Philosophy of AI
Vol. 2, 2026, 224-243
DOI: 10.18716/ojs/phai/2026.11709
©2026 by the author(s)

THE DARK SIDE OF TECHNOMORAL LIFE: TOWARDS A VICE FRAMEWORK FOR ARTIFICIAL INTELLIGENCE

Jennifer Chubb[1]*, *Ian Kidd*[2], *Joshua Forstenzer*[3]

[1] University of York, Department of Sociology, York, United Kingdom
[2] University of Nottingham, Department of Philosophy, Nottingham, United Kingdom
[3] University of Sheffield, Department of Philosophy, Sheffield, United Kingdom

*jennifer.chubb@york.ac.uk

Abstract Contemporary theorising about the ethical dimensions of Artificial Intelligence (AI) indicates at least two problematic tendencies. The first is a default focus on a narrow range of moral concepts, such as autonomy, justice, and fairness, to the occlusion of a wider range of concepts. The second is a 'bright siding' focus on concepts that track the positive dimensions of moral practice and experience, such as virtue, flourishing, solutionism, moral progress, and the good life. Taken together, the consequence is the production of large-scale moral frameworks that systematically fail to accommodate negative dimensions of moral life (Jobin et al., 2019). Two main symptoms are a one-sided focus on virtues and excellences at the expense of vices and failings, and an affirmation of the edifying potential of new technologies without due appreciation of their corruptive potential (Coeckelbergh, 2020, p. 59). Drawing on the example of higher education research and academic life, we suggest that a more comprehensive assessment of the moral possibilities of new technologies requires a greater inclusion of a negatively toned, vice-informed range of moral concepts, narratives, and sensibilities. We elucidate these points by providing examples of techno-vice in everyday life. This helps to clarify the concept of epistemic vices and the phenomenon of corruption, which, we argue, inform the concept of technomoral vices.

Keywords Technomoral Vice; Ethics of AI; Technology; Corruption

1. Introduction

Artificial Intelligence (AI) is no longer a looming force on the horizon; it is already an integral part of everyday life. From education and healthcare to environmental policy and entertainment, AI shapes how individuals learn, communicate, and make decisions. Its ubiquity and expansion raise well-rehearsed concerns about AI's propensity to marginalise (O'Neil, 2017), discriminate (Benjamin, 2023), and amplify social inequalities (Zimmermann, 2021). Advances in large language models (LLMs), capable of convincingly communicating complex information, intensify these fears, as do AI tools capable of altering audio-visual inputs in ways that undermine public trust and democratic processes (Matthews, 2022).

Public understanding of AI is shaped not only by technical developments but also by powerful cultural AI narratives (Cave & Dihal, 2020). Fictional representations of AI often oscillate between utopian promises and dystopian anxieties – the light and the dark side. Consider HAL's murderous tendencies in *2001: A Space Odyssey*, the bleak future presented in *The Matrix* trilogy, or the machine awakening in *Detroit: Become Human*. These stories foreground ethical dilemmas and tensions between human and artificial agents, often symbolising widespread fears of moral decay and human subjugation. Some portrayals are more hopeful, featuring virtuous AI agents – most notably, Data, the lovable artificial lifeform, in *Star Trek: The Next Generation* – but overwhelmingly, the dominant tropes present polarised narratives that signal malevolence over beneficence, or vice over virtue (Cave et al., 2020). Usually, these stories present AI as at least a potential external threat to humans: the AI's emerging selfhood and power present dangers to humanity that we may fail to take stock of until it is too late (the *Terminator* is probably the starkest example in this genre). As far-fetched as these narratives might seem, many of these stories shape public imagination, and through it, public policy (S. Dillon & Craig, 2021), as well as having a considerable effect on framing the public understanding of science (Fast & Horvitz, 2017). In a sense, the extreme dark sides of these technologies are amplified through a process that lures individuals away from positive or rational behaviour (virtues) toward harmful outcomes or irrational behaviour (vices). What was once consigned to the realms of science fiction is in many ways reflected in how we experience our everyday lives (Beukes et al., 2017). For instance, the Internet and technological developments, including AI, have played "a pivotal role in self-radicalization" through technological seduction (Alfano et al., 2018).

As a result, the scholarly field of AI ethics has proliferated, producing numerous frameworks and guidelines grounded in concepts like justice, autonomy, transparency, and fairness (Buhmann & Fieseler, 2021; Floridi et al., 2021). These frameworks often reflect a liberal moral orientation and a 'bright siding' focus on positive solutions and moral aspirations, such as well-being and "flourishing". However, we argue that this tendency to emphasise the promise of cultivating virtue and overlook the risks of entrenching vice presents a risk of moral and theoretical myopia (Hagendorff, 2022; Jobin et al., 2019). As some scholars argue, the "ethicisation" of AI can veer into tokenism or "ethics washing" (Dubber et al., 2020). Despite an increased focus on AI safety in both academic (Habli et al., 2020) and political circles, what is frequently overlooked, then, are the real and potential harms of using AI, especially the ways in which AI may influence us and, in particular, our habits, character, and moral development (Chubb, Missaoui, et al., 2022). Moreover, these are framed more in general ethical terms of harms (which lends itself to solutionist responses), rather than explicitly in terms of vice, perhaps intentionally by those in power (Chubb & Tutton, 2026).

Virtue ethicists have paid relatively little attention to vice in the context of AI. Despite some references (Brey, 2022; Coeckelbergh, 2021; Hagendorff, 2022; Vallor, 2016), vice, we observe, is often given only secondary importance[1], or rather abstract consideration, or is limited to rather narrow case studies, despite the prevalence of vicious conduct and vicious persons in human life (Battaly, 2018; Fritts & Cabrera, 2022; Kidd, 2018). To a degree, this repeats a broader tendency, within philosophical character ethics, to focus on virtues and the good life. Indeed, what we just called philosophical character ethics is more usually called 'virtue ethics', a term that signals the conceptual prioritisation of the good sides of moral

1.Indeed, the *Oxford Handbook of Philosophy of Technology*, a rather hefty tome, gives little attention to vice, with only the following cursory mention: "Vallor claims that particular technologies tend to promote the development of certain virtues and vices in users: virtues such as honesty and empathy, and vices such as dishonesty and carelessness." (Brey, 2022, p. 408).

character. When considering AI, we contend that this imbalance must urgently be redressed in the philosophy and sociology of AI.

Although there is a tendency to oversell the dislocation caused by the rate of change brought on by the emergence of any one technological innovation, a reasonably cautious attitude to the development of AI would demand that we take the transformational potential of AI more seriously than most other technological innovations. This is partly because of its breadth of application, partly because of its generative capacities, and partly because of its self-reinforcing tendencies. Taken together, these demand an urgent critical consideration of the key elements on which many other consequences ultimately rest.

We contend that a full ethical appraisal of the promise and peril of AI must include not only its capacity to enable our flourishing, or to wrong us (if it is a moral agent, then it can do the wrong thing), but also its potential to corrupt human agents. We take all three of these moral dimensions to be of great importance, but we are concerned that the risk of corruption is paid comparatively little attention. By 'corruption,' we mean specifically the manner in which certain social and material arrangements cause damage to the moral and epistemic character of individual agents (Kidd, 2021, 2022). Our project thus includes examining technomoral systems and cultures that are *vice-enabling* and *virtue-eroding*. Given AI's deep and growing influence, a precautionary approach to these moral risks is essential. To be clear, if the precautionary principle applies in general instances of technological development, it should apply even more strongly in the case of AI because its breadth of application, its generative capacities, its rate of change, and its self-reinforcing tendencies all point to a major risk of moral corruption for the public at large.

This article responds to that need by integrating perspectives from virtue ethics, especially those advanced by Shannon Vallor (2016), while pressing further into the neglected domain of *technomoral vices*. We argue that, alongside virtue, a robust ethical framework must also explore vice: not merely as the absence of virtue, but as a set of moral and epistemic failings and corruptive tendencies shaped by our interaction with technology. This article advocates for developing a wider and richer negatively toned range of moral concepts, narratives, and sensibilities about AI beyond the dominant binary opposing 'techno-optimists' to 'techno-pessimists' (Sacasas, 2011; Smith & Marx, 1994). In conducting moral inquiry, we think that we are well served by considering and understanding the nature and aetiology of relevant vices prior to engaging in any ameliorative pursuit. We contend that this is best done by theorising a vice framework, applicable in the philosophy of AI. We thus conceptualise technomoral vice and introduce a vice framework for AI, aiming to expand and rebalance the moral vocabulary applied to emerging technologies.

To motivate this discussion, we present the example of the use of AI in higher education (HE) research and knowledge production, showing how AI systems can cultivate vice. We first introduce the notion of vice and its defining characteristics before using the example to then foreground discussion on the corrupting potential of technomoral systems.

2. "The dark side"

The term 'vice' refers, in its most common sense, to a negative character trait. Cruelty, selfishness, dishonesty, recklessness, and dogmatism are standard examples of vices, relatively enduring character traits or dispositions. Historically, the study of vices has been a secondary accompaniment to the study of virtue, defined as relatively enduring and cross-situationally stable positive character traits, such as kindness, generosity, courage, and honesty. Over the last decade, philosophical attention has slowly expanded to include a sustained and specific focus on moral and epistemic vices (Battaly, 2015, 2019a; Kidd et al., 2020).

In *Vice Epistemology*, philosophers Kidd, Battaly, and Cassam provide accounts of epistemic vice, arguing that "a defining feature of the vices of the mind is that they make us bad thinkers – some of these are irksome, others dangerous" (Kidd et al., 2020, p. 1). In this way, a vice is a specific and important type of character failing. A distinction should be drawn between vices and failings, where the latter is the broader category. While all vices are failings, not all failings are vices. If I lack the skills needed to fulfil my professional duties, that is a failing, but it is not a negative character trait (Matthews, 2025). Something further is required to go from a mere failing to a vice: vices are failings of character, construed as a relatively temporally persisting and cross-situationally stable set of dispositions with affective, cognitive, and behavioural qualities that is integral to our sense of someone's identity.

In the philosophical literature, two positions are distinguished on the question of what vices are. A vice-monist maintains that vices are one kind of thing, the standard answer being that vices are character traits of individuals. A vice-pluralist, by contrast, maintains that vices can be different kinds of things. Recent candidates include vicious attitudes, vicious postures, and vicious thinking-styles (Cassam, 2019; Tanesini, 2021). In most philosophical character theory, the default position, since Aristotle, has been the monist position that vices are failings of individual character.[2] Another common conviction, also relevant to our subject, concerns the claim that vices are simply the 'opposite' of a virtue (that cowardice, say, is the absence of courage). Crerar (2018) labelled this the *inversion thesis* and demonstrated that it cannot account for, among other things, the sheer variety and psychological distinctiveness of vices. The virtue of humility, say, is 'opposed' by a range of vices – arrogance, haughtiness, *superbia*, hubris – complicating the claim that there is an easily-specifiable opposite (for other criticisms, see R. S. Dillon, 2012).

A further distinction is drawn between accounts of why vices are bad. One position is *vice-consequentialism*: vices are bad because they systematically tend to (i) produce bad effects and/or (ii) fail to produce good effects. Cassam, for instance, defends a form of this position called *obstructivism*: epistemic vices, such as dogmatism, closed-mindedness, and intellectual laziness, systematically obstruct the gaining, keeping, and sharing of knowledge (Cassam, 2016, 2019). Battaly calls these *effects-vices*, and Kidd adds that those which systematically tend to cause bad effects are *productive effects-vices* and those which systematically tend to obstruct good effects are *passive effects-vices* (Battaly, 2015; Kidd, 2021).

The other normative account is *vice-motivationalism*: vices are bad due to motivational defects. Tanesini (2021) argues that epistemic arrogance, for example, expresses a desire to assert superiority over other epistemic agents and signals an intolerance of criticism. Some epistemically vicious motivations include the desire to undermine the epistemic agency of others, to propagate false beliefs that serve one's interests, to lower the collective epistemic standards of one's community, and to create social conditions hostile to epistemic values. The ultimate motivation common to all epistemic vices is arguably a desire to resist what Zagzebski (2010) calls 'cognitive contact with reality.'[3] Crerar (2018) has distinguished two varieties of vice-motivationalism. Some vices register the presence of bad motivations and desires, call these *presence-vices.* The vice of cruelty, for instance, is bad because a cruel person desires to cause needless or undeserved suffering. Other vices register the absence of good motivations and desires, called *absence-vices* (Crerar, 2018; Tanesini, 2021). The vice of insouciance, likewise, is bad because it registers a failure to care, morally or epistemically, about

2.To be clear, some habitual failings do not, on our account, reach the standard of being a vice, because their aetiology points to merely circumstantial features of the environment in which the agent finds themselves.
3.Our account of epistemic vice hopes to remain agnostic about the realist/anti-realist debate.

persons or things about which one ought to care. One could, of course, also combine these consequentialist and motivationalist positions.

Contemporary philosophical vice theory, therefore, offers a triple pluralism, according to which there are *many vices* that can be *different kinds of things* that are appraised as *bad* for a range of reasons. We endorse this pluralism: given the variety of our moral and epistemic vices, we see no reason to stipulate monism about what vices are (or must be) or any reductive account of what is bad about them. There is also a third distinction, important to our account, namely the question of what can be *vice-bearers*. What kinds of things can have or possess vices? There are three main positions. In descending order of acceptance, these are *individual agents*, *collective agents*, and *institutions*. Everyone who accepts the reality of vices accepts that individual agents are vice-bearers, and more theorists defend the related idea that collective agents can possess vices, too (Astola, 2021; Baddorf & McKay, 2024; Byerly & Byerly, 2016; Lahroodi, 2007). Fewer theorists defend the idea that there can be institutional vices, insofar as institutions can be modelled as agents with stable dispositions to certain patterns of action and certain motivations (Fricker, 2010, 2020).

In our discussion, we endorse both ontological and normative pluralism about vices. There are many *vices*, which can be different *kinds of things* that are *bad* for different kinds of reasons. In the next section, we argue that technomoral systems can be *corrupting*, in the sense that interactions with them can be virtue-extirpating and vice-enabling, in ways which, unless they are resisted, can cause systematic damage to one's moral and epistemic character. In the case of AI, machine learning algorithms can develop patterned epistemic behaviours that mimic these obstructive traits. For instance, an AI system trained on biased or narrow data sets may repeatedly prioritise certain types of information or sources, leading to entrenched epistemic blind spots. Just as in humans, these algorithmic habits can become self-reinforcing, producing outputs that further entrench the system's epistemic limitations (Cheong et al., 2024). The use of AI in higher education research provides one such powerful example of how these technomoral systems can be corrupting.

3. AI and the corruption of academic research

Researchers are increasingly adopting AI tools to support their research, from literature reviews and data analysis to writing and summarisation (Chubb, Cowling, et al., 2022). As universities adopt AI tools for literature reviews, summarisation, and writing, scholars risk developing an over-reliance, an uncritical acceptance of the products of these technologies, and a growing detachment from and unfamiliarity with primary sources. Chubb, Cowling, et al. (2022) note that AI's promise of efficiency may, in fact, erode critical engagement and creativity. Whilst there is a paucity of literature focused specifically on HE research as a site of technomoral epistemic corruption, the effects of AI and algorithmic processes in education have long been documented as problematic. For instance, Uszkai claimed of such tools in education that: "algorithmic epistemic outsourcing raises all the relevant epistemic red flags: it can be unreliable, lead to gullibility, irrationality, dependence, and a decrease of epistemic autonomy" (Uszkai, 2017, p. 21). AI's role in research may compromise practices like critical close reading, interpretive diligence, and constructive disciplinary debate. Scholars may rely on AI-generated overviews without engaging primary sources, leading to a decline in intellectual rigour. For instance, if a researcher regularly uses generative AI to complete tasks quickly, they may gradually come to rely more and more on the AI to produce most of the work, while virtues such as curiosity, perseverance, and honesty may begin to erode. This kind of offloading (Gerlich, 2025; Lu et al., 2020) can be understood as passive corruption or the gradual cultivation of epistemic vices. In this case, these are traits that distort

or impede the pursuit of knowledge (Kidd et al., 2020). Going further still, by drawing on Cassam's (2019) concept of epistemic vice, we can identify negligence, overconfidence, and intellectual laziness not only in users but in system design and institutional uptake. These examples reveal the diverse ways AI can encourage epistemic or moral failings, from individual habits (impatience, arrogance) to collective phenomena (polarisation, bias). Yet, mainstream techno-ethics still largely sidelines detailed consideration of vice. To properly understand the moral stakes of AI, we must go beyond virtue to systematically examine how technologies shape, intensify, or normalise vice.

Kidd (2021) argues that epistemic corruption need not involve overt deception or malice. It often develops through habituation to practices that erode intellectual virtues such as honesty, diligence, and humility, resulting in a kind of epistemic laziness, where speed or ease comes to be overvalued at the expense of other values, such as accuracy and high-quality double-checking. There is emerging evidence indicating the duality of AI use in education. For instance, on the one hand, AI brings a promise of efficiency and productivity or critical thinking and some educators building tools that engage students in Socratic questioning vital for critical thinking (Kosmyna et al., 2025), its widespread use may subtly erode key epistemic virtues central to academic inquiry (Shaw, 2025) and these vices are less explored. Is there, then, a threat of technomoral corruption in academia arising from the use of AI in research?

Such uses can enable the cultivation of vices such as the following:

- intellectual laziness – deferring critical analysis to machine outputs;
- negligence – skipping due diligence in evaluating sources or arguments;
- overconfidence – accepting summaries without interrogation, assuming machine neutrality;
- unimaginativeness – allowing AI to narrow the range of questions asked or problems posed – implications for creativity (Manditereza & Chamboko-Mpotaringa, 2024).

As AI use in research becomes more prevalent, what is described here are not accidental misuses of technology but likely results of its use; in other words, they are systemically encouraged dispositions. Indeed, the perceived utility of AI technology might depend on one's exercise of these vices.

Take, for instance, the very structure of academic labour marked by metrics, deadlines, and performance pressures, which incentivises reliance on fast, streamlined, AI-assisted outputs (Chubb, Cowling, et al., 2022). This convergence of neoliberal HE and technological temptation creates fertile ground for moral and epistemic corruption (Forstenzer, 2018; Kidd et al., 2021). The danger is not just to individual scholars but to the epistemic character of research communities and especially to particular groups who may be disadvantaged, such as early-career researchers honing their craft in conducting literature reviews or humanities scholars (Chubb, Cowling, et al., 2022). Indeed, if academic cultures valorise speed, performativity, output volume, and citation counts (Beer, 2016; Wilsdon, 2015) over responsible, critical engagement and creativity, then AI systems that serve those values will accelerate the erosion of academic virtues such as truthfulness (Kidd et al., 2021). Further, there is a risk that overconfidence in the use of the tool reinforces beliefs because of a sycophantic AI (Sharma et al., 2023).

In this example, corruption occurs at three levels: design, system, and user. First, at the *design level*, AI tools used in academic research are often developed by commercial or institutional actors who prioritise functionality, speed, and market logics over epistemic integrity

(Kidd, 2022, 2023a). Developers may assume that LLMs are sufficiently accurate for scholarly use, despite issues that raise questions about hallucinated references, bias toward dominant academic voices, and lack of transparency in training data (H. Nguyen & Eger, 2024; Rodriguez-Saavedra et al., 2025). This may reflect intellectual arrogance in overestimating the epistemic reliability of AI tools without sufficient external validation, and it may also reflect, we argue, negligence in failing to anticipate or mitigate the epistemic risks posed by these tools.

Second, at the *system level*, AI tools often reinforce dominant academic paradigms. Citation recommendation engines, for example, may prioritise high-impact journals and English-language sources, filter out dissenting or non-mainstream perspectives, and fail to adapt to new or interdisciplinary research trends. This creates a form of algorithmic dogmatism, where the system resists epistemic diversity and reinforces existing academic hierarchies (Becher & Trowler, 2001).

Finally, at the *user level*, researchers and students or early-career academics may begin to use AI tools to generate summaries, paraphrase texts, or even draft arguments. They may even accept AI outputs without verification, reduce engagement with primary sources, or avoid critical dialogue. This could be seen to foster epistemic laziness, cognitive offloading, overreliance, and intellectual arrogance, as users assume that using AI makes them more efficient or knowledgeable than peers who engage in traditional research methods (Georgiou, 2025). These concerns are echoed in (Chubb, Cowling, et al., 2022), who show that while AI can support efficiency and interdisciplinarity, it also risks reinforcing bureaucratic pressures and metric-driven practice of the managerial university. The authors argue that AI may potentially undermine human creativity and critical engagement with more traditional 'scholarly norms' (Bucchi, 2015). The example of AI in higher education research aligns closely with the vice-based analysis presented here, in that one might argue that the institutional embrace of AI without sufficient epistemic scrutiny reflects negligence at the design level, dogmatism at the system level, and epistemic laziness at the user level.

In this article, we refer to this example of AI in HE research, using what we present as a preliminary framework of technomoral vices. In doing so, we show how AI systems may cultivate vice and character corruption and, in so doing, provide reasons for other researchers to pay closer attention to the darker ethical considerations of AI use in research and implications for critical thinking and epistemic authority (Zagzebski, 2012).

4. The corrupting potential of technomoral systems

Human character is clearly changeable over time; we can become more or less vicious and more or less virtuous. To develop that thought, one can turn to the concepts of *aetiology* and *corruption*. Aetiological sensitivity (Kidd et al., 2020) involves asking about the origins or causes of an agent's epistemic vices. Instead of simply identifying the vice of a person, one also tries to enquire into the developmental history of that person's vice, including the social and material conditions which may have facilitated its development and entrenchment. Medina (2013) refers to this as the “sociogenesis” of epistemic character. Kidd (2016) asserts that when attributing epistemic vices to an agent, we should inquire into the origins of their vicious character for several reasons. First, it can help us identify deficiencies in our collective systems of epistemic socialisation. Second, it helps us avoid essentialising assumptions about epistemic character, especially for members of marginalised groups who are often, unfairly, perceived as vicious 'by nature'. Third, it allows us to demonstrate justice and fairness to epistemically vicious agents, since the responsibility for the condition and development of our moral-epistemic character may not be wholly due to the agent's own

decisions. Fourth, effective diagnosis is crucial for effective improvement. This idea of aetiological sensitivity also complicates issues of blame and responsibility. Battaly (Battaly, 2019b) argues that vice epistemology has a "responsibility problem." For instance, an agent may not be responsible for acquiring their vices, though, at a certain point in their development, they may become responsible for retaining them, for instance, if they have been subjected to reasonable character criticism.

Character corruption occurs when an agent's moral-epistemic character is said to be damaged in one of two ways: either through *active corruption*, which is the erosion or extirpation of virtues, or through *passive corruption*, which is the development and exercise of vices and failings. Corruption involves destructive interactions with corruptors, which can be a variety of conditions, processes, social structures, or incentives that tend to erode virtues and feed failings. There are several modes of character corruption: the *acquisition* of new vicious dispositions, the *activation* of latent vicious tendencies, the *propagation* of a vice (increasing its scope), the *stabilisation* of a vice (increasing the reliability of its presence), and the *intensification* of a vice (increasing its strength). Experiences of corruption could involve a person coming to acquire a vice which was not previously a part of their character, for instance, or it could involve their extant vices becoming stronger, more stable, and more entrenched.

Character corruption is a dynamic and dialectical phenomenon that agents can try to counteract in various ways, retaining a role for resistant kinds of agency. If this is true, then explanations for personal failings must be multidimensional and sensitive to psychological, characterological, situational, interpersonal, and structural factors (Cassam, 2019; Kidd, 2023b). Character corruption should be studied empirically, as shown by Kidd et al.'s (2021) work on the corrupting structural features of UK higher education. These are *corruptors*. General kinds of *corruptors* include the absence or derogation of exemplars to model virtue, the valorisation of vicious people and actions, the rebranding of vices as virtues, the establishment of conditions that increase the exercise costs of virtues, the inclusion of conditions that expand the scope and severity of vicious actions, and the establishment of conditions that increase the incentives and rewards of vice (Kidd, 2022; Kidd et al., 2020). Some corruptors are only contingently corrupting. Self-auditing could facilitate dishonesty, but perhaps not in institutions where there are random spot-checks.

Our proposal here is that technological systems contain a variety of corruptors that, however unintentionally, can tend to facilitate the erosion or extirpation of individual virtues and excellences and the development or exercise of individual vices and failings (Chamma et al., 2025). This idea is already well-established within critical discourses about technology, where social media and algorithms, whilst a double-edged proposition (Dwivedi et al., 2018; Li et al., 2023), can often be seen to inflame our worst traits, to bring out the worst in us, and to foster echo chambers (Barberá, 2020; Terren & Borge, 2021). Further, epistemic bubbles result in affective polarisation, for instance, where there is a gradual gravitation toward certain views because of algorithmic behaviour (Beer, 2019; Bucher, 2018).

Although the terms 'echo chambers' and 'epistemic bubbles' are often used interchangeably, they should be carefully distinguished, since Nguyen (2020) describes how the bubbles are full of those not exposed to ideas from the other side because of technological mediation. This is in contrast to the chambers, which serve to collectively align those with the same views, actively distrusting the other side (Coady, 2024; Miller & Record, 2013). Further, there are also Hobbesian arguments one could draw on about the corrupting effects of anonymisation and online aggression, where if there are no sanctions or constraints, people, in turn, 'attack.' (Cooper, 2025, pp. 16–17).

As part of a framework of vice, there is then further need for theorisation about which aspects of technomoral systems are acting as corruptors and which vices they corrupt. One could also extend this to considering the extent to which corrupting effects are specific to certain groups and which features of technomoral systems are conducive to which vices and failings.

5. Virtue, vice, and technology

Virtue ethicists have increasingly argued that we can go beyond principles and rules to frame AI ethics through character-based approaches (Coeckelbergh, 2020). Shannon Vallor (2016) offers one of the most influential accounts, promoting the cultivation of "technomoral virtues", or habits of moral character needed for human flourishing in an age shaped by opaque and rapidly evolving technology. Influenced by Aristotelian thinking, Vallor argues that technology does not merely extend human action; it reshapes patterns of thought, behaviour, and ways of valuing:

> Ethics and technology are connected because technologies invite or afford specific patterns of thought, behaviour, and valuing; they open up new possibilities for human action and foreclose or obscure others (Vallor, 2016, p. 2).

The idea is that technologies are an extension of our existing moral capabilities and provide optimistic ways of extending ourselves in an opaque and changing world. As such, Vallor provides accounts of how virtuous conduct can be pursued. Vallor's virtues include care, civility, courage, empathy, flexibility, honesty, humility, justice, perspective, self-control, magnanimity, and technomoral wisdom.

Consider, for example, technomoral courage. Vallor describes this as a "reliable disposition toward intelligent fear and hope with respect to moral and material dangers and opportunities presented by emerging technologies" (2016, p. 131). As a further example, for Vallor, technomoral humility involves the "renunciation of the blind faith that new technologies inevitably lead to human mastery and control of our environment" (2016, pp. 126–7).

Whilst these accounts certainly have an intellectual pedigree, we want to question the *selection*, *definition*, and *priority* assigned to these technomoral virtues. The abundance of virtues has prompted Nancy Snow to warn of a "proliferation problem" (Snow, 2019). However, our central concern here is not so much with the specifics of Vallor's account of technomoral virtues as it is with a more general feature of her account: the tendency to overlook potential or real technomoral vices.

To be clear, we contend that Vallor's account of technomoral life offers a valuable map of the virtues we might cultivate in interaction with emerging technologies, and we recognise that she does mention the occurrence of technomoral vices. However, we propose that existing accounts of the technomoral move too quickly to propose solutions, and they pay comparatively too little attention to the darker terrain, those spaces where vices, corruption, and lives of moral failure take root. If as many suggest, the moral character of AI is fashioned by the ways it becomes *embedded in social life, in language, and in human culture*, then AI is likely to inherit and reiterate our pre-existing defects and unwarranted proclivities. There has already been well-documented evidence of this in the form of racist, sexist, or otherwise discriminatory algorithms (Benjamin, 2023; Noble, 2018; Verma, 2019). Moreover, we have good prima facie reasons to worry that digital technologies can have lasting negative effects on human well-being and behaviour, given the mounting empirical evidence of the impact of smartphones, social media apps, and other forms of digital entertainment on children (Chubb, Missaoui, et al., 2022).

In fact, in her more recent book, *The AI Mirror*, Vallor (2024), like many others, addresses many of the real and potential harms caused by AI but does not develop a full technomoral vice approach. Instead, Vallor cautions us against our own biases and prejudices being reflected back to us through AI. Ultimately, in the face of these risks, she urges her readers to develop the right kinds of virtues. As in her previous work, 'flourishing with' and even 'potentially against technology' is the focus of her argument. Rather, we think that moral agents and, indeed, public authorities, should consider not just *how* to flourish in a world dominated by AI, but how to account for the moral risk – the risk to our characters – of growing ever more reliant on AI in the first place.

Hence, we do not contend that all efforts to resist technomoral corruption are inevitably futile. Rather, we contend that we are better served by considering the nature of the technomoral challenges we face on their own terms, and without rushing to consider potential 'solutions'. Ultimately, we argue that a vice approach can provide the appropriate moral vocabulary to draft an inventory of such risks.

Why, then, are the risks associated with entrenched vices so often eclipsed in ethical discussions of AI? Several forces converge to dim them from view:

1. There is an understandable human impulse to avert our gaze from life's dark sides, especially when radical technological optimism promises easier emotional rewards.
2. Virtue-ethical frameworks themselves sometimes encourage a "bright-siding" rhetoric, foregrounding ideals of "flourishing" while relegating systemic failure to the margins.
3. Commercial actors eager to promote and entrench AI possess strong incentives to accentuate success stories and downplay talk of corruption or harm.
4. Culturally powerful "liberationist" narratives depict technology as an emancipatory force, further amplifying positive imagery at the expense of critique.
5. The cult of 'efficiency' that pervades much of the neoliberal and post-neoliberal institutional settlements makes the adoption of AI-powered tools ever more seductive, since it promises individual adaptation and maybe even social elevation in the face of near permanent *déclassement* (see Bauman, 2000) for an account of the social anxiety we refer to.
6. Finally, a pragmatic resignation: *AI is here to stay, so we might as well think well of it*. This can blunt critical attention to vice.

Yet these omissions are increasingly untenable: from biased hiring software (Albaroudi et al., 2024) to social-media recommender systems, public controversies daily illuminate the moral hazards that accompany AI. The AI chatbot Grok on the platform X, which started to refer to itself as "MechaHitler" provides one such example.[4] In light of this, a genuinely technomoral philosophy must therefore reckon not only with the possibilities for the cultivation of virtues but also with the possibilities that technologies can activate, entrench, and intensify vices.

Although there is comparatively less research focused on the dark side of technology, referring to vices and corrupting tendencies, as we have said, some scholars have gone some way towards addressing it. In addition to Vallor's account, importantly, Coeckelbergh (2020) engaged in limited but important reference to the vices of AI. For Coeckelbergh, technolog-

4.See BBC article; Musk says Grok chatbot was 'manipulated' into praising Hitler https://www.bbc.co.uk/news/articles/c4g8r34nxeno accessed from the BBC website on 7th October 2026.

ical developments "open up new possibilities for human action and foreclose [and] obscure others" (p. 2). There is a recognition of "blind faith that new technologies inevitably lead to human mastery and control of our environment" (pp. 126-7).

As stated in our introduction, dominant narratives, moreover, often focus on the risks that technology fosters vices and failings of character, often in the dramatic form of hostile robotics. More realistically, narratives of injustice, dishonesty, and unfairness dominate public discourse about AI futures, as issues of privacy and transparency, race, and equality become more prevalent (Cave et al., 2020). These issues are framed broadly in terms of harms which require refinement to reflect the full impacts of AI on society (Fearnley et al., 2025). When vice is specifically discussed, it is often in a secondary fashion, with comments made in passing, or remaining at a fairly abstract level, focusing on the nature of vice or studies of specific vices (e.g., Battaly, 2018; Cassam, 2016; Kidd, 2018). Any account of moral life that deploys virtue-concepts without also deploying vice-related concepts in a similarly systematic way is at risk of a sort of theoretical myopia.

From an empirical perspective, there is a plethora of examples that show how, for instance, social media can encourage vices, such as aggressiveness, dogmatism (Fast & Horvitz, 2016; Watson et al., 2025), and self-righteousness, as well as vicious collective phenomena (Brady, 2020). These include the dangers of social comparison (Lippitt, 2023), such as bullying and mobbing (Kircaburun et al., 2020), as well as polarisation (Alfano et al., 2018; Rost, 2016). And yet, at a theoretical level, philosophical discourses about technology rely heavily on 'liberationist' vs. 'enslavement' approaches, and these can be generalised to include edificationist and corruptionist framings of technology as drivers of virtues and vices.

The additions to Vallor's account centre on two key concerns: (a) the neglect of technomoral vices, and (b) the possibility that technomoral systems can be corrupting. These are not merely abstract or theoretical worries; both are integral to contemporary academic and popular discourse on technology and are supported by growing empirical evidence. Our intention is not to reject Vallor's framework but to expand and offer a corrective steer to it, restoring attention to the 'dark sides' of technomoral life that her virtue-focused account tends to underplay.

6. Vice-enabling and virtue-eroding AI in the everyday

We have argued that systems and platforms are never neutral (Gitelman, 2013); in many cases, they have the potential to cultivate vice. Public descriptions of these platforms often reveal a deep moral discomfort: The platform 'X' is often described as toxic, poisonous, or humanity at its worst (Gevers et al., 2022; Parent et al., 2019). Such language reflects widespread perception of the character-shaping harms of these technologies. This is not merely about how technologies function but what they *do to us*. The recommendation algorithms behind platforms, which are now a relatively rudimentary form of AI, can drive polarisation and erode civility, not only by surfacing extreme content but by reshaping the ways people interact, listen, and express themselves. The harm lies not just in outcomes, but in the cultivation of new habits and moral orientations, as well as the loss of old habits and capacities. Though these technological tools can foster connection and creativity, they reward attention-seeking, impulsivity, outrage, and manipulation, creating environments where vice is encouraged to thrive.

Children's interactions with conversational AI are another important example. Chubb, Missaoui, et al. (2022) show that voice assistants can affect children's development of moral care, patience, and civility. For instance, when digital agents respond instantly to commands,

children may grow less tolerant of real-world delays or human inconsistency. This can reinforce impatience or a sense of entitlement, subtly reshaping the moral expectations children bring into their human relationships. Anthropomorphism, the process of attributing human-like qualities to machines, adds another layer of complexity. Research has shown that children (and adults) form emotional attachments to AI agents, blurring moral boundaries. As Coeckelbergh (2021) argues, treating robots well may not cultivate virtue, but mistreating them can still reflect or reinforce vice. Our most pressing concern here is not so much what robots deserve (although this, too, is no doubt an important moral question, which will probably grow in significance as our confidence in AI's lack of sentience, consciousness, or personhood erodes if certain prognosticated breakthroughs occur), but what our treatment of them says about us and encourages in us.

If AI can help cultivate and express virtues, it can just as readily encourage failings of character. Indeed, in many cases, it already does. Moreover, dishonesty both on the part of AI systems and their users is rendered harder to detect and easier to indulge in, since generative AI can fabricate sources or mislead users, and people may use AI tools to plagiarise, misrepresent their knowledge, or construct disingenuous arguments. In both cases, the technology lowers the moral friction that would normally constrain deceptive conduct. Another significant concern is impatience. There is some neuroscientific evidence that interacting with systems that deliver instant results, feedback, and gratification trains users to expect the same from human interactions (White et al., 2024). This can corrode virtues such as patience, tolerance, and attentiveness, traits essential for civil discourse and deep engagement. AI can also breed dependency. As users increasingly outsource tasks to machines, from remembering appointments to generating content, their own cognitive and moral capacities risk degradation and atrophy (Fan et al., 2025; Risko & Gilbert, 2016; Sparrow et al., 2011). Over time, dependency becomes not just a convenience but a disposition and a reluctance or inability to act without technological mediation (Ahmad et al., 2023). Consider, for example, the loss of map-reading capacity caused by overdependence on navigational apps that can direct people from one place to another. Although for the most part this loss of capacity can be harmless, under conditions of duress or malfunction, it can result in a much more damaging inability to exercise sensible judgment, attend to one's environment, or follow road markings (Dahmani & Bohbot, 2020; Fajnerová et al., 2018).

Other common technomoral vices emerging from AI systems include deception (Park et al., 2024), kinds of cynicism (Matthews, 2022), carelessness (Gerlich, 2025), incivility, prejudice, and overconfidence. These are not theoretical risks; they are patterns already visible in our daily lives. Taken together, they suggest the need for a richer vice framework that tracks the moral downsides as well as acknowledges the potential of AI to, in many cases, foster virtues. [5] We have seen how AI is affecting our daily lives and drawn on the example of academic research as one key area of potential epistemic corruption. To complement virtue-based approaches to AI ethics, we argue for the development of a corresponding account of technomoral vice. First, we must account for some of the technomoral vices of AI.

7. Technomoral vices and AI

This section first sets out a preliminary list of technomoral vices and then considers how technomoral systems can corrupt.

5.See, for instance, as a counterargument, AI can be seen as an agent of change to support civility and respect in online discussions (Chakraborty, 2025).

7.1 A preliminary list of technomoral vices

- Intellectual laziness
- Dishonesty
- Impatience
- Dependency
- Deception
- Carelessness
- Incivility
- Prejudice
- Gossip
- Negligence
- Thoughtlessness
- Unimaginativeness
- Overconfidence

The list above is a non-exhaustive set of examples of the kinds of technomoral vices of AI. All but gossip and thoughtlessness are explored here. Technomoral vices are not just naturally occurring vices of character (although the naturally occurring varieties of these also exist). Instead, these technomoral vices are often embedded, encouraged, and amplified by technological systems with or without the intent of those who have originally designed the technological systems in question. A properly responsible ethical framework for thinking of AI must account not only for intentions, but more importantly for unintended outcomes over time. Crucially, this includes how interaction with these technologies diachronically shapes moral character, and careful assays of the varieties of moral and epistemic vices which might be facilitated by different ways of using and interacting with different kinds of technologies.

Corruption in technomoral systems

To understand the full moral impact of artificial intelligence, we must thus move beyond the binary of harm versus benefit, and consider a deeper ethical question: how do AI systems shape our moral character over time? More specifically: how might they *corrupt* us? In the context of technomorality, corruption refers to the erosion of technomoral virtues and the reinforcement or creation of technomoral vices. On our account, human character is not static; it is shaped and reshaped by our social and technological environments. As such, AI systems can play a powerful role in moral degradation, subtly reducing our tendency and our ability to act well, reason clearly, or relate to others ethically in a reliable fashion.

Drawing on contemporary philosophical vice theory, we distinguish between two forms of character corruption:

- *active corruption*: the erosion or extirpation of virtues and other excellences of character;
- *passive corruption*: the introduction, activation, development, reinforcement, or normalisation of vices and other failings of character.

These forms of corruption are not always dramatic or visible. They can be gradual, diffuse, and embedded in everyday interactions with technology. Indeed, Coeckelbergh (2020) remarks how AI is often invisible and how these effects may develop over time without our awareness. For instance, a user who regularly engages with an AI that rewards self-destruc-

tive, impulsive, extreme, or aggressive content may over time become more gullible, less caring, less patient, less reflective, or more cynical. In the case of teenage boys and young men, there is a growing body of evidence that suggests that the social media algorithms and generative AI (Wagner & Cetinić, 2026) amplify misogynistic voices and encourages a disregard for women's rights and autonomy (YouTube, for instance) (Döring & Mohseni, 2019; Singh et al., 2025). Again, this suggests the presence of similar technomoral vices. Once more in the academic context, a researcher who over-relies on AI-generated summaries may slowly drift towards a kind of intellectual passivity, where reading source material or complex scientific articles becomes too demanding (Noorbehbahani & Oyibo, 2026). This process involves a range of mechanisms, including the acquisition of new vices, the activation of latent tendencies, the propagation of existing flaws, and the stabilisation of undesirable traits.

In many cases, AI systems can unintentionally act as *corruptors*: features of a technological environment that erode virtue and enable vice. In other cases, designers may also set out to corrupt (Kidd, 2022). These corruptors can include:

- the absence or abuse of virtuous exemplars;
- the valorisation of vice (e.g., celebrating tech oligarchs who demonstrate arrogance, hubris, or manipulativeness);
- the rebranding of vices as virtues (e.g., overconfidence reframed as *confidence*, manipulation as *persuasion*);
- increasing the exercise costs of virtue, making honesty or patience harder to perform and sustain;
- increasing the rewards of exercising vices, such as greater visibility or success through polarising, deceptive, or lazy content.

Crucially, these processes may occur with little or no conscious intent on the part of the user or designer. They are often systemic, emerging from design incentives or structural logics of platforms and algorithms. Moreover, they can occur even if the person is aware of the negative effects on their moral and epistemic character.

The notion of aetiological sensitivity, drawn from Kidd and Medina, holds that when we attribute a vice to an agent, we should also ask *how* that vice was formed: Was it encouraged by social context? Incentivised by a platform? Rewarded by algorithmic feedback? What kinds of counter-corrupting resources were available? Such sensitivity matters, not to absolve individuals of responsibility, but to enable more effective moral critique.

Corruption, then, in this context, is a system of *processes* through which virtue is undermined, and vice is entrenched. Corruption of this kind is dynamic, interactive, and often covert. As such, our ethical analysis of AI must attend not only to how it instrumentally performs set tasks or what outcomes it generates, but also to how it changes *us,* our habits, our dispositions, and ultimately our moral character.

8. Conclusion

We have argued for the necessity of expanding AI ethics by offering a fuller account of technomoral vice and technomoral corruption. We see this as a corrective to two related tendencies in current work. First, a default to liberal moral frameworks, such as justice, fairness, and autonomy, that tend to foreground normative aspirations and the promise of moral enhancement. Second, a "bright siding" of technology that prioritises virtue, flourishing, and other positive moral concepts, while neglecting vice, corruption, and the other 'darker' concepts.

Vallor's technomoral virtue framework initiated a challenge to the first tendency, and our account of technomoral vice and corruption responds to the second. Moreover, we are better placed to make progress in developing a *technomoral character theory*, attuned to virtues and flourishing as well as vices and corruption (R. S. Dillon, 2012; Kidd, 2022; Kidd et al., 2020).

Building on Vallor's technomoral virtue framework, we have extended the conversation by explicitly attending to the vice-enabling and virtue-eroding potentials of AI systems. We have argued that AI functions not merely as an instrument of action but as a technosocial environment that shapes moral perception, conduct, and character. In this context, we introduced the concept of technomoral corruption, distinguishing between active and passive forms, and outlining the mechanisms by which vice is cultivated, valorised, or rendered morally invisible.

To substantiate our argument, among other things, we used the example of higher education research. Here, the uptake of generative AI into scholarly workflows from literature review to writing and dissemination provides a concrete example of epistemic and moral vice formation (Kosmyna et al., 2025). Vices such as intellectual laziness[6], overconfidence, and negligence are not idiosyncratic failures but systemically incentivised tendencies within an institutional logic that prioritises speed, efficiency, and measurable output. This example illustrates how vice can be cultivated not despite, but through the ordinary use of AI in high-trust, high-stakes domains. Foregrounding vice in AI ethics then allows for a richer and more critical moral vocabulary, one that recognises not only the normative possibilities of AI but also its capacity for moral deformation.

We have proposed that AI systems function not merely as tools for moral action, but as environments that actively shape technomoral character. They do so by enabling certain habits of perception, judgment, and conduct while foreclosing others. These environments can cultivate vice not only in the sense of individual moral failings, but as structural and systemic tendencies toward corruption.

This work raises important questions about epistemic responsibility in AI design and deployment. While humans are typically held accountable for their epistemic vices, the attribution of responsibility in AI contexts remains ethically and philosophically contested. Nonetheless, recognising the potential for vice-like dispositions in AI underscores the need for careful attention to the epistemic environments in which such systems are trained and deployed. In light of this, we argue for an expanded ethical vocabulary that explicitly theorises vice within the context of AI. The dominant emphasis on virtue, flourishing, and moral enhancement within AI ethics, while important, offers an incomplete account. Without an equal sensitivity to vice and to the conditions under which it is produced and sustained, ethical discourse risks both naivety and complicity. Attending to technomoral vice allows for a richer and more critical engagement with the moral life of technologies: one capable of interrogating not only what AI enables, but what it corrupts.

References

Ahmad, S. F., Han, H., Alam, M. M., Rehmat, M., Irshad, M., Arraño-Muñoz, M., & Ariza-Montes, A. (2023). Impact of artificial intelligence on human loss in decision-making, laziness, and safety in education. *Humanities and Social Sciences Communications*, *10*(1), 1–14.

6.The proliferation of various corruptors within techno-moral systems tends to have deleterious impacts on people's epistemic performance.

Albaroudi, E., Mansouri, T., & Alameer, A. (2024). A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. *AI*, *5*(1), 383–404. https://doi.org/10.3390/ai5010019

Alfano, M., Carter, J. A., & Cheong, M. (2018). Technological seduction and self-radicalization. *Journal of the American Philosophical Association*, *4*(3), 298–322.

Astola, M. (2021). Do Collective Epistemic Virtues have to be Scaled-Up Individual Virtues? *Social Epistemology Review and Reply Collective*, *10*(7), 40–47.

Baddorf, M., & McKay, N. (2024). A Theory of Collective Virtue. *Journal of Ethics and Social Philosophy*, *27*(3), 405–423. https://doi.org/10.26556/jesp.v27i3.2005

Barberá, P. (2020). Social media, echo chambers, and political polarization. In N. Persily & T. A. Tucker (Eds.), *Social media and democracy: The state of the field, prospects for reform* (pp. 34–55). Cambridge University Press.

Battaly, H. (2015). *Virtue*. Polity Press.

Battaly, H. (2018). Closed-mindedness and dogmatism. *Episteme*, *15*(3), 261–282. https://doi.org/10.1017/epi.2018.22

Battaly, H. (Ed.). (2019a). *The Routledge handbook of virtue epistemology*. Routledge.

Battaly, H. (2019b). Vice epistemology has a responsibility problem. *Philosophical Issues*, *29*(1), 24–36.

Bauman, Z. (2000). *Liquid Modernity*. Polity Press.

Becher, T., & Trowler, P. (2001). *Academic tribes and territories*. McGraw-Hill Education (UK).

Beer, D. (2016). *Metric Power*. Palgrave Macmillan.

Beer, D. (2019). The social power of algorithms. In *The social power of algorithms* (pp. 1–13). Routledge.

Benjamin, R. (2023). Race after technology. In *Social Theory Re-Wired* (pp. 405–415). Routledge.

Beukes, L., Robinson, K. S., Liu, K., Rajaniemi, H., Reynolds, A., & Bodard, A. (2017). Science fiction when the future is now. *Nature*, *552*(7685), 329–333. https://doi.org/10.1038/d41586-017-08674-8

Brady, W. J. (2020). Temporal increases in the expression of moral outrage on social media. *Science Advances*, *6*(48).

Brey, P. (2022). Understanding Engineering Design and Its Social, Political, and Moral Dimensions. In S. Vallor (Ed.), *The Oxford Handbook of Philosophy of Technology*. https://doi.org/10.1093/oxfordhb/9780190851187.013.24

Bucchi, M. (2015). Norms, competition, and visibility in contemporary science: The legacy of Robert K. *Merton. Journal of Classical Sociology*, *15*(3), 233–252.

Bucher, T. (2018). *If... then: Algorithmic power and politics*. Oxford University Press.

Buhmann, A., & Fieseler, C. (2021). Towards a deliberative framework for responsible innovation in artificial intelligence. *Technology in Society*, *64*, 101475.

Byerly, R., & Byerly, M. (2016). Collective Virtue. *The Journal of Value Inquiry*, *50*(1), 33–50.

Cassam, Q. (2016). Vice epistemology. *The Monist*, *99*(2), 159–180.

Cassam, Q. (2019). *Vices of the Mind: From the Intellectual to the Political*. Oxford University Press.

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

Cave, S., Dihal, K., & Dillon, S. (Eds.). (2020). *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford University Press.

Chakraborty, S. (2025). From Reactive to Proactive: AI's Role in Promoting Civility and Respect in Online Discussions. In *Ethical AI Solutions for Addressing Social Media Influence and Hate Speech* (pp. 25–36). IGI Global Scientific Publishing.

Chamma, D., Ahmadpour, N., Ishtiaque Ahmed, S., Jahan Mim, N., Zhang, W. Q., Bona, K., Sachathep, T., Horst, H., & Imad Harb, J. (2025). Epistemic Injustice in and through AI. *Proceedings of the 37th Australian Conference on Human-Computer Interaction*, 1020–1025. https://doi.org/10.1145/3764687.3767279

Cheong, M., Abedin, E., Ferreira, M., Reimann, R., Chalson, S., Robinson, P., Byrne, J., Ruppanner, L., Alfano, M., & Klein, C. (2024). Investigating Gender and Racial Biases in DALL-E Mini Images. *ACM Journal on Responsible Computing*, *1*(2), 1–20. https://doi.org/10.1145/3649883

Chubb, J., Cowling, P., & Reed, D. (2022). Speeding up to keep up: Exploring the use of AI in the research process. *AI & Society*, *37*(4), 1439–1457.

Chubb, J., Missaoui, S., Concannon, S., Maloney, L., & Walker, J. A. (2022). Interactive storytelling for children: A case-study of design and development considerations for ethical conversational AI. *International Journal of Child-Computer Interaction*, *32*, 100403. https://doi.org/10.1016/j.ijcci.2021.100403

Chubb, J., & Tutton, R. (2026). How the tech oligarchy colonizes our futures. *Science as Culture*, 1–13. https://doi.org/10.1080/09505431.2026.2688841.

Coady, D. (2024). Stop Talking about Echo Chambers and Filter Bubbles. *Educational Theory*, *74*(1), 92–107. https://doi.org/10.1111/edth.12620

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

Coeckelbergh, M. (2021). Does kindness towards robots lead to virtue? A reply to Sparrow's asymmetry argument. *Ethics and Information Technology*, *23*(4), 649–656. https://doi.org/10.1007/s10676-021-09604-z

Cooper, D. E. (2025). *Pessimism, Quietism, and Nature as Refuge*. Agenda.

Crerar, C. (2018). Motivational Approaches to Intellectual Vice. *Australasian Journal of Philosophy*, *96*(4), 753–766. https://doi.org/10.1080/00048402.2017.1394334

Dahmani, L., & Bohbot, V. D. (2020). Habitual use of GPS negatively impacts spatial memory during self-guided navigation. *Scientific Reports*, *10*(1), 6310. https://doi.org/10.1038/s41598-020-62877-0

Dillon, R. S. (2012). Critical Character Theory: Toward a Feminist Perspective on 'Vice' and 'Virtue.' In S.L. Crasnow & A. M. Superson (Eds.), *Out From the Shadows: Analytical Feminist Contributions to Traditional Philosophy* (pp. 83–106). Oxford University Press.

Dillon, S., & Craig, C. (2021). *Storylistening: Narrative evidence and public reasoning*. Routledge.

Döring, N., & Mohseni, M. R. (2019). Male dominance and sexism on YouTube: Results of three content analyses. *Feminist Media Studies*, *19*(4), 512–524. https://doi.org/10.1080/14680777.2018.1467945

Dubber, M. D., Pasquale, F., & Das, S. (2020). *The Oxford handbook of ethics of AI*. Oxford University Press.

Dwivedi, Y. K., Kelly, G., Janssen, M., Rana, N. P., Slade, E. L., & Clement, M. (2018). Social Media: The Good, the Bad, and the Ugly. *Information Systems Frontiers*, *20*(3), 419–423. https://doi.org/10.1007/s10796-018-9848-5

Fajnerová, I., Greguš, D., Hlinka, J., Nekovářová, T., Škoch, A., Zítka, T., Romportl, J., Žáčková, E., & Horáček, J. (2018). Could Prolonged Usage of GPS Navigation Implemented in Augmented Reality Smart Glasses Affect Hippocampal Functional Connectivity? *BioMed Research International*, *2018*(1), 2716134. https://doi.org/10.1155/2018/2716134

Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, *56*(2), 489–530.

Fast, E., & Horvitz, E. (2016). *Identifying Dogmatism in Social Media: Signals and Models*. https://arxiv.org/abs/1609.00425v1

Fast, E., & Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. *Proceedings of the AAAI Conference on Artificial Intelligence*, *31*(1).

Fearnley, L., Cairns, E., Stoneham, T., Ryan, P. M., Chubb, J., Iacovides, J., Iglesias Urrutia, C., McDermid, J., & Habli, I. (2025). *Risk of What? Defining Harm in the Context of AI Safety*.

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2021). How to design AI for social good: Seven essential factors. In *Ethics, governance, and policies in artificial intelligence* (pp. 125–151). Springer International Publishing.

Forstenzer, J. (2018). Teaching Excellence Framework, Epistemic Insensibility and the Question of Purpose. *Journal of Philosophy of Education*, *52*(3), 548–574.

Fricker, M. (2010). Can there be institutional virtues? In T. S. Gendler & J. Hawthorne (Eds.), *Oxford Studies in Epistemology* (Vol. 3, pp. 235–252). Oxford University Press.

Fricker, M. (2020). Institutional Epistemic Vices. The Case of Inferential Inertia. In I. J. Kidd, H. Battaly, & Q. Cassam (Eds.), *Vice Epistemology* (pp. 89–107). Routledge.

Fritts, M., & Cabrera, F. (2022). Fake News and Epistemic Vice: Combating a Uniquely Noxious Market. *Journal of the American Philosophical Association*, *8*(3), 454–475. https://doi.org/10.1017/apa.2021.11

Georgiou, G. P. (2025). *ChatGPT produces more 'lazy' thinkers: Evidence of cognitive engagement decline*. https://doi.org/10.48550/arXiv.2507.00181

Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*(1), 6.

Gevers, I., Markov, I., & Daelemans, W. (2022). Linguistic analysis of toxic language on social media. *Computational Linguistics in the Netherlands*, *12*, 33–48.

Gitelman, L. (Ed.). (2013). *Raw data is an oxymoron*. MIT Press.

Habli, I., Lawton, T., & Porter, Z. (2020). Artificial intelligence in health care: Accountability and safety. *Bulletin of the World Health Organization*, *98*(4), 251–252.

Hagendorff, T. (2022). A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology*, *35*(3), 55.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

Kidd, I. J. (2016). Charging others with epistemic vice. *The Monist*, *99*(2), 181–197.

Kidd, I. J. (2018). Deep epistemic vices. *Journal of Philosophical Research*, *43*, 43–67.

Kidd, I. J. (2021). A case for an historical vice epistemology. *Humana.Mente*, *14*(39), 69–86.

Kidd, I. J. (2022). From Vice Epistemology to Critical Character Epistemology. In M. Alfano, C. Klein, & J. Ridder (Eds.), *Social Virtue Epistemology* (pp. 84–102). Routledge.

Kidd, I. J. (2023a). Corrupted temporalities, 'cultures of speed', and the possibility of collegiality. *Educational Philosophy and Theory*, *55*(3), 330–342. https://doi.org/10.1080/00131857.2021.2017883

Kidd, I. J. (2023b). Multidimensionalism, Resistance, and the Demographic Problem. *European Journal of Analytic Philosophy*, *19*(1). https://doi.org/10.31820/ejap.19.1.4.

Kidd, I. J., Battaly, H., & Cassam, Q. (2020). *Vice epistemology*. Routledge.

Kidd, I. J., Chubb, J., & Forstenzer, J. (2021). Epistemic corruption and the research impact agenda. *Theory and Research in Education*, *19*(2), 148–167. https://doi.org/10.1177/14778785211029516

Kircaburun, K., Griffiths, M. D., & Billieux, J. (2020). A systematic review of the relationship between social media usage and cyberbullying. *Frontiers in Psychology*, *11*(Article 570431).

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X. H., Beresnitzky, A. V., & Maes, P. (2025). *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for an essay writing task*.

Lahroodi, R. (2007). Collective Epistemic Virtues. *Social Epistemology*, *21*(3), 281–297. https://doi.org/10.1080/02691720701674122

Li, X., Guo, X., & Shi, Z. (2023). Bright sides and dark sides: Unveiling the double-edged sword effects of social networks. *Social Science & Medicine*, *329*, 116035. https://doi.org/10.1016/j.socscimed.2023.116035

Lippitt, J. (2023). Kierkegaard, "the Public", and the Vices of Virtue-Signaling: The Dangers of Social Comparison. *Religions*, *14*(11), 1370. https://doi.org/10.3390/rel14111370

Lu, X., Kelly, M. O., & Risko, E. F. (2020). Offloading information to an external store increases false recall. *Cognition*, *205*, 104428. https://doi.org/10.1016/j.cognition.2020.104428

Manditereza, B., & Chamboko-Mpotaringa, N. (2024). Generative AI and its implications for higher education students' creativity. In *Impacts of generative AI on creativity in higher education* (pp. 177–192). IGI Global.

Matthews, T. (2022). Deepfakes, Intellectual Cynics, and the Cultivation of Digital Sensibility. *Royal Institute of Philosophy Supplement*, *92*, 67–85. https://doi.org/10.1017/S1358246122000224

Matthews, T. (2025). Vice, Skill, and the Non-Ideal. *International Journal of Philosophical Studies*, *33*(2), 234–257. https://doi.org/10.1080/09672559.2025.2482919

Medina, J. (2013). *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*. Oxford University Press. https://doi.

Miller, B., & Record, I. (2013). Justified belief in a digital age: On the epistemic implications of secret internet technologies. *Episteme*, *10*(2), 117–134. https://doi.org/10.1017/epi.2013.11

Nguyen, C. T. (2020). Echo chambers and epistemic bubbles. *Episteme*, *17*(2), 141–161.

Nguyen, H., & Eger, S. (2024). *Is there really a Citation Age Bias in NLP?*

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. In *Algorithms of Oppression*. New York University Press.

Noorbehbahani, F., & Oyibo, K. (2026). AI over-dependence and human cognitive decline: Hazards, evidence, and mitigation strategies. *Computers in Human Behavior Reports*, 101102.

O'Neil, C. (2017). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Parent, M. C., Gobble, T. D., & Rochlen, A. (2019). Social media behavior, toxic masculinity, and depression. *Psychology of Men & Masculinities*, *20*(3), 277.

Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. *Patterns*, *5*(5).

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rodriguez-Saavedra, M. O., Barrientos-Alfaro, A. R., Málaga-Dávila, C. P., Chávez-Quiroz, F. G., & Arguedas-Catasi, R. W. (2025). Ethics and biases in the use of ChatGPT for academic research. *International Journal of Innovative Research and Scientific Studies*, *8*(1), 874–885.

Rost, K. (2016). Digital Social Norm Enforcement: Online Firestorms in Social Media. *PLOS ONE*, *11*(6).

Sacasas, L. M. (2011, December 2). *Leo Marx on the sources of our technological pessimism. The Frailest Thing*. https://thefrailestthing.com/2011/12/02/leo-marx-on-the-sources-of-our-technological-pessimism/

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., & Perez, E. (2023). *Towards understanding sycophancy in language models*.

Shaw, D. (2025). *The digital erosion of intellectual integrity: why misuse of generative AI is worse than plagiarism*. AI & SOCIETY.

Singh, A., Sharma, D., & Singh, V. K. (2025). Misogynistic attitude detection in YouTube comments and replies: A high-quality dataset and algorithmic models. *Computer Speech & Language*, *89*, 101682.

Smith, M. R., & Marx, L. (Eds.). (1994). *Does technology drive history?: The dilemma of technological determinism*. MIT Press.

Snow, N. (2019). Proliferating virtues? a clear and present danger. In E. Grimi (Ed.), *Virtue Ethics: Retrospect and Prospect*.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776–778. https://doi.org/10.1126/science.1207745

Tanesini, A. (2021). *The mismeasure of the self: A study in vice epistemology*. Oxford University Press.

Terren, L., & Borge, R. (2021). *Echo chambers on social media: A systematic review of the literature*.

Uszkai, R. (2017). Fostering intellectual and ethical virtues in the age of Artificial Intelligence. *School Children and the Challenge of Managing AI Technologies*, *18*.

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Vallor, S. (2024). *The AI Mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.

Verma, S. (2019). Weapons of math destruction: how big data increases inequality and threatens democracy. *Vikalpa*, *44*(2), 97–98.

Wagner, L., & Cetinić, E. (2026). Perpetuating misogyny with generative AI: How model personalization normalizes gendered harm. *Big Data & Society*, *13*(3), 20539517261467361.

Watson, D. S., Mökander, J., & Floridi, L. (2025). Competing narratives in AI ethics: a defense of sociotechnical pragmatism. *AI & Society*, *40*(5), 3163–3185.

White, B., Clark, A., & Miller, M. (2024). Digital Being: social media and the predictive mind. *Neuroscience of Consciousness*, *2024*(1). https://doi.org/10.1093/nc/niae008

Wilsdon, J. (2015). *The metric tide: Independent review of the role of metrics in research assessment and management*. HEFCE.

Zagzebski, L. T. (2010). Exemplarist virtue theory. *Metaphilosophy*, *41*(1–2), 41–57.

Zagzebski, L. T. (2012). *Epistemic Authority: A Theory of Trust, Authority, and Autonomy in Belief*. Oxford University Press.

Zimmermann, A. (2021). Stop Building Bad AI. *Boston Review*, *64*(2), 100–110.