

# Dead Zone of Accountability: Why Social Claims in Machine Learning Research Should Be Articulated and Defended

Tianqi Kou<sup>1</sup>, Dana Calacci<sup>\*1</sup>, Cindy Lin<sup>\*2</sup>

<sup>1</sup>Pennsylvania State University

<sup>2</sup>Georgia Institute of Technology

## Abstract

Many Machine Learning research studies use language that describes potential social benefits or technical affordances of new methods and technologies. Such language, which we call "social claims", can help garner substantial resources and influence for those involved in ML research and technology production. However, there exists a gap between social claims and reality (*the claim-reality gap*): ML methods often fail to deliver the claimed functionality or social impacts. This paper investigates the *claim-reality gap* and makes a normative argument for developing accountability mechanisms for it. In making the argument, we make three contributions. First, we show why *the symptom* - absence of social claim accountability - is problematic. Second, we coin *dead zone of accountability* - a lens that scholars and practitioners can use to identify opportunities for new forms of accountability. We apply this lens to the claim-reality gap and provide a *diagnosis* by identifying cognitive and structural resistances to accountability in the *claim-reality gap*. Finally, we offer a *prescription* - two potential collaborative research agendas that can help create the conditions for social claim accountability.

## 1 Introduction

Machine Learning (ML) research tends to make claims about potential social benefits or technical affordances of new methods and technologies. For example, McCradden et al. (McCradden et al. 2023) have pointed out the lack of credible evidence in social claims about the efficacy or safety of ML-based medical care systems. In the domain of facial recognition, claims that ML models are "closely approaching human-level performance" (Taigman et al. 2014) or that they represent "convenient and non-intrusive means for a wide range of identification applications" (Cheng, Zhu, and Gong 2018) hype up the efficacy of these models. One stark counterexample is the false arrest of Robert Williams (Williams 2021) due to a facial recognition error, resulting in trauma for him and his family. Such cases demonstrate how ML models for facial recognition generally advertised as "convenient" or "reliable" have manifested the opposite of the claimed effects. Beyond individual cases, social claims about ML models' capability can have even broader harms.

<sup>\*</sup>These authors contributed equally.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

For instance, ML studies claiming that *Deep Neural Networks* can "detect sexual orientation from facial images" (Wang and Kosinski 2018) or infer gender from handwriting (Bi et al. 2019) and facial images (Kumar, Singh, and Kumar 2019) rest on reductive assumptions about sex, gender, and sexuality. By simply relying on these assumptions, these papers make implicit social claims. These flawed premises not only lack scientific rigor but also risk exacerbating systemic discrimination and psychological harm. These claims are socio-technical in nature but usually substantiated only by numerical evidence such as model performance.

Currently, there is a concerted effort within the AI ethics and ML community to develop accountability mechanisms for technology production. However, there is no accountability mechanism for when the advertised benefits or functionality in mainstream general-purpose or foundational ML research fail to manifest in practice. Namely, theorizing accountability for the communication of the impact of ML research that precedes technology production has thus far been largely ignored in the AI ethics community. In this paper, we shift the attention of AI accountability discourse from technology production to ML research production. We do this by identifying the reasons for this absence and providing potential solutions. In doing so, our analysis serves as a case study in applying the new accountability lens advanced in this paper.

### 1.1 Outline

**Section 2 - Symptom** describes the phenomenon of the *claim-reality gap* and shows why the existing critical discourse needs new treatments to address it. Section 2 also introduces the concept of the *dead zone of accountability* as a more general phenomenon and a tool to explain why the claim-reality gap exists and why it continues to be resistant to accountability. **Section 3** identifies three epistemic beliefs, and **Section 4** the epistemological foundations of ML research, that both resist accountability in the *claim-reality gap*. Sections 3 on cognitive resistance and 4 on structural resistance together **diagnose** that the *claim-reality gap* constitutes a *dead zone of accountability*. **Section 5 - Prescription** proposes two research avenues to create the conditions of possibility for accountability in the *claim-reality gap*.

## 2 Symptom: The Claim-Reality Gap and Dead Zone of Accountability

Deployed AI systems have led to abundant failures and harms (Kirchner and Goldstein 2020b,a; Kippin and Cairney 2022; Lecher 2018). The ML discourse is still hype-driven on both the building side and the critical side to different degrees. On the practitioners' side, overoptimism driven by computational innovation, building, and commercial profit has led to misuse and applications with little adequate testing. On the critical side, "Criti-hype" (Vinsel 2021), a genre of AI critique that focuses on identifying and mitigating ML harms from models that "work too well" (Raji et al. 2022), is detached from the reality that ML systems still often fail at basic tasks (Raji et al. 2022). The grandiosity of language in hype-driven discourse allows the proliferation of exaggerated, unsubstantiated, and non-replicable claims with little to no evidential support (Kou 2024; Birhane et al. 2022; Grill 2022), creating a gap between claim and reality. As exemplified in the introduction in (McCradden et al. 2023; Taigman et al. 2014; Cheng, Zhu, and Gong 2018; Williams 2021; Wang and Kosinski 2018; Bi et al. 2019; Kumar, Singh, and Kumar 2019), this gap is particularly stark when considering social claims - claims made in ML research that suggest social benefit or impact via ML (Kou 2024; Birhane et al. 2022). In this section, we first scope the claim-reality gap and clarify what it applies to and how it appears. We then discuss why it is beneficial to distinguish *the claim-reality gap* from the notion of *socio-technical gap*. Finally, we define the dead zone of accountability and clarify its scope of application.

### 2.1 Scope of the Claim-Reality Gap

When we point to the *claim-reality gap* as evidence of insufficient attention to social claims, we do not mean that such a gap exists in all ML research. For example, the Innovative Applications of Artificial Intelligence Conference (IAAI) has, for over three decades, fostered communities focused on "deployed applications with measurable benefits whose value depends on the use of AI technology."<sup>1</sup> However, this orientation is not mainstream. While premier ML conferences such as NeurIPS, ICML, KDD, NAACL, and EMNLP include application-oriented tracks - e.g., NeurIPS's "Applications," "Machine Learning for Sciences," and "Social and Economic Aspects of Machine Learning"; ICML's "Application-Driven Machine Learning"; KDD's "Applied Data Science Track"; and NAACL's or EMNLP's "Industry Track" - and while these venues make laudable efforts to develop high-quality benchmarks and datasets, papers in these tracks still tend to rely exclusively on benchmarking against standardized datasets. This is in stark contrast to IAAI papers, which typically provide more diverse supporting evidence for their social claims. Namely, despite a professed focus on application, the culture of treating model performance benchmarks as sufficient evidence still dominates ML research - yet such performance is far from enough to justify social claims (Kou 2024). Moreover, the *claim-reality gap* applies even to studies that do not make explicit

<sup>1</sup><https://aaai.org/conference/iaai/>

social claims, because, as we argue in Section 3, even "foundational research" focused solely on theory can bear implicit social claims that, in our view, should be made explicit.

### 2.2 Why We Should not Subsume the Claim-Reality Gap into the Socio-technical Gap

One way to view the *claim-reality gap*, and the social claims gap in particular, might be as a socio-technical gap, introduced by (Ackerman 2000) as the space "between social requirements and technical feasibility". Ackerman pointed out that human activities and social mechanisms are flexible, nuanced, and context-dependent but technical systems are usually rigid and unable to deliver flexibility and adaptability to reflect the social complexity of human activities. However, reducing the claim-reality gap to the socio-technical gap risks fostering two problematic tendencies.

First, when critics use the socio-technical gap to make ML systems more accountable, it suggests that with enough societal inputs, ML practitioners' faulty social claims will improve. This ideology of "improvement" is pervasive in software work culture (Bialski 2024), and explaining the claim-reality gap as stemming from the socio-technical gap only emphasizes a "culture of improvement": the more societal or human factors are being accounted for, the more ML practitioners can close the socio-technical gap, and hence, the claim-reality gap. As such, subsuming the claim-reality gap into the socio-technical gap all but guarantees its continued invisibility. Subsequently, the claim-reality gap will self-perpetuate. Grandiose claims in the earlier days of deep learning research used to put out erroneous social claims using "suggestive language" have become pervasive in ML because of the self-referential nature of ML research culture and the tendency for junior scholars to cite the older work of senior scholars (Lipton and Steinhart 2019).

Just like how earlier *science and technology studies* writings of how laboratories extend out into the social world through acquiring allies, citations, and resources in support of their research (Latour and Woolgar 1987), the accumulation and dissemination of grandiose claims create a kind of historical legitimacy - ML scientists can cite previous papers' social claims to gain epistemic authority and funding even if those benefits never materialized or had created the opposite effects. Even more worryingly, bolstering the legitimacy through the dissemination of false claims is turbocharged given the speed and reach of ML research production.

Second, it implicitly endorses a division of labor where functionality and value alignment are addressed primarily at the design and implementation phase and considered a downstream responsibility by default (with the exception of deployment-focused ML research). Given these risks, we echo D'Ignazio and Klein's question in considering data and feminism politics (D'Ignazio and Klein 2023) - "Why should we settle for retroactive audits...if we can design with a goal of co-liberation from the start?" In the same vein, we ask: why should we settle for intervening at sites of design or deployment if we can also intervene at the sites of ML knowledge production - that is, where ML research is

conducted and communicated before it is applied to the real world. This framing grounds our examination of the claim-reality gap and asks whether we can - and should - hold ML research itself accountable for the social claims it makes, even when it is further removed from applications than practitioners' work.

### 2.3 Dead Zones of Accountability

Critical computing scholars have produced useful recommendations that could potentially improve engagement in social claims. (Green and Viljoen 2020) stated that computer scientists' thinking during research design and evaluation is internal to the computational domain ("algorithmic formalism") and calls for "algorithm realism" and "porous thinking" to incorporate the broader context and fulfill their promises of enacting real-world impacts. Similarly, (Selbst et al. 2019) identified five "abstraction traps" when designing fair socio-technical systems, another set of conceptual tools that computer scientists can utilize to address the claim-reality gap. Other examples include (Laufer, Gilbert, and Nissenbaum 2023) where they reveal the normative commitments embedded in "optimization" in ML model training and (Hancox-Li and Kumar 2021) where they discussed the contradiction between feature importance ranking (a common ML technique) and feminist values. Although socio-technical thinking has been adopted among some (Caton and Haas 2024; Baker and Hawn 2022; Ehsan et al. 2021), they are still treated by most as external and rarely engaged in depth in mainstream ML research (Ashurst et al. 2022b,a; Kou 2024). For example, in subareas of ML such as transparency and interpretability, the framing becomes over-technical and tends not to serve the public (Corbett and Denton 2023).

As we will show in Sections 3 and 4, the sparse adoption of critical computing perspectives (resulting in the *claim-reality gap*) in mainstream ML research production is not coincidental but reflects what we term a *dead zone of accountability* which we define in the context of ML research as:

*aspects of the ML ecosystem - such as methodologies, tools, institutional norms, evaluation metrics, and power structures - that resist critical scrutiny and meaningful accountability.*

Although our conceptualization of *dead zone of accountability (DA)* is derived from *dead zone of imagination (DI)*, there are differences we want to highlight. To do that, we synthesize the definition of DI<sup>2</sup> to be the following:

*aspects of bureaucratic or institutional practices that appear mundane or meaningless on the surface but conceal deep power imbalances and structural violence.*

Unlike *DI*, which explicitly highlights structural violence and power imbalance, *DA* does not explicitly require structural violence to be present. This distinction is crucial when identifying a *DA*. Accountability is a normative concept presupposing rules, standards, and responsibilities; resistance

to accountability (which characterizes a *DA*) necessarily implies a power imbalance, since the legitimacy of the status quo and distribution of harms or benefits are contested. Therefore, thorough analyses of a phenomenon as a *DA* should incorporate, or be followed by, an examination of power dynamics. This paper does both for the *claim-reality gap*: characterizing it as a *DA* (Sections 3 and 4) and revealing its underlying power imbalance (Section 2.3).

### 2.4 Structural Violence in the Claim-Reality Gap

Graeber's analyses of the practice of filling in paperwork showed that healthcare seekers must navigate complex rules and manage cross-institutional communications to avoid being denied necessary care. Meanwhile, bureaucrats are not required to reciprocate this level of effort or understanding/imagination. "Filling paperwork" - may appear devoid of deeper meaning but, upon closer analysis, reveal underlying power imbalances. Specifically, the distribution of physical, emotional, and imaginative labor disproportionately burdens those disempowered, often with the threat of *structural violence* - losing access to health care.

The actors involved in the *claim-reality gap* typically include two sides. First, the ML research community - the "coding elites" (Burrell 2024) - with access to vast research funding and significant economic and social influence. Second, the impacted communities, regulatory agencies, think tanks, civic tech activists (such as technology auditors and investigative journalists), AI ethics scholars, and Human-Computer Interaction researchers. These communities work tirelessly to identify harms, uncover their causes, and propose solutions, including legal regulations, conceptual frameworks, tools, and community organizing to mitigate the damage caused by technologies rooted in the ML community's knowledge products. (Phan et al. 2022) For many, this effort constitutes their primary career or professional focus.

Communities outside of ML research must acquire technical literacy and navigate an unfamiliar world to understand and address the threats partially constituted by ML knowledge. Meanwhile, ML scientists-except for a rare minority who engage in cross-disciplinary or downstream work-rarely reflect on the experiences and efforts of those striving to make technologies function and/or less harmful. Their focus remains on unrealistic benchmarks detached from real-world contexts (Raji et al. 2021).

As David Graeber observed: "[T]he process of imaginative identification [is] a form of knowledge . . . within relations of domination, it is generally the subordinates who are . . . relegated the work of understanding how the social relations in question really work" (Graeber 2012). This illustrates a form of structural violence reflected in the current division of interpretive labor in ML research. Communities outside of ML research must exert tremendous efforts to understand the technical construction of and undertake socio-technical analysis of ML methods and technologies and develop solutions, otherwise they are subjected to harms that are inextricably connected to general-purpose ML research.

Although this uneven distribution of interpretive labor is not unique to ML and can be seen in other scientific disci-

<sup>2</sup>Graeber never gave an explicit definition

plines, it is particularly pronounced in ML research. Fields such as medicine often involve longer vetting and maturation periods before their knowledge has real-world impacts. In contrast, ML research frequently bypasses iterative testing and expansive evaluation, rapidly transitioning from being “exploratory” (Bouthillier, Laurent, and Vincent 2019; Herrmann et al. 2024) claims to real-world applications due to the low barriers to entry and the accessible computational resources.

Graeber made a moral case for revealing and addressing structural violence. To illustrate this point, Graeber uses a metaphor: a master whips a slave to enforce “unquestionable obedience” and establish “absolute and arbitrary power.” Once obedience is achieved, the master can remain completely oblivious to the slave’s experience, while the slave must deeply understand and navigate the master’s perception to survive. Graeber argues that scholars have a moral responsibility to interrogate “what ‘unquestioning’ actually means” and by extension, unveil the power dynamics between master and slave. Similarly, the structural violence this paper identifies propels scholars to identify zones such as the *claim-reality gap* and pose questions about the legitimacy of its existence.

Put differently, in the context of ML research and those subjected to their social claims, scholars critiquing such systems should also question (and not simply obey) not only the outcomes of these systems (as many have in the fields of algorithmic fairness and accountability) but also the social claims researchers make of the world.

## 2.5 Application of Dead Zone of Accountability

The broad phenomenon of the *claim-reality gap* is one example of a dead zone of accountability. We envision dead zones of accountability as applicable at both local levels (such as individual studies or subfields) and broader disciplinary contexts. Using this paper as an example, we identify the *claim-reality gap* as a *dead zone of accountability* by demonstrating that it is sustained by cognitive or cultural barriers - beliefs and assumptions (Section 3) - and structural or procedural barriers (Section 4) that are “hostile” to accountability within this gap.

In this paper, we are primarily concerned with revealing the *conditions of possibility* for accountability in the *claim-reality gap*<sup>3</sup>, and how the conditions of possibility thus far foreclose new and alternative forms of accountability in ML research.<sup>4</sup> We leave tracing causal links between social

<sup>3</sup>We use conditions of possibility in the Foucauldian sense (Foucault 2005). In his archaeological analysis of knowledge, Foucault uses *condition of possibility* to refer to the implicit rules, discursive formations, and epistemic constraints that shape knowledge production.

<sup>4</sup>Existing accountability mechanisms primarily target actors directly involved in the building and deployment of ML systems. By new and alternative, we refer to mechanisms such as social claim accountability that could link upstream ML researchers more closely to the real-world contexts in which their work may be used. At present, “accountability measures” for those producing general-purpose ML research typically take the form of impact statements or ethics checklists - tools that have been shown to be inadequate

claims and harm, and detailed recommendations of how to improve engagement with social claims, to future work.

By focusing on the *conditions of possibility*, we suggest that there are untapped forms of accountability that could emerge if ML research engaged with social claims more seriously. To create these conditions of possibility, we argue that AI ethics and ML research must challenge these limiting beliefs (in Section 3), resist *computational capture* (in Section 4.1), and strive toward a new form of epistemic authority in (Section 4.2).

Through the lens of *dead zone of accountability*, critical computing scholars and practitioners can pose questions about processes, actions, and entities that traditionally appear unaccountable - whether practically or normatively - such as the links between concrete harms and research culture, methodological practices, or computational tools and techniques. These inquiries offer valuable contributions to the *responsibility gap* discourse, which focuses primarily on tracing harms to stakeholders (Goetze 2022; Kou 2024). Labeling something a dead zone of accountability does not simply mean the absence of accountability mechanisms; it requires revealing why accountability is resisted and what practical or epistemological obstacles prevent it.

## 3 Cognitive Resistance to Accountability

In this section, we identify three flawed epistemic assumptions that make the *claim-reality gap* resistant to accountability.

### 3.1 “Making ML scientists Account for Social Claims Misunderstands Their Role Responsibilities”

One presumption may be that it is not within the “role responsibility” (Douglas 2009) of ML scientists to produce well-justified social claims. As Douglas explains, “Role responsibilities arise when we take on a particular role in society, and thus have additional obligations over and above the general responsibilities we all share.” (Douglas 2009) In the public and academics’ imagination, the primary responsibility of ML scientists’ role is to develop generalizable models that parallel or surpass human learning and problem-solving capabilities. Success in this role is measured through computational performance. Additional responsibilities, such as ensuring transparency, replicability, and objectivity, are framed as epistemic virtues, which are typically also operationalized computationally, through practices like sharing code and data, enabling others to replicate studies and reproduce model performance.

Under such a presumption, two corollary justifications might follow. First, asking ML scientists to take accountability for speculative social claims is seen as exceeding their role responsibilities because ML scientists, similar to other scientists, “are not trained to deal with social and political issues” (Pamuk 2024). As such, ML social claims are often dismissed by “experienced” readers - those familiar

and lacking in critical depth (Ashurst et al. 2022b; Kou 2024); more importantly, these measures are loosely enforced, recommended but not required

with widespread replication and functionality failures and value-alignment issues - as frivolous, unreliable, or erroneous. Social claims only attract scrutiny when they are tied to concrete, harmful applications, prompting actions from “fact-checking” communities such as auditors, regulators, civic tech activists, and critical computing scholars. Examples include the analysis of facial recognition technologies and algorithmic decision-making systems aimed at mitigating harm and improving functionality.

However, it is a mistake to assume that “experienced readers” represent the entirety of the audience of speculative social claims with widespread misuse and unwarranted hype. Similarly, accepting the status quo scope of “what they are best equipped to do” as sufficient and acceptable is equally flawed. On this note, it is crucial to recognize that when it comes to justifying social claims there is a middle ground between *doing it all* and *doing nothing at all*. Namely, this paper’s stance on accountability about social claims is not asking ML scientists to become social scientists or domain experts.

The second corollary justification under this presumption is: that the nature of scientific knowledge - being incomplete, uncertain, and fallible - means that ML scientists often cannot predict who will use their models or for what purpose. As such, regulating social claims through accountability is seen as infringing the nature of knowledge and, therefore can cause the epistemic harm of limiting the freedom of inquiry - a normative principle tied to the postwar funding ethos of science, which holds that unfettered scientific exploration benefits society collectively (Bush 2021; Douglas and Branch 2024). Under this view, scoping ML scientists’ responsibility within the computational domain and unburdening them with social reflections is deemed efficient and fair, allowing them to focus on what they are best equipped to do while preserving epistemic freedom (Douglas and Branch 2024).

However, it would be misguided to assume that regulating social claims will come at the expense of limiting free inquiry. History provides numerous examples where scientific progress has benefited from external social pressures on scientists’ conception of the social world. One salient example is (Epstein 1996)’s ethnographic and sociological study on AIDS activists’ impact on knowledge production in the scientific and medical communities. Amidst the AIDS crisis in the 1980s, AIDS activists contested the “widely accepted scientific approach...to design trials” (Pamuk 2024) - using homogenous populations disregarding sex and race. The activists rejected the use of homogenous populations as ill-suited to fulfill the practical purposes of AIDS research. As a result of social reflection, scientists developed two scientifically valid methodologies - the “pragmatic approach” that was aligned with clinical practices and the conventional “fastidious approach” (Epstein 1996). This example highlights how social considerations can be conducive to methodological diversity. To harness these benefits, we need to have suitable institutional structures or norms within science that effectively define and “administer” social considerations.

### 3.2 “Easy Solution! Let’s Erase Social Claims in ML Papers”

Some readers might view our stance on regulating speculative claims as excessive and propose a simpler solution: remove all social claims and any promissory tone from both applied and foundational ML papers and leave nothing to regulate. However, even if explicit social claims or subtle allusions to social benefits - whether regarding functionality or upholding social values - are eradicated from written texts, social claims will still be implicitly made. To evaluate the feasibility of erasing social claims, one must contextualize this proposal within ML culture. ML research does not exist in a vacuum; it operates within a complex network of communities, institutions, and relationships (Widder and Nafus 2023; Bannani-Taylor 2024). Currently, the ethos of ML culture is grounded in general applicability, functional utility, and “techno-optimism” (Campolo and Crawford 2020). These foundational aspects of ML culture cannot be extinguished by merely removing textual expressions of social claims.

Over the past decade, research agendas centered on tech ethics and social good have gained prominence in AI ethics and premier ML conferences. (Phan et al. 2022) For example, the 3.77% of NeurIPS submissions flagged for an ethics review in 2023 are bucketed into categories specifically concerned with concrete harms such as bias, privacy violations, and research integrity issues. (Mackey, Poland, and Chen 2023) This academic environment has fostered implicit expectations that researchers must navigate to secure funding, recognition, and citations. Researchers are now evaluated not only on their capacity for technical innovation (through formal peer review processes) but also on the moral valence and practical utility of their work (through general culture). Put simply, it is sexy and even extremely rewarding for ML researchers to create AI not only for social good but also for the social world.

This dual expectation of technical contribution and social impact has introduced two criteria through which ML research is interpreted and evaluated - first, is said ML research technically innovative? Second, is said ML research socially impactful? A published study might omit explicit references to social benefits, yet readers are likely to assume the work has the potential to deliver utility and/or ethical impact. For instance, tenure-track faculty in ML and AI tracks are evaluated on how impactful their research is. (Patterson, Snyder, and Ullman 1999) As a result, social claims - whether explicit or implicit - become deeply embedded in a hype-driven culture where both ethical alignment and practical utility are deeply intertwined and rewarded. Thus, the idea of fully disentangling and removing social claims from evaluations of a successful ML research output is not feasible because social claims can be inferred by reviewers, funders, or the public. Take ML-based science as an example (ML models or methods are introduced in other sciences to develop understandings of a phenomenon), the lack of awareness about the epistemological limitations of ML in scientific inferences (Humphreys 2020) has led to the production of “pseudoscience” (Andrews, Smart, and Birhane

2024) and “inaccurate findings”. People tend to take claims and numbers as “granted” (Grill 2022; Raji et al. 2022).

Thus, the suggestion to erase social claims from ML papers is neither feasible nor productive. Such an approach ignores the shifting cultural realities of the field, and how CS and ML professions are rewarded for making social claims. It also allows social claims to remain unacknowledged and unregulated while failing to address the negative impacts of its circulation and uptake.

### 3.3 “This is Too Hard, Let Practitioners Worry”

The third justification is grounded upon a sense of pessimism about developing accountability mechanisms for social claims. Even those sympathetic to the idea of regulating speculative claims often struggle to envision how this could be achieved.

Despite the added socio-political and ethical dimensions, accountability fundamentally rests on qualities and actions central to accounting - recording, documentation, and traceability (Cooper and Vidan 2022). To explore this further, consider (Bovens 2007)’s (2007) conception of accountability - “a relationship between an actor and a forum, in which the actor has an obligation to explain and to justify his or her conduct, the forum can *pose questions* and *pass judgment*, and the actor may *face consequences*” (italic added). Bovens’s (Bovens 2007) conception of accountability highlights answerability to harm as an essential component of accountability relationships.

The pessimistic view arises from the undeniable reality that we currently lack a clear framework for how ML scientists, trained to produce generalized models, should engage with the questions and demands of other social groups. There is no established direct and answerable “relationship” between the actor (ML research community collectively) and the forum (those impacted, positively or negatively, by their knowledge). This is unlike the extensively investigated relationship between tech designers and impacted communities under the Participatory AI movement (Young et al. 2024; Delgado et al. 2023). Consequently, the essential elements of accountability - *posing questions*, *passing judgments*, *addressing harm*, and *facing consequences* - are absent. ML scientists form a closed expert community that communicates in the computational domain, distinct from the language used by communities who articulate lived experiences.

Answerability implies an affinity toward being able to trace the causal chain from an entity/action to harms and having conceivable remediative actions (usually expressed as stakeholder actions and responsibilities) at the junctures of the causal chains - auditing of, legal regulation on, and participatory design of, technologies, etc. However, when moving upstream from ML technology to ML knowledge, we enter un “uncharted” territory - the pathway from speculative and generic social claims to tangible failures and harms becomes even more nebulous.

This lack of answerability and traceability often leads the above-characterized pessimism to morph into a normative argument against regulating speculative claims. Some may assert that “knowledge is just knowledge” and that

“dangers and ethical issues only arise when science is applied in technology” (Wolpert 1999). However, decades of analysis from science studies scholars like (Harding 2013) and (Keller 1982), etc., constituting counter-evidence to this stance, highlighting the inherently social nature of scientific knowledge and practice.

While we may not yet understand the precise pathway from a single social claim to concrete harm (unlike tracing harm to a specific technological design), this uncertainty does not justify inaction. Pessimism rooted in the unknown is merely intellectual laziness. Just because something is currently unknown does not mean it is unknowable - or that we should not attempt to understand it.

These three (likely non-exhaustive) faulty yet pervasive assumptions collectively position ML scientists within an “epistemic bubble” (Douglas and Branch 2024). By framing their primary contributions as computational and generalizable - detached from specific contexts or applications - ML scientists are afforded considerable freedom in crafting speculative claims, including their content, validity or whether or not to make them explicit. This freedom effectively absolves them of responsibility for the implications of these claims. With such an epistemic bubble encasing the dead zone of accountability, ML scientists continue to gain epistemic authority and funding through speculative social claims. Meanwhile, downstream communities, frustrated by the chaos and harm perpetuated by unwarranted ML hype, find themselves unable to contest the legitimacy of these claims. Instead, they are forced to accept the status quo as an unfortunate byproduct of the dead zone. These three popular epistemic assumptions are partially why ML researchers have not attempted to rectify the issue of absence of accountability in the *claim-reality gap*. The next section discusses the disciplinary logics of ML research that reinforces these beliefs.

## 4 Structural Resistance to Accountability

In this section, we introduce one key epistemological underpinning of ML research that causes its systematic resistance to accountability: *computational capture*. We define *computational capture* in the field of ML as:

*processes through which computational goals exclude or sideline social or practical goals through research practices, domain norms, or incentives structures*

In Section 4.1, we illustrate at least two ways in which *computational capture* manifests: through methodological rigidity and techno-elite control. In Section 4.2, we explore how ML’s current form of epistemic authority relates to these manifestations in ways that further reinforce computational capture.

### 4.1 Two Instances of Computational Capture

**Methodological Rigidity as Computational Capture**  
Computational capture can manifest in how problems are selected, studies are designed, tools are chosen, and success is measured, ultimately regimenting research into patterns that prioritize “SOTA-chasing” (Church and Kordoni 2022). As a result, the ML research process is characterized

by the Common Task Framework (CTF) (Lieberman 2010). (Donoho 2017) defined the CTF to have the following ingredients: (a) a publicly available training dataset, (b) a set of common tasks, and (c) a shared evaluation mechanism.

Under *computational capture*, the field tends to deem only "general" knowledge acceptable for the discipline, prioritizing justifying decontextualized model performance over studies connected to localized applications (Karl et al. 2024; Wagstaff 2012; Henderson et al. 2018). This limits or even punishes, methodological pluralism. For example, in the autoethnographic study by (Rubambiza et al. 2024), where computer scientists (computer network research more specifically) attempting to deviate from computational framing of innovation faced repeated rejections of their submissions as irrelevant and trivial for the field. This caused them to compromise their vision and presentation of "societal impact". Similarly as (Leonelli 2023) pointed out: "Researchers working [within] existing [research] repertoires (Ankeny and Leonelli 2016) may not value - or even consider - elements that are not already part of that repertoire. . . . when wanting to modify a [research] repertoire, researchers may face significant hurdles - in the shape of negative reviews, rejection by funding bodies and critical questioning by powerful peers."

The incremental and minuscule improvements in model performance made over shared tasks are proven to be unstable (Karl et al. 2024). Even when they are stable, these models fail to emulate human intelligence, often struggling with the most basic real-world tasks. General purpose ML research's state mirrors (McCarthy 1997)'s criticism of using computer chess as evidence of programs emulating human intelligence: "It was as if geneticists had focused their research efforts on breeding *Drosophila* (fruit flies) to race against each other. . . . [w]e would have some science, but mainly we would have very fast fruit flies." One might invoke Thomas Kuhn's (Kuhn 1997) conception of the scientific revolution and argue that ML research is simply in the phase of "normal science", and a paradigm shift will occur with time. However, as Paul Feyerabend (Feyerabend 1970) pointed out, a revolution requires knowledge and inquiries that challenge the presiding paradigm.

In this regard, methodological capture not only limits meaningful progress in innovation, aggravating the creativity statement (Cho 2024; Hill 2024), but also restricts the room for methodological pluralism that integrating social considerations requires.

**Techno-elite Control as *Computational Capture*** Techno-elites refer to powerful actors in ML ecosystem industry and academic ML research labs, funding bodies across industry and federal agencies. These actors' visions drive actions (e.g. assigning funding, setting conference aims and organization, determining publication and review standards, defining methodological rigor) that draw epistemic boundaries - centering ML research focused on producing state-of-the-art model performance on shared benchmarking datasets over meaningful engagement with social, ethical, or interdisciplinary considerations.

Not only does the computational capture constrain ML

within computational formalism, but it also betrays how the funders of ML research, who are "close to the source of knowledge" (mostly privatized funding from tech giants (Whittaker 2021)), monopolize the research paradigm for the preservation of existing paradigm and their benefits. This reality casts doubt on the prospects for a disciplinary revolution in ML research.

With the highly standardized and mathematized ML research process's main beneficiaries being the ML community, it aggravates the misalignment between the producers of ML knowledge and those impacted by it. Different communities have disparate assessments of the acceptability of uncertainty of knowledge (Pamuk 2024). ML scientists think of the variability of a model's performance across context or training iterations as an inherent statistical property whereas other communities prefer empirical evidence to give real-world meaning to such uncertainty. The choice of handling uncertainty requires not just an epistemic judgment, but also a moral one. Under computational capture, the decision over what is the right interpretation of this uncertainty is made by techno-elites. For example, misclassification of a few or a few thousand instances might have no impact on the model performance, leaving ML scientists unbothered, which can materialize into devastating and traumatizing effects in real life, as demonstrated in the case of Robert Williams (Williams 2021).

## 4.2 Inadequate Epistemic Authority Under *Computational Capture*

Under computational capture, ML scientists' epistemic authority rests on their expertise in designing models and evaluating performances against benchmark datasets. This subsection shows that such authority is increasingly insufficient for reflecting the diversity of current ML research. In today's context, a complementary form of epistemic authority - centered on practical efficacy - has become urgent. ML scientists should share the task of cultivating this new form of epistemic authority with practitioners who work at the sites of system building and deployment.

Science and Technology Studies scholars (Haraway 2013; Harding 2013) showed that objects of knowledge in science are historically constituted. The same object of study is represented differently across time, attributed varying scientific or cultural meanings, and associated with different effects. For example, objects of study (such as "rocks" and "electrons") in the natural sciences, though intuitively perceived to possess permanence and stability, are "socially constituted" (Harding 2013). Without delving into whether ML is science or engineering, it is crucial to stress that objects of inquiry in ML - usually mathematically constituted - are also socially shaped. This has broad implications for research culture beyond the objects of inquiry themselves.

ML has its early roots in AI research from the 1950s. At the time, computer scientists' artifacts were basic programs and rule-based games like chess, and the field held a scientific register with focuses on intelligence, learning, and understanding - qualities associated with Artificial General Intelligence (AGI). AI scientists in the earlier decades hoped that computer programs could demonstrate these qualities

(Minsky 1961; O'Regan 2013; Bennani-Taylor 2024). Current day ML research submissions also serves "scientific, engineering, and societal goals" (Blili-Hamelin et al. 2025) which can be observed from ML conferences' submissions, and usually these goals are mixed (Blili-Hamelin et al. 2025). It is important to note that even the most "basic" or "foundational" ML research is **not application-neutral - produced with no knowledge or intention for how it might be used**. ML scientists must respond to funders' needs to accrue funding as a community, therefore, unavoidably *not* application-neutral. For example, an analysis of U.S. Department of Defense (DoD) grant solicitations targeting AI researchers (Widder, Gururaja, and Suchman 2024) reveals how ML scientists and the DoD engage in mutual enlistment, and how "basic research" often obscures the fact that these projects serve militaristic agendas. The privatization of ML research by tech giants also indicates that basic research responds to research agendas set by private funders (Whittaker 2021). Put simply, AI has evolved from being "chiefly a scientific research subject during the twentieth century to a widely commercialized, [*politicized, and weaponized*] innovation" [italic texts are added] (Liu 2021) and it serves a broad set of goals that in ways render ML research *tools* rather than only scientific understanding.

The prominent goals of ML for science, technology, and social benefits call for a form of epistemic authority centered on practical efficacy, different from epistemic authority over model building and evaluating in lab settings. In other words, "predictive performance alone is not a good indicator of the worth of a [ML] publication" (Karl et al. 2024). However, one question remains, why should ML scientists cultivate this new form of epistemic authority instead of just practitioners who "specializes" in actualizing concrete goals.

Epistemic authority over practical efficacy means the capacity to adequately evaluate, explain, and justify how well an ML study serves the purported goals with considerations of intended contexts, operational constraints, and emergent society impacts, etc. There are two reasons for why ML scientists should cultivate this form of epistemic authority. First justification is ethical. The knowledge mainstream ML research produces is general but comes with a *generality debt* - "relying on the generality or flexibility of tools to postpone crucial engineering, scientific, or societal decisions" (Blili-Hamelin et al. 2025); this brings us back to Section 2.4, the postponing of difficult decisions will exacerbate the structural violence that is already great. Second justification is practical. ML scientists have unique influence over the choice of projects, model architectures and evaluation methods, infrastructural requirements, and the discursive dimension - shaping narratives and pre-empting misuse.

Sections 3 and 4 together are our diagnosis for the absence of accountability in the *claim-reality gap*, and the diagnosis indicates that both epistemic assumptions and ML research regime resist accountability, and these resistances render the *claim-reality gap* a *dead zone of accountability*. The final section will offer potential solutions to create conditions for developing accountability in the *claim-reality gap*.

## 5 Prescription: Social Claims Should be Articulated and Defended

This section outlines strategies to rectify the conditions that we identified throughout the paper to create possibilities for accountability in the *claim-reality gap*. In Section 5.1, we identify a barrier that can discourage epistemic reform in ML research and connect this to the discussion in Section 4.2 on epistemic authority as a means to mitigate such discouragement. In 5.2, we propose solutions to cultivate ML scientists with a new form of epistemic authority and create the conditions necessary for accountability.

### 5.1 Cultivating a New Form of Epistemic Authority

ML scientists risk certain losses if their research is held accountable for social considerations. Epistemically, their authority would be subjected to unfamiliar standards and contestation beyond relatively narrower computational evidence. Materially, justification of societal impact would face greater scrutiny, possibly threatening funding opportunities and diminished socio-economical status. As Porter (Porter 1996) observed:

"[S]cience is [made by] communities that are often troubled, insecure, and poorly insulated from outside criticism. Some of the most distinctive and typical features of scientific discourse reflect this weakness of community. . . [O]bjectivity in science is. . . partly a response to the. . . pressures." [p. 230]

The rigidity and control - characteristic of computational capture - serve to generate a shared discourse that unifies a "weak research community". These boundaries demarcate the field to resist external contestation, but engagement with the social breaks down these boundaries and incorporating knowledge from other domains can indeed be destabilizing (Burrell 2024).

Green and Viljoen (2020) argue that reform demands computer scientists recognize their positionality and develop the epistemic capacity to navigate sociopolitical dimensions. While we share this view, we emphasize that many ML scientists - especially those working on general-purpose models - care about social good, but do not primarily identify with the ML-for-social-good community as their main focus. Translating critical insight into systemic reform means acknowledging this heterogeneity. Reform efforts must recognize and accommodate this "epistemic diversity" within ML community<sup>5</sup>.

Given the disciplinary "insecurity", any effort to institutionalize socio-technical responsibility must acknowledge the pushbacks these reforms might face due to potential loss of resources and diminished epistemic authority. For this reason, we propose the *rigorous articulation of social claims* as a strategy likely to encounter less resistance as it affords ML scientists a new form of epistemic authority.

<sup>5</sup>In the context of the sciences, Leonelli (Leonelli 2023) identifies six broad categories to understand scientists' epistemic diversity: material, conceptual, methodological, infrastructural, socio-cultural, and institutional.

If social claims become the basis for a new form of epistemic authority, challenging computational capture would expand rather than diminish ML scientists' authority. Current reform tools - such as impact statements and ethics checklists - often cast sociotechnical concerns as limitations to be disclosed (Ashurst et al. 2022b). In contrast, when ML scientists explicitly ground their study in potential contexts, reason with limitations, uncertainty, and bias, they demonstrate the kind of intellectual rigor that defines good science. Efforts like these should be recognized and rewarded. Establishing this new epistemic authority means that computation alone does not determine the relevance or value of ML studies or the competence of ML scientists. Instead, they are evaluated for the quality of their reasoning and their ability to articulate both computational and social dimensions as claims - not simply for producing high, but often ambiguous, model performance scores (Bili-Hamelin et al. 2025).

The new form of epistemic authority mentioned requires specific arrangements in ML research practices. Epistemic authority is derived from expertise; assigning epistemic authority over practical efficacy implies expertise in proving practical efficacy. Just as model performance claims go through rigorous scrutiny, we propose that social claims should also undergo scrutiny. This brings us to the suggestion that social claims should be treated as knowledge claims that need to be articulated and defended with evidence, rather than being speculative or implicit. These knowledge claims should be justified with "sound reasoning and credible evidence". If a shared standard for such reasoning emerged, we could potentially design accountability mechanisms to hold ML researchers responsible for not specifying their social claims and/or for not providing sufficient evidence for them. Reform of this kind is within the would be internal to ML research instead of compartmentalizing tasks of computation and social in a separate basket such as ethical checklists or impact statements.

## 5.2 Two Collaborative Research Agenda of Developing Conditions for Social Claim Accountability

The goal suggested at the end of 4.1 is not easy to achieve because the maturing and reform of scientific norms and practices takes time. The Human-Computer Interaction (HCI) community has a history of developing tools or mechanisms to improve the rigor of scientific research, such as (Feger et al. 2019). In this section, we propose two collaborative research avenues that involve the HCI community to develop and implement evaluation mechanisms for social claims to create conditions of possibility for social claim accountability mechanisms.

**Identify Categories of Social Claims** ML papers frequently make broad social claims without specifying concrete applications. This vagueness is partly intentional, as it implies the generalizability of the research findings while acknowledging the difficulty of predicting all possible future uses of a method. The uncertainty around the balance between generality and specificity poses a challenge to mainstream ML researchers who focus on benchmarking perfor-

mances - they may struggle to scope social claims appropriately or to assemble suitable supporting evidence. While ML researchers are generally clear on the standards for justifying a computational contribution, there are few established norms or examples for substantiating more complex, context-dependent social claims.

To bridge this gap, a critical first step we suggest is developing an understanding of social claims by aggregating existing social claims from ML papers and deriving social claims from ML domain applications with the help of domain experts and other civic tech communities. Eventually, the social claim research community can 1) help develop social claim categories with case examples and 2) derive commonly used evidence to support different categories of social claims. This growing repository of social claims with case examples, and commonly used evidence (for each type of claim) can serve as pedagogical resources that can be integrated into ML scientists' training for socio-technical understanding of their work, and as a reference tool to substantiate their social claims when evaluating their methods and drafting manuscripts.

### Help ML Researchers Systemize Supporting Evidence

ML research workflow is characterized by the Common Task Framework (Donoho 2017) which emphasizes standardization of research process and reporting. However, such standardization is not observed in evidence presentation. Even for model performance claims - arguably the most rigorously evaluated aspect of ML research - there is typically no explicit representation of the relationship between a claim and its supporting evidence. Reviewers mainly rely on their trained judgment to evaluate its validity. However, this lack of transparency can obscure the reasoning process. To address this gap, HCI researchers can collaborate with ML researchers to identify common evidence in ML research. Common forms of evidence include benchmarking comparisons, cross-validation, ablation studies, A/B testing, deployment observations, and field feedback (Kou 2024).

Besides identifying key types of evidence, the HCI and ML communities could co-develop tools for documenting, organizing, and presenting evidence in ML papers. For instance, one approach could involve the use of diagrams that map individual claims to their supporting evidence, with each diagram representing a single claim. These visualizations could illustrate the logical connections between different types of evidence and the claims they support, offering a clear view of the study's inferential structure. Beyond improving clarity for fellow researchers, these tools would make ML research more accessible to less technical audiences, such as policymakers, AI auditors, and domain experts, who lack the specialized ML knowledge needed to parse technical content.

## 6 Conclusion - "Love Your Monsters"

This paper coins the term *dead zone of accountability* - a lens through which scholars can identify aspects of ML ecosystem that resist accountability. We demonstrated that our target case - the *claim-reality gap* - is a dead zone of accountability by showing the epistemic assumptions and ML

epistemological underpinnings that resist accountability for *claim reality gap*. We ended the paper with two collaborative research avenues that could create the conditions for developing accountability mechanisms for the *claim-reality gap*.

The AI ethics discourse underscores the importance of stepping out of *technological somnambulism* (Winner 1983) - sleepwalking through the development and deployment of technology without reflection. Yet in the general culture of ML research, this slumber has morphed into technological gaslighting ; the field is aware its creations rarely live up to or often fail their promises, yet it perpetuates unwarranted hype, voicing no concerns while reaping the rewards of inflated expectations. Bruno Latour reminds us in *Love Your Monsters* (2012) that sin is not in creation but in abandonment. ML scientists produce “context-free” models - but fail to care for the social worlds their knowledge reorders. The goal of ML research for societal impact and meaningful innovation can benefit from accountability for social claims, not evading them. On that note, we offer *articulation and justification of social claims* as one of many potential ways for ML research to evolve as situated actors who love and care for their knowledge products.

## Acknowledgements

In addition to his co-authors, Tianqi is especially grateful to his mentors: Daniel Susser, for providing feedback on an earlier draft of this work; David Gray Widder, for supporting his line of work and career development; and Leif Hancox-Li, for their generous commitment to improving Tianqi’s other writings and guiding his PhD studies.

## References

Ackerman, M. S. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Human-Computer Interaction*, 15(2-3): 179–203.

Andrews, M.; Smart, A.; and Birhane, A. 2024. The reanimation of pseudoscience in machine learning and its ethical repercussions. *Patterns*, 5(9).

Ankeny, R. A.; and Leonelli, S. 2016. Repertoires: A post-Kuhnian perspective on scientific change and collaborative research. *Studies in History and Philosophy of Science Part A*, 60: 18–28.

Ashurst, C.; Barocas, S.; Campbell, R.; and Raji, D. 2022a. Disentangling the components of ethical research in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2057–2068.

Ashurst, C.; Hine, E.; Sedille, P.; and Carlier, A. 2022b. Ai ethics statements: analysis and lessons learnt from neurips broader impact statements. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, 2047–2056.

Baker, R. S.; and Hawn, A. 2022. Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*, 1–41.

Bennani-Taylor, S. 2024. Infrastructuring AI: The stabilization of ‘artificial intelligence’ in and beyond national AI strategies. *First Monday*.

Bi, N.; Suen, C. Y.; Nobile, N.; and Tan, J. 2019. A multi-feature selection approach for gender identification of handwriting based on kernel mutual information. *Pattern Recognition Letters*, 121: 123–132.

Bialski, P. 2024. *Middle Tech: Software Work and the Culture of Good Enough*, volume 36. Princeton University Press.

Birhane, A.; Kalluri, P.; Card, D.; Agnew, W.; Dotan, R.; and Bao, M. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 173–184.

Blili-Hamelin, B.; Graziul, C.; Hancox-Li, L.; Hazan, H.; El-Mhamdi, E.-M.; Ghosh, A.; Heller, K.; Metcalf, J.; Murai, F.; Salvaggio, E.; et al. 2025. Stop treating AGI as the north-star goal of AI research. *arXiv preprint arXiv:2502.03689*.

Bouthillier, X.; Laurent, C.; and Vincent, P. 2019. Unreproducible research is reproducible. In *International Conference on Machine Learning*, 725–734. PMLR.

Bovens, M. 2007. Analysing and assessing accountability: A conceptual framework 1. *European law journal*, 13(4): 447–468.

Burrell, J. 2024. Automated decision-making as domination. *First Monday*.

Bush, V. 2021. Science, the endless frontier.

Campolo, A.; and Crawford, K. 2020. Enchanted determinism: Power without responsibility in artificial intelligence. *Engaging Science, Technology, and Society*.

Caton, S.; and Haas, C. 2024. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7): 1–38.

Cheng, Z.; Zhu, X.; and Gong, S. 2018. Surveillance face recognition challenge. *arXiv preprint arXiv:1804.09691*.

Cho, K. 2024. I sensed anxiety and frustration at NeurIPS’24.

Church, K. W.; and Kordoni, V. 2022. Emerging trends: Sota-chasing. *Natural Language Engineering*, 28(2): 249–269.

Cooper, A. F.; and Vidan, G. 2022. Making the Unaccountable Internet: The Changing Meaning of Accounting in the Early ARPANET. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 726–742.

Corbett, E.; and Denton, E. 2023. Interrogating the T in FAccT. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1624–1634.

Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23.

D’ignazio, C.; and Klein, L. F. 2023. *Data feminism*. MIT press.

Donoho, D. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4): 745–766.

- Douglas, H.; and Branch, T. 2024. The social contract for science and the value-free ideal. *Synthese*, 203(2): 40.
- Douglas, H. E. 2009. *Science, policy, and the value-free ideal*. University of Pittsburgh Pre.
- Ehsan, U.; Liao, Q. V.; Muller, M.; Riedl, M. O.; and Weisz, J. D. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–19.
- Epstein, S. 1996. *Impure science: AIDS, activism, and the politics of knowledge*, volume 7. Univ of California Press.
- Feger, S. S.; Dallmeier-Tiessen, S.; Woźniak, P. W.; and Schmidt, A. 2019. The role of hci in reproducible science: Understanding, supporting and motivating core practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6.
- Feyerabend, P. 1970. *Conversations for the Specialist*.
- Foucault, M. 2005. *The order of things*. Routledge.
- Goetze, T. S. 2022. Mind the gap: autonomous systems, the responsibility gap, and moral entanglement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 390–400.
- Graeber, D. 2012. Dead zones of the imagination: On violence, bureaucracy, and interpretive labor: The Malinowski Memorial Lecture, 2006. *HAU: journal of Ethnographic Theory*, 2(2): 105–128.
- Green, B.; and Viljoen, S. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 19–31.
- Grill, G. 2022. Constructing Certainty in Machine Learning: On the performativity of testing and its hold on the future.
- Hancox-Li, L.; and Kumar, I. E. 2021. Epistemic values in feature importance methods: Lessons from feminist epistemology. In *proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 817–826.
- Haraway, D. 2013. Situated knowledges: The science question in feminism and the privilege of partial perspective 1. In *Women, science, and technology*, 455–472. Routledge.
- Harding, S. 2013. Rethinking standpoint epistemology: What is “strong objectivity”? In *Feminist epistemologies*, 49–82. Routledge.
- Henderson, P.; Islam, R.; Bachman, P.; Pineau, J.; Precup, D.; and Meger, D. 2018. Deep reinforcement learning that matters. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.
- Herrmann, M.; Lange, F. J. D.; Eggensperger, K.; Casalicchio, G.; Wever, M.; Feurer, M.; Rügamer, D.; Hüllermeier, E.; Boulesteix, A.-L.; and Bischl, B. 2024. Position: Why we must rethink empirical research in machine learning.
- Hill, F. 2024. 200bn Weights of Responsibility.
- Humphreys, P. 2020. Why automated science should be cautiously welcomed. *A Critical Reflection on Automated Science: Will Science Remain Human?*, 11–26.
- Karl, F.; Kemeter, L. M.; Dax, G.; and Sierak, P. 2024. Position: Embracing Negative Results in Machine Learning. *arXiv preprint arXiv:2406.03980*.
- Keller, E. F. 1982. Feminism and science. *Signs: Journal of women in culture and society*, 7(3): 589–602.
- Kippin, S.; and Cairney, P. 2022. The COVID-19 exams fiasco across the UK: four nations and two windows of opportunity. *British Politics*, 17(1): 1–23.
- Kirchner, L.; and Goldstein, M. 2020a. Access Denied: Faulty Automated Background Checks Freeze Out Renters—The Markup.
- Kirchner, L.; and Goldstein, M. 2020b. How automated background checks freeze out renters. *The New York Times*, 28.
- Kou, T. 2024. From Model Performance to Claim: How a Change of Focus in Machine Learning Replicability Can Help Bridge the Responsibility Gap. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, 1002–1013.
- Kuhn, T. S. 1997. *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago.
- Kumar, S.; Singh, S.; and Kumar, J. 2019. Gender classification using machine learning with multi-feature method. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, 0648–0653. IEEE.
- Latour, B.; and Woolgar, S. 1987. *Laboratory life: The construction of scientific facts*. Princeton university press.
- Laufer, B.; Gilbert, T.; and Nissenbaum, H. 2023. Optimization’s neglected normative commitments. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 50–63.
- Lecher, C. 2018. What happens when an algorithm cuts your health care. *The Verge*, 21(3).
- Leonelli, S. 2023. *Philosophy of open science*. Cambridge University Press.
- Liberman, M. 2010. Obituary: Fred Jelinek. *Computational Linguistics*, 36(4): 595–599.
- Lipton, Z. C.; and Steinhardt, J. 2019. Troubling Trends in Machine Learning Scholarship: Some ML papers suffer from flaws that could mislead the public and stymie future research. *Queue*, 17(1): 45–77.
- Liu, Z. 2021. Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass*, 15(3): e12851.
- Mackey, L.; Poland, C.; and Chen, J. 2023. Reflections on the NeurIPS 2023 Ethics Review Process.
- McCarthy, J. 1997. AI as sport.
- McCadden, M. D.; Joshi, S.; Anderson, J. A.; and London, A. J. 2023. A normative framework for artificial intelligence as a sociotechnical system in healthcare. *Patterns*, 4(11).
- Minsky, M. 1961. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1): 8–30.
- O’Regan, G. 2013. *Giants of Computing*.
- Pamuk, Z. 2024. *Politics and expertise: How to use science in a democratic society*. Princeton University Press.
- Patterson, D.; Snyder, L.; and Ullman, J. 1999. Evaluating Computer Scientists and Engineers For Promotion and Tenure.

- Phan, T.; Goldenfein, J.; Mann, M.; and Kuch, D. 2022. Economies of virtue: the circulation of ‘ethics’ in Big Tech. *Science as culture*, 31(1): 121–135.
- Porter, T. M. 1996. *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton University Press.
- Raji, I. D.; Bender, E. M.; Paullada, A.; Denton, E.; and Hanna, A. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.
- Raji, I. D.; Kumar, I. E.; Horowitz, A.; and Selbst, A. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 959–972.
- Rubambiza, G.; Sengers, P.; Weatherspoon, H.; and Liu, J. 2024. Seam Work and Simulacra of Societal Impact in Networking Research: A Critical Technical Practice Approach. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*, 59–68.
- Taigman, Y.; Yang, M.; Ranzato, M.; and Wolf, L. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1701–1708.
- Vinsel, L. 2021. You’re Doing It Wrong: Notes on Criticism and Technology Hype.
- Wagstaff, K. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656*.
- Wang, Y.; and Kosinski, M. 2018. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of personality and social psychology*, 114(2): 246.
- Whittaker, M. 2021. The steep cost of capture. *Interactions*, 28(6): 50–55.
- Widder, D. G.; Gururaja, S.; and Suchman, L. 2024. Basic Research, Lethal Effects: Military AI Research Funding as Enlistment. *arXiv preprint arXiv:2411.17840*.
- Widder, D. G.; and Nafus, D. 2023. Dislocated accountabilities in the “AI supply chain”: Modularity and developers’ notions of responsibility. *Big Data & Society*, 10(1): 20539517231177620.
- Williams, R. 2021. I Did Nothing Wrong. I Was Arrested Anyway.
- Winner, L. 1983. Technologies as forms of life. In *Epistemology, methodology, and the social sciences*, 249–263. Springer.
- Wolpert, L. 1999. Is science dangerous? *Nature*, 398(6725): 281–282.
- Young, M.; Ehsan, U.; Singh, R.; Tafesse, E.; Gilman, M.; Harrington, C.; and Metcalf, J. 2024. Participation versus scale: Tensions in the practical demands on participatory AI. *First Monday*.