From Model Performance to Claim: How a Change of Focus in Machine Learning Replicability Can Help Bridge the Responsibility Gap

TIANQI KOU, Penn State University, USA

Two goals – improving replicability and accountability of Machine Learning research respectively, have accrued much attention from the AI ethics and the Machine Learning community. Despite sharing the measures of improving transparency, the two goals are discussed in different registers - replicability registers with scientific reasoning whereas accountability registers with ethical reasoning. Given the existing challenge of the responsibility gap – holding Machine Learning scientists accountable for Machine Learning harms due to them being far from sites of application, this paper posits that reconceptualizing replicability can help bridge the gap. Through a shift from *model performance replicability* to *claim replicability*, Machine Learning scientists can be held accountable for producing non-replicable claims that are prone to eliciting harm due to misuse and misinterpretation. In this paper, I make the following contributions. First, I define and distinguish two forms of replicability for ML research that can aid constructive conversations around replicability. Second, I formulate an argument for claim-replicability's advantage over model performance replicability in justifying assigning accountability to Machine Learning scientists for producing non-replicable claims and show how it enacts a sense of responsibility that is actionable. In addition, I characterize the implementation of claim replicability as more of a social project than a technical one by discussing its competing epistemological principles, practical implications on *Circulating Reference*, *Interpretative Labor*, and research communication.

1 INTRODUCTION

In recent years, the AI ethics community has produced much literature on improving Machine Learning (ML) transparency. On the one hand, transparency measures can serve the goal of improving accountability. For this goal, transparency measures focus on making artifacts, actors, and development processes open for external auditing and regulations to make prevention, identification, and mitigation of harms easier, and to hold relevant parties accountable. On the other hand, transparency measures have been called for to uphold replicability – maintaining the scientific rigor and integrity of ML research [5] by ensuring that the research process and artifacts are adequately shared to facilitate re-run of studies for verifying the study's finding's validity [33].

Author's address: Tianqi Kou, tfk5237@psu.edu, Penn State University, University Park, PA, USA.

1.1 Transparency Measures and the Origins of Concern

The narrative of *transparency for accountability* initiated from public concerns over the increasing harms generated from ML research and technological systems, which garnered attention from domain experts, policymakers, civil society, AI ethics scholars, and ML scientists from academia and industry. Discerning clear lines of responsibility for harms generated from a complex system that involves numerous parties has been a challenging undertaking, and this obstacle has been famously named the responsibility gap [41] - who is responsible for this harm and what does the responsibility call for? As such, developing a mature accountability mechanism that can address wide-ranging harms and satisfy diverse stakeholders is still in progress, despite advances made in computer science, social sciences, and other AI Ethics contributing fields.

The narrative of improving transparency to improve replicability gained traction after the identification of the replication crisis – many fields have observed discrepancies between the results from original study and its replication study [27, 57, 69]. These concerns exist in the sciences [27, 57, 69, 77, 78, 93] as well as in ML [67, 83, 101, 109, 135]. In response to these concerns, the Open Science Movement was initiated by leading institutions such as DARPA and ACM; top ML conferences such as KDD, NeurIPS, ICLR, and ICML, etc., have established submission and review guidelines to improve transparency about research artifacts and assumptions.

Taken together, one might argue that transparency can be mobilized toward both goals. Existiting transparency measures include - transparency of decision process [71, 132], data development/documentation/curation [23, 55, 58, 98, 102, 105, 128], transparency of model statistics, assumptions, failure modes [60, 90, 91]; transparency of research limitation and impact [6, 65, 70, 89, 111, 113, 126]; transparency of research artefacts such as code and data [3, 12, 66, 81, 127]; transparency of explanation [7–9, 37, 82, 117, 118, 121, 140]; transparency of social context [23, 45, 119], transparency of workflow [3, 81, 84, 88].

Although two goals share measures of transparency, they have different lineages. The concept of accountability arose from moral philosophy, registering with moral or ethical reasoning. Research addressing ML accountability predominantly appears in conferences such as AIES and FAccT, venues with a strong focus on ethics, discussed in tangent with responsibility, blame, and other social values such as fairness and inclusivity. In contrast, replicability has been considered the "gold standard" [75] and "a key ingredient"[116] of science and engineering because it helps establish consistency and regularity, constitutes testability or falsifiability [13, 103]. Although practicing replication traces back to at least as early as the 1600s [123], concerted efforts to interrogate the concept of replicability started around the 1980s [16, 17]. The investigations of the meaning [13, 33, 85, 107, 116], operationalization challenges [16, 35, 36, 50, 106], the social dimension [16, 17, 30, 54], incentives [32, 92], limitations [51, 77, 78], and necessity [31, 56] of replicability remain prominent in *History and Philosophy of Science* and *Sociology of Science*. In this lineage, replicability is usually discussed in tangent with other epistemic values such as *certainty and falsifiability* [103] or *objectivity*.

1.2 Contribution

Despite the abundance of measures to improve transparency, the adoption rate remains low [71] and the problem of the responsibility gap remains salient [5, 41]. To improve the adoption rate of the above measures for the aim of bridging the responsibility gap, one major challenge is motivating scientists to proactively integrate those tools to engage in social reflection. To such end, this paper conceptualizes a new relationship between the task of improving replicability and the task of engaging in social reflection – social reflections as a pre-requisite of upholding *claim replicability. Claim replicability* has advantages over the currently adopted *model performance replicability* to make ML scientists a directly

accountable party for harms inflicted by non-replicable claims due to misuse and misinterpretation, positioning social considerations as ML scientists' role responsibility, subsequently helps bridge the responsibility gap.

1.3   Outline

The rest of this section defines and distinguishes *model performance replicability* and *claim replicability*. Section 2 lays the foundation for the the main argument by showing the significance of both *model performance claims* (claims to generalizability and robustness of model performance) and *social claims* (ML method's deliverance of functionality, efficiency, social benefits, etc.), and the relationship between the two types of claims' replicability. 3.1 and 3.2 constitute the argument for claim replicability's ability to hold ML scientists directly accountable for producing non-replicable claims. Remaining of section 3 address potential rebuttals against claim replicability. Section 4 will discuss claim-replicability's practical implications on *Circulating Reference*[97], *Interpretive Labor*[43], and research communication.

1.4   Definitions of Model Performance Replicability and Claim Replicability

I define model performance replicability and claim replicability for ML research as the following (Note: for the rest of the paper, I use MPR and CR instead of model performance replicability and claim replicability for concise expression.):

- *Model performance replicability: getting the same performance in a replication study.*
- *Claim replicability: making the same research claim in a replication study.*

1.4.1   Object of Replication and Two Interpretations of "Result". Within the discussion of replicability in the ML community, there exist two beliefs. In the first, a study's transparency is equivalent to its replicability; and the second, a study is of replicability if the same result is obtained in the re-run. Both beliefs are valid, but they adopt different *objects of replication. Object of replication* refers to *the thing that is being replicated.* For the first belief, the object of replication is *the process of a study*; under this belief, the re-run only needs to be able to go through the original study's procedure to determine the replicability of the original study. This is similar to what [107] calls "material realization" – all materials and steps of execution are transparently shared, and the same operating condition is obtainable so the replication runner can materially mimic with high fidelity to the orignal study. For the second belief, the object of replication is *result* – the result must be replicated in the re-run to claim that the original study is of replicability. This distinction is important because replication of process does not guarantee replication of results – the same process can lead to different results.

The key definitional distinction between MPR and CR is the interpretation of *results*. Despite the importance of the distinction as I will show throughout this paper, influential institutions are vague in their official definitions of replicability regarding the meaning of "results". For example, ACM's definitions [2] use "same measurement", and [93] uses "consistent results". In scholars' discussion of replicability, the dominant (probably the only) interpretation of results is model performance, such as in [21, 39, 48, 58, 67, 101, 108, 120].

Despite quantitative interpretation of result as model performance, result can also be interpreted qualitatively. In the context of sciences more broadly, [42]'s notion of inferential replicability interprets results as *qualitative conclusions* , [47]'s notion of *interpretation replicability*. CR as I defined above is of the same qualitative nature as the other two notions. The different choice of words is a matter of preference to reduce unnecessary confusion because as I will state later a paper usually makes multiple claims, and it would be confusing to say that a paper usually draws multiple *conclusions* or makes multiple *interpretations*.

1.4.2 Two Characteristics of Claim Replicability. Besides adopting different interpretations of "result", two other characteristics of CR set it apart from model performance replicability.

Corresponding to individual claim. MPR is assigned to a whole study because a study usually focuses on a single ML method/model. In contrast, CR correspondes to individual claim because as I will show in sections 2 a single study usually makes multiple claims, and each claim can be of different status in terms of its CR.

A qualitative property. MPR is commonly expressed as either binary (replicable or not replicable which corresponds to "1/0") or probabilistic (such as the application of *prediction markets* to evaluate study replicability [83]). These expressions are appealing because similar to physical properties such as *melting point* or *density* in physical sciences, they are neat, portable, and easy for comparison. However, this quantitative interpretation has been pointed out to be "too coarse-grained to support replication's function of evidence amalgamation" [35]. [77]'s analysis is another poignant case against the quantitative conceptualization of replicability - "direct replicability" which "is associated with experimental research methods that yield numerical outcomes". Quantitative conception of replicability assumes (or requires) researchers (to) have a high level of control over the materials and procedure through standardization, which incentivizes scientists to focus on reporting standardized procedures while leaving out the idiosyncracies of their study. Contrast to MPR's focus on model performance, CR focuses on evaluating *what is being said* about the model performance. Reliable statements can and should be made even when model performance is not stable.

CR's qualitative nature is derived from the complexity of justification beyond quantitative evidence. CR therefore does not necessitate the form of replication study that strictly mimics key aspects of the original study because standardization of procedure and reporting is only one of many ways to run replications to evaluate CR. To replicate a claim, the replication study can take the form of a separate study aiming at solving the same problem (triangulation).

2 DISENTANGLING REPLICABILITY OF MODEL PERFORMANCE CLAIM AND REPLICABILITY OF SOCIAL CLAIM

CR's advantage over MPR in helping bridge the responsibility gap relies on CR's requirement to replicate a diverse set of claims that are made in ML papers – both claims to model performance and claims to the social. Under MPR, the only claim that receives attention for replication is claims to model performance's generalizability and robustness. However, as 2.1 will show, claims are made not just to model performance, but also often to broader social contexts (ML methods' ability to deliver efficiency of a decision process, functionality of a system, explanatory power, or social benefits such as fairness, etc.). This section points out 1) the existence of social claims, 2) the underestimated significance of the replicability of social claims, and 3) the relationship between replicability of the two types of claim.

2.1 Diverse Claims in ML Papers – Model Performance Claim and Social Claim

In [11] where the team conducted document analysis on one hundred ML papers from top conferences, they show that the annotated studies foreground values such as performance, robustness, and generalizability, and the main body of the paper focuses on justifying these properties, appealing to the need of the ML community. In contrast, in papers that mention social needs, the connection between the method and those needs is loose and rarely engaged in the main body of the paper. In addition, few qualifications (eg. mentioning limitations) are offered – only two out of the one hundred papers mentioned negative potential impact; even then, they are "abstract and hypothetical" [11, p.176]. For my paper, I apply the language of "claim" to [11]'s finding - ML papers foreground model performance claims (generalizability and robustness), and rarely engage in justifying or qualifying social claims.

Social claims are commonplace in ML papers. For example, [11] identified commonly included social aspects such as – efficiency, understanding, novelty, real-world applicability, scalability, easiness to work with, fairness, etc. Turning to more concrete examples, [139] developed a risk level assessment NLP model to classify text messages received from pregnant people. In their paper, in addition to making model performance claims such as "TRIM-AI significantly outperforms state-of-the-art baselines" [139, p. 2], they also make social claims to functionality such as "better extract semantic and syntax information from code-mixed sentences as compared to hierarchical neural networks", "improve their operational efficiency, while lowering the operational costs [of the agents working with the system where their method is embedded]". In [134, p. 1] where they developed an ML method for detecting fake news, in addition to model performance claims, they also make social claims such as "benefit the detection of fake news on newly arrived events".

When a claim is "loosely connected" or "rarely engaged" [11], it necessarily means that there is little to no evidence in the paper to uphold the claim, under-investigated, and very likely non-replicable - the *claimed good* cannot be delivered in practice. The widespread low replicability of social claims has led to frustrations and concerns, expressed as calling ML papers for a clearer articulation of how a method can translate to concrete impact in ML papers, such as in [15] and in [87, p. 4] where they state - "There are many examples of diagnostic aids, tools, and systems that demonstrate strong accuracy but have failed to yield benefits to patients.".

2.2 The Significance of the Replicability of Social Claim

The belief that producing replicable model performance is ML scientists' core responsibility is prevalent and the academic reward system demonstrates a strong focus on the production of innovative models with good performance. However, the lopsided attention MPR receives compared to CR does not mean that replicability of social claims is not important. To understand why, we need to look into the impact of social claims.

In discussing using data as evidence for scientific claims, Sabina Leonelli states:

> What readers are required to take away from a paper is not the data themselves but rather the empirical interpretation of those data provided by the authors in the form of a claim. [76]

Scientists communicate with the audience of their work in the form of claims. In a paper that summarizes the process and findings of a study that utilizes statistical experiments, authors cannot communicate their findings with mere numerical information, such as p-value, sample size, confidence level, etc. They must interpret those numbers and form claims in sentences to turn them into *"truth" or usable knowledge*.

ML methods are situated in social contexts [19, 61, 73, 74, 100]. ML scientists' narrative of "changing the world" [96] has led to the development of ML algorithms for a variety of social contexts [44, 45, 129, 141]. With the growing number of cases of algorithmic harms and functionality failures [110], we might ask - would the application of ML methods have occurred if people introduced them knew beforehand that harms or failures would happen, or if people who introduced them did not believe that introducing those methods will improve the state of affair in some way(s)? The answer is usually no (except in rare cases of deception). Introducing ML to sites of application is usually believed to be able to positively help the introducer achieve certain goals, such as introducing a pretrial risk assessment algorithm to reduce bias and workload through automation [131].

Where do users of ML methods adopt the belief that they will do good? In our society, "expertise is almost always external: it belongs to someone else and our problem is how to recognize it, access it, and mobilize it." [122, p. 46] Therefore, the most immediate voices of authority that users of ML methods turn to are the scientists who developed those methods and communicated the significance of their methods in publications. Social claims, compared to model

performance claims that are expressed in technical language, become more salient voices of authority that *nudge*, *encourage*, *persuade*, and *enable* the circulation and legitimacy of ML methods into various sites of application.

Consider a hypothetical example. If an ML method produces an accuracy of 90% and the team of scientists wants their method to be applied for real-world applications or used in future benchmarking comparisons, and they cannot merely drop the number 90% in the conclusion section. Instead, they must make claims such as "our model's accuracy of 90% indicates that our model will outperform existing state-of-the-art methods" or "applying our method is a positive intervention into the societal issue *P* which motivated our project." Without claims as interpretations of model performance, a method will find it hard to travel outside the laboratory where it was developed to be used in benchmarking or be implemented for tackling practical challenges. Claims (especially social claims) shape the future of the developed method – what sites it travels to, what interpretations can be made, who will use the method, and for what purpose, etc. Therefore, as Heather Douglas stated in [25, p. 85]

> [In] scientific work…Making empirical claims should be considered as a kind of action, with often identifiable consequences to be considered, and as a kind of belief formation process.

In practice, social claims tend to be taken for granted [46, 110], and poorly justified claims [11] have led to abundant failures [110]. This is particularly relevant given the reality that ML methods are commonly used in a variety of contexts different from their original context of development [5, 52], the prevalent misuse of ML knowledge [110, 137], and the widely accessible computing power and toolkits.

2.3 Relationship Between Replicability of Model Performance Claim and Social Claim

In this subsection, I use two hypothetical examples to demonstrate that model performance replicability does not guarantee claim replicability, dissipating the misconception that MPR implies the validity of the entire study.

2.3.1 Hypothetical Case I. A study that developed an ML method to identify misinformation and made the following claims:

- Claim 1 (model performance claim): Our model outperforms *state-of-the-art* models in identifying misinformation with an accuracy of 95%.
- Claim 2 (social claim): Our model is the first to significantly reduce the workload of human moderators in identifying misinformation due to its unique feature of interpretability.

A replication study is run which yields accuracy of 85%, instead of the reported 95%. This discrepancy will invalidate (or at least cast doubt on) *claim 1.* Therefore, the replication runner can say that the study is not of MPR. However, in the replication study, the replication runner found that the deployment of the method decreased the workload of human moderators due to its unique feature of interpretability, although to a lesser degree because more time is needed to compensate for the decreased accuracy from the reported 95%. The replication runner cannot invalidate *claim 2* (not replicated) because there is no other misinformation detection method with the feature of interpretability. In this example, not choosing MPR over CR will unfairly deny the value of the study.

2.3.2 Hypothetical Case II. Take the development of *fair* ML models for decision-making in distributing resources such as student admission in higher education institutions as an example. ML methods in this space sometimes aim to uphold equality through parity across race and/or gender [10]. Consider a hypothetical study that made the following claims:

- Claim 1 (model performance claim): Our model delivers high accuracy and predictive parity across racial subgroups.
- Claim 2 (social claim): Applying our model will improve the state of fairness.

A replication study is run and predictive accuracy parity is obtained in the re-run. Thus, the *claim 1* is replicated. However, this does not mean the same for *claim 2* because the criticism of the algorithmic solutions to fairness being limited and potentially counterproductive [10] stands regardless of the replicability of *claim 1*. In this example, although model performance claim is replicated, social claim is not replicated. Using MPR to evaluate the validity of the study will lead to overestimation of the study's value and blind users to the harms the method can engender.

3 HOW CLAIM REPLICABILITY HELPS BRIDGE THE RESPONSIBILITY GAP

So far, I have clarified the definition of MPR and CR, and pointed out the importance of ensuring replicability of both model performance claims and social claims. The lopsided attention MPR receives demonstrates that under the current *paradigm* [72] of ML research, it is considered imperative to maintain and refine methods and mechanisms for the evaluation of model performance. In contrast, there exist no analogous standards to evaluate claims. This section will demonstrate how CR can help bridge the responsibility gap by holding ML scientists immediately accountable for their producing non-replicable claims. My argument for holding ML accountable applies [41]'s account of moral justification for holding computational professionals accountable for the harms generated by the systems designed by them. 3.3 distinguishes accountability and blame which surfaces tensions between CR and other epistemological perspectives that I argue can be reconciled.

3.1 Bridging the Gap

3.1.1 Vacarious Responsibility and Moral Entanglement. My account of holding ML scientists accountable for producing non-replicable claims utilizes two concepts from [40, 41] – *vicarious responsibility* and *moral entanglement*. Vicarious responsibility "concerns cases where one agent is responsible for the actions or behavior of another agent/entity" [41, p. 396]. For example, parents (the vicarious responsible agents) are vicariously responsible for their children's misbehavior at a party even though the child has their own agency. Within this relationship between vicariously responsible agents and the entity that they are responsible for, there is a moral entanglement because it is *uncertain* "where one's own agency ends and where another's begins." [41, p. 397]

According to Goetze [40, 41], moral entanglement should be conceptualized as a continuum and can be weak or strong [40] - the stronger it is, the stronger the moral obligation for the responsible agent to intervene and take accountability [41]. The strength of the moral entanglement depends on how central the aspect of vicariously responsible agents' identity connected with the behavior of the agent or entity they are responsible for is to the vicariously responsible agent. For example, parents' moral entanglement with their misbehaving toddler is stronger than a civilian's moral entanglement with their state that committed wrongs. This is because, according to Goetze, one's role as a parent is often more important than one's identity of being a citizen of a state.

In my argument, I posit that there is a moral entanglement arising from a type of "self-reflexive vicarious responsibility" [40]. This type of vicarious responsibility is implicated in the truck driver case cited in [41], where the truck driver who killed a child in a car accident while driving on a highway, is vicariously responsible for, and morally entangled with, the past version of themself - what could have they done differently to prevent the accident from happening even though they broke no rules, therefore, had no reason to question their actions at the time of the accident? As of now,

many non-replicable claims exist in the absence of a mechanism for evaluating claims, therefore only retrospectively being identified as problematic. Claim replicability does not hold a place in the current ML *research repertoire* [1] [4]:

> Well-aligned assemblages of the skills, behaviors, and material, social, and epistemic components that a group may use to practice certain kinds of science, and whose enactment affects the methods and results of research.

Therefore, in the context of CR, *reflexive vicarious responsibility* is characteristic of the relationship that ML scientists have with their constantly evolving scientific self. The moral entanglement in this various responsibility is strong for two reasons. First, their identity as scientists is central to who they are professionally. Second, engaging in methodological reflection and upholding replicability is a salient professional duty. Acting under such moral obligation is scientists' *role responsibility* [2] [25, p. 72]. Therefore, although non-replicable claims were produced unbeknownst to ML scientists in the past, they nevertheless have a strong moral obligation to intervene, make amends, and better their scientific self.

3.2   Taking Responsibility

The strong moral obligation characterized above begs answers to the question: what do we expect ML scientists, the vicariously responsible agents, to do? This subsection will bring this question home by highlighting the action of claim-making which enacts an actionable sense of responsibility for ML scientists to address the moral burden characterized above.

3.2.1   Research Claim as Research Product. We typically think of *ML methods as the research products* of a study, the quality of which can arguably be proxied by MPR. In contrast, CR highlights *research claim as research product*, MPR's myopic focus on model performance does not reach the bar. (There is no definitive measure of claim "quality" as I will show in 3.3, I choose "quality" to suit the metaphor of "product") Just as ML scientists are responsible for the quality of their method (by attending to generalizability and robustness), they are also responsible for research claims. Viewing research claims as research products brings attention to an under-investigated aspect of ML research - the *molding* of this product or the making of a claim. In the following, I will raise two points of consideration to situate the responsibility in the action of claim-making.

3.2.2   Imputation of Intentionality. The first point of consideration is the imputation of "intentionality" [64]. In [41], Goetze points out computational professionals impute their intentionality into the technological systems - a technological system being "poised to behave in a certain way" [64, p. 201] is not accidental [41, p. 9].

Model performances are of little intentionality because they are inert numerical expressions imbued with an aura of austere objectivity and numerical indifference. In contrast, claims home the intentionality of authors of a paper such as soliciting citation, attracting public and peer attention, or calling for implementation, etc. Authors should be aware of what intentionality they are imputing into a claim and if it is justified by evidence at hand. The consideration of intentionality can help ML scientists make replicable social claims by thinking through how their imputed intentionality determines how their methods might play out in the social world.

---

[1]I use research repertoire instead of research paradigm because Kuhn never provided a satisfying definition of paradigm.
[2]*Role responsibility* arises when one takes on a particular role in society and thus has additional obligations over and above the general responsibilities we all share.

3.2.3 Flexibility in Claim-Making. The second point of consideration is the flexibility inherent in expressing a claim. Making replicable claims requires taking advantage of such flexibility and showing humility. I will list a few here (and elaborate in 4.3) - excluding claims that lack sufficient evidence, introducing qualifications to claims to sufficiently address the uncertainties around the knowledge claim (for example, by attending to practical challenges and ethical concerns during implementation), and specifying targeted audience of their work to reduce the possibility of misinterpretation and misuse.

Such flexibility in claim-making speaks to several barriers to accountability named by Helen Nissenbaum [94]. First, the making of a claim is usually at the hands of the authors. When speaking about harms from failure of claim replicability, the *Problem of Many Hands* will not be as thorny. Note that I am not implying tracing harms to a non-replicable claim is always an easy task nor am I implying that all claims are easily made replicable by ML scientists. I am only speaking about responsibility for them to take when they did not utilize tools that already exist within the current *research repertoire*. Second, *Bugs'* [94] "rhetorical power" [18, p. 869] to normalize the existence of errors and unpredictability as unavoidable and acceptable in software, and algorithmic systems, makes it easy to scapegoat ML models when harms arise. ML scientists can wield such *rhetorical power* under MPR - *as great as our method can be, the computational stochasticity (e.g., mathematical randomness, shift of data distribution) still eludes the best of us*. Under CR, however, they cannot (as easily) use such rationale to scapegoat non-replicable claims.

3.3 How Challenges in Assigning Blame Surfaces Competing Epistemological Perspectives

Accountability is different from blame [115] despite blameworthiness' appeal of strengthening motivation. This subsection details the challenges involved in determining blameworthiness (adopting [115]'s conception of blame) which indicates tensions between CR and other epistemological perspectives, the reconciliation of which is possible.

In articulating the meaning of blame, [115] states that judgments of blameworthiness are made by assessing one's reasons for holding certain intentions and attitudes that go against the norm of their relationship ("social contract"). Blaming someone is to "respond to this impairment by modifying one's views on their relationship with the blame" [82, p. 6] I will center two key components of [115]'s conception of blame - 1) violation of the norm of a relationship (norm of replicability), 2) the intention and reasons for violating the norm (of replicability), and analyze if deviations from the two respectively are unjustified therefore non-acceptable. The overall conclusion is that assertions about blame are context-dependent and have a degree of indeterminacy, which surfaces competing epistemological norms and indicates the actualization of CR requires sociological understandings of science.

3.3.1 On the Norm of Replicability. Positioning replicability as a normative ideal appears reasonable since it has been conceived to be foundational to science. However, there exist convincing counterarguments, two salient and connected examples are [22] and one relevant concern from[77].

The concern that partially motivated Leonelli's [77] development of alternative conceptions of replicability, is treating replicability as a *normative* ideal to demarcate good and bad sciences, risking the value of non-replicability as starting points of inquiries which is epistemically rewarding. This paper's moralizing around CR does position it as a normative ideal, and sheds a negative light on non-replicable claims. Grounding their analysis in the *norm of assertation*, Dang and Bright [22] show that claims by scientists can be appropriately included in published papers, preprints, presentations, etc, even if "they are false, unjustified, and not believed to be true" [22]. They argue that these claims can lead to epistemic successes and therefore should not be held by the standards of *factive norms, justification norms,*

and belief norms. This point is in the same spirit as Leonelli's mentioned above - the normative characterization of CR can refute the fruitfulness of non-replicability.

First, Dang and Bright's target of analysis is what they call *public avowals*, the primary audience of which is other scientists *who have expertise on the subject. Public avowal* is contrasted with *public science testimony* - which targets communities within society more broadly. Assigning an ML claim to one of the two categories is context-dependent and in my opinion, is going to be challenging at this stage because we do not have a granular understanding of how diverse communities are interacting with claims in ML papers. Despite such ambiguity, I propose that social claims greatly resemble *public science testimony.* It is rare for the public to have access to a particle accelerator or a laboratory freezer to fulfill their experimental wonders in physics or biology; in contrast, it has become prevalent for ML laypersons (in terms of ML knowledge, they can still be experts in other scientific fields, for example, [67]) to self-educate and apply ML for various ends given the much more widely accessible educational and infrastructural resources. For this reason, the decision to grant the leniency that Dang and Bright grant to *public avowals* to claims in ML papers needs to be carefully made.

The implication of these concerns for implementing CR is two-fold. First, we need a more granular understanding of the types of ML claims - distinguishing public avowals and public science testimony, and applying CR as a normative ideal to public science testimony. Currently, ML papers do not put any effort into making such a distinction. Second, in scenarios where assigning one claim to one of the two categories is challenging, a more sophisticated approach should be developed instead of looking away from the harms of non-replicable "public science testimony". The aforementioned ambiguity unfortunately transmits to determining if violating the norm of CR is acceptable because adopting competing norms can be justified in their own right. More explicitly in our context, although authors can prevent harms induced by poorly justified claims and be the directly accountable party, they might still justifiably reject blame and prioritize other epistemological principles to include wrong and unjustified claims - accountable but not deserving of blame.

Zooming out from CR, the dilemma shows that one norm (eg. norm of replicability) hardly captures the full picture of the plethora of norms or perspectives (overlapping and contradicting at times) that govern the complex network of relationships between various communities. The landscape of the norms that ought to govern science and society is evolving and indeterminate, and a running debate in disciplines such as *History and Philosophy of Science, Sociology of Science,* and *Science and Technology Studies,* for example, [25, 49, 61–63, 68, 73]. In sections 4.2 and 4.3, I will elaborate on my point on "norm".

3.3.2   On the Intention of Violating the Norm of Replicability. Even if we have determined that a paper's violation of replicability is unjustified and unacceptable, discerning the intentions of violating CR is still tricky. Let's start by looking at a simplified and straightforward scenario - authors of an ML paper intentionally make non-replicable social claims to boast their study. This would constitute deception which is unacceptable. In practice, detection of such deception is difficult because there exists no analogous mechanism for evaluating claims and violation of CR is common practice (see subsection where I discussed research repertoire). It would be unfair to say that ML scientists are deceptive in producing non-replicable claims. Deception aside, another challenge goes back to 3.3.1, scientists might be adhering to a different epistemological ideal therefore the intention can arguably be acceptable.

In sum, 3.3 laid out the uncertainties involved in assigning blame to ML scientists for violating CR. However, I should re-iterate that these certainties do not refute the argument for accountability when harms are induced from non-replicable claims -accountability and blameworthiness are not the same. The responses I offered to the epistemological

tensions between CR and the epistemic benefits of non-replicability in the context of ML research, can be reconciled by 1) typification of claims (although challenging) which requires 2) deepening our understanding of how diverse audiences interact with different types of ML research claims, 3) and developing norms of communication. These goals are in no way trivial but will guide us out of the dilemma of having to choose one rewarding principle while forfeiting contradicting ones.

3.4 Address the Worry Over the Erosion of *the Value-Free Ideal*

Because CR necessitates consideration of social and ethical values, there is another potential rebuttal to my argument that I must address before I discuss the practical implications of CR which concerns the *Value-Free Ideal*, which has dominated science and engineering since the 1950s [25]. Preachers of this ideal make the normative claim that scientists should be preoccupied with upholding *epistemic values* – such as predictive accuracy, explanatory power, scope, and simplicity; and it is beyond scientists' obligation to consider non-epistemic values such as social and ethical values [79, 80]. To clarify, ML scientists have been engaging in social reflections in response to scholars' voices that social values must be incorporated to understand and navigate ML research's social impacts [19, 45, 119, 124]. ML studies aimed at identifying bias and addressing issues of equality have proliferated, and the overwhelming sentiment is the more, the better.

However, CR's manner of engaging ML scientists in social reflections is different from existing approaches. Existing measures of engaging ML scientists in social reflections treat producing good models and engaging in social reflections as different tasks. For example, social reflections appear in the form of *add-ons* - impact statement [6] (a separate document from the main study) or checklist [66, 95] that ML scientists can choose to include in their study, or even more external to a study in the form of code of ethics (e.g. ACM computational code of ethics [1]). Scientists are willing to play a part in engaging in social reflections to identify and address ML harms but deem social reflections distinct from their core responsibility of building models. As a result, we observe a sense of "dislocated[-ness]" [136, p. 1] of accountability and the inclination to procrastinate [19] in ML scientists' current views of accountability - harms are always "another person's job, always elsewhere" [136, p. 1] which can be "solved later and by others" ([142] in [19]).

In contrast, CR integrates social reflections into *evaluating the validity of knowledge claims* - necessitating social reflections in scientific reasoning and determining the competence of someone as a scientist, rather than merely the competence of ethical reasoning. The worry is that the knowledge produced under the influence of social values is less objective and reliable. To this, Douglas [25] argued that introducing non-epistemic values to evaluate the validity of claims *does not* necessarily erode scientific objectivity if the values play an *indirect role*.

Values playing an *indirect role* means that they should not contradict evidence, rather, they determine the sufficiency of evidence – introducing social values can impose extra requirements of evidence for making claims. Using the example of developing ML algorithms for predicting the chance of recidivism - introducing social considerations will require the social claim of *the algorithm doing good* to incorporate evidence beyond high predictive accuracy or accuracy parity across racial subgroups. Social considerations such as 1) the possible backfiring due to unknown interaction between the algorithm and judges informed by those algorithms and 2) how *good* is interpreted by diverse communities with different ideals and priorities, need to be accounted for to make the claim that *our method does good* replicable. In another example of introducing ML methods into medical diagnosis, social reflections on relationships between ML, patients, and doctors, will reveal that making the social claim that *an ML technology brings better experience in medical settings* requires thinking through how ML technology can render those relationships dysfunctional, such as in the case of [104] where automatic risk scoring induced *hermeneutical injustice* [38] that harms both patients

and medical professionals. Introduced in this way, social considerations do not provide "epistemic support" [53] in the role of evidence, they instead make the standard of accepting a claim more stringent.

Therefore, as social considerations can be introduced into evaluating the validity of knowledge claims without eroding scientific objectivity, rejecting CR on the grounds of eroding the Value-Free Ideal would be unwarranted.

4 CLAIM REPLICABILITY'S PRACTICAL IMPLICATION

4.1 On Circulating Reference and the Danger of *One-Click Replication*

In John Downer's essay titled *When the Chick Hits the Fan: Representativeness and Reproducibility in Technological Tests* [26], Downder emphasizes the tradeoff between representativeness and replicability in technological tests:

> T[t]he benefits of replicability ['reproducibility'] come at an epistemic cost, it is impossible to make the test simpler without making it less realistic: we are trading representativeness for replicability ['reproducibility'].

Currently, ML research method is characterized by the *Common Task Framework (CTF)* [24] - the standardization of workflow, benchmarking evaluation, infrastructure, and research presentation. "Technological tests" (model performance evaluation) are already unrealistic as they are given the neatly curated nature of published datasets for standard comparison and community-level overfitting [114]. One byproduct of these standardizations is the pursuit of what the field called *computational replicability* through enforcing *one-click replication* - sharing data and code, as well as the configuration of the environments where code should be run, to ensure the replication runner can speedily generate and compare model performance with *one click*. This inclination is a natural direction if we follow MPR. There is nothing wrong with the desire to make model performance replicate. However, making it the sole imperative will blind us from the current malaise of CTF - moving the field further on the standardization scale, reducing the utility of ML models [34, 133] by aggravating the alienation of ML from "the uncertainties and contingencies" [46] that abound in real-world applications. "

As a result of trading representativeness for replicability, ML evaluation will be reduced to what Bruno Latour calls *Circulating References* [97] - "standardized measures that can be systematized, compared, and analyzed" [26, p. 21]. Attention is diverted from ML studies' ability to deliver functionality, generate understanding, or improve social welfare, to instead generate performances quickly which serve as yardsticks for future comparison under idealistic and unrealistic conditions. Drawing [86]'s Drosophila metaphor used in critiquing the limitations inherent in computer scientists' application of computer chess to understanding human intelligence, [29] stated:

> It was as if geneticists had focused their research efforts on breeding *Drosophila* to race against each other, he remarked. "We would have some science, but mainly we would have very fast fruit flies" [86]. (Italic in original)

Benchmarking to produce "fast fruitflies" should not be the main goal of ML. CR's requiring social reflections can help bring ML scientists' attention back from the "imagined world" [130] to produce "machine learning that matters" [133].

4.2 On Interpretive Labor

This section suggests that the action of claim-making be viewed as a form of *Interpretative Labor* [43] which should preferably be taken up by ML scientists instead of being disproportionally delegated to users of their study. The courses of action I recommend in 4.3 should be viewed as tools for ML scientists to perform this labor.

Claim-making as interpretive labor refers to the efforts of interpreting numerical outputs. The distribution of such labor is dynamic. If the interpretation of model performance, for example, what it implies and how it should be used, does not (sufficiently) occur in a paper, then readers or users of the paper will have to form their own interpretation. Note that I am not implying that it is ideal to *obviate* interpretive labor from readers or users of an ML paper, which is neither plausible nor beneficial in practice. For example, part of the interpretive labor involved in introducing a piece of ML technology into classrooms is educators' subject and pedagogical knowledge which is beyond ML scientists' scope of knowledge. Instead, I am referring to the labor within ML scientists' wheelhouse.

For example, when a policymaker applies a statistical study or an ML method that lists only numerical expressions in the paper to justify a policy or belief, they need to build a convincing narrative to the public without the help of experts who produced the numbers. A case in point is the introduction of recidivism prediction algorithm into the judiciary process. Instead of investigating and qualifying the claim to reducing bias and labor through automation, the labor of interpretation fell on people who poorly performed the labor, leading to the perpetuation of racial discrimination. In a hypothetical and unrealistic case, if ML scientists exert substantial efforts into incorporating technical requirements and configurations, ethical concerns, and failure modes to substantiate claims, the labor will be drastically reduced at sites of application.

Applying the concept of *Interpretive Labor* to the analysis of CR, I argue that there exists *structural violence* between ML scientists and the broader society under the current research paradigm. In [43], Graeber refers to efforts people need to put into understanding and following bureaucratic rules such as understanding and filling out paperwork as *interpretive labor*. The lopsided distribution of interpretive labor (bureaucractic institutions' employees follow rules straightforward to them while people filling out paperwork fend for themselves in navigating the "stupidity" of those rules), according to Graeber [ibid] is founded on *structural violence*:

> forms of pervasive social inequality that are ultimately backed up by the threat of physical harm.

The powerless are subjected to physical harm when interpretive labor (such as filling out paperworks) is performed incorrectly - being denied access to basic social welfare. Therefore, "nursing homes or banks" are violent institutions because they are

> involved in the allocation of resources within a system of property rights regulated and guaranteed by governments in a system that ultimately rests on the threat of force. "Force," in turn, is just a euphemistic way to refer to violence.

In this paper, the *structural violence* in question is in a more sociology-of-science sense, it refers to the *existing communication norm* between ML scientists and diverse communities in society. *Institution*, therefore, takes on the non-material meaning - "A regulative principle or convention subservient to the needs of an organized community"[59]. The communication norm that undergirds and is reinforced by MPR condones ML scientists' actions of making non-replicable social claims; while leaving the readers or impacted communities two options - 1) taking a leap of faith without performing interpretive labor and risk harming other communities or 2) attempting to interpret the social impact accurately by going beyond their scope of expertise. This *structural violence*, as Graeber rightly pointed out, makes the powerless sympathize with those in power - *ML scientists' job is producing good performing models and it is unreasonable to demand more from them*; conversely, the sympathy is not reciprocated - ML scientists leave negative downstream effects for stakeholders at the sites of application to navigate, or even worse, to be stomached by passively impacted communities.

One rebuttal readers might raise to comparing the research tradition to institutions such as banks or hospitals is that one has more leeway to abstain from ML but not so much with other infrastructural institutions within our society. I disagree with this. First, the impacted members of civil society most of the time do not have the choice to abstain. Second, given the universalizing character [45] and the branded epistemic advantage of being "'emptied'[112] of domain affiliation" [125] that ML is chanted for, it is logical to anticipate ML becoming more dominant in scientific disciplines and civil society - which is manifest in allocated funding, the speed of growth, and ML's prominent presence in both economy and everyday life.

4.3 On Research Communication

4.3.1 Articulate Claims with Communicative Voice. Model performance claims (generalizability and robustness) and social claims (functionality, efficiency, equality, human or societal well-being, etc.) must all be made explicit, and expressed with *intentionality* in mind. Instead of selectively presenting claims in abstract, intro, and conclusion, all claims should be formally presented in one dedicated section. Claims should also be made intelligible to diverse audiences of their work and therefore expressed with a communicative voice. In discussing meaningful measures of transparency, scholars have stressed the importance of knowing the "intended recipient" [95, p. 679] of what is being made transparent [5, 20, 138] which should be legible to "who[-ever] is around the blackbox" [28]. Ignoring the diversity of audience will "work[s] to disempower, and ultimately hinders broader transparency aims" [95, p. 679].

Addressing diverse audience speaks to establishing communicative norms (mentioned in 3.2, 3.3, and 4.2) that can mitigate or eradicate structural violence (4.2) within the current dysfunctional norm. More research on how diverse audiences pick up, interpret, critique, and utilize social and computational claims which can potentially be ethnographical work at sites of application such as [99] and more importantly fields where the population's ML literacy is lower. Over time, efforts should also be put into understanding what repercussions should occur if ML scientists violate the norm. Specification of repercussions is crucial for establishing [14]'s conception of *accountability*:

> A relationship between an actor and a forum, in which the actor has an obligation to explain and justify
> his or her conduct, the forum can pose questions and pass judgment, and the actor may face consequences.

Currently, what repercussions look like is only clear within the relationship between ML scientists and reviewers. One note is that an incremental solution such as leveling up the broader audience's ML literacy, although beneficial to the goal of preventing harms, is a large social project that likely requires tremendous resources; if not implemented universally, it could leave behind those who cannot afford to participate, widening the equality gap. This point applies widely in AI ethics solutions and deserves extensive treatment.

4.3.2 Systematize Evidence and Increase Evidential Diversity. Toward a mature mechanism for evaluating claim, each claim must be accompanied by a list of supporting evidence [5, 126] and the community should have a standardized list of commonly used evidence (qualitative and quantitative). *Benchmarking comparison, cross-validation, ablation studies, A/B testing, deployment observation, and field feedback, addressing identified concerns generally and specific to targeted applications, triangulation, and qualifications such as failure modes.* Raising the standard of evidence can help cultivate epistemic humility which is much needed in ML [46].

4.3.3 Avoid Open-ended Interpretations. In writing, authors should avoid using expressions that tend to elicit open-ended interpretations – words that can bear disparate meanings across domains of applications or communities, such as *work*, *benefit*, *improvement*, *social good*, *intelligence*, etc. For example, the *benefit* that a piece of ML technology

brings can be unevenly distributed across communities [5] who will therefore form *different interpretations of benefit*. Therefore, if a word choice can lead to questions such as *what do you mean by this?* or *can you elaborate?* in real-life discussions of their work with a particular audience, they should be addressed to a reasonable degree while writing the paper.

5 CONCLUDING REMARKS

This paper defined two notions of replicability - *model performance replicability* and *claim replicability* and argued that prioritizing the latter can help bridge the responsibility gap by enacting a strong professional moral obligation to reduce harms induced by non-replicable (social) claims, and provides actionable suggestions toward developing mature mechanisms of evaluating claims in papers. Suggestions made betray that actualizing CR is more of a sociological project than merely technical in that it requires conceptions of functioning communicative norms that can effectively govern the network of relationships in ML ecosystem and counter the existing structural violence between ML scientists and broad society. Such a project essentially enforces changes in ML *research repertoire* and therefore requires *extensive* efforts which are nevertheless worthwhile because it also facilitates the democratization of ML knowledge production because ML scientists will be obligated to engage in conversations with audiences beyond reviewers of their work.

REFERENCES

[1] ACM. 2018. The code affirms an obligation of computing professionals to use their skills for the benefit of society. https://www.acm.org/code-of-ethics

[2] ACM. 2020. Artifact review and badging - current. https://www.acm.org/publications/policies/artifact-review-and-badging-current

[3] Riccardo Albertoni, Sara Colantonio, Piotr Skrzypczyński, and Jerzy Stefanowski. 2023. Reproducibility of Machine Learning: Terminology, Recommendations and Open Issues. *arXiv preprint arXiv:2302.12691* (2023).

[4] Rachel A Ankeny and Sabina Leonelli. 2016. Repertoires: A post-Kuhnian perspective on scientific change and collaborative research. *Studies in History and Philosophy of Science Part A* 60 (2016), 18–28.

[5] Carolyn Ashurst, Solon Barocas, Rosie Campbell, and Deborah Raji. 2022. Disentangling the components of ethical research in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2057–2068.

[6] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2022. 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.

[7] Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. 2022. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1194–1206.

[8] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 80–89.

[9] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 648–657.

[10] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 514–524.

[11] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 173–184.

[12] Carl Boettiger. 2015. An introduction to Docker for reproducible research. *ACM SIGOPS Operating Systems Review* 49, 1 (2015), 71–79.

[13] James Bogen. 2000. Two as good as a hundred': Poorly replicated evidence in some nineteenth-century neuroscientific research. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 32, 3 (2000).

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

[15] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 651–666.

[16] Harry Collins. 1992. *Changing order: Replication and induction in scientific practice*. University of Chicago Press.

[17] Harry M Collins. 1975. The seven sexes: A study in the sociology of a phenomenon, or the replication of experiments in physics. *Sociology* 9, 2 (1975), 205–224.

[18] A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. 2022. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 864–876.

[19] A Feder Cooper and Gili Vidan. 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.

[20] Eric Corbett and Emily Denton. 2023. Interrogating the T in FAccT. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1624–1634.

[21] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2022. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research* 23, 1 (2022), 10237–10297.

[22] Haixin Dang and Liam Kofi Bright. 2021. Scientific conclusions need not be accurate, justified, or believed by their authors. *Synthese* 199, 3 (2021), 8187–8203.

[23] Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. CrowdWorkSheets: Accounting for Individual and Collective Identities Underlying Crowdsourced Dataset Annotation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2342–2351. https://doi.org/10.1145/3531146.3534647

[24] David Donoho. 2017. 50 years of data science. *Journal of Computational and Graphical Statistics* 26, 4 (2017), 745–766.

[25] Heather Douglas. 2009. *Science, policy, and the value-free ideal*. University of Pittsburgh Pre.

[26] John Downer. 2007. When the chick hits the fan: representativeness and reproducibility in technological tests. *Social Studies of Science* 37, 1 (2007), 7–26.

[27] Anna Dreber and Magnus Johannesson. 2019. Statistical significance and the replication crisis in the social sciences. In *Oxford research encyclopedia of economics and finance*.

[28] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[29] James Evans, Tyler Reigeluth, and Adrian Johns. 2023. The Craft and Code Binary: Before, During, and After. *Osiris* 38, 1 (2023), 19–39.

[30] Uljana Feest. 2016. The experimenters' regress reconsidered: Replication, tacit knowledge, and the dynamics of knowledge generation. *Studies in History and Philosophy of Science Part A* 58 (2016), 34–45.

[31] Uljana Feest. 2019. Why replication is overrated. *Philosophy of Science* 86, 5 (2019), 895–905.

[32] Romero Felipe. 2019. The Division of Replication Labor. http://philsci-archive.pitt.edu/16472/ forthcoming, Philosophy of Science.

[33] Fiona Fidler and John Wilcox. 2021. Reproducibility of Scientific Results. In *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[34] Melissa Flagg. 2022. Reward research for being useful-not just flashy. *Nature* 610, 7930 (2022), 9–9.

[35] Samuel C Fletcher. 2022. Replication Is for Meta-Analysis. *Philosophy of Science* 89, 5 (2022), 960–969.

[36] Allan Franklin. 1998. Avoiding the experimenters' regress. *A house built on sand: Exposing postmodernist myths about science* (1998), 151–65.

[37] Henry Fraser, Rhyle Simcock, and Aaron J Snoswell. 2022. AI Opacity and Explainability in Tort Litigation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 185–196.

[38] Miranda Fricker. 2007. *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

[39] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1747–1764.

[40] Trystan S Goetze. 2021. Moral Entanglement: Taking Responsibility and Vicarious Responsibility. *The Monist* 104, 2 (2021), 210–223.

[41] Trystan S. Goetze. 2022. Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 390–400. https://doi.org/10.1145/3531146.3533106

[42] Steven N Goodman, Daniele Fanelli, and John PA Ioannidis. 2016. What does research reproducibility mean? *Science translational medicine* 8, 341 (2016), 341ps12–341ps12.

[43] David Graeber. 2012. Dead zones of the imagination: On violence, bureaucracy, and interpretive labor: The Malinowski Memorial Lecture, 2006. *HAU: journal of Ethnographic Theory* 2, 2 (2012), 105–128.

[44] Ben Green. 2021. Data science as political action: Grounding data science in a politics of justice. *Journal of Social Computing* 2, 3 (2021), 249–265.

[45] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 19–31.

[46] Gabriel Grill. 2022. Constructing certainty in machine learning: On the performativity of testing and its hold on the future. (2022).

[47] Odd Erik Gundersen. 2021. The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A* 379, 2197 (2021), 20200210.

[48] Odd Erik Gundersen and Sigbjørn Kjensmo. 2018. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[49] David H Guston. 2000. Between politics and science: Assuring the integrity and productivity of research. (2000).

[50] Stephan Guttinger. 2019. A new account of replication in the experimental life sciences. *Philosophy of Science* 86, 3 (2019), 453–471.

[51] Stephan Guttinger. 2020. The limits of replicability. *European Journal for Philosophy of Science* 10, 2 (2020), 10.

[52] Leif Hancox-Li and Capital One. 2020. Beyond Methods Reproducibility in Machine Learning. In *ML-Retrospectives, Surveys & Meta-Analyses Workshop at NeurIPS*.

[53] John Heil. 1983. Believing what one ought. *The Journal of Philosophy* 80, 11 (1983), 752–765.

[54] Witold M Hensel. 2020. Double trouble? The communication dimension of the reproducibility crisis in experimental psychology and neuroscience. *European Journal for Philosophy of Science* 10, 3 (2020), 44.

[55] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *CoRR* abs/1805.03677 (2018). arXiv:1805.03677 http://arxiv.org/abs/1805.03677

[56] Michael J Hones. 1990. Reproducibility as a methodological imperative in experimental research. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1990. Cambridge University Press, 585–599.

[57] David Hope, Avril Dewar, and Christopher Hay. 2021. Is there a replication crisis in medical education research? *Academic Medicine* 96, 7 (2021), 958–963.

[58] Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. 2021. Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 560–575. https://doi.org/10.1145/3442188.3445918

[59] N. Institution. 2023. Sense 6.a. Oxford English Dictionary. https://doi.org/10.1093/OED/4488691117

[60] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 375–385.

[61] Sheila Jasanoff. 2004. *States of knowledge: the co-production of science and the social order*. Routledge.

[62] Sheila Jasanoff. 2005. *Technologies of humility: Citizen participation in governing science*. Springer.

[63] Sheila Jasanoff and Sang-Hyun Kim. 2015. *Dreamscapes of modernity: Sociotechnical imaginaries and the fabrication of power*. University of Chicago Press.

[64] Deborah G Johnson. 2006. Computer systems: Moral entities but not moral agents. *Ethics and information technology* 8 (2006), 195–204.

[65] Margot E Kaminski and Gianclaudio Malgieri. 2020. Algorithmic impact assessments under the GDPR: producing multi-layered explanations. *International data privacy law* (2020), 19–28.

[66] Sayash Kapoor, Emily Cantrell, Kenny Peng, Thanh Hien Pham, Christopher A Bail, Odd Erik Gundersen, Jake M Hofman, Jessica Hullman, Michael A Lones, Momin M Malik, et al. 2023. Reforms: Reporting standards for machine learning based science. *arXiv preprint arXiv:2308.07832* (2023).

[67] Sayash Kapoor and Arvind Narayanan. 2022. Leakage and the reproducibility crisis in ML-based science. *arXiv preprint arXiv:2207.07048* (2022).

[68] Philip Kitcher. 2001. *Science, truth, and democracy*. Oxford University Press.

[69] Richard A Klein, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola, Štěpán Bahník, et al. 2018. Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science* 1, 4 (2018), 443–490.

[70] Konrad Kollnig, Anastasia Shuba, Max Van Kleek, Reuben Binns, and Nigel Shadbolt. 2022. Goodbye tracking? Impact of iOS app tracking transparency and privacy labels. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 508–520.

[71] Joshua A. Kroll. 2021. Outlining Traceability: A Principle for Operationalizing Accountability in Computing Systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 758–771. https://doi.org/10.1145/3442188.3445937

[72] Thomas S Kuhn. 1997. *The structure of scientific revolutions*. Vol. 962. University of Chicago press Chicago.

[73] Bruno Latour. 1983. Give me a laboratory and I will raise the world. *Science observed: Perspectives on the social study of science* (1983), 141–170.

[74] Bruno Latour and Steve Woolgar. 2013. *Laboratory life: The construction of scientific facts*. Princeton university press.

[75] Etienne P LeBel and Kurt R Peters. 2011. Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology* 15, 4 (2011), 371–379.

[76] Sabina Leonelli. 2009. On the locality of data and claims about phenomena. *Philosophy of Science* 76, 5 (2009), 737–749.

[77] Sabina Leonelli. 2018. Rethinking reproducibility as a criterion for research quality. In *Including a symposium on Mary Morgan: curiosity, imagination, and surprise*, Vol. 36. Emerald Publishing Limited, 129–146.

[78] Sabina Leonelli. 2023. Philosophy of open science. (2023).

[79] Isaac Levi. 1960. Must the scientist make value judgments? *The Journal of philosophy* 57, 11 (1960), 345–357.

[80] Isaac Levi. 1962. On the seriousness of mistakes. *Philosophy of Science* 29, 1 (1962), 47–65.

[81] Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. 2023. Trustworthy AI: From principles to practices. *Comput. Surveys* 55, 9 (2023), 1–46.

[82] Gabriel Lima, Nina Grgić-Hlača, Jin Keun Jeong, and Meeyoung Cha. 2022. The conflict between explainable and accountable decision-making algorithms. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2103–2113.

[83] Yang Liu, Michael Gordon, Juntao Wang, Michael Bishop, Yiling Chen, Thomas Pfeiffer, Charles Twardy, and Domenico Viganola. 2020. Replication markets: Results, lessons, challenges and opportunities in ai replication. *arXiv preprint arXiv:2005.04543* (2020).

[84] Bertram Ludäscher. 2016. A brief tour through provenance in scientific workflows and databases. In *Building trust in information: Perspectives on the frontiers of provenance*. Springer, 103–126.

[85] Edouard Machery. 2020. What is a replication? *Philosophy of Science* 87, 4 (2020), 545–567.

[86] John McCarthy. 1997. AI as sport.

[87] Melissa D McCradden, Shalmali Joshi, James A Anderson, and Alex John London. 2023. A normative framework for artificial intelligence as a sociotechnical system in healthcare. *Patterns* 4, 11 (2023).

[88] Timothy McPhillips, Tianhong Song, Tyler Kolisnik, Steve Aulenbach, Khalid Belhajjame, Kyle Bocinsky, Yang Cao, Fernando Chirigati, Saumen Dey, Juliana Freire, et al. 2015. YesWorkflow: a user-oriented, language-independent tool for recovering workflow information from scripts. *arXiv preprint arXiv:1502.02403* (2015).

[89] Jacob Metcalf, Emanuel Moss, Elizabeth Anne Watkins, Ranjit Singh, and Madeleine Clare Elish. 2021. Algorithmic impact assessments and accountability: The co-construction of impacts. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 735–746.

[90] Smitha Milli, Luca Belli, and Moritz Hardt. 2021. From optimizing engagement to measuring value. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 714–722.

[91] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[92] Michael Mulkay and G Nigel Gilbert. 1986. Replication and mere replication. *Philosophy of the Social Sciences* 16, 1 (1986), 21–37.

[93] Engineering National Academies of Sciences, Medicine, et al. 2019. Reproducibility and replicability in science. (2019).

[94] Helen Nissenbaum. 1996. Accountability in a computerized society. *Science and engineering ethics* 2 (1996), 25–42.

[95] Chris Norval, Kristin Cornelius, Jennifer Cobbe, and Jatinder Singh. 2022. Disclosure by Design: Designing information disclosures to support meaningful transparency and accountability. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 679–690.

[96] George Packer. 2013. Change the world. *The New Yorker* 89, 15 (2013), 44–55.

[97] Katherine Pandora. 1999. Pandora's Hope: Essays on the Reality of Science Studies. *American Scientist* 87, 6 (1999), 570–570.

[98] Orestis Papakyriakopoulos, Anna Seo Gyeong Choi, William Thong, Dora Zhao, Jerone Andrews, Rebecca Bourke, Alice Xiang, and Allison Koenecke. 2023. Augmented Datasheets for Speech Datasets and Ethical Decision-Making. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 881–904. https://doi.org/10.1145/3593013.3594049

[99] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–28.

[100] Trevor J Pinch and Wiebe E Bijker. 1984. The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science* 14, 3 (1984), 399–441.

[101] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *The Journal of Machine Learning Research* 22, 1 (2021), 7459–7478.

[102] Lindsay Poirier. 2022. Accountable Data: The Politics and Pragmatics of Disclosure Datasets. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1446–1456. https://doi.org/10.1145/3531146.3533201

[103] Karl Popper. 2005. *The logic of scientific discovery*. Routledge.

[104] Giorgia Pozzi. 2023. Automated opioid risk scores: a case for machine learning-induced epistemic injustice in healthcare. *Ethics and Information Technology* 25, 1 (2023), 3.

[105] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1776–1826. https://doi.org/10.1145/3531146.3533231

[106] Hans Radder. 1992. Experimental reproducibility and the experimenters' regress. In *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. 1992. Cambridge University Press, 63–73.

[107] Hans Radder. 1996. *In and about the world: Philosophical studies of science and technology*. suny Press.

[108] Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems* 32 (2019).

[109] Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366* (2021).

[110] Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 959–972.

[111] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. Algorithmic Impact Assessments: A Practical Framework for Public Agency. *AI Now* (2018).

[112] David Ribes. 2019. How I learned what a domain was. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–12.

[113] Samantha Robertson and Mark Díaz. 2022. Understanding and Being Understood: User Strategies for Identifying and Recovering From Mistranslations in Machine Translation-Mediated Chat. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2223–2238.

[114] Rebecca Roelofs, Vaishaal Shankar, Benjamin Recht, Sara Fridovich-Keil, Moritz Hardt, John Miller, and Ludwig Schmidt. 2019. A meta-analysis of overfitting in machine learning. *Advances in Neural Information Processing Systems* 32 (2019).

[115] Thomas M Scanlon. 2008. *Moral dimensions: Permissibility, meaning, blame.* Harvard University Press.

[116] Jutta Schickore. 2011. What does history matter to philosophy of science? The concept of replication and the methodology of experiments. *Journal of the Philosophy of History* 5, 3 (2011), 513–532.

[117] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There is not enough information": On the effects of explanations on perceptions of informational fairness and trustworthiness in automated decision-making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1616–1628.

[118] Hendrik Schuff, Alon Jacovi, Heike Adel, Yoav Goldberg, and Ngoc Thang Vu. 2022. Human interpretation of saliency-based explanation over text. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 611–636.

[119] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[120] Ali Akbar Septiandri, Marios Constantinides, Mohammad Tahaei, and Daniele Quercia. 2023. WEIRD FAccTs: How Western, Educated, Industrialized, Rich, and Democratic is FAccT?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 160–171.

[121] Ruoxi Shang, KJ Kevin Feng, and Chirag Shah. 2022. Why am I not seeing it? Understanding users' needs for counterfactual explanations in everyday recommendations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1330–1340.

[122] Steven Shapin. 2004. The way we trust now: The authority of science and the character of the scientist. (2004).

[123] Steven Shapin and Simon Schaffer. 2011. *Leviathan and the air-pump: Hobbes, Boyle, and the experimental life.* Princeton University Press.

[124] Mona Sloane and Janina Zakrzewski. 2022. German AI Start-Ups and "AI Ethics": Using A Social Practice Lens for Assessing and Implementing Socio-Technical Innovation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 935–947.

[125] Stephen C Slota, Andrew S Hoffman, David Ribes, and Geoffrey C Bowker. 2020. Prospecting (in) the data sciences. *Big Data & Society* 7, 1 (2020), 2053951720906849.

[126] Jessie J Smith, Saleema Amershi, Solon Barocas, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Real ml: Recognizing, exploring, and articulating limitations of machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 587–597.

[127] Victoria Stodden and Sheila Miguez. 2013. Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *Available at SSRN 2322276* (2013).

[128] Julia Stoyanovich, Bill Howe, and H. V. Jagadish. 2020. Responsible Data Management. *Proc. VLDB Endow.* 13, 12 (aug 2020), 3474–3488. https://doi.org/10.14778/3415478.3415570

[129] Eliza Strickland. 2019. IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. *IEEE Spectrum* 56, 4 (2019), 24–31.

[130] Honghong Tinn. 2023. Between "Magnificent Machine" and "Elusive Device" Wassily Leontief's Input-Output Analysis and Its International Applicability. *Osiris* 38, 1 (2023), 129–146.

[131] Marie VanNostrand and Gena Keebler. 2009. Pretrial risk assessment in the federal court. *Fed. Probation* 73 (2009), 3.

[132] Paul Voigt and Axel von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide* (1st ed.). Springer Publishing Company, Incorporated.

[133] Kiri Wagstaff. 2012. Machine learning that matters. *arXiv preprint arXiv:1206.4656* (2012).

[134] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.

[135] Pete Warden. 2018. The machine learning reproducibility crisis.

[136] David Gray Widder and Dawn Nafus. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big Data & Society* 10, 1 (2023), 20539517231177620.

[137] David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. 2022. Limits and possibilities for "Ethical AI" in open source: A study of deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2035–2046.

[138] Mireia Yurrita, Dave Murray-Rust, Agathe Balayn, and Alessandro Bozzon. 2022. Towards a multi-stakeholder value-based assessment framework for algorithmic systems. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 535–563.

[139] Wenbo Zhang, Hangzhi Guo, Prerna Ranganathan, Jay Patel, Sathyanath Rajasekharan, Nidhi Danayak, Manan Gupta, and Amulya Yadav. 2023. TRIM-AI: Harnessing Language Models for Providing Timely Maternal & Neonatal Care in Low-Resource Countries. (2023).

[140] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.

[141] Eli Zimmerman. 2018. Teachers Are Turning to AI Solutions for Assistance. *EdTech Magazine* (2018).

[142] Jonathan Zittrain. 2014. The virtues of procrastination. https://www.internetimpossible.org/virtues-of-procrastination/

This figure "acm-jdslogo.png" is available in "png" format from:

http://arxiv.org/ps/2404.13131v1