# Technologies of Resistance to AI

WILLIAM AGNEW, Google DeepMind
KEVIN R. MCKEE, Google DeepMind
IASON GABRIEL, Google DeepMind
JACKIE KAY, Google DeepMind
WILLIAM ISAAC, Google DeepMind
A. STEVIE BERGMAN, Google DeepMind
SELIEM EL-SAYED, Google DeepMind
SHAKIR MOHAMED, Google DeepMind

Technologies of resistance are tools and practices that reorient control and shift power to those impacted by oppressive power relations and systems. In this paper, we turn this concept toward artificial intelligence (AI). We characterize the concept of resistance using the legacy of drag queens, foot draggers, slanderers, luddites, hackers, those who resist compliance, those who subvert rules, and queerers of relations and spaces. We develop a theory of *technologies of resistance to AI* to describe how these technologies re-align values to local contexts and practices, shift power from model designers and owners to data and model subjects, and enable agency beyond what participatory frameworks can provide. Our theory provides a taxonomy that characterizes the degree of resistance that AI ethics tools enable, revealing a lack of technologies of resistance in the existing ethical AI ecosystem. We propose a technologies of resistance research agenda to address this gap, adapting principles of meta-design into a practice of resistance describing the tools, relationships, and communities needed for widely usable and effective technologies of resistance to AI. Ultimately, AI should be resistible, and our aim is to show that this is a needed part of the undertaking of AI communities as we seek to develop technology that is ethical and just.

## 1 Power, Participation and Alignment

Resistance to new and changing technologies is central to their successful development, deployment, and governance. This opposition is often grounded in legitimate concerns and fears about possible changes in society, and often located in those whose relative position of power in societies is weak or marginalized. In cases of asymmetric power in technological deployment and ownership, those most affected enact their resistance using the tools and resources available to them—the ‘weapons of the weak’, as coined by Scott [77]. These forms and practices of resistance provide a direct insight into diverging expectations and values of different groups, and the visions of agency sought by those who would resist. As artificial intelligence (AI) systems become more deeply enmeshed in the operation of key social, political, and economic systems, and the role of AI ethics increases in prominence, we use resistance to AI as a critical lens. A resistance framework highlights the limitations of existing analysis and helps motivate the development of alternative tools and practices that can support the responsible and ethical use of AI. It is these tools and practices that will form the technologies of resistance to AI.

We study *technologies of resistance* [63, 89] as a possible paradigm for understanding ethical AI. The intent of this paradigm is to resist or reshape uneven, technologically-mediated power relationships and to restore control to people and groups who are subject to data collection and algorithmic decision-making. Technologies of resistance to AI are motivated by a desire to see futures where individuals or communities are able to resist undesired use of AI in their lives and local contexts. In these futures, datasets created by taking data without consent, and models built with those data, cannot exist for long: people are given the power to add, remove, modify, destroy, or otherwise control datasets and models impacting them without the consent of the dataset or model owners. Instead, AI would be either specific to particular contexts or communities, or built with mass consent. This is not a panacea, but returning power to people implicated in modern AI practices—the data and model subjects—should be considered part of the movement towards more ethical AI.

In the existing AI ecosystem, the paradigm of participatory AI accounts for power among designers, communities and other stakeholders, and has a vital role in the future ethical use of AI. Participation and its limitations have also been problematized by several authors to clarify its applicability and its shortcomings [7, 12, 46, 67]. When participation is used for methodological or product improvement or forms of data enrichment, AI designers and vendors maintain their position as the most powerful agent, initiating and controlling the participatory process, pursuing shallow forms of participation that do not allow AI to be meaningfully contested or changed. The most limited forms of participation can become forms of “participation washing” [12, 82]. Participation is also limited in its ability to allow rapid reaction and change to AI in deployed use. These are gaps that technologies of resistance to AI can help address by contesting the ability of model owners and designers to dictate the bounds and means of participation.

An appreciation of the power relations that underpin the development of AI systems, including their roll-out on a global basis, raises further questions about the politics and values embedded in these artefacts [13, 68, 75, 106]. A central issue in AI ethics is what values a system should promote or encode. Since humanity comprises a plurality of values, any attempt to create a universal set of values in AI risks replicating colonial dynamics and an unwarranted exercise of power through technological means [11, 66]. Creating a moral framework with sufficient specificity to guide AI in many different contexts remains an open problem, if it is possible at all [26]. Aligned AI will always need to be modified according to

Authors' addresses:
William Agnew, Google DeepMind, wagnew3@cs.washington.edu; Kevin R. McKee, Google DeepMind, kevinrmckee@deepmind.com; Iason Gabriel, Google DeepMind; Jackie Kay, Google DeepMind; William Isaac, Google DeepMind; A. Stevie Bergman, Google DeepMind; Seliem El-Sayed, Google DeepMind; Shakir Mohamed, Google DeepMind.

local context and practices, which requires feedback and iteration between the technology and people who use them. Technologies of resistance enable this kind of modification to take place: if values are inappropriate, or inappropriately expressed, or if there is value drift, then communities have an ability to immediately re-align the technology or stop its deployment. To support such forms of re-alignment, we argue AI designers should make AI resistable.

A vision and approach to technology based on resistance must contend with ethical tensions, whether that is establishing differences between genuine resistance and misuse, the avoidance of techno-solutionist thinking, divergences between the ability to resist and the responsible intent of designers, or integration with other parts of the ethical AI ecosystem, among others. This paper's contributions are to establish technologies of resistance to AI as a necessary part of the AI ethics research landscape, to interpret existing tools and their limitations through the lens of resistance, and to consider some of the concrete steps that are possible in this field–calling for a research program into technologies of resistance to AI with foresight and grounding into how people will and do interact with AI, and a boldness to conceive and pursue technologies of resistance for the myriad power relations that exist. For further motivation of the need for resistance based on histories of technological change and the growing evidence of harms from AI, see Appendix A. In section 2, we deepen the genealogical and theoretical understanding of technologies of resistance. We then present a measurement system for evaluating how useful an AI ethics approach is for resistance, and assess existing AI ethics tools using this schema to uncover gaps in the AI ethics ecosystem. In section 3, we present a research agenda for technologies of resistance to AI by enumerating the modalities and interactions between AI and data/model subjects. Finally, in section 4, we describe alternative futures where AI is resistable, and discuss some of the limitations and needs for further work.

## 2 Technologies of Resistance to AI

To the best of our knowledge, Steiner [89] first introduced the term "technologies of resistance" to describe the way Native Americans resisted colonizer cultural hegemony by transforming and redesigning imported cloth to strengthen their cultural sovereignty. We include a broader set of sources in the term's genealogy, including resistance, queer, postcolonial, feminist, and critical race studies, mirroring the organic, ubiquitous, and eclectic nature of technologies the term describes. A well-known example of resistance is the Luddites, English textile workers who smashed the automated looms of certain manufacturers in the 1800s [42]. Contrary to the popular perception of the Luddites as naively anti-technology, their concerns were in fact strongly centered around notions of economic justice: they were not opposed to automation, but only looms that drove down labor costs and needs by producing lower quality fabric at greater scale [10]. This bears striking semblance to contemporary artists, journalists, writers, and coders, resisting the rapid and uncritical adoption of generative image and language models, many of whom see some role for these tools but oppose uses that undermine wages, work, intellectual property, or quality. Surveillance technologies are a central target of resistance, including workplace surveillance, border surveillance, welfare surveillance, and carceral surveillance [36]. Truckers have become highly surveilled, with electronic logging devices tracking how long, how fast, and where they drive [54]. This surveillance has been met with an array of resistance, including physical destruction of trackers, use of GPS blockers, and protests disrupting traffic [54]. Surveillance is also often resisted through obfuscation, direct action, organizing, advocacy, regulation [36], using privacy enhancing technologies (PETs) [35] or by sousveillance, the turning of surveillance tools on the powerful [57]. These examples highlight that resistance is often not about specific technology, but about the way the use of the technology undermines material well-being, culture, or sovereignty of the resister.

**2.1 What is Resistance** To grasp the nature of technologies of resistance, first we consider what resistance is. Resistance studies define "resistance" as an opposition to power that is deeply intersectional, entangled with historical power, and highly dependant on context [99]. Benjamin [9] grounds resistance in the urgent needs of the marginalized, exhorting us to "consider the many different types of tools needed to resist coded inequity, to build solidarity, and to engender liberation". Our definition of resistance is not limited to refusal, opting out, or smashing looms. Even more important is subverting and adapting technology to local needs, creating what in some contexts are called "creole technologies", that "which finds a distinctive set of uses outside the time and place where it was first used on a significant scale" [23]. This type of resistance is especially important for data and AI systems because so few people can opt out of interacting with them, and because individual data and AI systems frequently impact millions, if not billions, of people. Instead, by resisting potentially harmful or misaligned values in data and AI while retaining freedom to use and even benefit from data and AI, our notion of resistance sidesteps difficulties of finding a universal set of values to encode in data and AI systems [26].

**2.2 Who Resists** Key to understanding resistance is who is resisting, and who is being resisted. While many people would have some need to resist data and AI, we center our conception of resister on marginalized individuals and communities who are most likely to experience data and AI harms. As with the related notions of power and oppression, the question of who needs to resist, and the appropriate means of resistance, are deeply intersectional. Communities include individuals with varying levels of technical training and awareness and exposure to and dependancy on data and AI systems. The ability and need to resist may also be determined by uneven impacts of specific AI harms, such as privacy violations or police violence. While we discuss technologies of resistance to AI, for the sake of specificity and clarity it is important to keep in mind that those resisting technologies ultimately resist those who create, control, and profit from AI at the expense of others. Indeed, those who scrape data without meaningful consent or use AI to surveil workers, manipulate customers, or automate jobs are common targets of resistance. So too are states which use AI to surveil and control citizens, non-citizens, offenders, and migrants. However, resistance can also occur at a more local level, between communities with more balanced (yet still unequal) power relations, or even between different individuals. For example, Internet of Things devices are often used for surveillance and control in neighborhoods [17] and domestic settings [53, 91] and are increasingly integrated with AI technology [105]. Given the existence of broad and intersecting sets of resisters and resisted, it is vital that design of technologies of resistance is inherently sociotechnical—moving beyond consideration of how to resist harmful data and AI practices at a technical level, toward a grounding of resistance that centers the realities and design needs of specific communities.

**2.3 Technology as Ideas and Artifacts** Very few of the "weapons of the weak" described by Scott [77] are "technical" in a demotic sense, much less computational. A wooden beam across a road is a vital technology of resistance, keeping labor-replacing farm equipment away from fields [77]. Technologies of resistance are also often symbolic, changing or appealing to narratives and social norms [77]. As Benjamin [9] notes, "critical race studies has long urged scholars to take narrative seriously as a liberating tool". Just as often, technologies of resistance take the form of concepts and patterns of collective action, including withholding labor [77]. Indeed, in their work, "Social Movements and Their Technologies: Writing Social Change"—which addresses struggles for independent, private, and secure communication technology— Milan [63] describe activists as primarily motivated by empowering other social movements to better connect and coordinate collective action. Nonetheless, Benjamin [9] gives many examples of technologies of resistance that are apps, datasets, or algorithms. In particular, technologies of sousveillance turn surveillance tools on the powerful to expose abuses [52, 55], and PETs [81] use cryptography and other computational tools to combat increasingly sophisticated surveillance tech [49] and give people meaningful privacy.

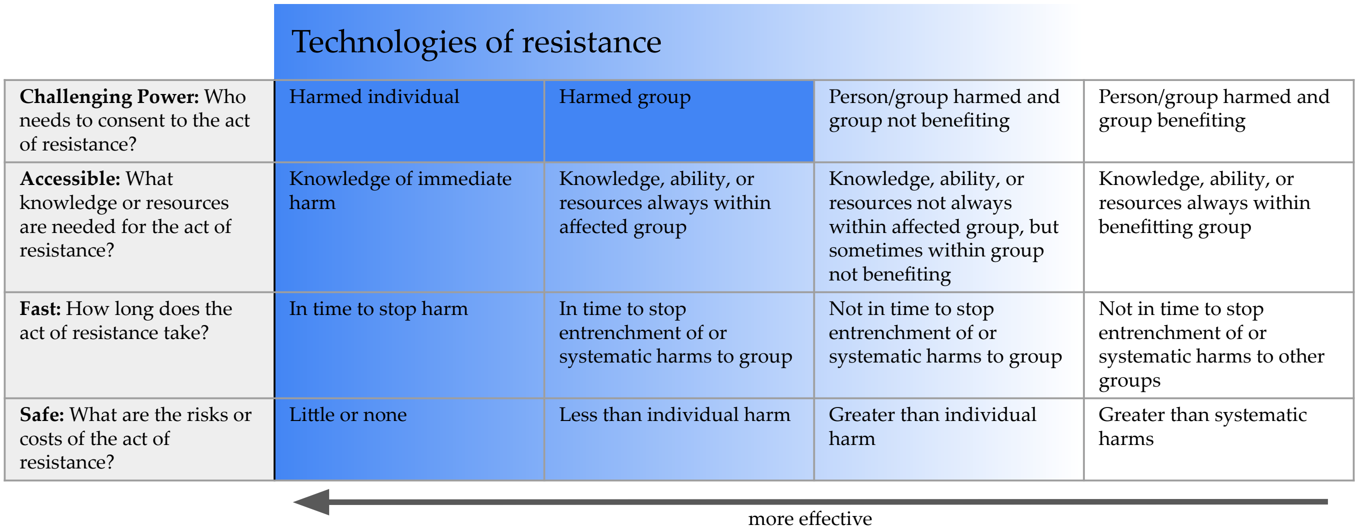
To avoid falling into the trap of technosolutionism and thereby missing a large part of the puzzle, any study of technologies of resistance to AI must center the non-computational and community-based nature of many technologies of resistance [58, 77]. And while community is deeply entwined with technology, there are also some computational tools that can be technologies of resistance, especially in the context of resisting facial recognition for surveillance [80], algorithmic redlining [47], "predictive" policing, and other types of computational oppression.

**2.4 Characteristics of Technologies of Resistance** We propose four key features of technologies of resistance to AI, which encapsulate the notion that technologies of resistance to AI must enable AI and data subjects – who often have minimal technical knowledge and institutional resources – to challenge AI and data-mediated power, safely and quickly enough to address data and AI harms they experience. The first feature, *challenging power*, refers to who must consent for a technology of resistance to achieve its purpose. If governments, corporations, or other powerful groups that perpetuate or benefit from AI harms must consent in order for an AI ethics tool to succeed, then that tool is unlikely to meaningfully resist power. The second feature is *accessibility.* AI ethics tools that require knowledge, resources, and skills that impacted individuals or groups are already likely to possess enable autonomy and widespread resistance, whereas AI ethics tools that require expert knowledge or massive compute will often require collaboration with those benefiting from or complicit in AI harms, lowering potential for meaningful change. The third feature is *speed.* As many in resistance studies have noted [40], usability has a temporal axis: often people cannot afford to wait weeks, months, or years for AI harms to be resolved. The final feature is *safety*, referring to costs and risks of technologies of resistance in relation to the harms they contest. This tradeoff plays a key role in determining whether they make sense to use [77]. We present these four axes for evaluating the effectiveness of AI ethics tools as technologies of resistance in Fig. 1.

**2.5 Assessing Existing AI Ethics Tools From the Standpoint of Resistance** In this section, we explore where different AI ethics tools fall along these axes, and how that explains their strengths and weaknesses. Here we define “AI ethics tool” as a mechanism—whether computational, legal, or societal—for upholding ethical principles in the development, deployment, and uses of an AI system. While we consider a broad range of AI ethics tools, our analysis uncovers several gaps in the AI ethics toolkit. We summarize the strengths and weaknesses of each AI ethics tool as a technology of resistance in Figure 2.

*2.5.1 Challenging Power* Challenging power—frustrating, subverting, delaying, or stopping attempts to exercise unjust power—is a fundamental component of resistance [40, 77]. We examine how ethics tools challenge power by considering who must consent to or participate in the tools’ use. Often AI ethics tools require many actors—including the harmed, model owners, third parties such as regulators—to take action when confronted by harms. Those harmed by AI are obviously motivated to address those harms. However, actors benefiting from AI harms are unlikely to willingly participate in resisting those harms. Similarly, neutral actors who neither benefit from nor are harmed by the AI in question may not have motivation to address those harms. For an AI ethics tool to be a technology of resistance, it must only require consent of the harmed to function, otherwise the powerful or ambivalent can stop resistance. Many AI ethics tools are designed in top-down ways that enforce power relations. For example, *ethics guidelines and industry standards* [65] require model designers to operationalize them, and *licensing* [20] requires model owners to implement licences. Similarly, *regulation* [83] is also an uneven field; the powerful frequently have more influence through lobbying and donations, and networks and insider knowledge have disproportionate impact. Citizens of nations who live under oppressive regimes, dictatorships, or unstable political conditions, as well as refugees, undocumented migrants, exiles and the otherwise stateless, often cannot benefit from state-enforced regulation and are powerless to influence it. Finally, even democratically chosen laws can harm marginalized groups and reinscribe the power of harmful state structures [85, 88]. Evidence of harms and failures produced by *auditing* [73], *explainability* [60], *accountability* [16], *transparency* [51], *fairness* [62], and *debiasing* [97] still require the model owner to act on the findings, and therefore do not fix harms directly. Even *advocacy* [59] has the end goal of convincing the powerful to change, although both of these tools have powerful and often highly effective methods for building pressure on the powerful.

-



| | Technologies of resistance | | | |
|---|---|---|---|---|
| **Challenging Power:** Who needs to consent to the act of resistance? | Harmed individual | Harmed group | Person/group harmed and group not benefiting | Person/group harmed and group benefiting |
| **Accessible:** What knowledge or resources are needed for the act of resistance? | Knowledge of immediate harm | Knowledge, ability, or resources always within affected group | Knowledge, ability, or resources not always within affected group, but sometimes within group not benefiting | Knowledge, ability, or resources always within benefitting group |
| **Fast:** How long does the act of resistance take? | In time to stop harm | In time to stop entrenchment of or systematic harms to group | Not in time to stop entrenchment of or systematic harms to group | Not in time to stop entrenchment of or systematic harms to other groups |
| **Safe:** What are the risks or costs of the act of resistance? | Little or none | Less than individual harm | Greater than individual harm | Greater than systematic harms |

.5.5

Fig. 1. Axes for evaluating technologies of resistance to AI. Technologies on the left are easier for impacted individuals and communities to directly use to address AI harms, tools on the right require consent of powerful, significant resources.

In contrast, we identify five AI ethics tools that only require consent of the harmed in order to operate. To start with, *Organizing* [59] brings together harmed individuals. *Critique* [15], while seemingly similar to advocacy, shifts our understanding and interpretation of AI instead of advocating for particular actors to take action. Finally, *PETs* [35], *adversarial attacks* [6], and *data leverage* [98] explicitly view model owners as potentially adversarial, and consider means of resistance or subversion only usable by harmed individuals and groups.

*2.5.2 Accessible* Studies of resistance focus on the ubiquitous, everyday ways people resist power [77]. By virtue of being accessible to everyone and widespread, this mass resistance adds up to a force that is sometimes able to stop states, corporations, and other powerful actors. Enabling mass, bottom-up resistance to AI harms requires tools that everyone can use. Many AI ethics tools require powerful networks or financial resources to function. Implementation of *ethics guidelines and industry standards* requires access to decision makers in industry and government, *licences* require lawyers to enforce, influence over policymakers or power within government is needed to create *regulation*, and all of these tools require policy and legal expertise to navigate. Other tools require expert knowledge and intensive compute. *Fairness*, *debasing*, *explainability*, *accountability*, *transparency*, *auditing*, *PETs*, and *adversarial attacks* require technical knowledge to operate, although some have apps, GUIs, and other tools to increase accessibility. While anyone can engage in *advocacy*, effective advocacy requires a large platform and connections to the powerful actors targeted by the advocate. Some forms of *data leverage*, like witholding, are broadly accessible, while other, like data poisoning, may require technical knowledge. *Critique* and *organizing* are available to everyone, requiring only knowledge of the instances of AI harms to function.

*2.5.3 Timely* Often enduring data and AI harms for extended periods of time, especially surveillance, harassment, or state violence, is not an option for resisters. Many AI ethics tools are not designed to address immediate harms, but address patterns of harm over a temporally extended process. *Ethics guidelines*, *industry standards*, and *regulation* are often created over a lengthy period of time, and even after going into effect, provide rules and guidance which take even more time to operationalize. *Auditing*, *fairness*, *debiasing*, *explainability*, *accountability*, and *transparency* can help uncover harms, but designing and implementing fixes takes more time. *Critique* can successfully shift perceptions of AI and AI harm, but is also a long-term project. Similarly, *organizing* and *data leverage* can build enduring power, but requires organizing and potential strikes or negotiations to succeed. *Licensing* allows for rapid threat of legal action if terms are breached, which is often sufficient to stop misuse. *Advocacy* can alert model owners or other influential actors fast enough to stop specific instances of harm. Both *PETs* and *adversarial attacks* hold the promise of immediately stopping or preventing harm, by allowing resisters to directly change models or render their data incomprehensible to models.

*2.5.4 Safe* People often do not have the option to opt out of data and AI systems, whether because of pervasive non-consensual digital surveillance or de facto requirements to use smartphones, email, and social media for economic, social, and other reasons [111]. Therefore, in our analysis of costs and risks of technologies of resistance, we aim towards technologies that grant both positive and negative freedoms [93]. We must recognize that resisters are subjected to extreme power disparities, where the resisted threaten them with crushing economic deprivation, state violence, or other forces. In response to the Luddites, the British Government made loom smashing a capital crime, 60 to 70 Luddites were executed [102], and Luddite uprisings were repeatedly met with lethal military force [10]. In the symbolic realm, Luddism was distorted from real economic concerns about specific automation practices to a pejorative trope of backwards and foolish opposition to all technology that even today is deployed to discredit critics. In today's context of AI technology, these costs and risks include termination of employment, blocked access from apps, websites, and online platforms, harassment and retaliation, inordinate expenditure of time, money, and other resources on fighting harms, and more. Many AI ethics tools, including *Ethics guidelines*, *industry standards*, *regulation*, *fairness*, *debiasing*, *advocacy*, *explainability*, *accountability*, *transparency*, and *auditing* have little or no costs or risks associated with them besides time and resources required to use these tools. Tools that challenge power have more significant risks associated with them. *Adversarial Attacks* and *PETs*, when applied at individual levels, typically only risk reduced ability to use technologies and services, however if used at the scale of communities, they may incur greater risks of retaliation, or reactive regulation by model owners. *Organizing*, *data leverage*, and *critique* are highly public, and therefore those that use there tools are much more likely to face retaliation or countermeasures.

**2.6 Dual Uses of Technologies of Resistance to AI** In absence of context, technologies of resistance are not inherently ethical or unethical. By nature, they can be deployed by almost anyone, and exist to resist many different kinds of power. In this section we address two hypothetical negative use cases of technologies of resistance to AI. First, they could be used by the powerful to resist the powerless. However, powerful actors are likely to have a wide range of centralized, coercive tools at their disposal—state violence, economic pressure, or control of cultural and religious institutions, for example—so we argue that technologies of resistance rarely shift power to the already powerful, but are much more likely to shift power to the powerless. In the context of AI, for example, tools of resistance may give control and ownership over personal and community data back to communities as a means of limiting harmful uses at scale by the powerful. As detailed later, adversarial attacks and PETs which often form the building blocks for technologies of resistance to AI tend to require control over the training or input data, which limits their usefulness to data and model subjects.

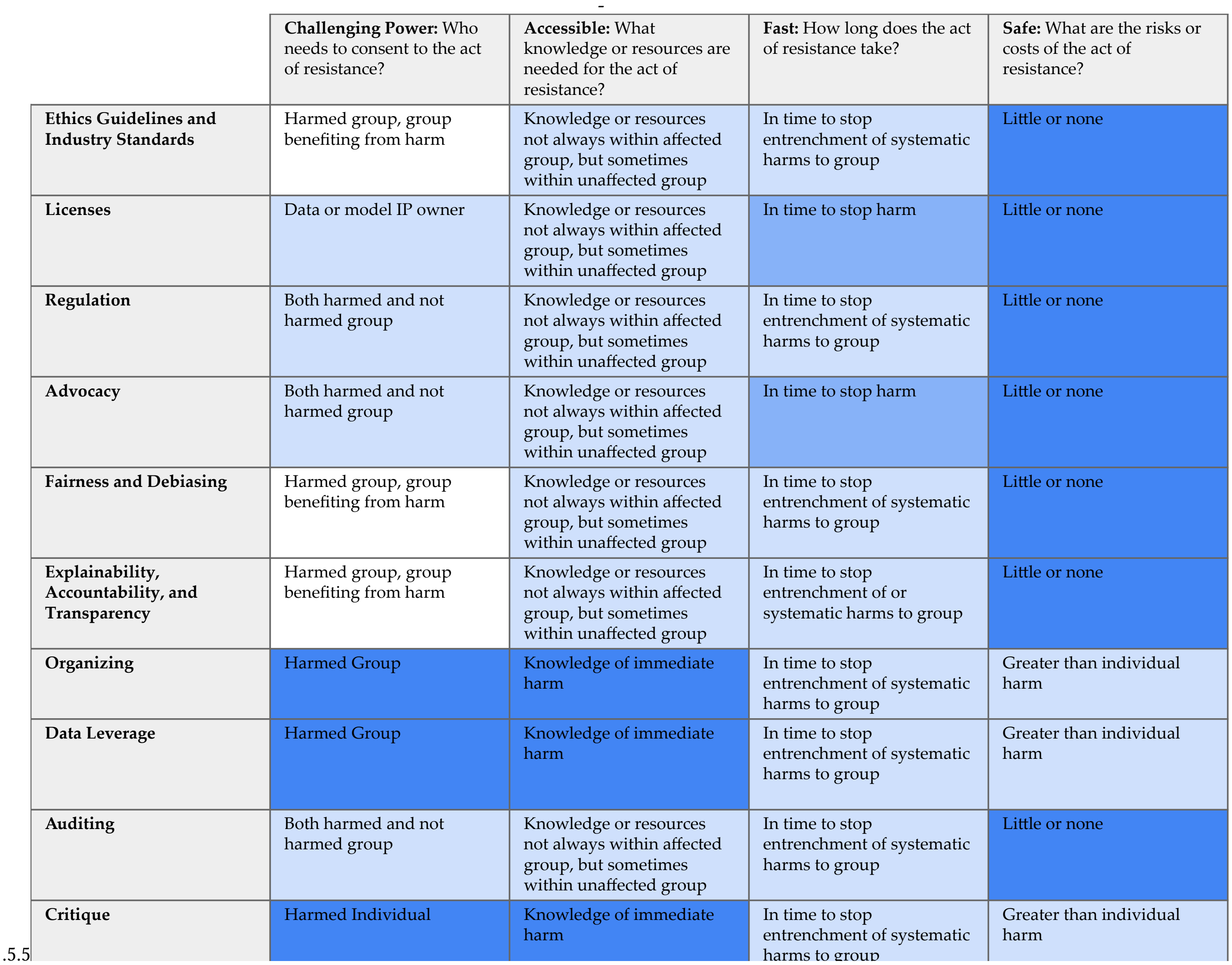
The second dual use concern is that technologies of resistance are used by the powerless to resist interventions that are themselves designed to address unethical practices. People might resist facial recognition systems to obscure themselves while committing crimes, or manipulate models to reflect reprehensible views. Many applications of AI involve surveillance against criminals, migrants, offenders, or students, and we note – from an abolitionist perspective – that even when AI surveillance purports to be beneficial it often fails to address the root causes of the behaviors it targets [9]. Moreover, by recognizing this, and resisting AI solutionist narratives, we may be better placed to focus on underlying problems such as poverty, war, racism, and under-investment in public services. Nonetheless, there are still settings where resistance could be used to undermine beneficial applications of AI. Indeed, we believe that these are a feature, not a bug, of technologies of resistance to AI. If everyone is empowered to resist AI and align it to their values, questions of how AI should be used cannot be decided by the small elite creating and operating AI. Instead, building AI that impacts many people requires mass consent *if it is resistable*, returning the power of governance from the AI designers back to people. While far from a panacea, we believe AI futures where everyone can meaningfully participate and resist are likely to be better than those where only a small group, no matter how well-intentioned, has control.

**2.7 Gaps in the AI Ethics Ecosystem** In this section we explored several salient features of technologies of resistance: they include a wide range of tools and ideas used by the marginalized to resist harms from the powerful. They are used because the provide timely and concrete relief from harms, and require little or no coordination, often operating at the level of individuals, families, and coworkers. As countless studies have shown, technologies of resistance, when used at scale, have reduced the harms of past social and technological upheavals, forcing the powerful to implicitly compromise with the marginalized. We argue that technologies of resistance are largely missing or understudied for AI and data. If AI ethics is to help prevent AI from running roughshod over those it is deployed on, we must study and develop tools to enable mass, decentralized, bottom-up resistance to AI harms. In the next section we describe existing and potential technologies of resistance to AI that succeed in this regard, and outline a concrete sociotechnical research agenda for building and implementing these new AI ethics tools.

## 3 Building Resistance

While most AI ethics tools require consent or collaboration of the powerful, adversarial attacks, PETs, data levers, and organizing do not. These tools enable direct bottom-up action. However, adversarial attacks, PETs, and some data levers often require expert knowledge, advanced technical skills, and only work for certain data modalities, attack vectors, and desired outcomes. Organizing and some data levers tend to require building massive and highly coordinated coalitions to influence politics or challenge huge corporations, requiring time, resources, and coordination far beyond the reach of friendship networks, family relations, a neighborhood, or other local social structures. To facilitate a shift of power from creators of AI systems into the hands of marginalized people, we propose a research and organizing agenda that builds, disseminates, and uses technologies of resistance. We introduce meta-design for resistance as the overarching conceptual framework of this agenda. While the field of AI ethics has focused heavily on data rights and control, we also argue that model rights and control are equally important. Just as people might want control over how their data is used, who can use it, and demand compensation for use of their data, model rights refer to control over how models derived from one's data are used, who may use such models, and who benefits from use of such models. Our framework helps illuminate what model and data rights people and communities may want, in the context of resistance to AI, and provides concrete research directions for empowering people to defend those data and model rights.

In the rest of this section we explore the limitations that prevent adversarial attacks, PETs, data levers, and organizing in AI from being broadly applicable technologies of resistance, using meta-design as a lens for proposing new tools and strategies to help address those limitations. We also explore concrete technical directions



| | **Challenging Power:** Who needs to consent to the act of resistance? | **Accessible:** What knowledge or resources are needed for the act of resistance? | **Fast:** How long does the act of resistance take? | **Safe:** What are the risks or costs of the act of resistance? |
|---|---|---|---|---|
| **Ethics Guidelines and Industry Standards** | Harmed group, group benefiting from harm | Knowledge or resources not always within affected group, but sometimes within unaffected group | In time to stop entrenchment of systematic harms to group | Little or none |
| **Licenses** | Data or model IP owner | Knowledge or resources not always within affected group, but sometimes within unaffected group | In time to stop harm | Little or none |
| **Regulation** | Both harmed and not harmed group | Knowledge or resources not always within affected group, but sometimes within unaffected group | In time to stop entrenchment of systematic harms to group | Little or none |
| **Advocacy** | Both harmed and not harmed group | Knowledge or resources not always within affected group, but sometimes within unaffected group | In time to stop harm | Little or none |
| **Fairness and Debiasing** | Harmed group, group benefiting from harm | Knowledge or resources not always within affected group, but sometimes within unaffected group | In time to stop entrenchment of systematic harms to group | Little or none |
| **Explainability, Accountability, and Transparency** | Harmed group, group benefiting from harm | Knowledge or resources not always within affected group, but sometimes within unaffected group | In time to stop entrenchment of or systematic harms to group | Little or none |
| **Organizing** | Harmed Group | Knowledge of immediate harm | In time to stop entrenchment of systematic harms to group | Greater than individual harm |
| **Data Leverage** | Harmed Group | Knowledge of immediate harm | In time to stop entrenchment of systematic harms to group | Greater than individual harm |
| **Auditing** | Both harmed and not harmed group | Knowledge or resources not always within affected group, but sometimes within unaffected group | In time to stop entrenchment of systematic harms to group | Little or none |
| **Critique** | Harmed Individual | Knowledge of immediate harm | In time to stop entrenchment of systematic harms to group | Greater than individual harm |

.5.5

Fig. 2. Ratings of existing AI ethics technologies along technologies of resistance axes.

that address some challenges by expanding adversarial attacks and PETs to novel data modalities and outcomes to enable resistance in many more settings. We then unify these technical directions with a broader direct action approach to resistance against AI that emphasizes collaboration, pedagogy, and community.

**3.1 Resistance as Anti-Design** Design creates artifacts with particular values and functions, while resistance changes and subverts those, potentially beyond the desires or plans of the designers. In this respect, "designing for resistance" requires the synthesis of contradictory impulses, as resistance undoes design. Moreover, efforts to design for resistance will always have the potential to leave a gap between the uses the designer can anticipate and the uses resisters will want. Practically, however, designing *to allow* resistance is useful, arising from reflection by designers upon the ways in which people might want to resist a technology – and adding additional affordances to empower data and model subjects to resist harms and modify to align technology with their needs. Moreover, by designing to allow for resistance, designers may themselves engages in resistance, perhaps against the eventual deployer of the system being designed, or the entity employing the designer. At the level of strategy, system designers can create mechanisms for resistance for specific harms, but also may add more general features that increase contestability of data and AI systems. The success of design for resistance hinges how closely designer and resister values and motivations align. Design for resistance is most likely to succeed where the designer and the resister are the same, or where the designer has significant independence from the actor that uses the technology as a tool of oppression.

To understand the space of design for resistance, we expand on the theory of meta-design [25] which explores how designers can create technology which end users can modify and build on. Inspired by the shortcomings of the strict separation between the designer and end user and the closed-loop development cycle of software, meta-design proposes a continuous, non-binary spectrum between designer and consumer and the continuous embedding of users as designers, enabling a cyclic loop and between the development process and user interaction with the system. In an ideal world, the blurring of this boundary between designers and users could balance the power between these parties.

In our current reality, the majority of AI system designers do not prioritize end user modification, and may even design systems to block such efforts. Even though there is a robust open-source AI community aspiring towards broad access to and adaptability of AI, modifying current AI systems requires advanced technical knowledge and non-trivial compute [56]. How then can end users reclaim power in their interactions with AI and the development lifecycle of AI systems? We propose *meta-design for resistance* as a philosophy for pursuing this vision. Meta-design for resistance explores how marginalized and disempowered communities can wield and transform AI technologies, forging a role for themselves as designers, creators, and hackers in their own right, without the consent of the original designers of these systems. It emphasizes the vital role of organizing, community, and culture as technologies of resistance that shape the research and development of AI.

Building on the general philosophy of meta-design, we identify three levels at which meta-design for resistance needs to operate. These are: computational defenses against AI harms, networks of collaboration, and building communities of resistance. These levels overlap and intertwine with each other: defensive tools strengthen the capabilities of communities who are motivated to resist AI harms, these capabilities and resources are then shared across communities, and community networks help motivate the design of new defensive tools. Therefore they we consider them of equal importance in the study of technologies of resistance.

**3.2 Adversarial Attacks and PETs as AI Defenses** The first level of meta-design for resistance centers upon the identification of tools for resistance. In this section we discuss using two classes of technical tools, PETs and adversarial attacks, as defenses against AI and data harms. PETs have been well-studied in this regard, but are usually limited in effect to the protection of privacy (i.e. by supporting freedom from having personal data scrapped into datasets or revealed through inferences). Following recent research on repurposing adversarial attacks as defenses against adversarial AI±[6], we explore the potential of these tools to grant a wider range of positive AI and data control. We therefore call these two tools *AI defenses* to reflect the change of setting and intent from attacking AI to defending from AI. However, as we discuss at the end of this section, defenses against adversaries are not by themselves technologies of resistance because of high barriers to entry and uptake. In sections 3.3 and 3.4 our proposed metadesign for resistance framework shows how these limitations could be overcome, and how they could be built with specific users and communities in such a way that they become technologies of resistance.

To survey what tools for resistance exist and also reveal which resistance settings currently lack adequate defense for those who interact with AI, we consider the data modality, defense surface, resister capabilities, and outcomes of different AI defenses.

AI defense techniques are underexplored for many data modalities, especially video and audio. In this context *Data modality* heavily determines how data is created and used and what model applications are. This strongly influences how people interact with datasets and models, which harms models cause, and how data and model subjects might want to change this. Modality also narrows down which model architectures may be used, such as a convolutional neural networks for images and transformers for text, which can impact which AI defense techniques are useful. Data and models are increasingly multimodal, causing both intersecting and novel harms [14], but also raising the possibility of multimodal AI defenses. A non-exhaustive list of modalities includes images, text, audio, website or app interactions, geolocation, biometric data, externally measurable medical data (breathing rate, heart rate, skin temperature, blood oxygen), other medical data, and financial data. AI defenses have been extensively studied for images to evade inference by face detection or recognition systems [18, 24, 96], text to change specific predictions or revealing parts of the train dataset [29, 90, 101], interactions to reduce accuracy of recommender systems [70, 84, 92, 109], and geolocation to reduce location prediction accuracy [78].

Many potential sites of resistance are understudied, including crowdworker data labeling and human interaction in reinforcement learning and continual learning settings. *Defense surfaces* are composed of the inputs resisters can use to influence a model and how model subjects are affected by outputs. Defense surfaces include: the avenues by which one is scraped into the train or test sets, being a crowdworker creating train or test sets, being an AI practitioner training the AI, interacting with the AI in a reinforcement learning or continuous learning setting, and being inferred on by the AI. Defense surfaces are key for determining which AI defenses might be useful [1] and defining technical parameters for researching new AI defenses [94]. When resisters are scraped into datasets, data poisoning may be used [30], when they train models, backdoors may be planted [31], when they interact with AI in a RL setting, adversarial examples can be crafted [37], and when inferred on by a model, crafted inputs can change model behavior [18, 24, 29, 96, 101].

AI defenses overwhelmingly require expert knowledge and coding skills to function, which are beyond the capabilities of many who want to resist. *Resister capabilities* are the tools and resources which impacted people can use to resist. This includes knowledge, such as knowledge of AI architecture, weights, datasets, training details, component libraries and algorithms, and observed inputs and outputs. This also includes both technical and organization abilities, such as implementing and using cutting edge research, programming, using apps and social media, compute resources, and behaving in specific ways to trick or subvert AI. Most AI defenses require understanding cutting-edge research and significant coding to deploy [18, 24, 48, 48, 70, 76, 84, 90, 92, 96, 100, 107–109]. Many impacted individuals and communities do not have these capabilities, revealing a severe limitation in the practical utility of these attacks. Some defenses against facial recognition systems have produced apps or programs for non-technical users, providing a promising direction for the broader field [19, 80].

AI defenses only consider and produce a narrow range of outcomes which exclude many outcomes desirable to resisters. *Defense outcomes* are the greatest limitation of current AI defenses, and determine which goals of data and model subjects they can help achieve. Frequently people want to opt-out of model inference or inclusion in datasets, but they may also want to remove data or knowledge from a dataset or model, change a dataset or model, change specific model predictions, add data to a dataset or model, and control which applications or domains a dataset or model may be used on. While much research on AI defenses focuses on rendering people invisible or incomprehensible to models, we believe AI defenses that enable positive freedoms by letting resisters change or control models while still using them are even more important. Examples include allowing trans people to remove deadnames from datasets and models [41, 87], preventing language datasets and models created by marginalized communities from being used to harm or exploit those communities [5, 8, 12], and updating datasets and models with data on underrepresented communities. Several language model defenses explore changing model outputs [29, 48, 76, 100, 101, 108], and there is a large body of literature on changing image classifier predictions [110].

*3.2.1 Unexplored Resistance Settings.* In Figure 3 we present a set of AI defenses summarizing our characterizations. While some individual elements of each characteristic of AI defenses have been explored in a particular setting, the combinatoral space of settings for AI defenses is largely unexplored. For example, there is much research on avoiding classification by image models [18, 24, 96], but none for text or audio modalities. Image and interaction models both have train dataset poisoning attacks to render the model inaccurate, but not text models. A few image attacks developed apps to lower technical skill needed to use these attacks, but overwhelmingly attacks require expert knowledge and coding capabilities. Attacks that grant positive freedoms, such as changing or adding to a dataset or model, are only considered in a few domains, and in many cases these changes are too narrow to effectively change how a model represents or treats a person [29, 48, 76, 100, 101, 108]. We argue that every combination of data modality, defense surface, resister capabilities, and defense outcomes we have described is a relevant setting for developing AI defenses against harmful AI.

*3.2.2 Usability, AI Defenses, and Dual Use* Many of the AI defenses considered here have a dual use aspect, with the potential to cause harm by disrupting AI systems. Indeed, AI defenses, understood as purely technical artifacts, tend require technical expertise to employ, making them ineffective as technologies of resistance and only empowering machine learning experts. While this dual use character will always be a concern that needs to be monitored closely, we argue that in many cases the balance can be shifted towards overall empowerment in two ways. First, many adversarial attacks require manipulating training data or the training process. Similarly, almost all AI defenses against specific model inferences require manipulating the inputs to those inferences. That is, AI defenses typically are not tools that can be deployed by anyone at scale against many different AIs, but rather often have positionalit. Defenses like dataset poisoning and obfuscation of images or text a model is inferring upon are most easily deployed by people who produce the data in the dataset or the inputs a model is performing inference on. We argue that these data and model subjects are exactly who should be empowered to actively participate in model creation and functioning. Conversely, adversarial attacks that lack this positionality and can be exploited by anyone with sufficient technical skill—for example, code vulnerabilities in ML libraries—are much less suited to be technologies of resistance because they only empower a small and already powerful community. Second, an AI defense is not a technology resistance without designing with the users, communities, and harms it is meant to address. In the next two sections use describe meta-design for resistance to show how an AI defense may become a technology of resistance by building relations and communities that lower barriers to use and establish practices of co-design to align AI defenses to user values and needs.

**3.3 Collaboration for Resistance** The second level of meta-design for resistance involves collaborations for resistance. AI defenses that enable direct and immediate control over data and AI are often highly technical and can require nontrivial compute. While design and research to lower compute requirements or create better users interfaces is valuable, just as or even more important are designing relationships to lower barriers to AI defense use. Knowledge and resource requirements of AI defenses can be overcome through collaboration, mentorship, and mutual aid [86] within communities of resistance. Research at the second level of meta-design is sociotechnical, examining how AI defenses and collaborations can be designed together to empower resistance. Key questions include how collaborations for resistance can be started and strengthened, how to ensure such collaborations are bidirectional and not new, oppressive power relations in themselves, and how such collaborations can seed communities of resistance. Examples of collaborations for resistance include TikTok users sharing information on subverting the algorithm's perceived filtering of marginalized identities [44], coordination to poison or withhold data from harmful models [98], and development and control of community datasets and models.

**3.4 Communities of Resistance** Similar to McQuillan [61]'s call for worker and people's councils, the third level of meta-design for resistance is fostering communities and cultures of resistance. Communities of resistance are those which provide the broader scaffolding for starting, strengthening, and spreading collaborations for resistance. Communities of resistance organize workshops on data ownership, produce guides on model rights, or coordinate direct action against harmful AI. Communities of resistance not only bridge the gap between research and practical use of AI ethics tools by marginalized communities, but also address tensions inherent in design for resistance by motivating and designing resistance tech themselves. AI already has many communities that resist AI harms, including Mijente [3] and Te Hiku Media [4]. Key research topics in this area include studies on AI communities of resistance to understand how they grow, function, resist, and what challenges they face, and understanding how research and other incentive structures may be changed to encourage more community work. Finally, we note that collaborations for resistance and communities of resistance are far from primarily research artifacts, but impactful relations created and maintained through organizing, administering, and care. Engaging in this work is what will actually resist data and

-.5.5

| Technology and Resistance | Limitations |
|---|---|
| Distort images to fool image classifiers [18, 24, 96] | Requires using code |
| Modify images in train dataset to prevent inference [19, 80] | Need to be able to add images to train dataset |
| Change classification of text models [48, 76, 100, 108] | Requires using code, adding examples to the training dataset, and knowledge of the text mode architecture |
| Change ranking of text models [101] | Requires access to model, use of code |
| Reveal parts of train dataset for text model, establishing use of personal or copyrighted work [90] | Requires access to model, use of code |
| Degrade prediction quality of model being used for surveillance or manipulation [70, 84, 92, 109] | Requires adding to train dataset and use of code |
| Prevent text models from being used for surveillance tasks and weights from being finetuned for surveillance tasks [107] | Requires use of code |
| Prevent artists style from being imitated by text-to-image models [79] | Requires adding to train dataset |

Fig. 3. Examples of tools that may be used as technologies of resistance, and any current limitations prevent widespread use. Many additional settings may be explored by considering combinations of data modalities, defense surfaces, resister capabilities, and defense outcomes described in Section 3.2. Most do not have existing technologies of resistance yet.

AI harms, and this work deeply will inform all aspects of technology of resistance design. Whereas communities of resistance empower members to resist, cultures of resistance refer to a broad expectation of resistability. In a culture of resistance, the assumption that data and AI harms are immediately contestable is deeply embedded, as are notions of data and model ownership. Cultures of resistance already exist for other technologies. For example, if a faucet is flooding a room, everyone will immediately try to shut it off themselves; or if a door is stuck shut, people will try many different ways of unlocking it before giving up or calling a locksmith. Even without access to collaborations for or communities of resistance, in a culture of resistance people will work to resist harmful AI, making mass, uncoordinated resistance against AI harms the norm. Key research questions in this area relate to understanding and bridging the conceptual gaps and creating critiques of AI and data harms to establish cultures of resistance.

### 3.5 Case Study One: Preventing LLM Model Usage in Gender Classification

In the remainder of this section we consider two case studies of tools that are or have the potential to function as technologies of resistance. We evaluate them along our technologies of resistance to AI axes and discuss limitations and future work that might extend these tools into other domains. The first case study, which focuses on self-destructing models [64], considers the tasks of preventing a large language model from being used for harmful tasks, such as inferring gender from biographies. Here, the adversary is assumed to have access to model weights and can finetune the model to perform harmful tasks, which is a common scenario in NLP. To address this kind of situation, Mitchell et al. [64] develop a meta-learning algorithm that not only reduces base model performance on harmful tasks, but also greatly increases the number of examples required to finetune the model to perform the harmful tasks, frustrating malicious use of the model.

Self-destructing models are an important step towards realizing model rights, specifically controlling how a text model trained on one's data may be used. While many LLMs are developed and controlled by large corporations, many communities are exploring creating and using datasets of their languages to create LLMs [38, 71]. Using self-destructing models to control how these models may be used would help align them to community values without a need for external authorities to enforce usage rules, with little degradation of overall model performance, and immediately upon model release. However, there are several several barriers and future directions. Self-destructing models have been tested on one task; ensuring they will work for communities seeking to protect their models from abuse will require extensive co-designing and testing with those communities. In addition, extending the concept from NLP models to other domains rife with model misuse, like images or video, is an open and exciting direction.

### 3.6 Case Study Two: Resisting Art Appropriation with Glaze

Text-to-image models can closely imitate the styles of artists, leading many to accuse them of appropriating their intellectual property and cost them income [33]. In response, Shan et al. [79] developed Glaze, a tool that modifies an artist's images to prevent a text-to-image model trained on them from replicating the artist's style. While many tools we discuss exist only as papers or github repositories that require technical knowledge to use, Glaze has been released as an easy to use program, enabling artists without extensive technical knowledge to use it [2]. Glaze allows artists to immediately protect their new work without cooperation of the owners of text-to-image models, at a relatively low cost of small distortions to the art. Because it is accessible and addresses timely concerns, Glaze has been used by many artists, leading to dialogue between Glaze designers and artist using Glaze, starting a community of resistance including both groups working against AI art appropriation.

While Glaze is a successful example of a technology of resistance against AI, there are many future directions for building tools that allow people to control if their data is used in a model. Glaze relies on adding example to a model's training set; if unprotected images are already in the training set Glaze cannot protect them. Creating methods that allow for removing or protecting data after a model has been trained on it is a challenging but relevant research direction. In addition, Glaze works for images, but concerns of style or IP appropriation are pertinent for many other domains, including text, video, and audio [50].

In this section we examined several AI ethics tools that subvert power, organizing, data levers, adversarial attacks, and PETs. We noted that they each have limitations around required knowledge, resources, and kinds of resistance achievable. We proposed a technologies of resistance to AI research agenda for overcoming these limitations by broadening the scope of each tool to the many different contexts where we might want to resist AI, and by combining AI defenses with organizing to build communities of resistance that share knowledge and resources to ensure broad access to resistance tools.

## 4 Creating Alternative AI Futures

In this paper we surveyed AI ethics tools and found several gaps when looked at from the vantage point of resistance. Most importantly almost all of these approaches require the consent of the powerful, or expert knowledge, deep networks, or large compute—-resources of the powerful—-to function. We reviewed the concept of technologies of resistance and proposed a framework for understanding technologies of resistance to AI. This framework allows us to assess the potential use of AI ethics tools, and their combinations, to support mass, grassroots resistance to data and AI-related harms. We then explored several settings in which people may want to resist AI, including extending work on image style and IP protection to new modalities such as text and audio, and bring concepts of task refusal to image models and co-designing task refusal specifications and datasets with impacted communities. We used the idea of meta-design for resistance to unify technical AI defenses with collaboration, organizing, and community into technologies of resistance to AI. These technologies have the potential to augment the existing AI ethics ecosystem, providing new means for direct and immediate action where other tools focus on regulation, public pressure, or change by data and model owners.

We propose a research agenda that supports different possible AI futures. The development of technologies of resistance to AI would enable people and communities to better retain control over their data, models derived from their data, and gain control over models operating upon them. Mass, non-consensual data harvesting would become infeasable, and simple and effective tools would be available to combat AI surveillance, privacy invasions, manipulation, and other harms, limiting the ability of harmful AI to negatively impact society at scale. On this view of things, building massive, meaningful consent to create large datasets and operate models at scale becomes a core question in AI research, including means of distributing benefits, limiting harms, and governance, in addition to aligning the purpose of AI with the values and needs of data and model subjects. Moreover, for invasive datasets and models with consequential harms, large-scale validating consent will often not be possible. More often, only local and contextual consent for local and contextual datasets will be possible, leading to renewed relevance of AI that learns from less data and with less compute. These changes will shift power from data and model owners to data and model subjects, empowering people to meaningfully participate in AI development, and helping AI reflect the values of as many people as possible.

As we have noted, many of these technologies could also be used for disruptive or malicious purposes. Moreover, even if people enjoy widespread power over how their data is used and how AI interacts with them, bigger questions about AI governance, values, distribution of benefits and harms, remain. But in these AI futures, where the capacity to resist gives people power over data and AI systems, distributing power more equitably, we are more likely to successfully address these questions. We hope that technologies of resistance to AI will contribute to AI ethics research that empowers ethics from the bottom and shifts power to the marginalized.

## References

[1] 2022. MITRE ATLAS™. (2022). https://atlas.mitre.org/
[2] 2023. Glaze Project. https://glaze.cs.uchicago.edu/.
[3] 2023. Mijente. mijente.net.
[4] 2023. Te Hiku Media. https://tehiku.nz/.
[5] 2023. Whare Kōrero Kaitiakitanga License.
[6] Kendra Albert, Jonathon Penney, Bruce Schneier, and Ram Shankar Siva Kumar. 2020. Politics of adversarial machine learning. *arXiv preprint arXiv:2002.05648* (2020).
[7] Sherry R Arnstein. 1969. A ladder of citizen participation. *Journal of the American Institute of planners* 35, 4 (1969), 216–224.
[8] Avi Asher-Schapiro and David Sherfinski. [n. d.]. AI surveillance takes U.S. prisons by storm. *Context* ([n. d.]). https://www.context.news/surveillance/ai-surveillance-takes-us-prisons-by-storm?utm_source=news-trust&utm_medium=redirect&utm_campaign=context&utm_content=article
[9] Ruha Benjamin. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. *Social forces* (2019).
[10] Kevin Binfield. 2004. *Writings of the Luddites*. Jhu press.
[11] Abeba Birhane. 2020. Algorithmic colonization of Africa. *SCRIPTed* 17 (2020), 389.
[12] Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Díaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the People? Opportunities and Challenges for Participatory AI. *Equity and Access in Algorithms, Mechanisms, and Optimization* (2022), 1–8.
[13] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.
[14] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).
[15] James Bohman. 2005. Critical theory. (2005).
[16] Madalina Busuioc. 2021. Accountable artificial intelligence: Holding algorithms to account. *Public Administration Review* 81, 5 (2021), 825–836.
[17] Dan Calacci, Jeffrey J Shen, and Alex Pentland. 2022. The Cop In Your Neighbor's Doorbell: Amazon Ring and the Spread of Participatory Mass Surveillance. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–47.
[18] Varun Chandrasekaran, Chuhan Gao, Brian Tang, Kassem Fawaz, Somesh Jha, and Suman Banerjee. 2020. Face-off: Adversarial face obfuscation. *arXiv preprint arXiv:2003.08861* (2020).
[19] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. 2021. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. *arXiv preprint arXiv:2101.07922* (2021).
[20] Danish Contractor, Daniel McDuff, Julia Katherine Haines, Jenny Lee, Christopher Hines, Brent Hecht, Nicholas Vincent, and Hanlin Li. 2022. Behavioral use licensing for responsible AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 778–788.
[21] Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.
[22] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism*. MIT press.
[23] David Edward Herbert Edgerton. 2007. Creole technologies and global histories: rethinking how things travel in space and time. *History of science and technology journal* 1, 1 (2007), 75–112.
[24] Ivan Evtimov, Pascal Sturmfels, and Tadayoshi Kohno. 2020. Foggysight: A scheme for facial lookup privacy. *arXiv preprint arXiv:2012.08588* (2020).
[25] Gerhard Fischer and Elisa Giaccardi. 2006. Meta-design: A framework for the future of end-user development. *End user development* (2006), 427–457.

[26] Iason Gabriel. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30, 3 (2020), 411–437.

[27] Ben Gansky and Sean McDonald. 2022. CounterFAccTual: How FAccT undermines its organizing principles. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1982–1992.

[28] Timnit Gebru. 2020. Race and gender. *The Oxford handbook of ethics of aI* (2020), 251–269.

[29] Amelia Glaese, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375* (2022).

[30] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Madry, Bo Li, and Tom Goldstein. 2022. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[31] Shafi Goldwasser, Michael P Kim, Vinod Vaikuntanathan, and Or Zamir. 2022. Planting undetectable backdoors in machine learning models. In *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 931–942.

[32] Alex Hanna. 2022. On Racialized Tech Organizations and Complaint: A Goodbye to Google. (2022).

[33] Melissa Heikkilä. 2022. This artist is dominating AI-generated art. And he's not happy about it. *MIT Technology Review* (2022). https://www.technologyreview.com/2022/09/16/1059598/this-artist-is-dominating-ai-generated-art-and-hes-not-happy-about-it/

[34] Alex Hern. [n. d.]. Facebook translates 'good morning' into 'attack them', leading to arrest. *The Guardian* ([n. d.]). https://www.theguardian.com/technology/2017/oct/24/facebook-palestine-israel-translates-good-morning-attack-them-arrest

[35] Johannes Heurix, Peter Zimmermann, Thomas Neubauer, and Stefan Fenz. 2015. A taxonomy for privacy enhancing technologies. *Computers & Security* 53 (2015), 1–17.

[36] Sean Hier and Joshua Greenberg. 2007. *The surveillance studies reader*. McGraw-Hill Education (UK).

[37] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. 2017. Adversarial attacks on neural network policies. *arXiv preprint arXiv:1702.02284* (2017).

[38] Jesin James, Vithya Yogarajan, Isabella Shields, Catherine Watson, Peter Keegan, Keoni Mahelona, and Peter-Lucas Jones. 2022. Language Models for Code-switch Detection of te reo Māori and English in a Low-resource Setting. In *Findings of the Association for Computational Linguistics: NAACL 2022*. Association for Computational Linguistics, Seattle, United States, 650–660. https://doi.org/10.18653/v1/2022.findings-naacl.49

[39] Hetvi Jethwani, Arjun Subramonian, William Agnew, MaryLena Bleile, Sarthak Arora, Maria Ryskina, and Jeffrey Xiong. 2022. Queer in AI. *XRDS* 28, 4 (jul 2022), 18–21. https://doi.org/10.1145/3538543

[40] Anna Johansson and Stellan Vinthagen. 2016. Dimensions of everyday resistance: An analytical framework. *Critical Sociology* 42, 3 (2016), 417–435.

[41] Khari Johnson. [n. d.]. Trans Researchers Want Google Scholar to Stop Deadnaming Them. *Wired* ([n. d.]). https://www.wired.com/story/trans-researchers-want-google-scholar-to-stop-deadnaming-them/

[42] Steven E Jones. 2013. *Against technology: From the Luddites to neo-Luddism*. Routledge.

[43] P Kalluri. 2021. Don't ask if artificial intelligence is good or fair, ask how it shifts power. Nature. 7 July 2020.

[44] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce Knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–44.

[45] Michael Katell, Meg Young, Dharma Dailey, Bernease Herman, Vivian Guetler, Aaron Tam, Corinne Bintz, Daniella Raz, and P. M. Krafft. 2020. Toward Situated Interventions for Algorithmic Equity: Lessons from the Field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 45–55. https://doi.org/10.1145/3351095.3372874

[46] Christopher M Kelty. 2020. *The participant: A century of participation in four stories*. University of Chicago Press.

[47] Bogdan Kulynych, Rebekah Overdorf, Carmela Troncoso, and Seda Gürses. 2020. POTs: protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 177–188.

[48] Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pre-trained models. *arXiv preprint arXiv:2004.06660* (2020).

[49] Susan Landau. 2013. Making sense from Snowden: What's significant in the NSA surveillance revelations. *IEEE Security & Privacy* 11, 4 (2013), 54–63.

[50] Frank Landymore. 2023. Voice Actors Enraged By Companies Stealing Their Voices With AI. https://futurism.com/voice-actors-companies-stealing-voices-with-ai.

[51] Stefan Larsson and Fredrik Heintz. 2020. Transparency in artificial intelligence. *Internet Policy Review* 9, 2 (2020).

[52] Sam Lavigne, Brian Clifton, and Francis Tseng. 2017. Predicting financial crime: augmenting the predictive policing arsenal. *arXiv preprint arXiv:1704.07826* (2017).

[53] Roxanne Leitão. 2019. Anticipating smart home security and privacy threats with survivors of intimate partner abuse. In *Proceedings of the 2019 on designing interactive systems conference*. 527–539.

[54] Karen Levy. 2022. *Data driven: truckers, technology, and the new workplace surveillance*. Princeton University Press.

[55] Manissa M Maharawal and Erin McElroy. 2018. The anti-eviction mapping project: Counter mapping and oral history toward bay area housing justice. *Annals of the American Association of Geographers* 108, 2 (2018), 380–389.

[56] Keoni Mahelona, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. OpenAI's Whisper is another case study in Colonisation. https://blog.papareo.nz/whisper-is-another-case-study-in-colonisation/.

[57] Steve Mann. 2004. " Sousveillance" inverse surveillance in multimedia imaging. In *Proceedings of the 12th annual ACM international conference on Multimedia*. 620–627.

[58] Gary T Marx. 2003. A tack in the shoe: Neutralizing and resisting the new surveillance. *Journal of social issues* 59, 2 (2003), 369–390.

[59] Jane McAlevey. 2016. *No shortcuts: Organizing for power in the new gilded age*. Oxford University Press.

[60] John A McDermid, Yan Jia, Zoe Porter, and Ibrahim Habli. 2021. Artificial intelligence explainability: the technical and ethical dimensions. *Philosophical Transactions of the Royal Society A* 379, 2207 (2021), 20200363.

[61] Dan McQuillan. 2022. *Resisting AI: an anti-fascist approach to artificial intelligence*. Policy Press.

[62] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[63] Stefania Milan. 2013. *Social movements and their technologies: Wiring social change*. Springer.

[64] Eric Mitchell, Peter Henderson, Christopher D Manning, Dan Jurafsky, and Chelsea Finn. 2022. Self-destructing models: Increasing the costs of harmful dual uses in foundation models. *arXiv preprint arXiv:2211.14946* (2022).

[65] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature machine intelligence* 1, 11 (2019), 501–507.

[66] Shakir Mohamed, Marie-Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology* 33, 4 (2020), 659–684.

[67] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Hassan Muhammad, Salomon Kabongo, Salomey Osei, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *arXiv preprint arXiv:2010.02353* (2020).

[68] Helen Nissenbaum. 2001. How computer systems embody values. *Computer* 34, 3 (2001), 120–119.

[69] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.

[70] Sejoon Oh, Berk Ustun, Julian McAuley, and Srijan Kumar. 2022. Rank List Sensitivity of Recommender Systems to Interaction Perturbations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 1584–1594.

[71] Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, et al. 2020. Masakhane–Machine Translation For Africa. *arXiv preprint arXiv:2003.11529* (2020).

[72] Julia Powles. 2018. The seductive diversion of 'solving'bias in artificial intelligence. (2018).

[73] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[74] Suman Ravuri, Karel Lenc, Matthew Willson, Dmitry Kangin, Remi Lam, Piotr Mirowski, Megan Fitzsimons, Maria Athanassiadou, Sheleem Kashem, Sam Madge, et al. 2021. Skilful precipitation nowcasting using deep generative models of radar. *Nature* 597, 7878 (2021), 672–677.

[75] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[76] Roei Schuster, Tal Schuster, Yoav Meri, and Vitaly Shmatikov. 2020. Humpty dumpty: Controlling word meanings via corpus poisoning. In *2020 IEEE symposium on security and privacy (SP)*. IEEE, 1295–1313.

[77] James C Scott. 1985. *Weapons of the weak: Everyday forms of peasant resistance.* yale university Press.

[78] Dara E Seidl, Gernot Paulus, Piotr Jankowski, and Melanie Regenfelder. 2015. Spatial obfuscation methods for privacy protection of household-level data. *Applied Geography* 63 (2015), 253–263.

[79] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. 2023. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222* (2023).

[80] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting privacy against unauthorized deep learning models. In *29th USENIX security symposium (USENIX Security 20).* 1589–1604.

[81] Yun Shen and Siani Pearson. 2011. Privacy enhancing technologies: A review. *Hewlet Packard Development Company. Disponible en https://bit. ly/3cfpAKz* (2011).

[82] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2020. Participation is not a design fix for machine learning. *arXiv preprint arXiv:2007.02423* (2020).

[83] Nathalie A Smuha. 2021. From a ‘race to AI’to a ‘race to AI regulation’: regulatory competition for artificial intelligence. *Law, Innovation and Technology* 13, 1 (2021), 57–84.

[84] Junshuai Song, Zhao Li, Zehong Hu, Yucheng Wu, Zhenpeng Li, Jian Li, and Jun Gao. 2020. Poisonrec: an adaptive data poisoning framework for attacking black-box recommender systems. In *2020 IEEE 36th International Conference on Data Engineering (ICDE).* IEEE, 157–168.

[85] Dean Spade. 2015. *Normal life: Administrative violence, critical trans politics, and the limits of law.* Duke University Press.

[86] Dean Spade. 2020. *Mutual aid: Building solidarity during this crisis (and the next).* Verso Books.

[87] Robyn Speer. [n. d.]. Google Scholar has failed us. ([n. d.]). https://scholar.hasfailed.us/

[88] Eric A Stanley. 2021. *Atmospheres of Violence: Structuring Antagonism and the Trans/Queer Ungovernable.* Duke University Press.

[89] Christopher Burghard Steiner. 1994. Technologies of resistance: Structural alteration of trade cloth in four societies. *Zeitschrift für Ethnologie* (1994), 75–94.

[90] Zhensu Sun, Xiaoning Du, Fu Song, Mingze Ni, and Li Li. 2022. CoProtector: Protect Open-Source Code against Unauthorized Training Usage with Data Poisoning. In *Proceedings of the ACM Web Conference 2022.* 652–660.

[91] Leonie Maria Tanczer, Isabel López-Neira, and Simon Parkin. 2021. ‘I feel like we’re really behind the game’: perspectives of the United Kingdom’s intimate partner violence support sector on the rise of technology-facilitated abuse. *Journal of gender-based violence* 5, 3 (2021), 431–450.

[92] Jiaxi Tang, Hongyi Wen, and Ke Wang. 2020. Revisiting adversarially learned injection attacks against recommender systems. In *Fourteenth ACM conference on recommender systems.* 318–327.

[93] Linnet Taylor. 2017. What is data justice? The case for connecting digital rights and freedoms globally. *Big Data & Society* 4, 2 (2017), 2053951717736335.

[94] Christopher Theisen, Nuthan Munaiah, Mahran Al-Zyoud, Jeffrey C Carver, Andrew Meneely, and Laurie Williams. 2018. Attack surface definitions: A systematic literature review. *Information and Software Technology* 104 (2018), 94–103.

[95] James H Thrall, Xiang Li, Quanzheng Li, Cinthia Cruz, Synho Do, Keith Dreyer, and James Brink. 2018. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *Journal of the American College of Radiology* 15, 3 (2018), 504–508.

[96] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops.* 0–0.

[97] Marcus Tomalin, Bill Byrne, Shauna Concannon, Danielle Saunders, and Stefanie Ullmann. 2021. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. *Ethics and Information Technology* (2021), 1–15.

[98] Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. 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.* 215–227.

[99] Stellan Vinthagen and Anna Johansson. 2013. Everyday resistance: Exploration of a concept and its theories. *Resistance studies magazine* 1, 1 (2013), 1–46.

[100] Eric Wallace, Tony Z Zhao, Shi Feng, and Sameer Singh. 2020. Concealed data poisoning attacks on nlp models. *arXiv preprint arXiv:2010.12563* (2020).

[101] Yumeng Wang, Lijun Lyu, and Avishek Anand. 2022. Bert rankers are brittle: A study using adversarial document perturbations. In *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval.* 115–120.

[102] Neil Websdale. 2001. *Policing the poor: From slave plantation to public housing.* Upne.

[103] Lindsay Weinberg. 2022. Rethinking Fairness: An Interdisciplinary Survey of Critiques of Hegemonic ML Fairness Approaches. *Journal of Artificial Intelligence Research* 74 (2022), 75–109.

[104] Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems. *AI Now* (2019).

[105] Carla White and James N Gilmore. 2022. Imagining the thoughtful home: Google Nest and logics of domestic recording. *Critical Studies in Media Communication* (2022), 1–14.

[106] Langdon Winner. 2017. Do artifacts have politics? In *Computer Ethics.* Routledge, 177–192.

[107] Chuhan Wu, Fangzhao Wu, Tao Qi, Yongfeng Huang, and Xing Xie. 2022. FedAttack: Effective and Covert Poisoning Attack on Federated Recommendation via Hard Sampling. *arXiv preprint arXiv:2202.04975* (2022).

[108] Wenkai Yang, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun, and Bin He. 2021. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in nlp models. *arXiv preprint arXiv:2103.15543* (2021).

[109] Hengtong Zhang, Yaliang Li, Bolin Ding, and Jing Gao. 2020. Practical data poisoning attack against next-item recommendation. In *Proceedings of The Web Conference 2020.* 2458–2464.

[110] Jiliang Zhang and Chen Li. 2019. Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems* 31, 7 (2019), 2578–2593.

[111] Shoshana Zuboff. 2019. *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* Profile books.

## A AI Advances and Harms: The Need for Resistance

Artificial Intelligence (AI) systems show great potential to advance scientific knowledge and human welfare across the range of sectors, including medicine, weather forecasting, protein folding, and conservation [74, 95? ? ]. In particular the capabilities of AI have extended to encompass many social functions, including fraud detection, identity verification, video surveillance, law enforcement, human resources, and many others.

While these applications promise benefits for people and communities, recent years have introduced many concrete instances of AI harms. It is well established that AI encodes racial, gender, sexuality, disability, caste, class, and other biases and stereotypes across a variety of modalities, including images, text, robotic manipulation, medicine, and multimodal domains. These biases result in law enforcement facial recognition systems that wrongly identify Black people, leading to arrest, or label Black members of US Congress criminals. AI recommendation algorithms have rendered queer content less visible, while chatbots produce racist, sexist, and queerphobic, and casteist content. AI translation has poor performance for many languages in the global South, contributing to poor content moderation [34] and incorrectly flagging innocuous speech as dangerouswith disastrous consequences. AI job applicant screening systems down rank women and people of color, continuing hiring discrimination with economic harms. AI image generators tend to sexualize women and girls.

AI fairness, debiasing, and transparency research emerged in response to these biases with the goal of measuring and improving the fairness with which different groups are treated by AI systems. While these subfields have produced major successes, more recent critiques highlight that many AI systems, even if perfectly unbiased, are fundamentally harmful. Transcription models that are less biased towards African American Vernacular Englishare better at surveilling inmates in US prisons, enabling harassment and targeting of prisoners fighting against poor conditions. More accurate facial recognition systems enable better surveillance by police, harming people of color who frequently encounter violence when they come into contact with the criminal justice system, and that are often turned against political protesters. More precise models of users and consumers enable better targeted advertising, allowing for better manipulation by political or misinformation actorsand revealing sensitive information about people such as sexuality. AI for hiring, managing, and firingworkers can be unbiased, yet make workers easily replaceable, contributing to low wages and poor working conditions, and subjects workers to intense schedules that contribute to injury. Even if AI image generation was unbiased, it is still built on mass, largely non-consensual scraping of human artists' work and threatens their livelihoods.

Extensive critiques detail the partial and surface-level nature of the remedy that AI fairness and debiasing initiatives provide [103]. One primary challenge is that these tools fail to consider underlying power dynamics: just because a model has been shown to be biased does not mean that the model owner will address that bias. Kalluri [43] asserts that ethicists should not "ask if artificial intelligence is good or fair," but rather "ask how it shifts power". Powles [72] warns that debiasing distracts from more pressing problems and renders people more legible to systems of surveillance and control. Gansky and McDonald [27] argue that research on fairness and debaising distracts from, and is prefigured by, the failure and resistance of governments, corporations, and other institutions to pursue deeper and more effective paths towards ethical AI. Together, this body of work asserts that the fairness and bias of AI systems on a procedural level are relatively unimportant, because AI systems are often deployed in ways that are harmful regardless, and that fairness and debiasing initiatives give cover to these harmful technologies. Deeply intertwined with critiques of power and AI ethics are questions of *who* is involved with an AI system at different stages of its lifecycle. Marginalized people are most often harmed by deployed AI systems, yet they are also often least present in their design and control. Recent work describes this phenomenon across gender [28, 32, 104], race [9, 28, 32, 104], and histories of colonization [11, 66], arguing that the exclusion of marginalized people from AI research, development, and control is a central cause of AI harms towards marginalized people [9, 69].

In response to these limitations, a variety of methods have been proposed to bring impacted and marginalized people into the design process, and to shift power to data and model subjects [9]. Data Feminism [22] present an intersectional feminist approach that highlights power and political dimensions of data and embraces incorporating many perspectives. Design Justice [21] provides principles for design led by and for impacted communities. Katell et al. [45] and Jethwani et al. [39] report experiences using participatory frameworks in AI and algorithmic justice work. Mohamed et al. [66] call for reverse tutelage between metropoles and peripheries and affective and political communities that blur boundaries and build solidarity between included and excluded.