Disentangling the Components of Ethical Research in Machine Learning

Carolyn Ashurst
The Alan Turing Institute
UK

Solon Barocas
Microsoft Research and Cornell University
USA

Rosie Campbell
OpenAI
USA

Inioluwa Deborah Raji
Mozilla Foundation and UC Berkeley
USA

ABSTRACT

While practical applications of machine learning have been the target of considerable normative scrutiny over the past decade, there is growing concern with machine learning *research* as well. Debates are currently unfolding about how the research community should develop its research agendas, conduct its research, evaluate its research contributions, and handle the publication and dissemination of its findings, among other matters. At times, these debates have been quite heated, with different actors adopting different positions on what it means to do machine learning research ethically. In this paper, we show that some of the disagreement owes to a lack of clarity about what ethical issues are at stake in machine learning research, how these issues—in particular, the concerns with research integrity, research process harms, and downstream consequences—relate to (or, more often, differ from) one another. We then explore which mechanisms are most appropriate for dealing with the different types of ethical issues, and highlight which ethical issues require more attention than they are currently receiving. Ultimately, we hope to foster more productive discussions about the responsibilities that the community bears in addressing the ethical challenges tied to machine learning research and how to best fulfil these responsibilities.

ACM Reference Format:
Carolyn Ashurst, Solon Barocas, Rosie Campbell, and Inioluwa Deborah Raji. 2022. Disentangling the Components of Ethical Research in Machine Learning. In *FAccT '22: Conference on Fairness, Accountability, Transparency, June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3531146.3533781

1 INTRODUCTION

Over the past decade, practical applications of machine learning (henceforth referred to as ML) have been the target of considerable normative scrutiny [11, 26, 50, 51, 70]. More recently, not just algorithmic deployments but ML *research* itself has come under similar scrutiny, with ethical concerns raised about everything from the data that the research community relies on to the handling and dissemination of research findings [9, 19, 21, 32, 33].

In response, the ML research community is currently grappling with its responsibilities, both in terms of its research practices as well as its part in developing technology whose deployments may pose significant risks of harm. New governance mechanisms have been implemented to encourage or mandate certain aspects of responsible research, including ethics codes [28], ethics boards [24], ethics and impact statements [48], and ethics review within both peer review [41] and funding applications [12]. However, there is still a lack of consensus around what the responsibilities of the research community are—or should be—and much confusion over what we mean by ethical issues in research, and which ethical oversight mechanisms are most appropriate.

Our high level goal in this paper is to alleviate some of this confusion in order to enable more effective discussion of these important issues and to enable more effective targeting of governance mechanisms for the adequate ethical oversight of ML research. To achieve this, we aim to disentangle the different ethical issues in ML research and the different mechanisms available to address them. In the following three sections, we provide a description of the different ethical concerns in ML research (which we name "components" of ethical considerations). We divide the components into three categories: *research integrity* (§2), which concerns reliable scientific findings; *mitigating research process harms* (§3), which concerns addressing researcher responsibilities to a range of stakeholders due to the design and execution of a research process; and *downstream consequences* (§4), which considers the potential impacts resulting from applications of the research. The components that comprise each category can be seen in Table 1. Following the description of these components, we discuss example governance mechanisms (§5) (henceforth referred to as "mechanisms"), before addressing common points of confusion and disagreement, highlighting which ethical concerns require more attention than they are currently receiving, and noting where there continue to be gaps in the mechanisms available to address them (§6).

We note that not all components of ethical consideration apply to all ML research. For example, not all research involves human research subjects, and consideration of downstream consequences is far more pertinent to some types of ML research than others (see §6.3). The components are not mutually exclusive, nor are they likely to be exhaustive. As the community continues to grapple with what it means to do ML research responsibly, we expect this taxonomy to evolve and expand.

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533781

Category	Components	Section
Research integrity	Not engaging in misrepresentation and fraud	2.1
	Reproducibility and replicability	2.2
	Addressing assumptions and limitations	2.3
Mitigating research process harms	Ethical obligations to research subjects	3.1
	Ethical treatment of workers who contribute to the research	3.2
	Consideration of research process impacts to non-participants	3.3
	Appropriate management and use of assets, including data	3.4
Downstream consequences	Research questions and problem formulations	4.1
	Ongoing consideration of downstream impacts	4.2
	Mitigation of possible downstream impacts	4.3

Table 1: The different components of ethical research in machine learning

1.1 Why clarity is needed

In Sections 2, 3 and 4 we provide a typology that disentangles different components of ethical research considerations. In this section, we take a step back to discuss the present state of confusion and why this disentanglement—and the additional clarity it provides—is needed.

First, there is a need for more effective dialogue about the research community's responsibilities. ML is a relatively new field, whose techniques have rapidly shifted from academic curiosities to widely deployed methods that impact lives daily. Many are urgently grappling with what ethical research means in this new context, and what the responsibilities of various actors should be. However, given the many components of ethical research, with their distinct causes, aims, and possible mitigation strategies, it is vital that those involved in (and subject to) this discussion are able to distinguish between these components, in order for conversations on community interventions to be productive. A first step in deciding the roles, responsibilities, and opportunities for different actors (such as individual researchers, publication venues, research labs, funders, or regulators), is an articulation of what the different issues and aims of ethical research are, and a clearer understanding of how these different concerns require a range of oversight mechanisms.

Second, it can be detrimental if governance mechanisms are implemented without a clear articulation of which aspects they are targeting or why [6, 56]. Lack of clarity in motivation can make it hard for those subject to a particular governance mechanism to know how to comply, or to understand why the mechanism may be worthwhile. In order to develop best practice and enabling tools to support researchers in complying with a particular requirement, those who might be in a position to create such tools also need to understand the aim of the requirement, and justify its introduction. Lack of clarity can also make it challenging to evaluate the effectiveness of certain mechanisms, and weigh them against alternatives.

Finally, in some cases, mechanisms that assist with one component of ethical consideration can have little effect or even worsen other components (even seemingly closely related components). For example, methods to protect the privacy of research subjects could enable research whose future use infringes the privacy of others, resulting in a negative downstream consequence. Confusion between components can make it hard to understand these trade-offs.

1.2 Related work

There have been a number of recent attempts to tease apart the ethical issues in ML research. Last year, the Partnership on AI (PAI) released a report on publication norms for responsible AI which separated out "research integrity", "research ethics", "research culture", "downstream consequences", and "broader impacts", disambiguating and defining the terms in the process [53]. Others have made similar distinctions between research ethics, on the one hand, and downstream consequences and broader impacts, on the other. For example, Bernstein et al. [12] propose that the research community adopt so-called ethics and society review boards (ESRs) alongside institutional review boards (IRBs), with the latter narrowly concerned with the ethical obligations that researchers have to research subjects and the former more broadly concerned with the responsibilities that researchers have to society at large. The Turing Way handbook to reproducible, ethical and collaborative data science [67] centres its discussion of ethical research around 'Research Integrity' and 'Responsible Research and Innovation' (RRI). The handbook suggests that research integrity is "inward-looking", focusing on how research findings are produced, while RRI can be viewed as "outward-looking", and concerned with how the public might perceive the harms of research. Our own survey of the relevant components relies on many of these same distinctions, drawing the boundaries between high-level categories slightly differently to stress important points of difference, while also offering a greater degree of granularity within each component.

A body of research has also emerged to evaluate the mechanisms that have already been proposed or adopted by the ML community in its efforts to grapple with the ethical issues posed by research [1, 6, 47, 56]). Unlike this work, we do not aim to assess the performance of a given mechanism, but aim to provide a landscape view of the different components of research ethics before discussing a range of example mechanisms. This allows us to evaluate which broad areas of ethical research have the least developed response, and discuss the ways in which governance mechanisms can be tied more explicitly to specific ethical concerns. If the goal is to effectively monitor and assess a range of risks, it is crucial to adopt a bird's eye view and analyze the complete ecosystem of interventions required.

Research integrity			
Component		Examples	Example governance mechanisms
Not engaging in misrepresentation and fraud	2.1	• Fabricating data • Falsifying results • Serious misrepresentation of findings • Conflicts of interest	• Codes of conduct • Peer review • Processes to support/protect whistle blowers • Formal censure and/or future restrictions/bans by publication venues, conferences, professional bodies, institutional home or funding agency • Legal action • Mandatory disclosure of conflicts of interest • Publication venue policies, e.g. not publishing work by actors who stand to benefit from the findings • Pre-registration • Making funding available for research that is independent of interested or possibly conflicting parties
Reproducibility and replicability	2.2	• Reproducibility, e.g. sharing data, code etc. • Replicability, including statistically sound performance claims • Reliability and validity issues • Producing trustworthy results • Proof/evidence of claims • Disclosure of complexity/compute requirements	• Pre-registration • Reproducibility badges • Reproducilibity checklists • Data publication requirements or code Release • Required disclosure of experimental details e.g. compute spending • Peer review, e.g. to scrutinise statistical claims • Funding scientific reproduction projects, including for research that would be prohibitively expensive for non-industry researchers
Addressing assumptions and limitations	2.3	• Disclosure of assumptions, and conditions under which claims hold, including generalisability • Quantifying uncertainty • Discussion of project limitations • Discussion of dataset limitations, including bias	• Standardized reporting practices, e.g. datasheets and model cards • Publication norms and requirements, e.g. mandating a discussion of limitations, error bars on charts, etc. • Peer review, e.g. to scrutinise evidence of claims, and to verify other publication norms and requirements • Research community norms

Table 2: Disentangling research ethics - research integrity.

2 RESEARCH INTEGRITY

Our first category, *research integrity*, refers to those aspects that ensure robust scientific findings. See Table 2.

2.1 Not engaging in misrepresentation and fraud

The integrity of scientific research depends on the honesty of researchers and accurate representations of their work. In particular, researchers who fabricate data for the purposes of being able to deliver favorable results are understood to have committed a serious ethical breach. Falsifying results is an equally serious transgression, though research findings can be misleading for a range of reasons, some more ethically charged than others. Researchers may engage in so-called p-hacking, in which they purposefully seek out patterns in data that can be falsely presented as statistically significant [60]. They can also cherry-pick examples that give an overly favourable impression of the research finding (e.g., uncommonly compelling examples to illustrate the output of generative models) [42]. Sometimes researchers' misrepresentations might be due to a lack of care in the execution of the research, rather that explicit attempts to mislead, as is the case with so-called HARKing (hypothesizing after the results are known) [39], which can result from conflating exploratory and confirmatory analyses. Finally, researchers may fail to report or publish negative results, which may lead others to develop overly optimistic ideas about scientific progress on a particular topic [62].

Researchers seeking to enhance their reputation or advance professionally may have their own reasons to engage in these activities, but conflicts of interest may play a part, too. For example, the research funding entity may want to see a particular result and thus

place pressure on researchers to deliver it. Or the funding entity may reserve the right to edit or censor research papers [23]. Many therefore view the failure to disclose any potential conflicts of interest that could influence the research findings or its presentation as a serious ethical transgression.

2.2 Reproducibility and replicability

The reproducibility ethos asserts that consistent findings from independent investigators are the primary means by which scientific evidence accumulates for or against a hypothesis [40]. Claerbout and Karrenbach [18] classify research as *reproducible* when "[a]uthors provide all the necessary data and the computer codes to run the analysis again, recreating the results", and *replicable* when "[a] new study arrives at the same scientific findings as a previous study, collecting new data (with the same or different methods) and completes new analysis" [18, 67]. Stodden [63] provides a summary of what is needed to achieve these objectives: (i) Computational reproducibility: when details are provided about code, software, hardware and implementation details; (ii) Empirical reproducibility: when details are provided about non-computational empirical scientific experiments and observations. In practise this is enabled by making the data and data collection process freely available; and (iii) Statistical reproducibility: when detailed information is provided about the choice of statistical tests, model parameters, threshold values, etc. This includes pre-registration of study design to prevent p-value hacking and other manipulations [63, 67].

A particularly acute threat to ML reproducibility is the general lack of the reporting of computational complexity or compute requirements. Experiments involving significant compute may be prohibitively expensive for all but the most well-funded researchers. Likewise, certain research topics may depend on specialist hardware that is simply not available to other researchers. Recent efforts to compel researchers to include measures of compute in their publications are designed to not only give some impression of the environmental impact of their research (as we'll discuss in 3.3), but to also give others an indication of whether they would even be able to attempt to reproduce the study [59].

Closely related is the adequate provision of proofs and the evidence of claims. In the same way that code release can help others reproduce project outcomes and verify experimental claims, rigorous proofs (for example) allow others to reproduce the logic behind a result, and verify theoretical claims. To address this concern, researchers must ensure the reliability of empirical evidence, high quality experimental design as well as the soundness and validity of proofs. The adequate citing of prior research is also a meaningful consideration.

2.3 Addressing assumptions and limitations

Some features of the academic environment can encourage researchers to overstate the utility of their research and to gloss over its drawbacks. For some, ethical research means disclosing assumptions and limitations so that readers can appreciate the true significance of the work. For example, the NeurIPS 2021 Checklist specifies that "[c]laims in the paper should match theoretical and experimental results in terms of how much the results can be expected to generalize" and that "[t]he paper's contributions should be clearly stated in the abstract and introduction, along with any important assumptions and limitations" [49]. Assumptions can pertain to the data (e.g. properties of its distribution or the relationship between the data and the situation that the dataset purports to reflect), to the method (e.g. that metrics used for optimization or evaluation reflect the true goals implied by the claims), and to the context of use (e.g. regarding how a system will be used in practice or how use of the system could impact future training data).

These assumptions can affect the reported generalisability of the research findings—that is, whether the results and methods apply to situations that differ from those used within the research process. This may include the ability to generalise to rare data points, to apply to data collected under different conditions (such as at a different place or time), or for the system to perform well if used by others or for different purposes. These concerns are sometimes expressed as concerns about external validity. Beyond threats to generalisability, limitations might also manifest in the robustness of the results or method (including adversarial robustness); in the method's vulnerabilities to particular types of error, failure modes, or biases; in the data, memory or compute required for the research or the application of its findings (thereby impacting reproducibility and real-world use); in the interpretability of models as well as findings; and even within the privacy and security risks [5].

An important class of limitations are those relating to datasets. Historically, this has meant adequate treatment and reporting of noise, missing or inferred values or data points, and different types of (statistical) biases, such as selection bias. Many commonly used datasets have been found to inadequately represent marginalised or vulnerable groups and to encode normatively troubling inequalities. The use of such datasets in research can skew research directions, result in distorted results and conclusions [61], and result in a wide range of discriminatory harms if models do get deployed, especially in high stakes settings. In recognition of this, researchers are increasingly being encouraged to investigate and disclose societal biases within their datasets, and to use more representative datasets where possible [10].

3 MITIGATING RESEARCH PROCESS HARMS

Our second category is *mitigating research process harms*. Here we consider harms caused by the research process to various stakeholders, beyond those that may be caused by unreliable findings.

3.1 Ethical obligations to research subjects

It has long been recognised that those conducting research on human subjects have a duty of care to those subjects [22, 38, 43]. For example, in 1981 the Common Rule established requirements for biomedical and behavioural research with human subjects in the US, based on the Belmont Principles. The Common Rule includes conditions for IRB approval (which is required for all US federally-funded research), which includes criteria relating to informed consent, risks to subjects, and equitable selection of subjects [31]. While experiments involving interacting with human participants are relatively rare in ML, using data about people, including identifiable and private information, is common. Ethical obligations to such research subjects often includes not only obtaining informed and voluntary consent, but minimising risks to the participants, including risks

Mitigating research process harms			
Component		Examples	Example governance mechanisms
Ethical obligations to research subjects	3.1	• Treatment of individuals participating in the research study • Treatment of those whose data is used within the research process	• Institutional review boards (IRBs) • Compulsory training • Complaints processes for research subjects • Processes to support whistle blowers • Formal censure by institutional home • Restrictions on government funding to institutions that permitted research violating Belmont Principles • Reputational incentives • Legal action
Ethical treatment of workers who contribute to the research	3.2	• Working conditions, including consideration of harmful data content • Labour conditions and compensation, particularly for data workers	• Ethics review boards • Labour standards set by funders, publication venues, or research institutions • Labour laws • Worker organising, including through unions • Worker feedback • Disclosure of payment and working condition details in publication
Consideration of research process impacts to non-participants	3.3	• Wider impacts of the research process, e.g. from testing products outside of the lab • Impacts from disseminating research outputs, e.g. the creation/use of datasets containing offensive content • Impacts to the targets of research studies • Environmental impacts of the research process	• Institutional review boards • Ethics review boards • Ethical standards for field experiments • Safety standards • Complaints processes • Whistle blower processes and protection • Legal action • Publication requirements, e.g. mandatory disclosure of environmental impact • Tools such as emissions calculators
Appropriate management and use of assets, including data	3.4	• Respect for data protection and other privacy laws • Respect for intellectual property and contract law • Appropriately citing data sources, code contributors, etc. • Documenting data provenance and curation	• Legal requirements and action (e.g. around use of personal data) • Asset documentation standards/guides (e.g data sheets) • Publication requirements and peer review • Codes of conduct • Formal censure or future restrictions/bans by publication venues, conferences, professional bodies, institutional home or funding agencies • Reputational costs

Table 3: Disentangling research ethics - mitigating research process harms.

to the privacy and security of their data. As Metcalf and Crawford [45] show, however, applying established principles from research ethics to computational research is not without controversy, in large part because of disagreement about whether research involving data collected as a matter of course even constitutes research with human subjects.

3.2 Ethical treatment of workers who contribute to the research

Recently, there has been growing concern for the treatment of the workers who contribute to ML research. This includes consideration for the working conditions and compensation of the scholars undertaking the research – especially the historically marginalised and those vulnerable to exploitation – but it also includes those hired to facilitate the research process in other ways. For instance, crowd workers, through platforms such as Mechanical Turk [3], have come to play a particularly important role in the data labelling and cleaning processes for ML research. Scholars have shown that these workers are often subject to poor working conditions and compensation [29, 36] and may be exposed to distressing content, such as racist and misogynistic text and labels as well as pornographic or violent images [14, 20, 57].

3.3 Consideration of research process impacts to non-participants

In many cases, the research process can also pose threats to people that are not even considered part of the research project. When research is conducted in the field rather than the lab, it may be challenging to ensure that only those who have voluntarily consented to participate in the research project are the ones that face any heightened risk of being harmed. Although this consideration for field research is much more common in the social sciences[44], ML can also involve research very much along these lines, as is the case with research that tests autonomous vehicles by deploying them on real roads, thereby placing everyone in proximity to the car at risk [13]. Similar expectations might apply also to research involving malware or weapons, whose unintentional release or misfiring might pose threats to computer security and physical safety, much in the same way as bio-hazards for research in the physical sciences. Finally, the environmental impact of the computations that go into ML research can be non-trivial and this impact definitely imposes on people who are not formally part of that process. An OpenAI report estimated that, since 2012, the compute needed to train the largest AI models doubles every 3.4 months[4]. It has been claimed that training a BERT language model has similar carbon emissions to a trans-American flight [65]. Calls for "Green AI" advocate for ML research that takes environmental costs into account. While Schwartz et al. [59] are also concerned with the environmental impact of downstream application of these methods or practical application of these models, the environmental impacts of the research process alone are sufficiently significant to provoke ethical concern.

3.4 Appropriate management and use of assets, including data

Whenever research uses existing assets, such as datasets, code, models, researchers must ensure (and often demonstrate) their right to use these assets. This can include the disclosure of licensing agreements, terms of service, or copyright agreements for datasets and tools used, and appropriate citing of prior research, datasets, and code. When research uses datasets involving data about individuals, researchers must ensure compliance with appropriate privacy or data protection laws. For example in the UK, data containing personal information (roughly, data from which individuals can be directly or indirectly identified) [35] or special category data (such as data concerning health, sexual orientation or racial origin) [34] is subject to UK GDPR [66, 68]. Despite the regulations surrounding the use of data, there remain many controversies around the privacy of those whose information appears in datasets [30]. Scholars have argued that many current safeguards (such as licenses, dataset retraction, and current dataset management practices) are inadequate, and that programme committees and dataset creators and users can do more [54].

4 DOWNSTREAM CONSEQUENCES

While *mitigating research process harms* concerns impacts resulting from the research itself, *downstream consequences* is concerned with the (potential) impacts from future use of the research. Traditionally, research ethics has concentrated on the risks posed to research participants; in contrast, *downstream consequences* asks researchers to take a broader perspective, and ask how research outputs (such as data, methods, models) could (or will) be used, and what their societal impacts could be. Ideally researchers would take potential societal impacts into account throughout the research process. Indeed, this is the aim of so-called responsible research and innovation (RRI) [52], a term that has emerged largely in Europe to describe the need for research to takes societal impacts into account throughout the research process. For example, in the UK, the Engineering and Physical Sciences Research Council (EPSRC)—a government funding body—has developed an 'AREA' framework that states that researchers should Anticipate, Reflect, Engage, and Act [25] at each step of the research process, from choosing research directions and questions (§4.1), to taking a participatory approach to review decisions made throughout the process (§4.2), to implementing improvements and mitigations in light of potential risks and harms (§4.3). As the work on RRI also makes clear, downstream consequences can include both the impacts that result from applications that work as intended as well as those that result from errors, accidents, or misuse and abuse.

4.1 Research questions and problem formulations

Researchers frequently take societal impacts into account when choosing their research questions because they often want to generate knowledge that will have positive practical impact. In this regard, downstream consequences are a common consideration in setting a research agenda and choosing problems to work on. Given the risks posed by certain research directions, though, there is growing pressure on researchers to also consider the potentially

Downstream consequences			
Component		Examples	Example governance mechanisms
Research questions and problem formulations	4.1	• Choosing research questions and directions with impacts in mind • Research that should not be done	• Ethics and society review (ESR) boards • Conference tracks, calls for papers and competitions to incentivise certain research questions • Peer review, to reject papers with a significant potential for harm • Field building: promoting/funding fields and individual projects that aim to improve societal outcomes • Participatory approaches
Ongoing consideration of downstream impacts	4.2	• Evaluation of methods based on societal impacts • Reflection and disclosure of possible downstream impacts • Disclosure of appropriate use	• Ethics and broader impact statements • Model cards, to disclose appropriate use • The development of techniques to identify/measure downstream impacts (e.g. fairness or privacy metrics) • Participatory approaches
Mitigation of possible downstream impacts	4.3	• Implementation of mitigations for individual research • Developing new methods to identify, measure, and mitigate possible downstream harms	• Ethics review boards • Field building: promoting/funding fields and individual projects that aim to develop new methods to identify, measure and mitigate downstream harms • Developing policy/laws to prohibit/control harmful downstream applications

Table 4: Disentangling research ethics - downstream consequences.

negative practical impacts of their research—and to reconsider or abandon research questions that are more likely to harm society than benefit it, or that will disproportionately harm some groups more than others [8]. Even when certain research questions do not seem to pose a threat of harm, researchers may still be expected to take into account whose interests and welfare would be advanced in answering these research questions; the benefits of research might be quite unevenly distributed. For this reason, there are serious concerns about who is able to participate in the process of steering a field's research agenda, especially when research might be supported by the state. These concerns are increasingly common in the ML research community, where a number of controversial studies have been called into question not only for their lack of scientific merit, but for their lack of moral legitimacy. For example, a large set of scholars have called for the retraction of research that sought to predict criminality from people's appearance, arguing that the research was premised on discredited ideas that have been used to perpetuate injustice in the past and that would seriously risk perpetuating injustice in the future [19]. Likewise, the out-sized role that technology companies play in ML research has sparked recent debate about its influence over the research questions that get asked [69].

4.2 Ongoing consideration of downstream impacts

Beyond the choice of research questions, researchers might be expected to reflect on the potential downstream consequences of their specific research findings, both as they are obtaining preliminary results and once they arrive at their final conclusions. For example, various choices made in developing the research design for a study may not only threaten the validity of a study's findings, but lead to limitations that—unless properly understood by those relying on the research—give rise to avoidable negative outcomes in practice. Researchers might be viewed as derelict in their ethical duties if they fail to anticipate how their work might be misinterpreted in ways that might lead to serious downstream harms. The same might be true of failures to adequately reflect on how vulnerable certain methods might be to errors or mistakes, especially if these could again lead to serious harms when the methods are adopted in

practice. These all suggest the need for researchers to take responsibility for ensuring adequate disclosure to head off inappropriate applications of their research.

It is not uncommon for research findings to have many possible uses in practice, including those that were not the use that was initially imagined by those who undertook the research. This is similar to the challenge of 'dual use' technology in which a technological innovation developed with beneficial goals in mind can also be put to unrelated and more troubling use. These concerns are especially salient in the case of ML research as many advancements in the field are general methods that can be applied to a broad set of real-world tasks. Indeed, sometimes a welcome research breakthrough could be abused for malicious purposes, such as criminal activity [17]. These possibilities may again place pressure on researchers to consider how they go about their research and the publication of their findings. For example, OpenAI, in recognition of the fact that its large language model could be used to automate the generation of "deceptive, biased, or abusive language at scale" initially gated the release of the model, limiting access to people that were expected to abstain from such harmful uses [58].

These same considerations might carry over to the way that researchers evaluate their own work. Researchers often have a choice in how to evaluate the techniques they develop, e.g. which metrics and other criteria to use. While performance metrics such as accuracy have been traditionally favoured, consideration of societal impacts can result in other evaluation methods being preferred. For example, fairness considerations may encourage disaggregating results by subgroup, as well as using fairness metrics to diagnose methods. The use of such methods at the end of research can improve transparency about potential societal impacts. The use of such methods throughout the research process can also influence which methods get taken forward, resulting in research outputs which better mitigate certain harms.

4.3 Mitigation of possible downstream impacts

In addition to disclosing potential risks, researchers may also be expected to reflect on, develop, and share possible mitigation strategies to deal with these risks. As those closest to the research, researchers may sometimes be well positioned to both identify and implement mitigations. These mitigations could be technical interventions, organisational procedures, or governance strategies. In computer security, for example, there are norms around the ethical disclosure of vulnerabilities that might put particular systems and people at risk. Researchers are frequently expected to alert those at risk, provide guidance on how to address it (if the researchers have been able to ascertain how to do so), and leave time for those changes to be made prior to publication. Researchers are thus understood to bear an ethical burden to help mitigate the potential risks that would be created in publishing their research. Similar expectations might be extended to ML researchers. Of course, the mitigation of downstream harms is not the sole responsibility of researchers. Legislation, regulation, and other controls are needed to mitigate potential harms, as we will discuss in §5 and §6.

5 OVERVIEW OF EXISTING GOVERNANCE MECHANISMS

In this section, we describe some *existing* example governance mechanisms that target different components of ethical research, organised by the stakeholder in a position to intervene. We further group these existing opportunities for intervention into the categories of agenda setting, training, rules and requirements, review, incentive mechanisms, complaint mechanisms, and sanctions. We do this to map out the range of mechanisms currently deployed today, and map them to specific ethical considerations and actors. See also Tables 2, 3 and 4 for examples organised by component.

We discuss several actors, but focus on research institutions and publication venues as primary actors. Research institutions can involve academic settings or industry labs, as well as other organisations involving a supported research team. Publication venues can act as final gate-keepers to set the standard for research community participation and check that these accepted standards are met. In the ML field, these venues are often conferences, rather than journals.

5.1 Mechanisms targeting research integrity

Research integrity aims to promote reliable science. Responsibility for ensuring that best practice is followed predominantly falls to the scientific community itself, as represented by following actors, who may deploy the mentioned mechanisms as part of a broader approach to ethical oversight and overall governance.

Mechanisms executed by research institutions.

- *Agenda setting*: These institutions can direct internal funding to set priorities for the involved researchers, including encouraging and investing in research on reproducibility.
- *Training*: Many research institutions involve an educational opportunity - to train employees or students to produce high quality research and adhere to expectations of research integrity.
- *Complaints mechanisms*: Internal processes to support whistle blowers, in the case of fraudulent behaviour, and allow for a way to identify internal integrity issues.
- *Sanctions*: Ethical expectations can also be enforced through the threat of censure, restrictions, and banning in the case of identified or reported fraudulent behaviour.

Mechanisms executed by publication venues.

- *Agenda setting:* Conference tracks, workshops, and calls for papers can be used to incentivise work on identifying limitations, or reproducing past studies [55].
- *Rules:* Codes of conduct can make expectations explicit. Mandatory publication requirements, such as the release of code, including a discussion of limitations, or declaring conflicts of interest can also increase overall research integrity. Mandating the use of standardize reporting practices such as datasheets [27], data nutrition labels [64] and modelcards [46], can also ensure disclosure of limitations etc.
- *Peer Review:* Peer review is used as the main oversight mechanism, both to check mandatory requirements, as well as to

verify that established norms and standards have been met, such as adequate presentation of assumptions. This is also a key approach to assessing that adequate evidence of claims and findings has been provided, such as rigorous proofs, adequate citing, and rigorous use of statistics when presenting results.

- *Checklists & Artifact Badges:* Mechanisms to improve transparency can also incentivise aspects of research integrity, such as artifact badging and checklists. Artifact badges are a publicly visible flag to show whether accepted papers have satisfied optional reproducibility requirements. For example, ACM recommends the use of badges under three categories, following an artifact review process: Artifacts Evaluated, Artifacts Available, and Results Validated [2]. Checklists ask authors to declare answers to a range of questions. For example, the NeurIPS conference adopted a checklist in 2021, including questions such as "Did you include the code, data, and instructions needed to reproduce the main experimental results...?" [49]. The questions served as prompts for researchers to think about these ethical considerations, and provide transparency to reviewers, and, if published, to readers.
- *Sanctions:* Publication venue sanctions include censure, restrictions and banning as punishments for fraudulent behaviour.

5.2 Mechanisms targeting research process harms

Research process harms aim to protect individuals, rather than just promote robust science. As a result, the involved governance mechanisms fall on a range of actors, including those outside the research community.

Mechanisms executed by legislators and regulators. To protect those both within and outside the research process, certain legal safeguards should be set to protect workers, establish safety standards for field research, protect intellectual property and personal data, as well as provide avenues for legal action in the case of harm.

Mechanisms executed by research institutions and funders.

- *Training:* Including compulsory training for researchers regarding human subjects research, as well as broader ethical concerns from research can improve awareness and adherence to these expectations.
- *Review:* As discussed in §3.1, institutional review boards (IRBs) have been developed to protect the wellbeing of research and data subjects for certain types of research. As well as making IRB approval a condition of funding, restrictions on funding (particularly government funding) can be placed on institutions found to violate research ethics principles. Complementing IRBs, ethics boards can be used to ensure that protections are in place for a wider pool of stakeholders. For instance, as part of Stanford University's *Ethics and Society Review* (ESR), funding applicants write a statement about potential societal impacts and research process harms, which is reviewed by a panel[12]. Similarly, the technology company Microsoft implemented the "Research Ethics Submission System" (RESS) since early 2013, which operates as an internal review board [16].
- *Complaints processes:* Complaints processes for research subjects, and processes to support whistle blowers are key interventions for raising awareness about unethical research practice.
- *Sanctions:* Formal censure and restrictions also apply in this category if research is found to violate research ethics principles.

Mechanisms executed by publication venues. As with research integrity, publication venues can also provide additional influence and checks to address research process harms.

- *Agenda setting:* Publication venues can influence community inquiry into ethics-related topics through the introduction of new tracks, calls, and competitions.
- *Rules:* These include mandatory disclosure of environmental impacts or worker compensation; or rules that set wage minimums for workers.
- *Review:* Ethics review boards can be used to ensure that societal risks have been adequately mitigated.
- *Sanctions:* Formal censure and future restrictions or bans can also apply in the case of violations.

5.3 Mechanisms targeting downstream consequences

Downstream consequences aims to mitigate harms from research applications. Since future use likely involves other actors, as well as much uncertainty, responsibilities are less clear cut, and there is little consensus at present. Once research reaches deployment, legal mechanisms are complemented by self-governance from the appropriate industries and organisations to address ethical issues, with advocacy groups, civil society, and scholars playing a crucial role in understanding and raising awareness of impacts. Whilst still in the research stage, the main actors will again be research institutions, funders, and publishers.

Mechanisms executed by research institutions and funders

- *Agenda setting:* Research labs and funders can play a role in choosing beneficial research projects. Establishing, promoting or funding subfields and individual projects that aim to improve societal impacts could do a lot to push for beneficial outcomes for the field.
- *Rules:* Policies regarding research that they will not fund or participate in. Mandatory requirements can also include impact statements as part of funding applications, to encourage researchers to reflect on future consequences.
- *Review:* As discussed in §5.2, ethics boards can be used to discuss societal impacts, including downstream consequences. For example, the Stanford ESR process asks applicants to consider "who will be impacted by the technology once it leaves the lab".

Mechanisms executed by publication venues

- *Agenda setting:* Conference tracks, calls and competitions can incentivise certain research questions or framings, including mitigation methods and research for beneficial applications. Conferences can also play a role in field building.
- *Rules:* Guidance on thinking through broader impacts can be implemented through ethics codes. For example, the ACM code of ethics recommends to "give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks" [28]) and mandatory ethics statements such as the NeurIPS 2020 broader impact requirement [48], and the NeurIPS 2021 checklist questions on potential negative societal impacts [49], also prompt this form of reflection.
- *Review:* Ethics review boards can be used to deliberate on whether certain research projects should be published, if it may cause unacceptable risks.

6 MECHANISMS IN NEED OF FURTHER DEVELOPMENT

Broadly speaking, the three categories represent how thinking about scientific conduct has developed over time. Research integrity aims to support the rigour of the scientific method, with roots going back centuries [15]. Codes for the rights of human research subjects can be traced back to the 1948 Nuremberg Code and 1979 Belmont Principles [22, 38]. In recent years, consideration of the impacts of research on other stakeholders (e.g. data subjects) has received increasing attention. For many fields, including ML, the extent to which researchers should consider the societal impacts of future use of their work is a more recent question. As such, there is a more established set of best practices for research integrity, and to an extent mitigating research process harms, compared to downstream consequences.

6.1 Research integrity

In general, mechanisms for research integrity are well developed, at least in theory. For example, peer review has been developed over several centuries to ensure adequate standards. That said, there are weaknesses. The replicability crisis in various fields is well known [7]. Within ML specifically, the use of large datasets means that unintended p-hacking and HARKing is a particular risk. Also, ML does not have a strong tradition of hypothesis testing or randomized control trials (RCTs). Instead, the current major paradigm is the use of benchmarks, which can lead to incremental improvements that are not replicable, since they may overfit to particular datasets. Cherry picking of datasets and illustrative examples is also a concern. Large compute requirements can also inhibit the reproducibility of large models. Disclosure of conflicts of interest is standard practice in many areas. However, an issue of particular importance to ML is the influence of private firms, who control significant resources for research. While current conflict of interest norms may be sufficient for some fields, whether they are sufficient for ML remains an open question. Overstating claims and understating limitations continues to be a concern, though this is not unique to ML (though some fields, such as medical research, have more established and more rigorous standards for verifying clinical claims). There are aspects of particular relevance to ML, such as understanding the limitations of large datasets, and questions around generalisability. We therefore suggest that innovative thinking is required by the community concerning the prevailing paradigm of benchmarks, as well as careful reflection by the researcher community, civil society and government over the role of private firms in influencing research.

6.2 Mitigating research process harms

As discussed, duty of care to human subjects has been a concern for decades, and thus IRBs and similar mechanisms are well established. Somewhat less well developed is extending this to data subjects. Though IRBs have been extended to data subjects, not all research involving data generated by or about people is subject to IRB approval. Even less well developed is consideration for the welfare of data workers and non-participants. Legal requirements are well established for IP, contract law, and the use of personal data. However, as the nature, availability, and uses of data continue to evolve, data regulation requires ongoing development. Less well developed are mechanisms targeting problematic datasets in research (e.g. datasets containing bias or offensive content), although some publishers are starting to develop policies for their use. While analysis and disclosure of complexity and compute are standard practice in some subfields of computer science, they are not typically demanded of ML papers, perhaps because of the recent emphasis on empirical rather than theoretical results. Whether researchers should disclose complexity, compute, and environmental costs remains an open question. We recommend that research organisations and publishers review their policies regarding protections for a wider class of stakeholders (e.g. data workers) and around disclosure of compute, and that scholars continue to develop best practice regarding problematic datasets, with publication venues adapting policies accordingly.

6.3 Downstream consequences

While some mechanisms targeting downstream impacts have been developed (e.g. ethics statements), these have not been widely adopted, and there is little consensus over whether or how they should operate [37, 56] and how effective they would be. The question of when these are appropriate, and how to make them proportionate according to different types of research, remains an open question. The level of responsibility may depend on (i) where research lies on the theory-application spectrum, and its proximity to deploymentand (ii) where research lies on the incremental-innovative spectrum (research that may enable very new real-world applications has a higher burden than research that incrementally improves existing methods). While consideration of downstream consequences is essential for certain types of research, we need to avoid placing a disproportionate burden on researchers to compensate for a lack of laws and regulations directly governing deployed systems. We suggest that the community continue to reflect on best practice regarding downstream consequences, in particular regarding which mechanisms are effective and proportionate for different types of research.

7 CONCLUSION

In this paper, we propose a typology to disentangle the different components of ethical research in ML. We discuss some of the governance mechanisms developed to address these issues and identify areas for which mechanisms are less well developed. Our hope is that this work will help bring much-needed clarity to discussions about ethical research in ML and enable more effective development and targeting of governance mechanisms.

ACKNOWLEDGMENTS

We thank the participants of the "Navigating the Broader Impacts of AI Research" workshop at the 2020 Neural Information Processing Systems conference for the feedback and engagement that heavily informed this article.

REFERENCES

[1] Grace Abuhamad and Claudel Rheault. 2020. Like a Researcher Stating Broader Impact For the Very First Time. *Navigating the Broader Impacts of AI Research Workshop at the 34th Conference on Neural Information Processing Systems* (2020).

[2] ACM. 2021. Artifact Review and Badging – Version 2.0. Retrieved January 21, 2022 from https://www.acm.org/publications/policies/artifact-review-badging

[3] Amazon. [n.d.]. *Amazon Mechanical Turk*. Retrieved January 21, 2022 from https://www.mturk.com/

[4] Dario Amodei and Danny Hernandez. 2018. AI and Compute. Retrieved January 21, 2022 from https://openai.com/blog/ai-and-compute/

[5] Carolyn Ashurst, Markus Anderljung, Carina Prunkl, Jan Leike, Yarin Gal, Toby Shevlane, and Allan Dafoe. 2020. A Guide to Writing the NeurIPS Impact Statement. Retrieved January 21, 2022 from https://medium.com/@GovAI/a-guide-to-writing-the-neurips-impact-statement-4293b723f832

[6] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2021. AI Ethics Statements - Analysis and Lessons Learnt from NeurIPS Broader Impact Statements. *arXiv preprint arXiv:2111.01705* (2021).

[7] Monya Baker. 2016. 1,500 Scientists Lift the Lid on Reproducibility. *Nature News* 533, 7604 (2016), 452.

[8] Solon Barocas, Asia J Biega, Benjamin Fish, Jedrzej Niklas, and Luke Stark. 2020. When Not to Design, Build, or Deploy, An Interactive Discussion at the ACM FAT* 2020 Conference. https://when-not-to-build.github.io/

[9] Emily Bender. 2019. Is There Research That Shouldn't Be Done? Is There Research That Shouldn't Be Encouraged? Retrieved January 21, 2022 from https://medium.com/@emilymenonbender/is-there-research-that-shouldnt-be-done-is-there-research-that-shouldn-t-be-encouraged-b1bf7d321bb6

[10] Samy Bengio, Kate Crawford, Jeanne Fromer, Iason Gabriel, Amanda Levendowski, Deborah Raji, and Ranzato Marc'Aurelio. 2021. Ethics Guidelines. https://neurips.cc/public/EthicsGuidelines

[11] Ruha Benjamin. 2019. Race After Technology: Abolitionist Tools for the New Jim Code. *Social Forces* (2019).

[12] Michael S Bernstein, Margaret Levi, David Magnus, Betsy A Rajala, Debra Satz, and Charla Waeiss. 2021. Ethics and Society Review: Ethics Reflection as a Precondition to Research Funding. *Proceedings of the National Academy of Sciences* 118, 52 (2021).

[13] Sarah Bird, Solon Barocas, Kate Crawford, Fernando Diaz, and Hanna Wallach. 2016. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning*.

[14] Abeba Birhane and Vinay Uday Prabhu. 2021. Large Image Datasets: A Pyrrhic Win for Computer Vision?. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1536–1546.

[15] Dorothy Bishop and Eoin Gill. 2020. Robert Boyle on the Importance of Reporting and Replicating Experiments. *Journal of the Royal Society of Medicine* 113, 2 (2020), 79–83.

[16] Anne Bowser and Janice Y Tsai. 2015. Supporting ethical web research: A new research ethics review. In *Proceedings of the 24th international conference on world wide web*. 151–161.

[17] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The Malicious Use of Artificial Iintelligence: Forecasting, Prevention, and Mitigation. *arXiv preprint arXiv:1802.07228* (2018).

[18] Jon F Claerbout and Martin Karrenbach. 1992. Electronic Documents Give Reproducible Research a New Meaning. In *SEG technical program expanded abstracts 1992*. Society of Exploration Geophysicists, 601–604.

[19] Coalition for Critical Technology. 2020. Abolish the #TechToPrisonPipeline. Retrieved January 21, 2022 from https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16

[20] Kate Crawford and Trevor Paglen. 2019. Excavating AI: The Politics of Training Sets for Machine Learning. Retrieved January 21, 2022 from https://excavating.ai

[21] Emily Denton and Timnit Gebru. 2020. Tutorial on Fairness Accountability Transparency and Ethics in Computer Vision at CVPR 2020. https://sites.google.com/view/fatecv-tutorial/

[22] DHEW. 1978. The Belmont Report. Retrieved January 21, 2022 from https://videocast.nih.gov/pdf/ohrp_belmont_report.pdf

[23] Christoph Ebell, Ricardo Baeza-Yates, Richard Benjamins, Hengjin Cai, Mark Coeckelbergh, Tania Duarte, Merve Hickok, Aurelie Jacquet, Angela Kim, Joris Krijger, et al. 2021. Towards intellectual freedom in an AI Ethics Global Community. *AI and Ethics* 1, 2 (2021), 131–138.

[24] EMNLP. 2020. Call For Papers. Retrieved January 21, 2022 from https://2020.emnlp.org/call-for-papers

[25] EPSRC. 2020. Anticipate, Reflect, Engage and Act (AREA). Retrieved January 21, 2022 from https://epsrc.ukri.org/research/framework/area/

[26] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92.

[28] Don Gotterbarn, Amy Bruckman, Catherine Flick, Keith Miller, and Marty J Wolf. 2017. ACM Code of Ethics: A Guide for Positive Action. *Commun. ACM* 61, 1 (2017), 121–128.

[29] Mary L Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley From Building a New Global Underclass*. Eamon Dolan Books.

[30] Jules. Harvey, Adam. LaPlace. 2021. *Exposing.ai*. Retrieved January 21, 2022 from https://exposing.ai

[31] HHS. 2018. Code of Federal Regulations - Title 45 Public Welfare CFR 46 - 2018 Common Rule. https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/revised-common-rule-regulatory-text/index.html

[32] Kashmir Hill. 2014. Facebook Manipulated 689,003 Users' Emotions For Science. Retrieved January 21, 2022 from https://www.forbes.com/sites/kashmirhill/2014/06/28/facebook-manipulated-689003-users-emotions-for-science/?sh=5fd1818b197c

[33] Jeremy Hsu. 2019. Microsoft's AI Research Draws Controversy Over Possible Disinformation Use. Retrieved January 21, 2022 from https://spectrum.ieee.org/microsofts-ai-research-draws-controversy-over-possible-disinformation-use

[34] ICO. 2018. Special Category Data. Retrieved January 21, 2022 from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/lawful-basis-for-processing/special-category-data

[35] ICO. 2018. What is personal data? Retrieved January 21, 2022 from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/key-definitions/what-is-personal-data

[36] Lilly C Irani and M Six Silberman. 2013. Turkopticon: Interrupting Worker Invisibility in Amazon Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 611–620.

[37] Khari Johnson. 2020. NeurIPS Requires AI Researchers to Account for Societal Impact and Financial Conflicts of Interest. Retrieved January 21, 2022 from https://venturebeat.com/2020/02/24/neurips-requires-ai-researchers-to-account-for-societal-impact-and-financial-conflicts-of-interest

[38] Jay Katz. 1996. The Nuremberg Code and the Nuremberg Trial: A Reappraisal. *Jama* 276, 20 (1996), 1662–1666.

[39] Norbert L Kerr. 1998. HARKing: Hypothesizing After the Results are Known. *Personality and social psychology review* 2, 3 (1998), 196–217.

[40] Jeffrey T Leek and Roger D Peng. 2015. Opinion: Reproducible Research Can Still Be Wrong: Adopting a Prevention Approach. *Proceedings of the National Academy of Sciences* 112, 6 (2015), 1645–1646.

[41] Hsuan-Tien Lin, Maria Florina Balcan, Raia Hadsell, and Marc Aurelio Ranzato. 2020. What We Learned From NeurIPS 2020 Reviewing Process. Retrieved January 21, 2022 from https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2020-reviewing-process-e24549eea38f

[42] Jennifer M Logg and Charles A Dorison. 2021. Pre-registration: Weighing Costs and Benefits for Researchers. *Organizational Behavior and Human Decision Processes* 167 (2021), 18–27.

[43] Jharna Mandal, Srinivas Acharya, and Subhash Chandra Parija. 2011. Ethics in Human Research. *Tropical parasitology* 1, 1 (2011), 2.

[44] Rose McDermott and Peter K Hatemi. 2020. Ethics in Field Experimentation: A Call to Establish New Standards to Protect the Public From Unwanted Manipulation and Real Harms. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30014–30021.

[45] Jacob Metcalf and Kate Crawford. 2016. Where are Human Subjects in Big Data Research? The Emerging Ethics Divide. *Big Data & Society* 3, 1 (2016), 2053951716650211.

[46] 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.

[47] Priyanka Nanayakkara, Jessica Hullman, and Nicholas Diakopoulos. 2021. Unpacking the Expressed Consequences of AI Research in Broader Impact Statements. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)* (2021).

[48] NeurIPS. 2020. Getting Started with NeurIPS 2020. Retrieved January 21, 2022 from https://medium.com/@NeurIPSConf/getting-started-with-neurips-2020-e350f9b39c28

[49] NeurIPS. 2021. NeurIPS 2021 Paper Checklist Guidelines. Retrieved January 21, 2022 from https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist

[50] Safiya Umoja Noble. 2018. *Algorithms of Oppression*. New York University Press.

[51] Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[52] Richard Owen, Phil Macnaghten, and Jack Stilgoe. 2012. Responsible Research and Innovation: From Science in Society to Science For Society, With Society. *Science and Public Policy* 39, 6 (Dec. 2012), 751–760. https://doi.org/10.1093/scipol/scs093 Publisher: Oxford Academic.

[53] PAI. 2021. Managing the Risks of AI Research. https://www.partnershiponai.org/responsible-publication-recommendations

[54] Kenneth L Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons From 1000 Papers. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[55] 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. *Journal of Machine Learning Research* 22 (2021).

[56] Carina EA Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. Institutionalizing Ethics in AI Through Broader Impact Requirements. *Nature Machine Intelligence* 3, 2 (2021), 104–110.

[57] Katyanna Quach. 2020. MIT Apologizes, Permanently Pulls Offline Huge Dataset That Taught AI Systems To Use Racist, Misogynistic Slurs. Retrieved January 21, 2022 from https://www.theregister.com/2020/07/01/mit_dataset_removed

[58] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever. 2019. Better language models and their implications. *OpenAI Blog https://openai. com/blog/better-language-models* 1 (2019), 2.

[59] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green AI. *Commun. ACM* 63, 12 (2020), 54–63.

[60] Uri Simonsohn, Leif D Nelson, and Joseph P Simmons. 2014. P-curve: A Key to the File-drawer. *Journal of experimental psychology: General* 143, 2 (2014), 534.

[61] Ana-Maria Simundic. 2013. Bias in Research. *Biochemia medica* 23, 1 (2013), 12–15.

[62] Fujian Song, Sheetal Parekh, Lee Hooper, Yoon K Loke, J Ryder, Alex J Sutton, C Hing, Chun Shing Kwok, Chun Pang, and Ian Harvey. 2010. Dissemination and Publication of Research Findings: An Updated Review of Related Biases. *Health technology assessment* 14, 8 (2010), 1–220.

[63] Victoria Stodden. 2014. 2014 : What Scientific Idea is Ready For Retirement? https://www.edge.org/response-detail/25340.

[64] Julia Stoyanovich and Bill Howe. 2019. Nutritional Labels for Data and Models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering* 42, 3 (2019).

[65] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3645–3650.

[66] The Data Protection Act 2018. https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted

[67] The Turing Way Community. 2021. *The Turing Way: A Handbook For Reproducible, Ethical and Collaborative Research*. https://doi.org/10.5281/zenodo.3233853

[68] UK Data Service. [n.d.]. The Data Protection Act and GDPR. Retrieved January 21, 2022 from https://ukdataservice.ac.uk/learning-hub/research-data-management/data-protection/data-protection-legislation/data-protection-act-and-gdpr

[69] Meredith Whittaker. 2021. The steep cost of capture. *Interactions* 28, 6 (2021), 50–55.

[70] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future At the New Frontier of Power*. Profile books.