What Can AI Ethics Learn from Anarchism?

AI ethics is experiencing two crises: It is disconnected from communities being impacted by AI and largely funded by and dependent on tech companies profiting from harms. Drawing on anarchist ideas, AI ethicists have recently started building tools to challenge this status quo. What else can AI ethics learn from anarchism?

By William Agnew

DOI: 10.1145/3665594

OPEN ACCESS

AI ethics is facing several crises. AI ethicists and institutions are often disconnected from those facing AI, data, and algorithmic harms. Too frequently our field learns about AI harms only after they occur and those people experiencing them organize and advocate for themselves. Rather than anticipating harms before they occur and working to mobilize and share resources with those experiencing harms, our field is often another institution that harmed communities must convince to help them. At worst, the deep entanglements of our field with the very institutions and corporations causing harms make AI ethicists an obstacle to impacted communities seeking justice.

The deep entanglement of AI ethics with industry and governments causing AI harms underscores the second crisis AI ethics faces. Even when AI ethicists are aware of AI harms, they often lack the power to address them. AI ethicists are largely employed by tech companies rushing toward building and deploying AI or they are at universities training students to work for those companies, receiving funding from those companies or industry-friendly government agencies. Dependance on institutions that prioritize technical AI development limits the range of solutions AI ethicists can advocate for. Not developing or decommissioning AI or datasets is rarely an option, and even AI ethics methods like dataset audits, debiasing, and feedback from impacted communities are limited in scope, duration, and depth of impact by profit and industry growth incentives.

A CALL TO ACTION

AI researcher Pratyusha Ria Kalluri charts a path forward by calling for AI ethics to focus on how AI systems shift power instead of whether they are fair or unfair to better assess and address the harms they cause [1]. Here we conceptualize power as the possession of agency and the ability to impact others' agencies [2]. Kalluri argues, while values like fairness, accountability, and transparency are desirable, they do not encompass everything we want. It is possible for an AI system to have these values and still be highly unethi-

Image by Azer Mess / Shutterstock.com

cal. Instead, rather than AI developers choosing what values AI should have and imposing them on everyone else, AI should improve, or at the very least not infringe upon, the agency of communities it is operating on. By increasing people's agency, we allow them to effect changes and live in ways aligned with their values and needs rather than choosing those values for them.

This call has been met in several ways. First, there has been a rapidly growing interest in participatory AI methods, or means of developing and governing AI, that effectively take into account the opinions of communities impacted by AI, thereby shifting power from AI developers to impacted communities. These methods have seen growing adoption in industry and academia as means for aligning outputs of generative models with user values and auditing or red teaming AI models. However, participatory methods in practice have been critiqued for "participation-washing," or giving the appearance of seeking meaningful feedback from impacted communities while, in reality, providing them with only very limited power over AI development.

Second, a growing number of researchers are calling for and building sociotechnical tools to enable resistance toward AI. Vincent et al. and McQuillan have proposed different organizations and strategies data owners and workers can use to collectively bargain with their data, for instance, threatening to withhold data from advertising companies [3, 4]. Das and Kulynych explored repurposing adversarial attacks on AI systems into defenses against unwanted AI use [5, 6]. Shan et al. created a protection tool to prevent facial recognition models from recognizing protected faces [7]. Shan et al. introduced Glaze, which protects artists by subtly modifying their art, so AI is unable to learn to mimic their styles [8]. Shan et al. also proposed Nightshade, which enables artists to resist AI art theft by subtly poisoning their images to decrease the performance of any AI image generator trained on them [9]. These tools shift power from AI companies to data owners and creators by allowing them to take back control over their data.

APPLYING ANARCHIST PRINCIPLES TO AI

These two current trends in AI ethics research, participation, and resistance, share much in common with anarchism. At a high level, anarchism is an opposition to coercive authority, including patriarchy, white supremacy, and imperialism. Anarchism encompasses a broad range of political thought and movements that have shaped and been shaped by many of the ideas underlying AI ethics. Despite this, AI ethics rarely

explicitly engages with anarchism. A search of the proceedings of FAccT and AIES, two top AI ethics conferences, returns only two papers with the word "anarchism," one a passing negative mention and the other as a description of the political valence of sentences in a natural language processing dataset. Anarchist thought closely aligns with current focuses in AI ethics on power, participation, and resistance, and indeed, many of these concepts and analyses have roots in anarchist thought. Given this close alignment, I argue AI ethicists should engage more with anarchism, echoing arguments others have made in HCI [10]. For the remainder of this article, I summarize several key concepts in anarchism that I believe would benefit AI ethics if they received wider consideration and deeper engagement. While no ideas are universal, I believe these core anarchist concepts point toward interesting possible futures of AI ethics.

The first anarchist principle is rejecting merely expanding individual freedoms, but rather focusing on dismantling institutions, such as capitalism or the police, that are responsible for the limitations of these freedoms in the first place. AI ethics frequently uncovers harms caused by algorithms and data used by powerful institutions, including biased criminal recidivism prediction algorithms, biased facial recognition algorithms leading to wrongful arrests of Black people, and predictive policing algorithms leading to disproportionate policing of historically overpoliced communities. However, AI ethics often stops at arguing these algorithms should be made less biased, or that algorithms and AI should not be used in certain settings. This reflects an implicit view that institutions should be reformed and their encroachment on individual freedoms reduced. Despite close study of the harms algorithms used by institutions, and by extension, the institutions themselves cause, AI ethicists rarely call for the abolition of police, border patrols, child protective services,[1] and other institutions.

By increasing people's agency, we allow them to effect changes and live in ways aligned with their values and needs rather than choosing those values for them.

AI ethics research is overwhelmingly conducted to critique or reform these institutions, but following the anarchist focus on dismantling rather than reforming institutions, we could imagine AI ethics research that explicitly studies and advances abolitionism. This AI ethics research could help examine and build policing alternatives, restorative justice mechanisms, or other abolitionist methods, asking the same critical questions around fairness, transparency, accountability, and effectiveness, but for revolutionary, rather than reformist, ends.

The second anarchist principle is prefigurative direct action for change. Direct action spans a wide variety of tactics united by not needing consent or permission from the powerful to take place and by not relying on changing the powerful to be effective. This stands in contrast to tactics like protests aimed at political leaders or legal action against unjust laws. Indirect actions have led to many foundational victories, but distinctions between direct and indirect actions are often fuzzy. Anarchists often critique indirect actions as requiring buy-in from powerful people and institutions, often perpetuating the very harms the actions are meant to address and as easy to co-opt since they are implemented through the institutions causing the harms. Direct actions, like mutual aid, the practice of sharing resources and labor within a community, address harms without support from powerful institutions. By building organizations and practices that are not dependent on support from such institutions, direct actions can better resist attempts to dismantle them by those very institutions. This points to what prefigurative direct actions are: Direct actions that not only address harms without permission or support from powerful institutions, but also help create new, radical futures where these powerful institutions are not needed. Community gardens are one such example; not only do they feed people without the need for agribusiness or for-profit grocery stores, but they create a small piece of a future where food production and distribution are not largely controlled by small numbers of institutions for profit. Recent AI ethics research on resistance represents a foray into direct action that should be expanded. Missing from even many of these resistance projects, however, is the ability to prefigure different, better futures. Existing tools can enable refusal and opting out. However, more remains to be done. What futures we should strive toward and what prefigurative direct actions we can take to move closer remain ill-defined for AI and art and many other areas impacted by AI.

The third anarchist principle is decentralization. Contrary to some misconceptions, anarchism is not a rejection of organization, but an emphasis on horizontal, decentralized forms of organization that do not allow too much power or authority to accumulate with any individual or organization. Decentralized organizations can form around mutual aid, disaster recovery, or other purposes top-down organizations fulfill. By limiting the distance between decision-makers and those being impacted by decisions, a decentralized organization at its best has increased accountability, transparency, and resistance to corruption. Decentralization has the potential to address the disconnects between academics and industry professionals with power in AI ethics and communities experiencing AI harms. Decentralization has already been used with some success. Queer in AI, an organization that works to reduce AI harms to queer people, has adopted a decentralized structure to improve the representation and engagement of different queer communities. Despite (or because of) being decentralized,

1 For further reading, see Dorothy Robert's *Torn Apart: How the Child Welfare System Destroys Black Families—And How Abolition Can Build a Safer World* (Basic Books, 2002).

Queer in AI is a highly active and coordinated organization, putting together well over 20 workshops, scores of socials, and managing a budget of nearly $100,000 each year, all while growing more diverse by empowering new organizers. Decentralization could also be applied to the design and governance of datasets and AI themselves. Under the current status quo, a powerful central state or corporation scrapes a massive dataset without consent and then trains a large model for their own purposes. In contrast, decentralized AI creation would require each data subject to provide active and informed consent for inclusion in datasets and give each data subject meaningful control over (including benefits from and the right to withdraw data from) AI created with their data.

The fourth anarchist principle is building solidarity and sharing resources with people everywhere in the struggle against abuse of power. Anarchist movements deeply care about struggles far from their local contexts. Calls for global days of action and proactive attempts to build solidarity across and power with wide arrays of communities are common in anarchist spaces. For example, an anarchist meeting against globalization in 1998 "included not only anarchist groups and radical trade unions in Spain, Britain and Germany, but a Gandhian socialist farmers' league in India (the KRRS), associations of Indonesian and Sri Lankan fisherfolk, the Argentinian teachers' union, indigenous groups such as the Māori of New Zealand and Kuna of Ecuador, the Brazilian Landless Workers' Movement, a network made up of communities founded by escaped slaves in South and Central America" [11]. AI ethics is relevant to a large and growing number of communities, including overpoliced populations, truckers, visual artists, musicians, voice actors, data workers, educators, prisoners, social media users, the workers and communities in AI supply chains, and many more people. Yet AI ethics rarely proactively engages with these populations to understand their needs, concerns, and experiences. AI ethics that embrace solidarity would actively build community with people being impacted by AI and seek to share our resources—knowledge, connections, prestige, and funding—with them to address the harms they are experiencing and help them build liberatory AI futures. This would help address one of the key challenges AI ethics faces by shifting where AI ethicists are located and who they are accountable to by moving from industry and academia to impacted communities.

The fifth anarchist principle is radical imagination. AI ethics largely emerged from the critique of existing AI systems. This work was incredibly valuable for building the field and creating awareness of the harms AI can cause and remains important for exposing and challenging abuses and bad practices. However, much of AI ethics remains rooted in understanding what the future should not be rather than imagining what the future should be. This lack of a positive (used here in the sense of constructive rather than good) vision limits the appeal and coherency of AI ethics. Capitalist and TESCREAL ideologies both present positive visions of the future [12]. Though these visions are rightly critiqued as unrealistic snake oil rooted in extractive practices and eugenics, they dominate imaginations of which futures are possible or likely, stymying any action or organizing for different, better futures. AI ethics should work to imagine what futures we want. AI ethics research should include, if not primarily focus on, how and if AI can help us reach those futures. Despite many potential pitfalls, there is significant potential in AI for social good, and especially in the development of new scientific and engineering tools. However, the relation AI ethics has with AI for social good is primarily one of critique. AI ethics should continue holding AI for social good accountable, but we should also work with these areas to imagine and build better futures.

These five anarchist principles–dismantling harmful institutions, prefigurative direct action, decentralization, building solidarity, and radical imagination–are not new to AI ethics, nor are they comprehensive of or unique to anarchism. However, I argue much of AI ethics does not reflect these principles, and even the strands that have adopted some principles would benefit from deeper engagement with the long histories of struggle and theoretical work in anarchism. I believe AI ethics can avoid repeating many mistakes and grow into a field more closely aligned with the ideals of justice and liberation through careful interdisciplinary study of AI ethics and anarchism.

References

[1] Kalluri, P. Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583, 7815 (2020),169–169.

[2] Young, I. M. *Justice and the Politics of Difference. The New Social Theory Reader*, 2nd edition. Routledge, 2020, 261–269.

[3] Vincent, N. et al. Data leverage: A framework for empowering the public in its relationship with technology companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21). ACM, 2021, 215–22; https://doi.org/10.1145/3442188.3445885

[4] McQuillan, D. *Resisting AI: An Anti-Fascist Approach to Artificial Intelligence*. Bristol University Press, 2022.

[5] Das, S. Subversive AI: Resisting automated algorithmic surveillance with human-centered adversarial machine learning. Resistance AI Workshop at NeurIPS, vol. 4. 2020.

[6] Kulynych, B. et al. POTs: Protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (FAccT '20). ACM, 2020, 177–188; https://doi.org/10.1145/3351095.3372853

[7] Shan, S. et al. Fawkes: Protecting privacy against unauthorized deep learning models. In *Proceedings of the 29th USENIX Conference on Security Symposium* (SEC '20). USENIX Association, 2020, Article 90, 1589–1604.

[8] Shan, S. et al. Glaze: Protecting artists from style mimicry by text-to-image models. In *Proceedings of the 32nd USENIX Conference on Security Symposium* (SEC '23). USENIX Association, 2023, 2187–2204.

[9] Shan, S. et al. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. arXiv:2310.13828 [cs.CR]. 2023.

[10] Keyes, O., Hoy, J., and Drouhard, M. Human-computer insurrection: Notes on an anarchist HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI '19). ACM, 2019, 1–13; https://doi.org/10.1145/3290605.3300569

[11] Graeber, D. The new anarchists. *New Left Review* 13 (Jan/Feb 2002); https://newleftreview.org/issues/ii13/articles/david-graeber-the-new-anarchists

[12] Torres, E. P. The acronym behind our wildest AI dreams and nightmares. Truthdig. June 15, 2023; https://www.truthdig.com/articles/the-acronym-behind-our-wildest-ai-dreams-and-nightmares

Biography

William Agnew (he/they) uses research and organizing to challenge power and technologies that concentrate power, and empower marginalized communities over tech, data, and AI impacting them. Agnew is a CBI Postdoc Fellow at CMU studying AI ethics, critical AI, community mobilization, and 3D vision.

1528-4972/24/06 $15.00