *19*(3), 227–252. https://www.jstor.org/stable/2265395

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy & Technology*, *33*(4), 685–703. https://doi.org/10.1007/s13347-020-00415-6

Coeckelbergh, M. (2024). *Why AI Undermines Democracy and What To Do About It.* Polity.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology*, *12*(3), 209–221. https://doi.org/10.1007/s10676-010-9235-5

Cohen, J. (1989). Deliberation and Democratic Legitimacy. In D. Matravers & J. Pike (Eds.), *Debates in Contemporary Political Philosophy: An Anthology* (pp. 342–360). Routledge. https://philpapers.org/rec/COHDAD-2

Cohen, J. (2004). Minimalism About Human Rights: The Most We Can Hope For? *Journal of Political Philosophy*, *12*(2), 190–213. https://doi.org/10.1111/j.1467-9760.2004.00197.x

Cohen, J., & Fung, A. (2021). Democracy and the Digital Public Sphere. In L. Bernholz, H. Landemore, & R. Reich (Eds.), *Digital Technology and Democratic Theory* (pp. 23–61). University of Chicago Press. https://doi.org/10.7208/chicago/9780226748603.003.0002

Coleman, D. (2019). Digital Colonialism: The 21st Century Scramble for Africa through the Extraction and Control of User Data and the Limitations of Data Protection Laws. *Michigan Journal of Race and Law*, *24*(2), 417–439. https://doi.org/10.36643/mjrl.24.2.digital

Collins, P. H. (2002). *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge.

Costanza-Chock, S. (2018). Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science*. https://doi.org/10.21428/96c8d426

Couldry, N., & Mejias, U. A. (2019). Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media*, *20*(4), 336–349. https://doi.org/10.1177/1527476418796632

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press.

Crawford, K., & Joler, V. (2019). Anatomy of an AI System. *Virtual Creativity*, *9*(1–2), 117–120. https://doi.org/10.1386/vcr_00008_7

Creel, K., & Hellman, D. (2022). The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Canadian Journal of Philosophy*, *52*(1), 26–43. https://doi.org/10.1017/can.2022.3

Crenshaw, K. (1991). Mapping the Margins: Intersectionality, Identity Politics, and Violence against Women of Color. *Stanford Law Review*, *43*(6), 1241–1299. https://doi.org/10.2307/1229039

Danaher, J. (2016). The Threat of Algocracy: Reality, Resistance and Accommodation. *Philosophy & Technology*, *29*(3), 245–268. https://doi.org/10.1007/s13347-015-0211-1

Danaher, J. (2022a). Freedom in an Age of Algocracy. In S. Vallor (Ed.), *The Oxford Handbook of Philosophy of Technology* (pp. 250-272). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190851187.013.16

Danaher, J. (2022b). Tragic Choices and the Virtue of Techno-Responsibility Gaps. *Philosophy & Technology*, *35*(2), 26. https://doi.org/10.1007/s13347-022-00519-1

Dauvergne, P. (2022). Is artificial intelligence greening global supply chains? Exposing the political economy of environmental costs. *Review of International Political Economy*, *29*(3), 696–718. https://doi.org/10.1080/09692290.2020.1814381

Devine, F., Krasodomski-Jones, A., Miller, C., Lin, S. Y., Cui, J.-W., Marnette, B., & Wilkinson, R. (2023). *Recursive Public: Piloting Connected Democratic Engagement with AI Governance*. Recursive Public. https://vtaiwan-openai-2023.vercel.app/Report_%20Recursive%20Public.pdf

Dewey, J. (1927). *The Public and Its Problems: An Essay in Political Inquiry*. Alan Swallow. https://www.jstor.org/stable/10.5325/j.ctt7v1gh

Dickey, C. (2025). A.I. Griefbots Are Just Our Latest Attempt to Talk to the Dead. *New York Times*. https://www.nytimes.com/2025/07/10/style/00death-spiritualism-talking-to-dead.html

D'Ignazio, C., & Klein, L. F. (2020). *Data Feminism*. MIT Press. https://doi.org/10.7551/mitpress/11805.001.0001

Donaldson, S. & Kymlicka, W. (2011). *Zoopolis: A Political Theory of Animal Rights*. Oxford University Press.

Douek, E. (2021). Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability. *Columbia Law Review*, *121*(3), 759–834. https://www.jstor.org/stable/27007631

Edelman, J. (2023, August 29). Democratic Fine-Tuning. *Lesswrong*. https://www.lesswrong.com/posts/ncb2ycEB3ymNqzs93/democratic-fine-tuning

Eggert, L. (2025). Autonomised harming. *Philosophical Studies*, *182*(1), 1–24. https://doi.org/10.1007/s11098-023-01990-y

Enoch, D., Spectre, L., & Fisher, T. (2012). Statistical Evidence, Sensitivity, and the Legal Value of Knowledge. *Philosophy & Public Affairs*, *40*(3), 197–224. https://doi.org/10.1111/papa.12000

Estlund, D. (2012). Introduction. In Estlund, D. (Ed.), *The Oxford Handbook of Political Philosophy* (pp. 3-20). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376692.013.0000

Ewing, B. (2019). Prior Convictions as Moral Opportunities. *American Journal of Criminal Law*, *46*(2), 283–332. https://heinonline.org/HOL/P?h=hein.journals/ajcl46&i=291

Falbo, A., & LaCroix, T. (2022). Est-ce que Vous Compute? Code-Switching, Cultural Identity, and AI. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14264

Farrell, H., & Han, H. (2025, August 1). AI and Democratic Publics. *Knight First Amendment Institute*. http://knightcolumbia.org/content/ai-and-democratic-publics

Fish, S., Gölz, P., Parkes, D. C., Procaccia, A. D., Rusak, G., Shapira, I., & Wüthrich, M. (2024). Generative Social Choice. *Proceedings of the 25th ACM Conference on Economics and Computation*, 985. https://doi.org/10.1145/3670865.3673547

Fisher, S. A., & Howard, J. W. (2024). Ambiguous Threats: "Death to" Statements and the Moderation of Online Speech Acts. *Journal of Ethics and Social Philosophy*, *28*(2), 208–229. https://heinonline.org/HOL/P?h=hein.journals/jetshy28&i=216

Fisher, S. A., Howard, J. W., & Kira, B. (2024). Moderating Synthetic Content: The Challenge of Generative AI. *Philosophy & Technology*, *37*(4), 133. https://doi.org/10.1007/s13347-024-00818-9

Forestal, J. (2021). *Designing for Democracy: How to Build Community in Digital Environments*. Oxford University Press.

Fraser, N. (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text*, *25/26*, 56–80. https://doi.org/10.2307/466240

Fraser, N. (1995). From Redistribution to Recognition? Dilemmas of Justice in a "Post-Socialist" Age. *New Left Review*, *I/212*, 68–93.

Gabriel, I. (2022). Toward a Theory of Justice for Artificial Intelligence. *Daedalus*, *151*(2), 218–231. https://doi.org/10.1162/daed_a_01911

Gauthier, D. (1987). *Morals by Agreement*. Oxford University Press. https://doi.org/10.1093/0198249926.001.0001

Green, B. (2025). Comparative AI Ethics Between Silicon Valley and the Vatican: Divergent Foundations, Convergent Initiatives, and "How-to" Ideas for Discussing and Developing Technology Ethics. In *Social and Ethical Considerations of AI in East Asia and Beyond* (pp. 51-74). Cham: Springer Nature Switzerland.

Gray, M. L., & Suri, S. (2019). *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Harper Business.

Griffin, J. (2008). *On Human Rights*. Oxford University Press.

Grimmelmann, J. (2014). Speech Engines. *Minnesota Law Review*, *98*(3), 868. https://doi.org/10.24926/265535.1199

Gudiño, J. F., Grandi, U., & Hidalgo, C. (2024). Large language models (LLMs) as agents for augmented democracy. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *382*(2285), 20240100. https://doi.org/10.1098/rsta.2024.0100

Gunkel, D. J. (2018). *Robot Rights*. The MIT Press.

Habermas, J. (1989). *The Structural Transformation of the Public Sphere: An Inquiry into a Category of Bourgeois Society*. The MIT Press.

Habermas, J. (1996). *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Polity.

Han, S. J. (2025). The Question of AI and Democracy: Four Categories of AI Governance. *Philosophy & Technology*, *38*(2), 1–26. https://doi.org/10.1007/s13347-025-00904-6

Heawood, J., & Peter, F. (2023). Political Legitimacy and the News Media: Four Normative Models of the Political Role of the News Media. In C. Fox & J. Saunders (Eds.), *The Routledge Handbook of Philosophy and Media Ethics* (pp. 83–95). Routledge. https://doi.org/10.4324/9781003134749-9

Hedden, B. (2021). On statistical criteria of algorithmic fairness. *Philosophy & Public Affairs*, *49*(2), 209–231. https://doi.org/10.1111/papa.12189

Hellman, D. (2025). Algorithmic Fairness. Edward N. Zalta & Uri Nodelman (eds.), *The Stanford Encyclopedia of Philosophy (Fall 2025 Edition)*. https://plato.stanford.edu/archives/fall2025/entries/algorithmic-fairness/

Hern, A. (2024). Digital recreations of dead people need urgent regulation, AI ethicists say. *The Guardian.* https://www.theguardian.com/technology/article/2024/may/09/digital-recreations-of-dead-people-need-urgent-regulation-ai-ethicists-say

Heyd, D. (2009). A Value or an Obligation? Rawls on Justice to Future Generations. In A. Gosseries & L. H. Meyer (Eds.), *Intergenerational Justice* (pp. 167–188). Oxford University Press.

Heyward, C. (2017). Ethics and Climate Adaptation. In S. M. Gardiner & A. Thompson (Eds.), *The Oxford Handbook of Environmental Ethics* (pp. 474–486). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199941339.013.42

Himmelreich, J. (2023). Against "Democratizing AI." *AI & SOCIETY*, *38*(4), 1333–1346. https://doi.org/10.1007/s00146-021-01357-z

Himmelreich, J. (2023). Should We Automate Democracy? In *The Oxford Handbook of Digital Ethics*, edited by Carissa Véliz. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198857815.013.33

Hindman, M. (2018). *The Internet Trap: How the Digital Economy Builds Monopolies and Undermines Democracy*. Princeton University Press.

Hiskes, R. P. (2009). *The Human Right to a Green Future: Environmental Rights and Intergenerational Justice*. Cambridge University Press.

Hoffmann, A. L. (2019). Where fairness fails: Data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society*, *22*(7), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

Howard, J. W. (2024). The Ethics of Social Media: Why Content Moderation is a Moral Duty. *Journal of Practical Ethics*, *11*(2), 33–52. https://doi.org/10.3998/jpe.6195

Howard, J. W. (2026a). Moderation by Machine. In E. Heinze, N. Alkiviadou, T. Herrenberg, S. Parmar, & I. Tourkochoriti (Eds.), *Oxford Handbook on Hate Speech*. Oxford University Press.

Howard, J. W. (2026b). *Setting Fire to Reason: The Ethics of Free Speech*. Princeton University Press.

Huq, A. Z. (2020). A Right to a Human Decision. *Virginia Law Review*, *106*(3), 611–688. https://virginialawreview.org/articles/right-human-decision/

Johnson, R. L., Pistilli, G., Menédez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). *The Ghost in the Machine has an American accent: Value conflict in GPT-3* (No. arXiv:2203.07785). arXiv. https://doi.org/10.48550/arXiv.2203.07785

Jørgensen, R. (2022). Algorithms and the Individual in Criminal Law. *Canadian Journal of Philosophy*, *52*(1), 61–77. https://doi.org/10.1017/can.2021.28

Jørgensen, R. (forthcoming). Protect, Serve, Predict? *Politics, Philosophy and Economics*. https://doi.org/10.1177/1470594x251379074

Jungherr, A., & Rauchfleisch, A. (2025). Artificial Intelligence in deliberation: The AI penalty and the emergence of a new deliberative divide. *Government Information Quarterly*, *42*(4), 102079. https://doi.org/10.1016/j.giq.2025.102079

Jungherr, A. (2023). Artificial Intelligence and Democracy: A Conceptual Framework. *Social Media + Society, 9(3)*. https://doi.org/10.1177/20563051231186353

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 169–169. https://doi.org/10.1038/d41586-020-02003-2

Kasirzadeh, A. (2022). Algorithmic Fairness and Structural Injustice: Insights from Feminist Political Philosophy. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 349–356. https://doi.org/10.1145/3514094.3534188

Kasirzadeh, A., & Gabriel, I. (2023). In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy & Technology*, *36*(2), 27. https://doi.org/10.1007/s13347-023-00606-x

Kay, J., Kasirzadeh, A., & Mohamed, S. (2025). Epistemic Injustice in Generative AI. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 684–697). AAAI Press. https://dl.acm.org/doi/10.5555/3716662.3716722

Kelly, E. I. (2018). *The Limits of Blame: Rethinking Punishment and Responsibility*. Harvard University Press.

Kiener, M. (2022). Can we Bridge AI's responsibility gap at Will? *Ethical Theory and Moral Practice*, *25*(4), 575–593. https://doi.org/10.1007/s10677-022-10313-9

Kira, B., & Howard, J. W. (forthcoming). Reduce or Remove: Demotion, Content Moderation, and Human Rights. *Law & Philosophy*.

Kirby, N. (2022). Institutional Integrity: Its Meaning and Value. *Ethical Theory and Moral Practice*, *25*(5), 809–834. https://doi.org/10.1007/s10677-022-10330-8

Kleinberg, J., Mullainathan, S., and Raghavan, M. (2016). Inherent Trade-Offs in the Fair Determination of Risk Scores. doi:10.48550/ARXIV.1609.05807

Kong, Y. (2026). What Is the Point of Equality in Machine Learning Fairness? Beyond Equality of Opportunity. *ACM J. Responsib. Comput.*, *3*(1), 4:1-4:25. https://doi.org/10.1145/3766539

Konya, A., Qiu, Y. L., Varga, M. P., & Ovadya, A. (2022). *Elicitation Inference Optimization for Multi-Principal-Agent Alignment*. https://openreview.net/forum?id=dz8i-yzXeVg

Konya, A., Schirch, L., Irwin, C., & Ovadya, A. (2023). *Democratic Policy Development using Collective Dialogues and AI* (No. arXiv:2311.02242). arXiv. https://doi.org/10.48550/arXiv.2311.02242

Kreps, S. & Kriner, D. (2023). How AI Threatens Democracy. *Journal of Democracy, 34(4),* 122–31.

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, *60*(4), 3–26. https://doi.org/10.1177/0306396818823172

Landemore, H. (2024). Can Artificial Intelligence Bring Deliberation to the Masses? In R. Chang & A. Srinivasan (Eds.), *Conversations in Philosophy, Law, and Politics* (pp. 36–69). Oxford University Press. https://doi.org/10.1093/oso/9780198864523.003.0003

Lazar, S. (2024). Legitimacy, Authority, and Democratic Duties of Explanation. In D. Sobel & S. Wall (Eds.), *Oxford Studies in Political Philosophy* (Vol. 10, pp. 28–56). Oxford University Press. https://doi.org/10.1093/oso/9780198909460.003.0002

Lazar, S. (2025). Governing the Algorithmic City. *Philosophy & Public Affairs*, *53*(2), 102–168. https://doi.org/10.1111/papa.12279

Lazar, S. (2026). *Connected by Code: How AI Structures and Governs the Ways We Relate*. Oxford University Press.

Lazar, S., & Manuali, L. (2026). *Using LLMs to Enhance Democracy Minds & Machines 36*, 12. https://doi.org/10.1007/s11023-026-09767-y

Lazar, S., & Stone, J. (2024). On the site of predictive justice. *Noûs*, *58*(3), 730–754. https://doi.org/10.1111/nous.12477

Le Bui, M., & Noble, S. U. (2020). We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings, and Ethics of Fairness. In M. D. Dubber, F. Pasquale, & S.

Das (Eds.), *The Oxford Handbook of Ethics of AI* (p. 0). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190067397.013.9

Lechterman, T. M. (2022). The Concept of Accountability in AI Ethics and Governance. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds.), *The Oxford Handbook of AI Governance* (pp. 164–182). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.10

Lechterman, T. M. (2024). The Perfect Politician. In D. Edmonds (Ed.), *AI Morality* (pp. 53–63). Oxford University Press. https://doi.org/10.1093/oso/9780198876434.003.0006

Leopold, D. and Stears, M. (2008). (Eds.) *Political Theory: Methods and Approaches.* Oxford University Press. https://doi.org/10.1093/oso/9780199230082.001.0001

Liao, S. M. (2020). The Moral Status and Rights of Artificial Intelligence. In Liao, S. M. (Ed.) *Ethics of Artificial Intelligence* (pp. 480–504). Oxford University Press. https://doi.org/10.1093/oso/9780190905033.003.0018

Liao, S. M. (2010). The Basis of Human Moral Status. *Journal of Moral Philosophy 7(2),* 159-179. https://doi.org/10.1163/174552409X12567397529106

Lin, T.-A. (2024). "Democratizing AI" and the Concern of Algorithmic Injustice. *Philosophy & Technology*, *37*(3), 103. https://doi.org/10.1007/s13347-024-00792-2

Lin, T.-A., & Chen, P.-H. C. (2022). Artificial Intelligence in a Structurally Unjust Society. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14191

Lovett, A., & Zuehl, J. (2022). The Possibility of Democratic Autonomy. *Philosophy & Public Affairs*, *50*(4), 467–498. https://doi.org/10.1111/papa.12223

Luccioni, A. S., Strubell, E., & Crawford, K. (2025, June). From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate. In *Proceedings of the 2025 ACM conference on fairness, accountability, and transparency* (pp. 76-88).

Maldonado-Torres, N. (2007). On the Coloniality of Being: Contributions to the Development of a Concept. *Cultural Studies*, *21*(2–3), 240–270. https://doi.org/10.1080/09502380601162548

Matthias, A. (2004). The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata. *Ethics and Information Technology*, *6*(3), 175–183. https://doi.org/10.1007/s10676-004-3422-1

McClure, E., & Wald, B. (2022). Algorithmic Microaggressions. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14276

McKinney, S. (2024). Integrating Artificial Intelligence into Citizens' Assemblies: Benefits, Concerns and Future Pathways. *Journal of Deliberative Democracy*, *20*(1). https://doi.org/10.16997/jdd.1556

Mendoza, G. B. (2023, October 31). Can we use AI to enrich democratic consultations? *Rappler*. https://www.rappler.com/technology/features/generative-ai-use-enriching-democratic-consultations/

Meyer, L. (2021). Intergenerational Justice. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/justice-intergenerational/

Mignolo, W. D. (2012). *Local Histories/Global Designs: Coloniality, Subaltern Knowledges, and Border Thinking*. Princeton University Press.

Mikhaylovskaya, A. (2024). Enhancing Deliberation with Digital Democratic Innovations. *Philosophy & Technology*, *37*(1), 3. https://doi.org/10.1007/s13347-023-00692-x

Milano, S., & Prunkl, C. (2025). Algorithmic profiling as a source of hermeneutical injustice. *Philosophical Studies*, *182*(1), 185–203. https://doi.org/10.1007/s11098-023-02095-2

Miller, D. (2022). Doing Political Philosophy. In Butt, D., Fine, S., and Stemplowska, Z. (Eds.), *Political Philosophy, Here and Now: Essays in Honour of David Miller* (pp. 232–248). https://doi.org/10.1093/oso/9780198807834.003.0015

Miller, E. (2021). Amplified Speech. *Cardozo Law Review*, *43*(1). https://papers.ssrn.com/abstract=4014849

Mintz-Woo, K. (2021). A Philosopher's Guide to Discounting. In M. Budolfson, T. McPherson, & D. Plunkett (Eds.), *Philosophy and Climate Change* (pp. 90–110). Oxford University Press. https://doi.org/10.1093/oso/9780198796282.003.0005

Mohamed, S., Png, M.-T., & Isaac, W. (2020). Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy & Technology*, *33*(4), 659–684. https://doi.org/10.1007/s13347-020-00405-8

Mollema, W. J. T. (2025). A Taxonomy of Epistemic Injustice in the Context of AI and the Case for Generative Hermeneutical Erasure. *AI and Ethics*, 5(5), 5535-5555. https://doi.org/10.1007/s43681-025-00801-w

Moosavi, L. (2020). The decolonial bandwagon and the dangers of intellectual decolonisation. *International Review of Sociology*, *30*(2), 332–354. https://doi.org/10.1080/03906701.2020.1776919

Mouffe, C. (1993). *The Return of the Political*. Verso Books.

Mouton, M., & Burns, R. (2021). (Digital) neo-colonialism in the smart city. *Regional Studies*, *55*(12), 1890–1901. https://doi.org/10.1080/00343404.2021.1915974

Muldoon, J., & Wu, B. A. (2023). Artificial Intelligence in the Colonial Matrix of Power. *Philosophy & Technology*, *36*(4), 80. https://doi.org/10.1007/s13347-023-00687-8

Narayanan, A. (2018). *Translation Tutorial: 21 Fairness Definitions and Their Politics* [Tutorial]. Retrieved October 3, 2023, from https://facctconference.org/static/tutorials/narayanan-21defs18.pdf

Ngcamu, B. S. (2023). Climate change effects on vulnerable populations in the Global South: A systematic review. *Natural Hazards*, *118*(2), 977–991. https://doi.org/10.1007/s11069-023-06070-2

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Noorman, M., & Swierstra, T. (2023). Democratizing AI from a Sociotechnical Perspective. *Minds and Machines*, *33*(4), 563–586. https://doi.org/10.1007/s11023-023-09651-z

Nothias, T. (2025). An intellectual history of digital colonialism. *Journal of Communication*, *75*(5), 385–397. https://doi.org/10.1093/joc/jqaf003

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

O'Neil, C. (2017). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

Oyedemi, T. D. (2019). Global Digital Capitalism: Mark Zuckerberg in Lagos and the Political Economy of Facebook in Africa. *International Journal of Communication*, *13*, 17–17. https://ijoc.org/index.php/ijoc/article/view/8774

Oyedemi, T. D. (2021). Digital coloniality and 'Next Billion Users': The political economy of Google Station in Nigeria. *Information, Communication & Society*, *24*(3), 329–343. https://doi.org/10.1080/1369118X.2020.1804982

Palomo Hernández, N. (2025). Towards Automating Deliberation? The Idea of Deliberative Democracy Embedded in Google's Habermas Machine. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *8*(2), 1951–1960. https://doi.org/10.1609/aies.v8i2.36687

Quijano, A. (2007). Coloniality and Modernity/Rationality. *Cultural Studies*, *21*(2–3), 168–178. https://doi.org/10.1080/09502380601164353

Rawls, J. (1971). *A Theory of Justice*. Harvard University Press.

Rawls, J. (1999). *The Law of Peoples: With "The Idea of Public Reason Revisited."* Harvard University Press. https://doi.org/10.2307/j.ctv1pncngc

Rawls, J. (2005). *Political Liberalism* (Expanded edition). Columbia University Press.

Reich, R., Sahami, M., & Weinstein, J. M. (2021). *System Error: Where Big Tech Went Wrong and How We Can Reboot*. Harper.

Renzo, M. (2025). Manipulation and Practical Agency. *Free & Equal: A Journal of Ethics and Public Affairs*, *1*(1). https://doi.org/10.16995/fe.17722

Ricaurte, P. (2019). Data Epistemologies, The Coloniality of Power, and Resistance. *Television & New Media*, *20*(4). https://doi.org/10.1177/1527476419831640

Risse, M. (2023). *Political Theory of the Digital Age: Where Artificial Intelligence Might Take Us*. Cambridge University Press. https://doi.org/10.1017/9781009255189

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press.

Robillard, M. (2018). No Such Thing as Killer Robots. *Journal of Applied Philosophy*, *35*(4), 705–717. https://doi.org/10.1111/japp.12274

Ryberg, J. (2020). Risk-Based Sentencing and Predictive Accuracy. *Ethical Theory and Moral Practice*, *23*(1), 251–263. https://www.jstor.org/stable/45294785

Ryberg, J. (2024). Criminal Justice and Artificial Intelligence: How Should We Assess the Performance of Sentencing Algorithms. *Philosophy & Technology*, *37*(1). https://doi.org/10.1007/s13347-024-00694-3

Ryberg, J., & Roberts, J. V. (2022). *Sentencing and Artificial Intelligence*. Oxford University Press.

Sagan, C. (1973). *The Cosmic Connection*. Anchor Press / Doubleday.

Sahebi, S., & Formosa, P. (2024). Artificial Intelligence (AI) and Global Justice. *Minds and Machines*, *35*(1), 4. https://doi.org/10.1007/s11023-024-09708-7

Santoni de Sio, F., & Mecacci, G. (2021). Four Responsibility Gaps with Artificial Intelligence: Why they Matter and How to Address them. *Philosophy & Technology*, *34*(4), 1057–1084. https://doi.org/10.1007/s13347-021-00450-x

Scheier, B. and Sanders, N. E. (2025). *Rewiring Democracy: How AI Will Transform Our Politics, Government, and Citizenship*. MIT Press.

Schulz, W. F., & Ramen, S. (2020). *The Coming Good Society: Why New Realities Demand New Rights*. Harvard University Press.

Seger, E., Ovadya, A., Siddarth, D., Garfinkel, B., & Dafoe, A. (2023). Democratising AI: Multiple Meanings, Goals, and Methods. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 715–722. https://doi.org/10.1145/3600211.3604693

Shany, Y. (2023). Digital Rights and the Outer Limits of International Human Rights Law. *German Law Journal*, *24*(3), 461–472. https://doi.org/10.1017/glj.2023.35

Shelby, T. (2022). *The Idea of Prison Abolition*. Princeton University Press. https://doi.org/10.2307/j.ctv2ks6t26

Sidgwick, H. (1907). *The Methods of Ethics*. Macmillan.

Simmons, A. J. (2015). Human Rights, Natural Rights, and Human Dignity. In R. Cruft, S. M. Liao, & M. Renzo (Eds.), *Philosophical Foundations of Human Rights* (pp. 138–152). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199688623.003.0007

Simpson, T. W., & Müller, V. C. (2016). Just War and Robots' Killings. *The Philosophical Quarterly*, *66*(263), 302–322. https://doi.org/10.1093/pq/pqv075

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Stemplowska, Z. & Swift, A. (2012). Ideal and Nonideal Theory. In Estlund, E. (Ed.), T*he Oxford Handbook of Political Philosophy* (pp.373–390.) Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376692.013.0020

Stewart, H., Cichocki, E., & McLeod, C. (2022). A Perfect Storm for Epistemic Injustice: Algorithmic Targeting and Sorting on Social Media. *Feminist Philosophy Quarterly*, *8*(3/4). https://doi.org/10.5206/fpq/2022.3/4.14291

Sultana, F. (2022). The unbearable heaviness of climate coloniality. *Political Geography*, *99*, 102638. https://doi.org/10.1016/j.polgeo.2022.102638

Sunstein, C. R. (2017). *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.

Ta, R., & Lee, N. (2023). How Language Gaps Constrain Generative AI Development. *International Journal of Comparative Studies in International Relations and Development*, *9*, 48–52. https://doi.org/10.48028/iiprds/ijcsird/ijcsird.v9.i1.03

Tadros, V. (2011). Punishment and Duty. In V. Tadros (Ed.), *The Ends of Harm: The Moral Foundations of Criminal Law* (pp. 264–292). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199554423.003.0012

Táíwò, O. (2022). *Against Decolonisation: Taking African Agency Seriously*. Hurst.

Talisse, R. B. (2020). *Overdoing Democracy: Why We Must Put Politics in its Place*. Oxford University Press. https://doi.org/10.1093/oso/9780190924195.001.0001

Taylor, I. (2021). Who Is Responsible for Killer Robots? Autonomous Weapons, Group Agency, and the Military-Industrial Complex. *Journal of Applied Philosophy*, *38*(2), 320–334. https://doi.org/10.1111/japp.12469

Taylor, I. (2025). Representative Robots: Can AI Systems Act in Our Name? *Free & Equal: A Journal of Ethics and Public Affairs*, *1*(1), Article 1. https://doi.org/10.16995/fe.18626

Tessler, M. H., Bakker, M. A., Jarrett, D., Sheahan, H., Chadwick, M. J., Koster, R., Evans, G., Campbell-Gillingham, L., Collins, T., Parkes, D. C., Botvinick, M., & Summerfield, C. (2024). AI can help humans find common ground in democratic deliberation. *Science*, *386*(6719), eadq2852. https://doi.org/10.1126/science.adq2852

Tigard, D. W. (2021). There Is No Techno-Responsibility Gap. *Philosophy & Technology*, *34*(3), 589–607. https://doi.org/10.1007/s13347-020-00414-7

Tubaro, P., Casilli, A. A., & Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data & Society*, *7*(1), 2053951720919776. https://doi.org/10.1177/2053951720919776

Valentini, L. (2012). Ideal vs. Non-ideal Theory: A Conceptual Map. *Philosophy Compass, 7(9),* 654-664. https://doi.org/10.1111/j.1747-9991.2012.00500.x

von Schirach, F. (2021). *Jeder Mensch*. Luchterhand Literaturverlag.

Vredenburgh, K. (2022). The Right to Explanation. *Journal of Political Philosophy*, *30*(2), 209–229.

Wong, P. (2025). Confucian 'Trustworthy AI': Diversifying a keyword in the ethics of AI and governance. In *Social and Ethical Considerations of AI in East Asia and Beyond* (pp. 3-14). Cham: Springer Nature Switzerland.

Worth, T. (2024). Ready or not, the digital afterlife is here. *Nature.* https://www.nature.com/articles/d41586-025-02940-w

Yang, Z. (2024). Deepfakes of your dead loved ones are a booming Chinese business. *MIT Technology Review.* https://www.technologyreview.com/2024/05/07/1092116/deepfakes-dead-chinese-business-grief/

Young, I. M. (2002). *Inclusion and Democracy* (1st edition). Oxford University Press.

Young, I. M. (2010). *Responsibility for Justice*. Oxford University Press.

Zimmermann, A., & Lee-Stronach, C. (2022). Proceed with Caution. *Canadian Journal of Philosophy*, *52*(1), 6–25. https://doi.org/10.1017/can.2021.17