AI and Ethics (2025) 5:2257–2262
https://doi.org/10.1007/s43681-024-00552-0

ORIGINAL RESEARCH

Overcoming AI ethics, towards AI realism

Michele Murgia[1]

Received: 23 May 2024 / Accepted: 8 August 2024 / Published online: 19 August 2024

Abstract
Discussions about artificial intelligence invariably include a nod to ethics. AI ethics has permeated the growing discourse surrounding AI, leading to numerous frameworks and principles intended to guide ethical design. This widespread surge in AI discussions, both academic and public, underscores a significant gap in normative political theory, a gap that urgently needs addressing. Although AI ethics as applied moral philosophy has been criticised as decontextualising AI or as outright useless, there remains a profound lack of understanding the proper political normativity of AI. The critique of AI ethics typically focuses only on feasibility concerns or moral harms, approaches that fail to capture the normative sources from which AI as a political phenomenon draws. The result is a depoliticisation of AI, risking further mystification and giving AI providers the means to justify illegitimate power relations. By leveraging the recent realism-moralism debate in normative political theory, I aim to show that the realist tradition can be the unexpected corner from which we can study these consequences and suggest a substantively different approach to AI in future, moving from AI ethics to AI realism.

Keywords AI · Realism · Normative political theory · Ethics

1 AI ethics and the problem of moralism

Artificial intelligence technologies are increasingly being deployed across various sectors, from law enforcement to education, bringing potential detriments into sharper focus. This use, however, is not without consequences. These are now predominantly addressed through appeals to 'moral principles' – AI ethics – ubiquitous from universities to EU legislation to the Vatican. In the AI ethics discourse we can roughly identify three AI ethics camps, though their boundaries are blurry and fluid: (i) corporatist AI ethics, (ii) reformist AI ethics and (iii) AI justice. Let us take a closer look at the differences between these camps before moving on to their common denominator: moralism.

(i) Corporatist ethics are what I call the use of broad ethical frameworks, for example by public organisations, corporations, or programmers to temper the (perceived) dangers of AI technologies. These frameworks typically contain principles noting terms such as fairness, privacy, and explicability. Importantly, these frameworks are meant not to rock the boat but rather to align with pre-existing interests. Ethics are here thus subsumed under existing practices.

This brings us to (ii) reformist ethics. Reformist ethics are for an important part a criticism of corporatist ethics. The critique can be easily summarised as corporatist ethics being 'ethics-washing': only appearing ethical while actually engaging in unethical practices [2]. Reformist ethics take the ethical principles professed by corporatist ethics and hold them up to scrutiny (see [10, 11, 12]. It is important to note that they analyse and criticise these principles from an ethical approach themselves and look for inconsistencies or shortcomings based on those principles. They might therefore share the belief in the importance of something like 'fairness' with corporatist ethics, but simply believe corporatist ethics is not actually employing the notion of fairness (adequately enough).

Another important criticism is that ethical principles are not isolated, meaning that the way technologies are socially and culturally embedded must be taken into account when thinking about ethical AI [15]. For this reason, AI problems cannot be solved in a technical vacuum, and unethical AI as a system follows from an unethical environment. The vacuum argument proves an inconvenience to the corporatist ethics view because it implies change rather than business as usual. It is against this backdrop that big-tech companies

✉ Michele Murgia
michele.murgia@nhlstenden.com; m.e.j.murgia@gmail.com

[1] NHL Stenden, University of Applied Sciences, Leeuwarden, Netherlands

such as Meta, X, and to an extent Microsoft have disbanded their entire AI ethics teams [4].

The criticisms outlined above are accompanied by the conversely positive approach to apply reformist ethics to the design of AI models. This also includes formulating new ethical lines of thought that are primarily meant to supplement the AI ethics discourse. Again, in reformist ethics, these new lines of thought still base themselves on ethical notions such as fairness, transparency, and accountability.

(iii) AI justice turns away from corporatist and ostensibly reformist ethics. It shares with the latter an emphasis on the point that AI ethical principles do not exist in an isolated vacuum. It also partly shares the notion that these principles are meaningless: key terms in corporatist ethics frameworks are brimming with ambiguity and vagueness. This fog of meaning facilitates finding a suitable definition for these terms that readily conforms to existing practices and covers up potential moral harms [5, 25]. Yet the critique that these principles are meaningless also extends to reformist ethics. In "The Uselessness of AI Ethics," Munn [17] argues that these principles remain at a highly abstract level and due to their ambiguity become incoherent. Moreover, they are toothless, meaning not enforceable. Here the issue of feasibility and more broadly operationalisation comes to the fore alongside moral harm. This also features in Mittelstadt [16], for whom AI ethical principles are "vague, high-level principles, and value statements which promise to be action-guiding, but in practice provide few specific recommendations and fail to address fundamental normative and political tensions embedded in key concepts." It is here where things become interesting.

For Munn, in citing Mittelstadt, agrees that "work on AI ethics has mainly produced [these kinds of principles] [my emphasis]." Both Munn and Mittelstadt, however, clearly concern themselves solely with criticising corporatist ethics. "Work on AI ethics" here means exactly those big and vague frameworks and not, say, academic non-corporatist work on AI ethics. This reduction of AI ethics generally to corporatist ethics specifically when critique is involved is often the case within the AI ethics discourse. Munn's own proposal to strive for AI justice rather than AI ethics is, in fact, based on this reduction. Let me briefly explain how this comes to pass.

The idea that AI ethical principles are meaningless can also be extended to reformist ethics. When reformist ethics criticise the moral notion of fairness within corporatist ethics by saying that the notion does not align with existing practices, they can be urged to give further substantive body to their own notion. For they do believe that fairness is a legitimate ethical principle. Once that happens, I believe reformist ethics will be hard-pressed to approximate principles that are not meaningless. More precisely, they will have trouble formulating principles that pay heed to context and provide good reasons as to why these principles are justified, ridding them of incoherency and incongruous aims. This tension between the absence of meaning in one ethical domain and the difficulty of providing it in its stead gives rise to proposals such as those entailed by AI justice. Justice is here a term meant to expand, as Munn says, "the ethical scope of inquiry and intervention." The necessity to turn away from AI ethics, however, is predicated on the reduction of AI ethics to corporatist ethics. This enables AI justice to be posited against a "de facto turn to ethical principles." Once we unveil the underpinnings of this juxtaposition, AI justice does not seem very different from reformist ethics.

Nor does the difference lie in feasibility concerns. All ethical camps outlined above consider the feasibility of ethical principles and their aims to some lesser or greater extent. Whether it be about the operationalisation of principles and subsequent translation into practices, the lack of enforceable principles, or a focus on existing or required enforcement mechanisms, one cannot say in absolute terms that any one of these camps eschews talking about feasibility constraints.

The distinction between reformist ethics and AI justice can be rather sought in ideology critique. AI justice looks at how ideology shapes what we mean by 'AI' and is more firmly rooted in critical theory than reformist ethics. It therefore scrutinises power relations and how they constitute the AI discourse, including AI ethics. This indeed broadens the scope of AI ethical inquiry and intervention in a non-trivial way. It is however born from the same considerations as reformist ethics, its impetus is morality and moral harm.[1] But these moral commitments do not offer good grounds to make normative claims about the political nature of AI. In fact, AI ethics risk exacerbating precisely the problems they warn about, such as the regulatory capture by the tech industry under the guise of corporatist ethics. My point here echoes the critique of ethics-first approaches to politics articulated by realists in recent political-theoretical debates. A closer look at this critique can further inform us of the problem of political moralism in AI ethics and point us towards alternative sources of normativity proper to AI politics. Let us turn to what realists exactly understand by 'political moralism'.

[1] Here I draw on Sankaran's [21] account of the 'new ideology critique' within Anglo-American philosophy and Rossi and Aytac's [1] reflections on its moral premises. AI justice as ideology critique similarly understands AI ideology as flawed insofar as it contributes to moral harm.

2 Ethics is dead politics

Coined by Bernard Williams [26], political moralism consists of deriving political prescriptions from pre-political moral ideals such as autonomy, fairness, equality, happiness, or justice. As Rossi and Sleat [20] note, these values are pre-political in two ways: '[i] they are taken to float free from the forces of politics, [ii] and they are assigned a foundational role insofar as they have antecedent authority over the political and determine or exhaust the appropriate ends and limits of politics'. For realists, this approach fails to adequately grasp politics as an autonomous sphere of human activity with its own separate norms. Realism broadly posits that explicating moral ideals to decide on political questions does not yield an understanding sensitive to the peculiarities of politics ([8]: 8). Across ethics and political philosophy, indeed across general thought about politics, we encounter these fruits of moral philosophy that take the moral as causal and proper normative source of the political. Yet for realists, politics cannot simply be applied moral philosophy.

The claim that politics is a distinct and autonomous sphere that cannot be reduced to ethics comes in different varieties. Realism harbours strong Machiavellian claims insisting that moral values, unlike political values, are ill-suited to derive normative political judgements from, as well as weaker claims that may accommodate morality as a normative source in politics but nevertheless assert that the political is a distinct realm and therefore irreducible to it. My point here is not to argue for any particular version but rather to draw attention to the broad realist critique of moralism and show how it can provide a more adequate venue to engage with AI as a political phenomenon. For whatever the version, there is a shared understanding that morality is a poor foundation for politics.

Perhaps some of the most discussed concerns and features relevant for political normativity among contemporary realists can be traced back to Thomas Hobbes. In his works, Williams identifies a fundamental political question that must be continuously 'solved' so that we may deal with other political matters. This comes down to securing 'order, protection, safety, trust, and the conditions of cooperation'—order and stability. The risk of putting much stock in this line of thought is leaning into the idea that might is right. What typically attracts realists to the notion of a first political question is, however, not this conservative slant but rather a way of thinking about the political apart from the moral. Williams himself introduced the 'Basic Legitimation Demand' (BLD) as a solution to the Hobbesian question: order and stability are necessary but not a sufficient condition for legitimate political power [26]. For Williams, the demands of legitimacy are in fact part and parcel of politics. To answer the first political question, an account of political power cannot simply be successful domination but must meet some kind of justificatory standard. Legitimacy as a political norm serves as such a standard by which the exertion of political power can be justified.

Williams' emphasis on legitimacy as being integral to the practice of politics is what makes him a realist. In the same vein, Williams thought that historical and social context must be considered when thinking about specific legitimation demands. What counts as sufficient reasons to meet those demands has to make sense in particular contexts. It is clear at this point that realism at the outset wants to pay close attention to the playing out of politics. But we're not out of the woods yet. To complicate things further, Williams adds a final, crucial qualification to his account of legitimacy: the 'critical theory principle'. Like much of critical theory, the principle functions as an immanent critique of the legitimacy of a political order. The acceptance of a justification cannot be produced by the coercive power in question. Anything else would render legitimacy as self-justification of the powers that be. While realists think about political norms such as prudence, security, and trust, recently Williams' notion of legitimacy seems to have gripped their hearts and minds the most. This is because it provides a novel way of thinking about the justification of political power that is not based on moral pre-commitments; a shift from the predominant focus within political theory on justice to another norm by which we might understand politics [23].

As Williams himself notes, the critical theory principle makes meeting the BLD very difficult (see also [22]). At stake here is proving and justifying an actual belief in legitimacy. 'Radical realists' are especially focused on this issue. They argue that our beliefs in legitimacy are often inaccurate and distorted, for example through past power relations [1, 3, 18]. As a partial remedy, they advocate for a (non-moral) epistemic approach in the form of an empirically informed ideology critique. This is meant to uncover held notions and intuitions as illusory and therefore epistemically flawed. If it's flawed, it's not justified; if it's not flawed, then it has shown the normative salience of the belief and is thereby (broadly) justified. In terms of ideology critique, one can think of the Marxist notion of false consciousness in which a subordinate class acts in accordance with the ideology of the ruling class, thereby not acting in its own interest, but also about the way the self-affirming intuitions of that same ruling class are shaped.

The above discussion of realism points towards three variants as outlined by Rossi [19]. (i) Ordorealism, which in line with the Hobbesian question prioritises the establishing of order; (ii) contextual realism, focusing on the contextual bounds of political power, such as the distinction between the personal and political; (iii) radical realism,

revolving around critiquing the intertwinement of power and knowledge. Importantly, realism and the non-moral political normativity it is concerned with involve some fidelity to facts, which differ in relevance per realist version. Ordorealists are most concerned with facts about providing feasible solutions to the question of order and stability, contextual realists with interpreting the points and purposes of relevant practices, and radical realists with facts about power relations [19]: 643). Radical realism has most notably no need for feasibility constraints but is rather concerned with an epistemic normativity meant to distinguish between acceptable and unacceptable legitimation stories. What remains now is to take stock of these different versions of realism vis-à-vis moralism and discern what kind of AI realism proves valuable.

3 AI realism

First, let us establish the consequential distinctions between realism and moralism. While realism broadly pays attention to feasibility concerns and contextual bounds, these are not necessarily excluded from moralist views. Moralist approaches like non-ideal theory take different facts and contexts into account as well. This more clearly applies to daily life in which thinking from the vantage point of moral norms does not preclude thinking about different contexts. The crucial difference is that the relevance of those facts and contexts must make sense with regards to pre-political, overarching moral commitments. This is the object of disagreement with realism and as a result realism pays heed to other (kinds of) facts. In sum, again, the main point of realism is that the political domain has a distinct normativity.

For clarity's sake it is worth pointing out that like the moralist camps outlined above, so too are the lines between ordorealism, contextual realism and radical realism blurry. Methodologies overlap, facts spill over and theoretical commitments mingle. These are family resemblances but distinguished at the point where separate families seem to form. Moreover, what realism holds in common with moralism is that it does not float free from ideology. It is not as though adopting realism means one escapes the stories that ripple and muddle our epistemic waters. In fact, the arguably ordorealist account by Gyulai and Ujlaki [9] of AI regulation rather uncritically buys into the idea of artificial general (super) intelligence. The more one is committed to facts relevant to providing order, such as those concerning feasible enforceability as in the case of these authors, the higher the chance of status quo bias. Here the result is buying into AI hype. All approaches however self-described therefore deserve the rigour of a critical view. That being said, moralism deserves its limelight due to its strong presence in political thinking. After all, if something is ideologically suspicious it is morality.

AI realism must therefore breathe radical realism. Without an epistemic filter based on a non-moral approach, one cannot distinguish between acceptable and unacceptable AI. The salience of such a filter is clear for AI design, development and employment. It is also relevant to meeting relatively new challenges posed specifically by generative AI, which co-narrates (legitimation) stories that are artefacts in and of themselves [3]. At the same time the broad realist attention to alternative political norms and contextual bounds chimes with the notion that AI does not happen in a vacuum. These kinds of critical insights benefit from realism insofar as it opens up new ways of thinking about the political nature of AI.

Take for instance AI value sensitive design (AI VSD). VSD is an approach aiming to integrate values into the design of technologies and it is increasingly being applied to AI. The basic VSD idea is that design without explicit attention for norms and values runs the risk of developing undesirable technologies [6]. VSD emphasises the context in which technology is embedded and the salience of respective values in that context. This has led to the worry that VSD, and therefore also AI VSD, amounts to a reduction to the values and interests of stakeholders, thus rendering AI design a reflection of stakeholder preferences [13].[2] Here we have an analogy with the self-justification of political power and hence a return to the legitimacy of AI technologies. Accordingly there is a need for an explanatory standard for justification. Yet, as is the case with most if not all AI design, this standard is sought in line with 'AI for social good', meaning a set of ethical principles [14, 24]. AI VSD provides us with much-needed work on the contextualisation of AI design and operationalisation of values, but due to its moralist underpinnings stops short at formulating desirable ends and norms that are (radically) politically salient. The consequence is buttressing AI design with means that are inadequate. In the worst case these ethics-first approaches are therefore complicit in why AI design is vacuous in the first place. At best they are insufficient.

What remains is an AI realism that builds on a non-moral epistemic approach to discern and approximate legitimate AI systems. Now, my detractors would point out that notwithstanding the supposed rigour of this approach, it is an evaluative one nonetheless. The question that accordingly rises is how it provides room for positive theorising. We might in fact get bogged down in negative critique without offering alternatives. These points and many more have been dealt with in realist literature more extensively than I can

[2] See Friedman et al. [7] for a broader sense of the VDS debate on i.a. issues related to accounting for power.

do so here, so I will forego responding in depth. Yet in my following final notes I hope to tacitly touch on them.

Radical realism is indeed characterised by critique but that does not mean it is not normative. It is a negative normativity if you will; what Rossi calls a form of genealogy that debunks or vindicates beliefs in legitimacy and political practices. In other words it can enable us to make normative political judgements. Of course, once beliefs in legitimacy are debunked or vindicated what will or should happen remains an open question. It is in this open space that ensues where further (theoretical) commitments play a key role and where AI realism can draw on the insights and facts pertaining to political normativity, bringing theory into the fold of politics proper. AI realism is thus not just a viable position because it politicises AI in its evaluative judgements, it is a viable *and preferred* position because it renders the open question of what ought to happen part of politics itself.[3]

The realist outlook enables us to formulate new norms and species of critique relevant to AI. A corresponding sensitivity to a non-moral epistemic approach can actually inform and change the design of AI by thinking along these norms. The future of AI realism thus lies in the conceptualisation, operationalisation and evaluation of realist norms in AI. Crucial to this purpose is taking cues from philosophy of technology and (machine) hermeneutics. This might for instance require rethinking parts of contemporary realism, such as the use of the pejorative 'illusory' concept of ideology in favour of an affective or productive one in order to integrate philosophical-technological insights. In this way AI realism can contribute to the broader realist debate. It holds a promise for kindling political thought needed to overcome moralisms. It is high time we make good on that promise.

Declarations

Conflict of interests The corresponding author states that there is no conflict of interest.

References

1. Aytac, U., Rossi, E.: Ideology critique without morality: a radical realist approach. Am. Political Sci. Rev. **117**(4), 1215–1227 (2022). https://doi.org/10.1017/S0003055422001216
2. Bietti, E.: From ethics washing to ethics bashing: a moral philosophy view on tech ethics. J. Soc. Comput. **2**(3), 266–283 (2021). https://doi.org/10.23919/JSC.2021.0031
3. Coeckelbergh, M.: Time machines: artificial intelligence, process and narrative. Philos. Technol. **34**, 1623–1638 (2021). https://doi.org/10.1007/s13347-021-00479-y
4. De Vynck, G., Oremus, W.: As AI booms, tech firms are laying off their ethicists. Washington Post. (2023). Available at: https://www.washingtonpost.com/technology/2023/03/30/tech-companies-cut-ai-ethics/. Accessed 21 May 2024
5. Floridi, L.: Translating principles into practices of digital ethics: five risks of being unethical. Philos. Technol. **32**, 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x
6. Friedman, B., Kahn, P.H., Jr.: Human values, ethics, and design. In: Friedman, B. (ed.) The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies, and Emerging Applications. L. Erlbaum Associates Inc (2002)
7. Friedman, B., Harbers, M., Hendry, D.G., Van den Hoven, J., Jonker, C., Logler, N.: Eight grand challenges for value sensitive design from the 2016 Lorentz workshop. Ethics Inf. Technol. **23**(1), 5–16 (2021). https://doi.org/10.1007/s10676-021-09586-y
8. Geuss, R.: Philosophy and Real Politics. Princeton University Press, Princeton (2008)
9. Gyulai, A., Ujlaki, A.: The political AI: a realist account of AI regulation. Inf. Társadalom. (2021). https://doi.org/10.22503/inftars.XXI.2021.2.3
10. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. Mind. Mach. **30**(1), 99–120 (2020). https://doi.org/10.1007/s11023-020-09517-8
11. Héder, M.: A criticism of AI Ethics Guidelines. Inf. Társadalom. (2020). https://doi.org/10.22503/inftars.XX.2020.4.5
12. Hatamleh, O., Tilesch, G.: Betweenbrains: Taking back our AI Future GTPublishDrive. Dr George Tilesch (2020)
13. Jacobs, N., Huldtgren, A.: Why value sensitive design needs ethical commitments. Ethics Inf. Technol. **23**(1), 23–26 (2018)
14. Jacobs, N., Huldtgren, A.: Why value sensitive design needs ethical commitments. Ethic. Inf. Technol. **23**, 23–26 (2021). https://doi.org/10.1007/s10676-018-9467-3
15. Lauer, D.: You cannot have AI ethics without ethics. AI Soc. **1**(9), 21–25 (2021). https://doi.org/10.1007/s43681-020-00013-4
16. Mittelstadt, B.: Principles alone cannot guarantee ethical AI. Nature Machine Intelligence. **1**, 501–507 (2019)
17. Munn, L.: The uselessness of AI ethics. AI Ethics. **3**, 869–877 (2023). https://doi.org/10.1007/s43681-022-00209-w
18. Prinz, J., Rossi, E.: Political realism as ideology critique. Crit Rev Int Soc Pol Phil **20**(3), 334–348 (2017). https://doi.org/10.1080/13698230.2017.1293908
19. Rossi, E.: Being realistic and demanding the impossible. Constellations **26**(4), 638–652 (2019). https://doi.org/10.1111/1467-8675.12446
20. Rossi, E., Sleat, M.: Realism in Normative Political Theory. Philosophical Compass. **9**(10), 689–701 (2014). https://doi.org/10.1111/phc3.12148
21. Sankaran, K.: What's new in the new ideology critique? Philos. Stud. **177**, 1441–1462 (2020). https://doi.org/10.1007/s11098-019-01261-9
22. Sleat, M.: Bernard Williams and the possibility of a realist political theory. Eur. J. Polit. Theory **9**(4), 485–503 (2010)
23. Sleat, M.: Legitimacy in Realist Thought. Political Theory **42**(3), 314–337 (2014). https://doi.org/10.1177/0090591714522250
24. Umbrello, S., Van de Poel, I.: Mapping value sensitive design onto AI for social good principles. AI Ethics. **3**, 283–296 (2021). https://doi.org/10.1007/s43681-021-00038-3

[3] A point one can similarly find in Foucault's project of 'ethos' and Derrida's aporetic account of the political.

25. Wagner, B.: Ethics as an escape from regulation: from 'Ethics-Washing' to ethics shopping? In: Being Profiled, pp. 84–89 (2018). https://doi.org/10.2307/j.ctvhrd092.18
26. Williams, B.: In the Beginning Was the Deed: Realism and Moralism in Political Argument. Princeton University Press, Princeton (2005)