Resistance as a Technomoral Virtue for Our Times

Nitya Mandyam
University of Cambridge

This paper critically examines differing responses to the harms of Artificial Intelligence (AI) and explores the ethical dimensions of resisting AI. It draws on case studies of three distinct AI-related harms - bias in language models, the environmental impact of large AI models, and job displacement due to automation, to define what constitutes an act of resistance. Applying a virtue ethical framework to the question of resistance, I posit that resistance is a technomoral virtue in our times and can act as a creative, generative, and dynamic principle in producing technologies that engender human flourishing.

Author's Note: The author would like to acknowledge the significant proliferation of data centres and the simultaneous rise in resistance towards them from local communities, activists, and policymakers, since creating this manuscript in late 2024. The data centre resistance highlighted in this paper at Cerrillos, Santiago, Chile (2020-24) is a forerunner of a multitude of similar resistances since that are occurring across the globe, with data centres becoming a focal point for AI resistance. A small but significant number of these conflicts have culminated in the imposition of moratoriums and bans on data centre development by local governments.

Keywords: Resisting AI, Harm Reduction, Virtue Ethics, AI Extractivism, Technomoral Virtue

Introduction

Recent methods of producing Generative Artificial Intelligence (AI) have raised concerns about harms ranging from the individual to societal, including threats to privacy and autonomy, amplification of bias in training data, rising unemployment due to automation, copyright law violation, power imbalances caused by consolidation of data and computer resources, and environmental damage caused by increased energy consumption demand. While AI engineers, researchers and legislators have sought harm reduction to mitigate some of these concerns, activists favour resistance (Narayanan, 2023). The primary argument against the harm reduction approach is that it risks legitimising unjust practices and outcomes. Resisting AI, on the contrary, can impede technological progress and the benefits it could bring. This paper focuses on the resistance piece of this debate and uses the virtue ethical framework to examine the following questions: When are acts of resistance justified? What kinds of resistance are unethical?

Section 1 of this paper lays out the different dimensions of the debate with concrete examples illustrating responses to three specific AI harms as important case studies: bias in language models, the environmental footprint of large AI models and job displacement due to automation. Aided by these examples, I will define what constitutes an act of resistance in AI, a definition that will be revisited later and refined to understand when resistance is virtuous. Section 2 examines this issue through the different normative ethical frameworks and lays out an argument for why virtue ethics is most suited to help resolve this question. Section 3 examines the virtues demonstrated by different parties in acts of AI resistance. Using philosopher Shannon Vallor's framework of technomoral wisdom (Vallor, 2016), it adjudicates whether these are holistically unvirtuous acts or steps toward resisting an unvirtuous ecosystem of AI production.

Based on these analyses, I present this essay's primary argument in Section 4: that resistance is a necessary technomoral virtue for our times. I reframe resistance as a creative and dynamic player in honing technologies that can prevent unsafety and act as a guiding principle in disrupting the structural dynamics that produce contemporary AI. I present resistance as a creative, constructive and generative principle for building moral machines in an unvirtuous ecosystem and envision the concrete impacts of resistance on current and new technologies.

1. Resistance as an Alternative to Harm Reduction

1.1. Some Examples of AI Harms

What constitutes technological harm can be deeply subjective, depending on the morality of the time, the metaethics of the people in question, and the epistemic stance of the arbitrator. Instead of attempting a universal definition of what constitutes "AI harm", this section relies upon some tangible scenarios where there is broad consensus on the presence of harm. Three well-acknowledged problematic side effects of contemporary AI are:

The Perpetuation of Bias in Language Models: This involves the different forms of oppression (among them racism, misogyny, transphobia and classism) inherent in the training data of these models and their subsequent persistence in the text generated by them (Bender et al, 2021, Gallegos et al, 2024).

Sustainability Concerns: The massive amounts of computing power required to pre-train, store, fine-tune, and infer from Generative AI models necessitate an increase in the consumption of environmental resources (A. S. Luccioni et al, 2024).

Displaced Employment due to Automation: It has been speculated that as AI models improve, employers will seek to cut costs by replacing human labour with AI, which adversely affects workers and increases economic inequality (Hatzius et al, 2023).

1.2. Harm Reduction

Harm reduction seeks to redress these harms by working within the system of contemporary AI production through technical and non-technical interventions. Technical interventions involve algorithmic approaches implemented while creating and deploying a model. Non-technical interventions can take the forms of human oversight, building and adhering to ethical codes within tech corporations, legislation by local and international authorities and regulations implemented by nation-states. Applying these initiatives to the three harms described above:

Mitigating Bias: A few techniques used to counter bias in language models are data curation prior to pretraining via toxicity filtering performed by annotators, Reinforcement Learning based on Human Feedback (RLHF) to align models to human preferences before model deployment (Lindström et al, 2024), and countering more downstream harms post-deployment by providing clear disclaimers allowing for user feedback.

Sustainability Initiatives: Researchers concerned about the sustainability implications of Generative AI models seek to quantify the environmental toll caused by their training and deployment and then find technical ways of minimising it. Techniques such as quantisation and low-rank adaptation reduce the computational power required for storing and fine-tuning models (Dettmers et al, 2023). A non-technical solution is to build data centres in areas powered by renewable energy. More recently, researchers have suggested drawing upon existing mechanisms of green policy, such as energy ratings, not unlike light bulbs or automobiles (Luccioni et al, 2024).

Reframing Jobs: Just as the industrial revolution created new job categories, AI purveyors predict that after a possible lull in employment opportunities, new jobs will emerge around managing Generative AI technologies, even as there is rising scepticism about this stance (Acemoglu et al, 2022).

1.3. Acts of Resistance

If harm reduction approaches are characterised by cooperation and constructive course corrections, acts of resistance are inherently non-cooperative and often seek to halt the course of implementation of a technology. Three singular events from recent history can be highlighted as being acts of resisting the harms mentioned at the beginning of this section:

The Firing of AI Researcher Timnit Gebru (2020): Two years before ChatGPT took the world by storm in late 2022, an AI Researcher at Google, Dr. Timnit Gebru, coauthored a paper (Bender, 2021), along with colleagues and academics outside of Google advocating for a halt to the development of large language models (LLMs) on the grounds of implicit bias. This suggestion directly contradicted Google's AI goals and would eventually lead to their firing.

Chilean Data Centre Protests (2020-24): (Arellano et al, 2020) Google had planned to build a Data centre in Cerrillos, a commune in Santiago, Chile, that would require 169 litres of water per second to cool its servers. Several parts of Chile have been severely drought-ridden in the past decade. Santiago, being an aquifer, was expected to source water for many regions in the country. Residents were baffled to see the project approved and resisted it via community organising through a group known as Mosacat. The resistance has delayed the project significantly, culminating in a Chilean court pulling the approval granted to Google in February 2024.

The Writers Guild of America (WGA) Strike (2023): 2023 saw two significant strikes by unionised writers (WGA) and actors (SAG-Aftra) demanding better wages within Hollywood's notoriously unequal studio system for nearly six months from May to November (Silberling, 2023). AI was at the heart of two negotiating points in the strikes - the first was a demand to use Generative AI as a tool and not a replacement within the industry, and the second was to stop copyright misuse of artists' creations.

1.4. Defining Resistance

The three case studies considered in this paper were chosen to reflect how acts of resistance can occur at chronologically different phases of AI development - before, during and after AI development. The Chilean environmental activism can be seen as resisting the planning and resourcing phase of AI development, Timnit Gebru's defiance can be considered as demonstrating resistance during AI development and the WGA strikes are reflective of resisting AI usage post-deployment.

The *resistors* include diverse players such as critics of AI (often tech workers themselves), frontline workers, people whose livelihoods are threatened by automation, anti-capitalists, marginalised communities more directly impacted by biases in AI systems and environmentalists. Additionally, the forms of resistance are very diverse. For instance, Timnit Gebru's defiance was a case of workplace insubordination leading to the termination of their employment. Data centre activism in Chile or Uruguay has taken the form of community organising and protesting via demonstrations and referendums. As a negotiating tactic, the SAG-Aftra and Writers Guild strikes disrupted everyday work routines, causing significant economic discomfort to the strikers and their bosses.

While these are all very different forms of resistance, they share a throughline: to disrupt "business as usual". Often, the goal is to cause sufficient discomfort to a tech company to incentivise them to change their actions. However, this can and often does come at the cost of disrupting unexpected parties, such as company workers or technology consumers. Resistance also comes at a cost to the person resisting because, by its very definition, they are counteracting a force larger than themselves.

2. Ethical Perspectives on AI Harms

2.1. Questions of Proportionality and Threshold

Many supposed harms of AI are often themselves subject to fierce debates about the extent of their harmfulness. This debate, while slightly different from the one this paper tries to address, is nonetheless entangled with the question of resistance. Resistance can be reasonably perceived as an excessive response if the harms are small or slow enough to take effect that an intervention (technical or otherwise) can mitigate it down the line. Regulation via legislation or building guardrails as and when problems crop up post-deployment can, in turn, seem ineffectual for those whose livelihoods or dignity are affected by immediate harm. In other words, disagreement on the *proportionality of harms* is a definitive factor driving the two approaches here.

A consequentialist approach that can empirically evaluate the proportionality of harm caused would presumably be an apt choice to resolve this debate. This is a worthy and meaningful endeavour that many technologists have and should continue to undertake, but it does not address this debate holistically. A crucial missing piece in this approach is the extent to which power, relationality and epistemic views shape the development of sociotechnical systems like AI. In his talk titled "Resistance or Harm Reduction?", Prof. Arvind Narayanan corroborates this view (Narayanan, 2023): *"Computer scientists, in particular, overwhelmingly prefer the harm reduction approach. It is a better match for our technical skills; we understand the details of these products, and we can help improve them. It is also a better match for the fact that we often receive money from tech companies, and we are perhaps uncomfortable with the kind of positions that resistance involves taking and the kinds of things that it involves doing."*

Many harm reduction approaches within AI already tend to have a utilitarian flavour. An action-oriented inclination towards analysing outcomes, seeking tangible optimal solutions, and working within institutional constraints to maximise profit suits the venture capital-driven corporate climate within which contemporary AI is created. In such a culture, resistance is intrinsically counterproductive in the way that productivity has been defined and rewarded. For a consequentialist to endorse resistance, two logical requirements must be met - that the harms outweigh the benefits of a particular technology and that the long-term outcome of resistance will maximise utility. The burden of proof here relies on predicting the course of a technology, and this is a difficult task even for the technically proficient!

Another way to approach responses to AI harms is to identify quantifiable thresholds that can be used to aid technical development. Deontological approaches find their place in this debate in harm reduction approaches through technical interventions by way of hard-coding machine ethics via programmable rules (Hooker and Kim, 2018). In developing the proper rules to mitigate harm, quantifying impact and drawing suitable thresholds based on that becomes essential. The primary difficulty here is that this process requires cooperation with the corporations responsible for creating AI products to demonstrate transparency and proactiveness in quantifying these harms. Companies prefer opacity about their products' inner workings to stay abreast of their competitors and secrecy about possible harms to avoid unsettling their consumers (Whittaker, 2021).

Additionally, the most common critique of deontology, which is that it falls short in its ability to handle genuine moral dilemmas (Timmermann, 2013) is of special relevance to the resistance debate in AI as it is challenging to ascribe overarching imperatives to different players especially when competing interests create diverse moral dilemmas. Deontology is nonetheless valuable in narrower contexts when one can clearly define the parameters of the moral conundrum. For instance, one could use a deontological framework to adjudicate whether data workers who perform tasks such as annotation, toxicity-filtering and fact-checking might be justified in organised action against their employers (Mathenge, 2024) depending on their compensation, working conditions and the psychological effects of the work.

2.2. A Question of Virtue?

There are two key ways this debate is complex to resolve using consequentialist and deontological frameworks. The first is a blind spot in both ethical frameworks to account for how power dynamics and relationality shape the course of technologies and, consequentially, the different responses towards them. The second is the inherent difficulties in predicting the extensive social effects of AI systems, a phenomenon referred to as "*acute technosocial opacity*" by virtue ethicist Shannon Vallor (Vallor, 2016) .

Vallor argues that the pace of technological development in a globalised world order makes it impossible to gauge our futures or even assess where we are now. Techno-optimists tend to claim that we are "*approaching the next dizzying explosion of technosocial progress.*" At the same time, techno-pessimists prefer to imagine that we are "*teetering on a precipice awaiting a calamitous fall*", and the consensus on the present moment remains fuzzy (Vallor, 2016). This fuzziness, in turn, makes figuring out the right thing to do, i.e., finding an ethical solution, a difficult task. Vallor proffers that we move away from seeking solutions by optimising imagined outcomes or attempting to build robust ethical rules. Instead, she urges us to think about the kinds of moral character, i.e., the *virtues*, that can promote human flourishing in the face of technosocial opacity.

Virtue ethics views moral character in agents as not an inherent or static trait but rather something developed and nurtured through consistent practice and habituation. It is not concerned with arbitration based on outcomes (consequentialism) or adherence to predetermined rules (deontology). It does posit that actions that come from a place of moral habituation will likely lead to good outcomes and prescribes many of the values that deontology does. However, it is more interested in reorienting the actions of moral agents (human or machinic) towards behaviours that promote human flourishing.

Virtue ethics thus allows for a dynamic, flexible, and context-dependent approach to resolving the debate around harm reduction and resistance. Harm reduction and resistance are both responses to the distinct anxieties created by the acute technosocial opacity generated by contemporary AI. Moving away from the impossible task of answering questions about proportionality or threshold of harms needed to resist AI prima facie, the coming section will examine the moral behaviours of the diversity of players involved in building and deploying AI, mitigating its harms, and resisting it, through a virtue ethical lens.

3. A Virtue Ethical Lens on Acts of Resisting AI

3.1. Moving from a Technoneutral to a Technomoral Ethos

"Technologies neither actively determine nor passively reveal our moral character, they mediate it by conditioning, and being conditioned by, our moral habits and practices." (Vallor, 2016)

In her seminal work, Technology and the Virtues, philosopher Shannon Vallor draws parallels between three global ethical traditions, Aristotelian, Confucian and Buddhist, to develop a unified and pluralistic virtue ethical framework. Infusing this framework with principles drawn from feminist ethics, as well as a sound understanding of the nature of technological development, Vallor lays out a taxonomy of *technomoral virtues* to guide human flourishing in the face of technosocial opacity. These values are honesty, self-control, humility, justice, courage, empathy, care, civility, flexibility, perspective, magnanimity and technomoral wisdom.

Vallor contends, "*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) The word *technomoral* reflects the inseparable and intricate ways our morality and technology inform and influence one another. It implicitly presents a counterpoint to a deterministic view of technological development because as long as there are many technomoral choices at every step in building a technology, we are not under an imperative to march towards a singular technological predestination.

3.2. On the Accumulation of Moral Debt in Generative AI

If we accept that there are technomoral choices en route to assembling a technology, Vallor contends that we accrue "moral debt" when we put off dealing with moral problems (Renieris, 2022). In software engineering, when the speed of delivery is prioritised over writing robust code, systems begin to accrue "technical debt". In more old school terms, choosing the "easier way out" has consequences. Moral debt is not dissimilar and is specifically very evident in a sociotechnical system like Generative AI.

Consider the plurality of resources that facilitate building and deploying a model like GPT-4: financial investment from the Global North, highly skilled technical workers engaged in building testing neural network architectures within OpenAI, vast swathes of digital information scraped with and without consent that serve as training data, labourers in the Global South who perform toxicity filtering and data annotation, the rare earth metals that power microchips within Graphical Processing Units (GPUs) and the water resources needed to cool data centres performing computation. Making ethical compromises in accumulating these resources to prioritise cost-cutting or faster delivery of AI systems, accrues moral debt. The downstream harms are more complex to mitigate as any structural injustices overlooked in the building and deployment phase will persist.

3.3. Insubordination, Activism and Strikes: The Morality of Resistance

Revisiting the examples of acts of resistance provided earlier in this paper, we can now take a deeper look at what was being resisted there. In the first example, two well-respected AI researchers at Google, Timnit Gebru and Margaret Mitchell, advocated for reconsidering Large Language Models (LLMs) as the course of action to pursue in AI development (Bender et al, 2021). Their examination of bias in training data and the environmental costs of pre-training led them to conclude that this was an ethical dead-end. Their subsequent disagreement with Google and their eventual departure from the company were subjects of newsworthy headlines. There is, however, little consensus on the sequence of events between the authors and the company. (Hao, 2020)

In the second example, an activist group in Chile named "Mosacat" led a campaign that very recently resulted in a temporary overturning of the approval for Google to construct a Data centre in an aquifer at Cerrillos, Santiago. Dr Sebastián Lehuaedé (Marx, 2024) studies data centre activism and says the following about their strategies: "*They went to street markets, they did workshops with the local community, they put up posters in bus stops in order to inform the community about this situation.*" The community's water needs and the dent the data centre's water consumption would create had yet to be analysed when the local environmental authority approved the project, a case of clear and wilful political negligence.

These examples are vastly different cases that resist different aspects of AI production and deployment. While Google seems to be the common denominator, it is not unique in the behaviour that engendered the resistance. Endorsing Dr. Gebru's research (Bender et al, 2021) would be sufficient cause to pause the primary course of AI development the company was pursuing and lose its edge in a hyper-competitive market. In the issue of building a resource-intensive structure in a drought-ridden country, it is worth asking: Is it morally incumbent on Google to inform locals about the amount of water required by a data centre? Is it fair that the moral burden of research and communication falls on a corporation when local politicians/businesspeople are more than willing to compromise on their community's needs? This situation reflects the complex nature of building Generative AI and reinforces Vallor's stress on considering what it would mean to make technomoral choices.

The other common denominator in these acts of resistance is the resistors' courage, care and magnanimity. Whether one agrees with Timnit Gebru's perspectives on LLMs technically or ethically, there is undeniable courage in going against one's self-interest to advocate for others. Similarly, activists in Chile demonstrate care and magnanimity in their willingness to educate and empower local populations unaware of the extent to which their lives might be affected by the construction of the data centre.

4. Reframing Resistance as a Generative Virtue

4.1. Resistance as a Technomoral Virtue

A virtual ethical lens sheds light on our current AI ecosystem as product of complex interactions between multiple agents, technologies, and systems. An ecosystem that incentivises entities within to prioritise their interests over human flourishing does not demonstrate technomoral wisdom. In an unvirtuous ecosystem, resistance automatically emerges as a corrective principle to reorient within our technomoral futures. Resistance is then a virtue for our times for tech workers, activists, venture capitalists, and non-tech-savvy consumers alike.

French philosopher Paul-Michel Foucault famously posited that resistance exists wherever there is power and can never be external to it. Resistance as an inherently oppositional idea might make it sound less like a virtue and more like a behaviour one might want to avoid. But philosopher Joanna Oksala reinterprets Foucault's view to underscore resistance as a dynamic and creative principle to organise ourselves with: *"... resistance against* *domination and the normalising effects of power/ knowledge must advance on multiple fronts. It consists of creative transformations of the self, communal forms of counterconduct, and critical interrogation of our present.*" (Lawlor, 2014)

Many of the traditional virtues contain a seed of oppositional energy: courage only makes sense in the presence of fear, humility counters hubris, and flexibility opposes rigidity. Virtues have an inherently relational quality, and resistance is no different. Vallor states that "*Any plausible virtue ethic must be rooted in a more or less stable range of human moral capacities, and in the absence of some radical alteration to our basic psychology, there is no reason to think that humans will suddenly acquire a wholly new repertoire of moral responses to the world.*" (Vallor, 2016) Resistance satisfies this requirement as it has existed as long as humans have sought power. Without engaging in creative acts of resistance, we would have never won many of the human rights we have today against forces of systemic oppression. When faced with technologies that risk entrenching patterns of oppression within AI systems, a phenomenon Vallor refers to as the "*AI Mirror*", resistance becomes a technomoral virtue for our times.

Author and historian of science and technology Jean-Baptiste Fressoz has written extensively about histories of technological risk and argues that "*what in history of technology is traditionally put in the footnotes, under the label of 'opposition' or 'resistance' to progress turns out to be essential for the 'shaping' of safer technological systems.*" (Fressoz, 2007) He bases this conclusion on multiple case studies about technological controversies such as deforestation due to development, side-effects of vaccination, and pollution due to chemical industries, where resistance shaped safer and more robust technological outcomes, thus corroborating Oksala's argument for resistance as a necessary principle for our advancement.

4.2. When is Resistance Unvirtuous?

Virtue ethics prioritizes context and nuance over universality and no virtue is thus a categorical imperative. Resistance can only be deemed a virtue if it can contribute to eudaimonia meaningfully, and this requires a stance against all forms of oppression. The Greek translation of "resistance" is antistasi (or "anti-fascism" more literally), a word the was strongly associated with the Greek National Resistance during World War II, one of Nazi-occupied Europe's most robust resistance movements. Resistance becomes unvirtuous, however, the moment it coopts the oppressor's strategies of violence, apathy, coercion and hubris. History is a testament to movements against oppression that are often an amalgamation of unconsidered, reactionary and violent forms of resistance and revolutionary, liberatory and non-violent forms of resistance. Specifically, movements of technological resistance in the 19th century can be considered to contain both.

There has been renewed scholarly interest in examining movements of technological resistance during the Industrial Revolution in Europe. Journalist Brian Merchant's work (Merchant, 2023), for instance, reclaims the Luddite movement in Great Britain in the early 1800s from a maligned technophobic and regressive characterization and recasts it as an informed labour struggle by an impoverished working class in a techno-feudal state that sought to deprive workers of their wages by using automation as an excuse. While the Luddites undoubtedly engaged in violent resistance and embraced their monikers as machine-breakers, Merchant traces the arc of systematic oppression faced by the Luddites for decades before the resistance turned violent, including a deprivation of livelihoods and sometimes, actual lives of mill workers. Merchant points to the chilling parallels between the callousness of the feudal class to the mill workers
in the 19th century and the apathy of tech companies to gig workers or data workers in the Global South.

Critical scholarship on AI has noted the undercurrents of colonialism (Muldoon and Wu, 2023), extractivism (Hogan, 2024) and eugenics (Chan, 2025) present in the mechanics of AI development today. This presents a case for looking into the modes of resistance that have engendered long-lasting positive change in struggles against the same. History is full of cautionary tales of revolutions that have turned violent and yet, there are inspiring examples of predominantly nonviolent resistance movements to be found in the last century in the anti-colonial resistance in India, anti-Apartheid struggles in South Africa and the civil rights movement in the United States. Some slogans from these movements that can guide us in characterising resistance as a virtue would be the Sanskrit word "*Satyagraha*", which translates to "*holding firmly to truth*" or the Nguni slogan, "*Amandla Ngewethu*", or, "*Power to the people!*"

Virtue ethics' insistence on context, nuance and presence mandates that different technomoral virtues acquire temporal importance based on what is most called for at that moment. Resistance is thus a virtue for our times *at this moment in human history*, and I posit that a thawing of this moral apathy would cause it to fade into the background to assume a quiet watchful eye that only springs into action when oppressive dynamics enter the picture. Until such a time, the tangible examples of ethical AI resistance laid out in the following section are worth contemplating.

4.3. Dimensions of Ethical AI Resistance

In adherence to Vallor's call for virtues to be rooted in a stable moral disposition, it is worth examining what modes of ethical AI resistance can look like today. The table below presents a brief yet diverse set of examples of movements and actions in the world of technology that would qualify as the same.

Table 1: Examples of ethical AI resistance

The digital degrowth movement	Political/Environmental	AI Accelerationism
Unionization/ Labour organizing	Economic/Sociopolitical	Oppressive power dynamics in tech companies
Grassroots activism	Environmental/Decolonial	Extractivist logics in AI development
Antitrust legislations	Legal	Monopolisation of AI development
Refusal/ Opting out	Individual	Surveillance logics in AI development
Strikes/ Non-cooperation	Collective	AI hype, AI misuse, Livelihood deprivation
Adversarial interoperability[1]	Commercial	Control of technological access
Whistleblower protection	Employment Rights	Corporate opacity, retaliatory practices

While this table is by no means exhaustive, there is nonetheless a striking absence of what resistance could look like in technical product development in AI. Many of the examples above fall under Blair Attard-Frost's "*AI Countergovernance*" framework (Attard-Frost, 2023), an alternative to traditional AI governance. Arguing that today's AI technologies "*are all by-products of a failed AI governance regime and the logics of its constituent organisations*," countergovernance is less concerned with technical intervention and more with opposing the political and economic ecosystems that foster them. To imagine what is possible in AI in the technical dimension when it is nurtured by technomoral choices grounded in resistance every step of the way, we can look towards a rare technological application, Signal, the encrypted messaging service.

Built by a non-profit foundation that seeks to provide surveillance-free messaging networks to everyone, developing Signal required a complete non-reliance on existing software and building entirely from the ground up, as the most widely used software packages of today, by default, collect and store metadata (Whittaker and Lund, 2023). Signal is the safe application of choice for movements across the globe that are resisting authoritarian regimes, protesting genocides or organising climate activism. Signal's commitment to a surveillance-free ethos is evident in all facets, from conception to funding to deployment and further, in its continued refusal to cave to demands from nation-states that do not permit surveillance-free communication (Greenberg, 2024) providing us with a valuable roadmap to build AI grounded in the virtue of resistance.

Conclusion

This paper has examined the ethical dimensions of resistance to AI through a virtue ethics framework. While harm reduction approaches operate within existing systems to mitigate AI's negative effects, resistance represents a necessary corrective force in an ecosystem that mirrors historic patterns of oppression. By reframing resistance as a technomoral virtue rather than mere opposition, we recognize its generative potential to shape safer, more ethical technological systems. In confronting AI's challenges, resistance provides a pathway to technologies that enhance rather than diminish human dignity and flourishing.

References

Acemoglu, D., Autor, D., Hazell, J., & Restrepo, P. (2022). Artificial Intelligence and Jobs: Evidence from Online Vacancies. *Journal of Labor Economics*, 40(S1), S293–S340.

Arellano, A., Cifuentes, L., & Ríos, C. (2020). The Dark Areas of the Environmental Evaluation that "Blindly" Authorized the Google Megaproject in Cerrillos.

Attard-Frost, B. (2023, December 16). AI Countergovernance.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

Chan, A. S. (2025). Predatory Data: Eugenics in Big Tech and Our Fight for an Independent Future. University of California Press.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient Finetuning of Quantized LLMs (arXiv:2305.14314). arXiv.

Doctorow, C. (2019, October 2). Adversarial Interoperability. Electronic Frontier Foundation.

Fressoz, J. (2007). Beck Back in the 19th Century: Towards a Genealogy of Risk Society. *History and Technology*, 23(4), 333–350.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey (arXiv:2309.00770). arXiv.

Greenberg, A. (2024). Signal Is More Than Encrypted Messaging. Under Meredith Whittaker, It's Out to Prove Surveillance Capitalism Wrong. Wired.

Hao, K. (2020, December 3). A Leading AI Ethics Researcher Says She's Been Fired from Google. *MIT Technology Review*.

Hatzius, J., Briggs, J., Kodnani, D., & Pierdomenico, G. (2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). Goldman Sachs.

Hogan, M. (2024). The Fumes of AI. Critical AI, 2(1).

Hooker, J. N., & Kim, T. W. N. (2018). Toward Non-Intuition-Based Machine and Artificial Intelligence Ethics: A Deontological Approach Based on Modal Logic. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 130–136.

Lawlor, L., & Nale, J. (Eds.). (2014). Resistance. In The Cambridge Foucault Lexicon (1st ed., pp. 432–437). Cambridge University Press.

Lindström, A. D., Methnani, L., Krause, L., Ericson, P., Troya, Í. M. de R. de, Mollo, D. C., & Dobbe, R. (2024). AI Alignment through Reinforcement Learning from Human Feedback? Contradictions and Limitations (arXiv:2406.18346). arXiv.

Luccioni, A. S., Jernite, Y., & Strubell, E. (2024). Power Hungry Processing: Watts Driving the Cost of AI Deployment? The 2024 ACM Conference on Fairness, Accountability, and Transparency, 85–99.

Luccioni, S., Gamazaychikov, B., Hooker, S., Pierrard, R., Strubell, E., Jernite, Y., & Wu, C.-J. (2024). Ratings—So Why Can't AI Chatbots? *Nature*.

Marx, P. (2024, February 6). How to Stop a Data Center. *Disconnect*.

Mathenge, R. (2024). Data Workers Organizing – The African Content Moderators Union. In: M. Miceli, A. Dinika, L. Dingebacher, C. Salim Wagner, and K. Kauffman (eds), The Data Workers' Inquiry. Creative Commons BY 4.0.

Merchant, B. (2023). Blood in the Machine: The Origins of the Rebellion Against Big Tech. Little, Brown.

Muldoon, J., & Wu, B. A. (2023). Artificial Intelligence in the Colonial Matrix of Power. *Philosophy & Technology*, 36(4), 80.

Narayanan, A. (2023, October 27). Resistance or Harm Reduction?

Nguyen, A., & Mateescu, A. (2024). Generative AI and Labor: Power, Hype, and Value at Work. *Data & Society Research Institute*.

Renieris, E. (2022). Virtue Ethics and Technomoral Futures [Broadcast].

Silberling, A. (2023, September 26). The Writers' Strike is Over; Here's How AI Negotiations Shook Out.

Timmermann, J. (2013). Kantian Dilemmas? Moral Conflict in Kant's Ethical Theory. *Archiv Für Geschichte Der Philosophie*, 95(1), 36–64.

Vallor, S. (2016). Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting. Oxford University Press USA.

Whittaker, M. (2021). The Steep Cost of Capture. *Interactions*, 28(6), 50–55.

Whittaker, M., & Lund, J. (2023, November 16). Privacy is Priceless, but Signal is Expensive. Signal.