A Conceptual Framework for Ethical Evaluation of Machine Learning Systems

Neha R. Gupta[1], **Jessica Hullman**[2], **Hari Subramonyam**[3]
[1] Carnegie Mellon University
[2] Northwestern University
[3] Stanford University
nehargupta@cmu.edu, jhullman@northwestern.edu, harihars@stanford.edu

Abstract

Research in Responsible AI has developed a range of principles and practices to ensure that machine learning systems are *used* in a manner that is ethical and aligned with human values. However, a critical yet often neglected aspect of ethical ML is the ethical implications that appear when *designing evaluations* of ML systems. For instance, teams may have to balance a trade-off between highly informative tests to ensure downstream product safety, with potential fairness harms inherent to the implemented testing procedures. We conceptualize ethics-related concerns in standard ML evaluation techniques. Specifically, we present a utility framework, characterizing the key trade-off in ethical evaluation as balancing information gain against potential ethical harms. The framework is then a tool for characterizing challenges teams face, and systematically disentangling competing considerations that teams seek to balance. Differentiating between different types of issues encountered in evaluation allows us to highlight best practices from analogous domains, such as clinical trials and automotive crash testing, which navigate these issues in ways that can offer inspiration to improve evaluation processes in ML. Our analysis underscores the critical need for development teams to deliberately assess and manage ethical complexities that arise during the evaluation of ML systems, and for the industry to move towards designing institutional policies to support ethical evaluations.

Introduction

Machine learning (ML) model evaluation typically focuses on estimating errors of prediction or estimation via quantifiable metrics. Given the increasing size and complexity of ML systems, comprehensive evaluations should ideally be multifaceted. For example, evaluations of large ML systems may include several methods, including A/B testing on live populations, adversarial testing to produce undesirable outputs, and comprehensive audits documenting outputs. Potential ethical harms of ML systems have gained increasing attention in the broad Responsible AI community. However, even when evaluation metrics are expanded beyond performance to include factors like fairness, privacy loss, or other harms induced by the machine learning system, this is often focused on the ethical harms of the released system, overlooking possible harms incurred during the machine learning development lifecycle itself. This is problematic because evaluation approaches do have the potential to cause ethical harm during evaluation. In a noteworthy example, Tesla's autonomous vehicle live testing systems on real roadways in California, has been widely criticized for being involved in various crashes (Nayak, Laing, and Hull 2022).

How should practitioners evaluate large complex systems with potentially unknown ethical harms across the engineering lifecycle, including during the evaluation process? We provide a conceptual framework that casts the primary trade-off in ethical evaluation decision-making as balancing the goal of optimizing for information gained in an evaluation, against the possible ethical harms that are induced.

Based on our sketch of this fundamental problem that practitioners face, we identify a series of challenges that can cause practitioners to stumble in selecting ethical evaluation practices. We illustrate these challenges using real-world examples of machine learning evaluations that encountered them. Then, we draw parallels between these challenges and evaluation practices in domains other than machine learning, to explore potential mitigation techniques. Together, our conceptual framework and characterization of challenges are intended to stimulate discussion among researchers and evaluation teams on how to balance information gain with potential ethical harm, and to motivate future exploration of policies or best practices for machine learning evaluation.

Related Works

Ethical AI

A growing body of literature discusses properties that ethical machine learning systems should inherently possess, and provides principles and guidelines for testing (Jobin, Ienca, and Vayena 2019; Zhang et al. 2020; Martínez-Fernández et al. 2022). Broadly speaking, the ethical values identified by prior work include: (1) **Non-maleficense**, which measures the extent to which the evaluation workflows and outcomes do not inflict harm or injuries on any individual or population (Mehrabi et al. 2021). (2) **Privacy**, which as a value refers to the principle of protecting personal and sensitive information from unauthorized access, use, or exposure during the entire ML lifecycle (Liu et al. 2021). (3) **Fairness**, in which the goal is to achieve equitable treatment and outcomes for all individuals, ensuring that the

benefits and burdens of AI systems are distributed justly across diverse populations (Mehrabi et al. 2021), (4) **Cultural Sensitivity**, which involves designing algorithms and models that are attuned to and respectful of cultural differences, ensuring that they do not perpetuate stereotypes, biases, or insensitivities (Jo and Gebru 2020). (5) **Sustainability**, which ensures that models are developed in a manner that is environmentally responsible, economically feasible, and socially equitable (on Artificial Intelligence 2019; Lacoste et al. 2019). (6) **Societal Impact**, which refers to ensuring that the models contribute positively to societal well-being, address social issues, and do not harm individuals or communities (Birhane et al. 2022). We emphasize that this is not a comprehensive listing of all ethical values that a machine learning system may seek to accomplish. We provide examples of these categories when characterizing challenges in designing ethical evaluations in later sections.

ML System Evaluation Practices

An *evaluation* is the process by which practitioners detect differences between desired and actual model behavior (Zhang et al. 2020), through empirical assessment of model properties (Shevlane et al. 2023). A growing body of work creates more comprehensive methods with which to evaluate systems, rather than providing a singular empirical metric or set of metrics. Some evaluation methods can be conducted *pre-deployment*, such as A/B testing or live testing. Other mechanisms are used *post-deployment*, such as bug bounty challenges, and provide infrastructure to support stakeholder feedback. A particular evaluation process may involve choosing one or many evaluation metrics to measure. These decisions are critical because they impact actions that are taken post-evaluation to improve system capabilities. They also may be associated with the potential for ethical harm incurred in the evaluation process, or after product release. *We define an ethical evaluation as an evaluation that does not sacrifice ethical values in its implementation, and attempts to forecast downstream ethical harms across the product lifecycle.*

Finally, by considering how to conceive of the value of information about model performance gained through an evaluation, our work is related to data valuation. Prior theoretical work in machine learning and related fields studies the value of data for purposes like explainability (e.g., the Shapley framework (Ghorbani and Zou 2019)), data markets and incentivizing collaboration in ML (Castro Fernandez 2023; Sim et al. 2020), and value of accurate or improved prediction for goals like treatment assignment or welfare maximization (Liu et al. 2024; Perdomo 2023).

Ethical Evaluation Model

Motivation

The economics discipline has a long history of creating highly simplified models of complex real-world processes to assist with predicting the consequences of actions. Abstracting away non-essential features of the complex real-world permits systematic reasoning.

For example, economic policy-makers concerned with pricing wheat might use a simplified model that includes the costs to the farmer while abstracting away other potentially relevant characteristics, such as soil quality and his educational background (Friedman 1953). Our conceptual model of the key trade-off in ML practitioners' evaluation decision-making focuses on the value of information gain relative to ethical harms. However, rather than contributing new theoretical results, our goals are epistemological: to prompt reflection on what it would mean to select the best evaluation in a way that accounts for potential ethical harms induced in evaluation. By conceptualizing the idea of an optimal balance between competing concerns in designing ML evaluations, our framework is meant to highlight difficult questions that largely remain un-navigated in the literature and practice of ML evaluation design, rather than to imply that a normative evaluation design is easy to identify. Below, we discuss the implications of components of the model, including the acceptability of some of the assumptions made for the sake of this model, issues that arise due to differing aims, and the subjectivity of variables in further sections.

Model Properties

ML teams select from a space of possible evaluations. An evaluation is a protocol for assessing and measuring a model's performance against a set of defined criteria or benchmarks, including specification of which information to collect and how. ML development teams face various considerations when planning evaluations that involve complex decisions across evaluation scope, context, and effect (Zhang et al. 2020; Riccio et al. 2020; Song et al. 2022); prior work has described how practitioners can suffer from a "paradox of choice" when it comes to deciding how to perform evaluation (Goel et al. 2021). We represent the space of possible evaluations under consideration by a team as $A = \{a_1, a_2, \ldots\}$ where a is an evaluation decision (e.g., evaluation method, metrics, sample selection, etc.).

$$\textit{choose some } a \in A$$

The utility of an evaluation approach depends on the relative value of information gained, ethical harms, and resource costs. The fair ML literature has represented decisions about model choice in ML in a utility framework, where models provide utility as a function of costs and benefits (Corbett-Davies et al. 2017; Corbett-Davies and Goel 2018; Chohlas-Wood et al. 2021). For example, in Corbett-Davies et al. (2017), the authors conceptualize 'immediate utility' reflecting the costs and benefits of a fair decision by a policymaker in the setting of pre-trial bail release decisions. A utility framing is also used in Hutchinson et al., to illustrate the task of evaluating an ML model's suitability for use in a specific application ecosystem.

Our conceptualization similarly draws on a utility framework common in statistical decision theory (Savage 1972; Steele and Stefánsson 2015; Von Neumann and Morgenstern 2007), but expands this to a broader view applied to decisions made by teams evaluating ML models or systems. The

proposed utility function implies that each evaluation decision option the team is considering can be compared along a single dimension (utility) along which they can be ranked.

We conceive of three categories of inputs to the utility function. The first concerns the value of information gain. The information learned from a test, regarding differences between desired and actual model behavior (Zhang et al. 2020), implies a gain from the evaluation process. Information gained in conventional ML performance evaluation includes estimates of how well a trained model generalizes to new samples from the same distribution the model was trained on, as well as measures of the robustness of the model, i.e., the degree to which the model or system maintains its correctness and performance under varying conditions and inputs, including invalid or adversarial inputs (Tjeng, Xiao, and Tedrake 2017; Zhang et al. 2020). When the evaluation produces information about the model performance along ethical dimensions, the information gained may come in the form of the expected magnitude or frequency of harms upon deployment.

The second component is ethical harms of the evaluation. Often overlooked, ethical issues incurred during evaluations, or downstream ethical issues not adequately predicted (and consequently encountered after deployment), can diminish the acceptability of the results, and the overall value and utility of the information derived from the evaluation.

The third component measures the material resources required to conduct an evaluation. Teams often consider options for reducing the costs of tests via methods such as test prioritization, in which inputs generated for tests are limited to inputs that are most indicative of problematic behavior (Zhang et al. 2020). Costs can take several forms. For instance, monetary constraints may restrict data collection abilities including the number of labelled data annotations procured for supervised ML models (Liao, Kar, and Fidler 2021; Goel and Faltings 2019). Cost constraints through labor force availability and time constraints can shift teams towards using automated software testing (Dustin, Garrett, and Gauf 2009). These resource constraints can challenge responsible model development (Hopkins and Booth 2021).

Consequently, we represent the utility of an evaluation decision as having these three inputs, with the information gain representing gains to utility and potential ethical harm and material cost representing decreased utility:

$$u(a) \;=\; (\textit{information gain} \;-\; \textit{ethical harm} \;-\; \textit{cost})$$

Information gain, ethical harm, and cost are forecasts. Before conducting an evaluation, a team cannot precisely predict the information gained about model performance, potential ethical harm, or exact material costs involved. Consequently, information gain, ethical harm, and cost as predicted values, which we represent as expectations over relevant sources of randomness. The fact that these quantities must be predicted emphasizes the uncertainty under which evaluation decisions are necessarily made.

These values are represented as expectations over distribution of possible values. Estimating these distributions is a fundamental part of the challenge in selecting an evaluation. Any particular evaluation involves a sample of instances for which evaluation data is gathered. Many evaluations sample from a population of participants. Complications arise as models can differ in social impact across groups, notably underrepresented groups (Hutchinson et al. 2022). Thus, the estimate of expected "ethical harm" requires averaging harm across several individuals who experience disparate impact from the models. Moving forward, we abbreviate information gain as $IG(a)$, and ethical harms as $EH(a)$:

$$u(a) = \mathbb{E}(\textit{IG}(a)) - \mathbb{E}(\textit{EH}(a)) - \mathbb{E}(\textit{cost})$$

Ethical harm can be decomposed by distinct ethical values. It is important to differentiate between various ethical values in our model, because it has been established that there are situations where some models may benefit a particular ethical value at the cost of another. For example, a privacy and fairness trade-off affects some ML models (Pujol and Machanavajjhala 2021). While interactions between ethical concerns may exist, for simplicity we think of $\mathbb{E}(EH(a))$ as representing a weighted sum of various ethical values, so $\mathbb{E}(EH(a)) = \sum_j w_j \mathbb{E}(EH_j(a))$ where j paramaterizes the ethical values discussed in Section 2. These weights can represent differing ethical priorities of teams or regulatory requirements on particular values.

$$\mathbb{E}(EH(a)) = \sum_j w_j \mathbb{E}(\textit{EH}_j(a))$$

The best evaluation method has the highest utility. We represent the optimal choice of evaluation practice for the team as the decision with the highest utility. An evaluation approach a equals the optimal decision a^* if it is the utility-maximizing decision, indicated as:

$$a^* = \textit{argmax}_{a \in A} U(a)$$

The final form of the utility model can be written as:

$$a^* = \textit{argmax}_{a \in A} \mathbb{E}(\textit{IG}(a)) - \sum_j w_j \mathbb{E}(\textit{EH}_j(a)) - \mathbb{E}(\textit{cost}) \tag{1}$$

In Table 1, we reiterate the key properties of practitioners' decision-making when selecting an ethical evaluation practice. We accompany these properties with guiding questions for the ML industry to explore, in order to move towards prioritizing ethics in evaluations. Combining these properties into Equation 1, we see that a best evaluation practice is chosen after considering the information gained from the practice, potential ethical harms, and is limited by the resources available.

The conceptual framework above provides a high-level sketch of how ethical harms associated with machine learning evaluation can affect the overall utility derived from the evaluation process. We now discuss how common challenges practitioners face in the process of selecting an evaluation practice raise questions about whether the best evaluation has been chosen.

Utility Framing	Explanation	Question for ML community
choose $a \in A$	Selecting an evaluation out of a set of options	*How can we ensure that practitioners are carefully weighing a range of options for evaluation?*
$u(a) =$ *information gain* $-$ *ethical harm* $-$ *cost*	Utility is composed of information gained, ethical harms, and costs	*How can we ensure ethical harms introduced in evaluation are considered?*
$u(a) = \mathbb{E}(IG(a)) - \mathbb{E}(EH(a)) - \mathbb{E}(cost)$	Information gained, ethical harm, and costs are in expectation	*How can we ensure practitioners account for potential ethical harms and issues in estimating harms incurred in evaluation?*
$\mathbb{E}(EH(a)) = \sum_j w_j \mathbb{E}(EH_j(a))$	Ethical harm is composed of weighted ethical values	*Which specific ethical values are impacted by evaluations? How might regulatory requirements for particular ethical values impact the choice of evaluation?*
$u(a^*) = argmax_{a \in A_c} U(a)$	The best evaluation has the highest utility	*How do practitioners who do consider ethical harms define the best evaluation framework and then compare between options?*

Table 1: Properties for a utility model framing the costs, benefits, and resource constraints for a team's decision-making.

Discussing the interactions between components of the framework, the philosophical challenges, and practical challenges that can arise serves two purposes. First, these challenges do appear in practice. We illustrate the nature of ethical harms resulting from real-world evaluation practices in order to make concrete the sorts of consequences that appear in selecting evaluation practices. We selected the examples below using a broad search across scholarly publications and news media related to ethical issues that arise in ML, with a specific focus on those that can affect evaluation. We focused on identifying instances that varied in evaluation practice, ethical values at risk, and context. We also prioritized examples that provided clear insights into potential incurred harms, or direct evidence of ethical harms.

Secondly, we discuss real-world evaluations to motivate exploration of mitigation strategies within the ML evaluation industry. We accompany each common challenge with an existing mitigation strategy from ethically-motivated evaluations in domains other than ML. Other fields have established regulatory and administrative systems that help them balance the tradeoffs which arise in evaluation practices, or have informal best practices. These potential mitigation strategies can guide future discussion and move the ML industry towards balancing compliance with ethics.

Our discussion is distilled into the notation from the notation from the utility model in Table 2.

Issue 1: Aggregating Over Populations Masks Group and Individual Differences

Taking the expectation of "ethical harm" aggregates over individuals and groups. Just as a particular value for a model error metric (like accuracy) or a point estimate (like an estimated average treatment effect) can admit numerous solutions that vary at the level of the individual units or groups (e.g., (Coston, Rambachan, and Chouldechova 2021; Gelman, Hullman, and Kennedy 2023; Marx, Calmon, and Ustun 2020)), aggregating ethical harms over different individuals can lead evaluators to overlook individual or group-specific concerns. For example, two evaluation protocols may be expected to result in the same level of ethical harm to participants, while differing greatly in how harm is distributed over the specific participants or groups of participants.

Example in ML Evaluations: Medical AI Device Testing. Researchers have raised concerns regarding evaluation practices of FDA-approved medical devices. In an analysis of 130 devices, 93 did not have multi-site assessment, meaning many were evaluated at one site, which may have limited geographic diversity. This includes 54 high-risk devices, and devices affecting a range of body areas (chest, breast, heart, head, other) (Wu et al. 2021).

The evaluation of medical AI devices on limited samples of the population is a form of the general practice of 'data splitting', where practitioners partition a population and monitor the performance of the model within a slice of data. This requires careful decision making on the choice of the slice (Chen et al. 2019). In this case, researchers criticize single-site assessment of medical AI device testing because the relationship between the performance information that is gained and performance in the broader population is unclear. Deploying a method evaluated on a narrow slice of the intended population can yield unintended biases in performance of the device on underrepresented groups post-deployment (Wu et al. 2021). The FDA has noted these ethical concerns, calling for greater transparency in testing and improved monitoring of algorithmic bias (Wu et al. 2021).

Mitigation Example from Analogous Domain: Representation in Clinical Trials.

The lack of representation in clinical research has been studied in contexts outside of medical devices. Statistical adjustment techniques, like population-weighted sampling and post-stratification, are common mitigation strategies in the causal inference literature.

Poor evaluation choices can mask heterogeneity in health needs, leading to downstream harms. For example, not having information on certain subsets of a population may ulti-

Challenge in balancing evaluation considerations	Example from ML evaluation practices	Mitigation strategies in analogous domain	Open question for ML community
Ethical harm, $\mathbb{E}(EH(a)$, combines expected ethical harm of many individuals or groups who may be impacted differently	Medical AI device testing done on a single geographic site as a form of 'data splitting' can yield low external validity and low utility to diverse populations	The medical community offers best practices to broaden representation in clinical trials and widen external utility through information campaign	How can we ensure $EH(a)$ considers consider individual and group specific impacts of an evaluation?
Subjective interpretations of the presence and magnitude of $\mathbb{E}(EH)$	Facebook NewsFeed randomized experiment was criticized by some as causing social harm or unfairness during evaluation	IRBs act as a centralized agency and 3rd party overseeing the approval of placebo control trials prior to deployment with a focus on ethics of the evaluation process	How can we ensure decision makers are able to estimate $\mathbb{E}(EH)$ despite conflicting opinions?
Utility balances tradeoffs between future $\mathbb{E}(IG(a))$ from the evaluation and potential ethical harms, $\mathbb{E}(EH(a))$	Adversarial testing is popular due to its anticipated information gain. But, ethical harms have been established in adversarial testing. For example, guidelines for the data labelling step establish cultural insensitivity or social harm, and other cases of adversarial testing in the physical domain illustrate privacy losses.	US NEPA guidelines require documentation and consideration of environmental harm when proposing federal actions, but is criticized because the documentation is not required to be a primary decision factor. A recommendation to improve environmental impact assessments includes follow-up monitoring.	How can the ML industry identify and regulate the consideration of particular ethical harms?
Difficulties in comprehensive risk assessments, including unknown probability of ethical harm, $\mathbb{E}(EH(a))$ at time of decision	Microsoft Tay was released for live testing following offline user studies and stress-testing. However, unknown vulnerabilities were not revealed in offline testing, and these led to ethical harm through cultural insensitivity during live testing.	US NRC conducts probabilistic risk assessments on nuclear power plants, focusing on distributions and likelihoods of risks. They impose safety standards to account for uncertainties in the distributions	How can we motivate better and more careful assessments of potential ethical harms during the evaluation process?
$u(a)$ having a negative relationship with $\mathbb{E}(cost)$ can lead to fewer evaluations than preferred by regulators	Tesla autonomous live testing has been involved in crashes in test driving, causing ethical harm through the value of social harm. Critics recommend raising standards for offline testing, which would require corporations to invest more resources to evaluations	US court decisions in the 1960s held manufacturers responsible for crashworthiness of vehicles, motivating manufacturers to use costly ATDs in testing in offline testing prior to release.	How can we ensure practitioners devote more resources to evaluations?
$u(a)$ does not capture the importance of downstream decisions ex-post evaluations	ML applications in education have been criticised for only evaluating model accuracy, rather than the use of models through impact on students in an intervention	Financial regulators impose stress-testing standards, including scenarios with sequences of decisions	How can we ensure that evaluation decisions are downstream actionable, in the face of considerable uncertainty?

Table 2: A summary of challenges in selecting ethical evaluations that are implied by the framework provided. The reality of each of these challenges in the machine learning industry is illustrated by providing an example for each. We also provide an example of a mitigation strategy in an analogous domain to motivate a discussion of lessons for the ML industry. This is not meant to be a comprehensive listing of issues in creating ethical evaluations, but allows us to explore key questions that could motivate avenues to move towards ethical AI through practitioner or regulatory action.

mately result in a lack of access to effective interventions for some groups, because sufficient information was not available to obtain treatments. This can potentially compound effects of health disparities, and increase costs (Bibbins-Domingo, Helman et al. 2022).

Proposed solutions to widen the inclusivity of clinical trials tend to imply that more resources must be expended in evaluation to improve the value of the information that is gained. These include tailoring recruitment materials with language that emphasizes available support and the value of participant involvement, and providing transportation for participating in a trial (Clark et al. 2019).

Open Question for ML community: How can we develop evaluation selection guidelines that motivate evaluators to consider individual and group specific impacts of an evaluation design?

Issue 2: Disagreement on Presence or Magnitude of Ethical Harms

When facing a decision regarding the best possible evaluation practice, estimating and agreeing upon expected ethical harm, $\mathbb{E}(EH)$, is a challenge. Some work in the machine learning and ethics literature argues that a universally acceptable function ranking ethical outcomes does not exist, and that impartiality is simply an ideal (Card and Smith 2020). Practitioners' interpretations of an ethical harm may differ between team members, or with the general public. Furthermore, ethical impacts are considered hard or even impossible to quantify, making it a challenge to prioritize them in metrics-driven development environments (Ali et al. 2023). This is distinct from the potential issue above, in that the magnitude may be similar across demographic groups but still difficult to agree upon.

Example in ML Evaluations: Facebook NewsFeed Randomized Experiment. For one week in January of 2012, Facebook ran a large scale randomized experiment on over 600,000 unknowing participants, randomizing the NewsFeed ranking algorithm they saw. The NewsFeed was the primary mechanism through which individuals saw their friends' content (Kramer, Guillory, and Hancock 2014), and relied on a machine learning algorithm optimized for user behavior while incorporating many weights (Meyer 2015).

This experiment, as with A/B testing more generally, is useful in allowing platforms to observe model performance in complex interactive systems, in order to establish the superiority of one approach over the other with a high degree of statistical certainty. The experiment yielded information gain that contributed to internal evaluation of the feed algorithm, as well as broader scientific understanding of *emotional contagion*, the impact on an individual's emotion based on exposure to friends' emotions (Kramer, Guillory, and Hancock 2014). While not necessarily intended to motivate direct action, this kind of knowledge gain presumably informs the platform's future design strategy. Hence, the information gained carried some utility to Facebook. This utility diminished, however, by a backlash among the public and some scholars who perceived it as imposing ethical harm on some users who received the negatively manipulated feed. Interestingly, the public backlash conflicted with what was suggested in some prior studies, that users react negatively to others' *positive*, envy-inducing content (e.g., (Krasnova et al. 2013)). Critics decried the practice of randomly placing some individuals in a treatment intervention that they expected (posthoc) to yield less positive emotions, arguing this produced ethical harm through the values of social harm and unfairness. Some critics called for federal agency investigations, and drew parallels to regulations enforced by Internal Review Boards (IRBs) (Meyer 2015). Others however have pointed out that these critiques, like other critiques of platform feed manipulations, exhibit a common "A/B illusion" (Meyer 2015) where a randomized experiment comparing two policies or treatments (A and B) with an unknown rank order in terms of some measure of quality, is deemed less appropriate than simply implementing either A or B for everyone (Meyer et al. 2019). Hence, the NewsFeed experiment, as an evaluation, led to highly contrasting views on whether inappropriate ethical harm was incurred.

Ultimately, the response to the perceived ethical harms can be thought of as a loss of utility, one that may have been prevented by a different evaluation design. For example, an alternative design might ask users if they want to opt into experimentation. However, this introduces the potential for less information gain, since selection biases come into play, illustrating the difficulty of balancing these concerns when what constitutes a harm can be contested.

Mitigation Example from Analogous Domain: Placebo-Controlled Trials and IRB regulations. Clinical research, including placebo-controlled trials, are regulated by IRBs (Polonioli et al. 2023; Food and Administration 1998). A/B tests using placebo-controlled trials may be ethically questionable as they deny some participants access to treatment. However, due to the advantages, such as a rigorous test of efficacy, medical practitioners have established guidelines for when placebo trials are appropriate (Millum and Grady 2013). Some of these principles include granting permission when no proven treatment exists for the disorder being studied, and when patients are exposed to at most "temporary discomfort or delay in relief of symptoms." (Miller and Brody 2002; Millum and Grady 2013).

The role of the IRB is to evaluate the ethics of a proposed evaluation (Polonioli et al. 2023). In our conceptualization, this is analogous to an external evaluation approving that $\mathbb{E}(EH)$ meets standards prior to the experiment proceeding. It has been proposed that corporations running online experiments launch internal IRBs or consider external IRBs (Polonioli et al. 2023). There have also been calls to increase the transparency of A/B testing or for platforms offering A/B testing to provide ethics training to practitioners to assist their evaluation design (Jiang, Martin, and Wilson 2019). Both of these aim to improve ethical impact estimates prior to deploying live experiments.

Open Question for ML Community: How can evaluators estimate ethical harms in ways that allow for potentially conflicting opinions on the presence or magnitude of harms?

Issue 3: Difficulty of Balancing Future Gains in Utility Against Immediate Ethical Harms

The conceptualization we propose requires balancing expected ethical harm with expected information gain. Even if ethical harm is established (as previously discussed in issues 1 and 2), situations may exist where teams believe it is permissible to ignore potential ethical harms that could occur in evaluation because the information gained through the process could lead to a more socially beneficial downstream ML system. In the absence of attempts to more carefully weigh concerns against each other, it is easy for model developers to engage in wishful thinking that minimizes more direct and immediate ethical harms incurred in evaluation under the guise of more abstract expected long-term benefits. In such cases, a regulatory requirement on ethical values in evaluation outside of ML could potentially mitigate issues.

Example in ML Evaluations: Adversarial Testing. Adversarial testing in evaluation has become increasingly popular in machine learning. Adversarial testing can be done either prior to the release of models or on released models as part of an iterated deployment process (Google 2023; Shevlane et al. 2023). In this process (also called 'red-teaming'), practitioners intentionally seek out cases where models can behave in ways that would be undesirable. A common manual approach involves individually devising malicious inputs to provide to models, and inspecting the corresponding outputs. This direct intervention approach aims to allow engineers to uncover model failures or vulnerabilities, and identify corrective steps (Google 2023; Shevlane et al. 2023).

However, adversarial testing can introduce ethically harmful impact to practitioners, through the value of harmfulness or cultural insensitivity. This can occur during the model output labeling step, which, according to the Google documentation, "necessarily involves looking at troubling and potentially harmful text or images, similar to manual

content moderation" (Google 2023). Potential ethical issues in content labeling have been established in other literature (Barbosa and Chen 2019). A risk is that the ethical value of harmfulness is underweighted as a concern when performing adversarial testing, relative to the more nebulous value of the anticipated information gain and subsequent anticipated safety gains of a system.

Another example of potential ethical harm in adversarial testing occurs in computer vision, where researchers test adversarial physical characteristics. Examples include t-shirts that evade computer vision systems that count individuals, or eyewear that evades facial identification systems. These studies have been critiqued for having limited samples when physical testing relative to digital testing. A further critique is the ethical harm of potential privacy losses, such as not blurring faces for the adversarial testers (Albert et al. 2020).

The unclear value of the information gain and potential ethical harms of adversarial testing are in tension with the rise in enthusiasm around adversarial testing. This enthusiasm is even backed by policy; there was recently an executive mandate to establish guidelines and require corporate reporting of performance (Biden 2023) for companies developing foundation models. Mandates for adversarial testing should be balanced with consideration towards their potential ethical harms. Future work should explore data valuation frameworks for identifying adversarial examples.

Mitigation Example in Analogous Domain: US government Environmental Impact Standards and values. The 1970 National Environmental Policy (NEPA), is a government regulation that was created to ensure ethical values are considered in evaluations. This requires possible environmental effects of actions to be considered and documented by federal agencies. Federal agencies must begin an environmental review process before their final decisions are made, in which they aim to determine if their proposed actions have causal relations to significant environmental effects. NEPA does not mandate the most environmentally sound alternative be chosen in any decision, but requires organizations have knowledge of the impact of decisions. Among other organizations, the Environmental Protection Agency works on overseeing NEPA (on Environmental Quality Executive Office of the President 2021).

Enforcing documentation and consideration of ethical harms is not universally accepted as a useful practice. Critics say NEPA is weak due to the lack of accurate ex-ante predictions on environmental impact and lack of follow-up monitoring (Karkkainen 2002), surfacing concerns about the difficulty of estimating ethical harms. Furthermore, many categorical exclusions exist, initially introduced to acknowledge that not all governmental actions pose environmental risk. Now, critics say this is abused as a loophole to bypass review (Fox 2023). A final criticism is that environmental reviews have become time and cost-consuming, and can delay critical clean energy projects, indicating clear dysfunction (Meyersohn 2023). This implies that such evaluations have a higher resource cost that is not well balanced by the value of the information gained. Reform efforts balance the need for action and ethical value evaluations, and recommend fewer exclusions, and adaptive mitigation strategies (Meyersohn 2023; Karkkainen 2002; Fox 2023)

Overall, the field of sustainability has established mandates that environmental analysis be considered in a decision-making process. This is analogous to mandating that the set of ethical values, j, include sustainability. The specific mechanics of the policy are critical to ensuring ethical development is successfully accomplished, as illustrated by critics of NEPA.

Open Question for ML Community: How can ML application industries identify and regulate the consideration of particular ethical harms?

Issue 4: Difficulties in Comprehensive Risk Assessment in Real-World Environments

A relevant challenge posed by the consequentialism framework of ethical decision-making processes is that forecasting future ethical well-being and harms across many hypothetical worlds is difficult (Card and Smith 2020). The expectation of $EH(a)$ has to aggregate ethical harm over expected sources of randomness (e.g., stemming from unknown baseline risks of offensive content in content moderation, or unknown, potentially adversarial user behavior after model deployment). This is intensive for practitioners to think about when making decisions, as they may do what they can to prevent harm and vulnerabilities but still experience unanticipated results.

Example in ML Evaluations: Microsoft Tay Chatbot Testing. In 2016, Microsoft launched a chatbot, Tay, live online. This online live testing and release came after an offline development lifecycle, during which they conducted user studies, and stress-tested the bot under various conditions, to ensure the bot had positive interactions. They hoped releasing Tay online on Twitter would allow them to reach a larger userbase to learn and improve it (Lee 2016). However, within 24 hours, it was removed from Twitter, because a vulnerability in the model led to inappropriate words and images from the chatbot (Wolf, Miller, and Grodzinsky 2017).

When the team performed offline testing on Tay, the information they gained led them to believe that it was ready for online interactions through live testing and release with the public on Twitter. At the time of release, the team did not have awareness of the vulnerability that was later exploited to cause harm (Lee 2016), through the ethical values of cultural sensitivity, unfairness, and harmfulness. This illustrates the way in which offline user studies were unable to provide sufficient information gain, and in which the anticipated ethical harms were miscalculated. This could be mitigated by better anticipating future risks.

Mitigation Example in Analogous Domain: Risk Assessment in Nuclear Power Plant Safety.

The US Nuclear Regulatory Commission (NRC) oversees US nuclear power plants, ensuring they "operate with minimal risk to public health and safety". They use probabilistic risk assessment techniques that explore likelihood of risks, offering a comprehensive approach including potential initiating events, their respective frequencies, and uncertainties. When the distributions of potential outcomes they see have inadequate uncertainties, they impose defenses and safety

margins that account for that uncertainty (U.S. Nuclear Regulatory Commission 2024).

The NRC notes that there are difficulties in estimating ethical harms. Rather than focusing on the expected values, they focus on estimating the entire distribution of ethically harmful outcomes in their evaluations. This way, they are able to ensure that safeguards exist against a range of potentially harmful outcomes. The approach of the NRC to ensuring that difficulties in estimating outcomes are accounted for is beyond simply improving their risk assessment models; In the face of uncertainty, they impose sufficiently strict safety margins and defenses to account for the uncertainties (U.S. Nuclear Regulatory Commission 2024).

Open Question for ML Community: How can we motivate firms to better assess potential ethical harms arising from real-world interactions between AI systems and users or environments, during the evaluation process?

Issue 5: Insufficient Resources for Evaluations

As discussed above, cost constraints can challenge responsible model development. For example, many generative machine learning models are trained on large, widely available datasets that are believed to be domain-general, then fine-tuned on many small datasets due to the cost of obtaining high-quality, domain-specific data. If sufficient resources aren't devoted to domain-specific testing, the performance observed in an evaluation might appear to be sufficient, but the model might fail dramatically once deployed. Hence, cost constraints can lead to overestimation of the value of information gained. With regulatory or social norms that penalize ethical harms, practitioners can be motivated to devote more resources to investing in ethical evaluation processes to improve systems.

Example in ML Evaluation: Tesla Autonomous System Live Testing. California permits autonomous vehicle manufacturers to evaluate autonomous driving systems on public roads (DMV 2022). One of the autonomous vehicle systems utilizing this program is the Tesla Autopilot system, which, during training and testing on public roads with other drivers, was involved in several tragic fatal crashes (Siddiqui, Lerman, and Merrill 2022).

The value of information gained in live testing autonomous vehicles has been established in prior literature. In general settings, live testing provides information about performance given real-time rare events. Driving involves many rare events, which could confuse an autonomous navigation system (Ackerman 2017). Real-world roads contain phenomena that is difficult to simulate in synthetic environments, such as inclement weather, variable road conditions, aggressive or erratic behavior from other cars, and foreign objects (Nidhi Kalra 2016). Therefore, testing on live roads allows engineers to study instances of systems returning manual control to the human driver (Banerjee et al. 2018). Studying these interventions allows teams to identify failures in perception or control systems (Wang et al. 2020). Additionally, teams can receive feedback from passengers of autonomous vehicles, to learn more about the smoothness of the ride and identify issues with service (D'Onfro 2018).

However, ethical harms have occurred with live autonomous vehicle testing, leading critics to note the underregulation of autonomous vehicle testing, such that corporations might rely on live testing before conducting sufficiently safer offline tests. Some critics advocate for raising standards such that corporations are required to devote more resources to offline testing. For example, proposals include vision tests regarding abilities to recognize surroundings, including cars and pedestrians, prior to live testing (Claybrook and Kildare 2018), similar to testing norms for allowing people to operate vehicles. Taken together, the need for real world testing, while additionally adhering to the proposed raised standards for offline tests, would substantially increase the resources needed for evaluations.

Mitigation Example in Analogous Domain: Crash Test Reconstruction. Regulators and automotive engineers administer tests to evaluate the "crashworthiness" of different vehicles. This is motivated by decisions made by American courts in the the late 1960s, when they began to find automotive manufacturers liable for passenger injuries when elements of the car exacerbated the harms experienced by passengers. Regardless of whether accidents were caused by human error, courts argued that the statistically inevitable nature of car accidents meant that manufacturers carried a duty to minimize the consequences of such accidents (Choi 2019).

Crashworthiness is evaluated through crash tests. In a crash test, researchers place a crash dummy in a vehicle, and remotely drive the vehicle into a barrier in order to simulate a crash (Engber 2006). Sensor readings from the vehicle and the crash dummies –along with camera footage of the crash – are then studied to determine the crashworthiness of the vehicle. These tests can be expensive: estimates of the cost of crash dummies range between roughtly $100,000 and $1 million USD (Automology 2021; Hall 2015; Ferris 2022).

Thus, evaluating vehicle crashworthiness requires engineers to navigate a tradeoff between the cost of a crash test and the information it provides. Automotive companies have historically been incentivized to invest resources in crash testing based on standards, such as legal penalties for failure to do so, and consumer preferences for safe vehicles.

Open Question for ML Community: How can we motivate practitioners to devote sufficient resources to evaluations despite the need for resource costs to offset the value of the information gain?

Issue 6: Impact of Evaluations Depends on Downstream Actions

Conceiving the value of information gained in an evaluation can be as challenging as forecasting expected ethical harm. This difficulty arises because the information obtained is often instrumental to subsequent decision processes rather than being valuable in isolation. The value of information gained from an evaluation can be conceptualized in several ways. Evaluators may simply be interested in determining whether the estimated performance falls within an acceptable range. If it does not, the information becomes an input into a subsequent decision problem where the team

must consider what actions should be taken to improve the model's performance or adjust its deployment context.

To expand upon our notation introduced earlier, at the time of designing an evaluation the team can only estimate the information gain given a choice of evaluation a, which we denote $\widehat{IG(a)}$. We denote the realized information gain from a as $IG(a)$. We use $\mathcal{T}$ to denote the set of possible actions that can be taken on the model to improve the ML system after the evaluation is completed. On completing evaluation a, the team chooses the best post-evaluation action that yields the highest utility:

$$t^* = argmax_{t \in \mathcal{T}} U(t(IG(a))$$

The *evaluation gain* can be thought of as the gain in utility from taking post-evaluation t^* compared to taking no action, denoted as t_0:

$$EG(a) = \mathbb{E}(U(t^*(IG(a)) - U(t_0))$$

Ideally, $EG(a)$ could be used in the computation of a^* in Equation 1 in lieu of $IG(a)$, as follows:

$$a^* = argmax_{a \in A_c} \mathbb{E}(EG(a)) - \sum_j w_j \mathbb{E}(EH_j(a)) - \mathbb{E}(cost) \tag{2}$$

but as illustrated here, computing EG contains a number of additional challenges (choosing the utility-maximizing $t^*(\cdot)$, which is dependent on the observed $IG(a)$), and this introduces considerable uncertainty.

Example in ML Evaluations: Education Algorithms. ML algorithms experience a rapid growth in education domains, with applications such as predicting student performance and dropout risk (Lakkaraju et al. 2015), or evaluating post-secondary admissions. The recent rapid growth of machine learning techniques in education suffers from questions regarding whether these techniques support education principles and goals. Critics highlight ethical concerns with these algorithms, noting negative impacts on historically marginalized students (Liu et al. 2023).

Critics argue that evaluations of these models often prioritize predictive accuracy over their ability to inform effective educational interventions. To better translate predictions into interventions, one recommendation is to frame products as causal inference problems testable through methods like A/B testing (Liu et al. 2023). Incorporating knowledge of potential post-development actions into model creation can lead to more targeted interventions (Liu et al. 2024).

Mitigation Example in Analogous Domain: Stress Testing in Financial Regulation.

To ensure the stability of financial institutions, financial regulators, such as the International Monetary Fund and the US Federal Reserve Bank, enforce various "stress testing" frameworks, in order to assess vulnerability and ensure the stability of macroeconomic conditions in the face of plausible, abnormal shocks, such as major changes to exchange rates, or large credit defaults that reduce anticipated cash flows (Blaschke et al. 2001; Federal Reserve Board 2024).

The IMF enforces specific types of evaluations that include scenarios with sequences of decisions. However, regulators note some limitations, including that these requirements impose significant resource costs and expertise by involved parties and suffer computational complexities or data availability (Blaschke et al. 2001). In our framing, this illustrates an example of an evaluation that is performed while investigating downstream options, but notes significant costs to doing so.

Open Question for ML Community: How can teams ensure that their evaluation decisions are downstream actionable in the face of considerable uncertainty and additional downstream decision-making?

Discussion

The concept of choosing an evaluation to maximize utility, defined as a sum over expected information gain, ethical harm, and resource costs, encapsulates how we might idealize ethical evaluation. However, the challenges we discuss to this framing illustrate selecting a good evaluation design in practice happens under significant uncertainty, and disagreement, around how to anticipate information gain, ethical harm, costs, beyond what constitutes these quantities in the first place.

According to our economic analogy, there is no politically agreed upon optimal social welfare function for aggregating utility across different individuals' preferences. The existence of subjectivity and ethical value judgements are broadly agreed in the economic literature to be inevitable in scientific analysis. Facing this difficulty, analysts typically proceed in the exercise of examining the consequences of various valuation judgments (Samuelson et al. 1983). The decision-makers in a machine learning evaluation practice must also reflect on a range of consequences prior to making final decisions, with the goal of reconciling as much as possible the impacts across a combination of concerns. By discussing the consequences of real-life scenarios where value judgements were problematic and mitigations from analogous domains, accompanied by questions for the evaluation industry to use while reflecting on their options, our conceptualization aims to prompt recognition of complex and nuanced values that arise in evaluation decisions.

The status quo approach to evaluation in research prioritizes sharing the results of evaluations. The pervasive sharing of code, data, and results has been called "frictionless reproducibility" (Donoho 2024) and used to explain the recent success of machine learning in the world, but a downside is prioritizing the results of evaluations—specifically, the production of point estimates of performance—over richer detail about the evaluation process and how it was selected. Our work highlights how evaluation choices implicate tradeoffs between information gains and potential ethical harms under uncertainty, an under-recognized issue in machine learning development.

Recommendations From the Model

Our discussion suggests two broad directions that the software industry could take toward improving decision-making

around evaluation trade-offs. First, issues 1 through 3 are likely to benefit from developing external review systems, similar to recommendations made for machine learning auditing. Best practices for external audits recommended by Raji et al. (2020) and Raji et al. (2022) include external oversight boards with data access, accreditation for auditors, and registries of ongoing audits. We echo these recommendations and encourage the evaluation industry to similarly move towards exploring effective external review boards. This would be most useful when considering an evaluation that impacts individuals outside of controlled lab experiments, when participants have not consented to participate in the evaluation. In ML evaluation, concerns regarding impact on non-consenting study participants are typically recorded internally or audited externally ex-post the evaluation practice. Instead, we support the community taking a more proactive stance and moving toward designing third-party review boards to plan evaluation practices.

An independent regulatory body that develops comprehensive risk-assessment frameworks for AI technology could also be beneficial, if these frameworks are able to enforce and capture the potential consequences of AI systems in diverse, unpredictable environments. For instance, the European Union's AI Act introduces a risk-based approach to AI regulation, which could inform the development of ethical risk assessment frameworks for AI evaluation (Veale and Zuiderveen Borgesius 2021; Novelli et al. 2024).

Secondly, we believe that internal decision-making can benefit from further reflection and resources in order to alleviate potential issues 4 through 6. Teams may need to be incentivized to focus more deeply on potential downstream harms (issue 4), adjust their resource allocations toward ethical practices (issue 5), and focus on selecting evaluations that are actionable, linking evaluation outcomes to specific improvement strategies (issue 6).

Resource constraints are relevant because they can impact the ability to refine evaluation techniques to ensure minimal likelihood of ethical harm. Incentives for private companies need to align toward greater investment in ethical internal evaluations. Reviews of internal AI system audits reveal that when recording concerns, internal stakeholders often prioritize regulators' or customers' issues over those of impacted communities, especially if these populations are systematically neglected or underrepresented. This is potentially due to conflicts of interest or efforts to reduce liability risks (Raji et al. 2022). It could be analogously possible that evaluations are also prioritizing regulators or customers. Our recommendation is to promote legal or social incentives that encourage corporations to invest in ethical evaluation practices.

Developing a system of governance that dictates approval for evaluations would require working with a wide range of stakeholders beyond practitioners, including legal experts, regulators, and various institutes that currently engage in AI policy. Future work could use case studies to carefully detail evidence of how specific evaluation missteps could have been prevented, and explore options for investigation prior to undergoing the evaluation processes. Practitioners and regulators could be interviewed or surveyed to understand specific weaknesses in their valuation of ethical values. Taken together, further research can allow the ML ethics community to move towards better-planned evaluations.

Alternative Conceptualizations

Our conceptualization of ethical evaluation selection is just one possible framing among many. While we chose it as the most versatile in that it takes as input predicted values of the terms rather than binary information about whether certain thresholds are passed, evaluations in practice may sometimes be better described by alternatives.

For example, an alternative conceptualization is to weigh cost explicitly against the other terms. Then, the choice of evaluation is limited to selection within a set of options that are not expected to exceed some maximum allowable cost; i.e., choose $a \in A_c$ where $A_c \subset A$ and for all $a \in A_c, cost(a) \leq max(budget)$. Another framing is concerned with ensuring that expected harm is below some threshold, t_e, denoted $\mathbb{E}(EH(a)) \leq t_e$. This approach, which corresponds to a "checklist" of potential ethical implications, corresponds to the approach some AI ethicists observe in industry, albeit with mixed feelings on the formalization of ethics in this way (Ali et al. 2023).

Our conceptualization emphasizes that evaluation designs are selected under significant uncertainty about the potential value of the information gained and ethical harms and other costs incurred. One issue that arises in practice is a "cold start" problem, where prior to running any evaluation, a team may feel unprepared to estimate the relevant terms. Addressing the dynamic aspect of evaluation decisions, where some initial evaluation is designed under low information, then subsequent evaluations designed as follow-up conditional on the results, is likely to be important in practice. When no model evaluation has yet been run, teams may benefit from considering similar models, if available, from other applications or described in the research literature. When choosing subsequent evaluations, teams should weigh the expected information gain against the current knowledge state.

Conclusion

We have discussed potential ethical harms due to AI systems that occur due to decisions made in the evaluation process. To separate and categorize various issues in evaluations, we conceptualize the decision problem faced by practitioners when selecting an evaluation. Our conceptualization frames a primary trade-off between the value of information gained in evaluation and the ethical harms and costs of evaluation incurred. We reference best practices for effective evaluations in analogous domains, as well as recommendations made by the machine learning audit research community, to discuss interventions that could improve ethics of evaluations, such as external reviews or devoting additional resources. Our work contributes to the conversation about the need for the machine learning ethics community to focus on deliberately designing evaluations in the development lifecycle to prevent harm from machine learning systems.

Acknowledgements

We thank the reviewers for their feedback on the paper. We also thank Neel Guha at Stanford University and Yannis Katsis at IBM Research for inputs on the model and examples from analogous domains. This work is supported by the National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award #2229873.

References

Ackerman, E. 2017. Autonomous Vehicles vs. Kangaroos: the Long Furry Tail of Unlikely Events Self-driving cars in Australia are preparing to handle kangaroos, but what about autonomous cars everywhere else? https://spectrum.ieee.org/autonomous-cars-vs-kangaroos-the-long-furry-tail-of-unlikely-events.

Albert, K.; Delano, M.; Penney, J.; Rigot, A.; and Kumar, R. S. S. 2020. Ethical Testing in the Real World: Evaluating Physical Testing of Adversarial Machine Learning. arXiv:2012.02048.

Ali, S. J.; Christin, A.; Smart, A.; and Katila, R. 2023. Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 217–226.

Automology. 2021. Why Good Crash Test Dummies Cost Over US$1 Million Each. https://www.automology.com/why-a-good-crash-test-dummies-cost-over-us1-million-each/.

Banerjee, S. S.; Jha, S.; Cyriac, J.; Kalbarczyk, Z. T.; and Iyer, R. K. 2018. s. In *2018 48th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 586–597.

Barbosa, N. M.; and Chen, M. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12.

Bibbins-Domingo, K.; Helman, A.; et al. 2022. Improving Representativeness in Clinical Trials and Research: Facilitators to Recruitment and Retention of Underrepresented Groups. In *Improving Representation in Clinical Trials and Research: Building Research Equity for Women and Underrepresented Groups*. National Academies Press (US).

Biden, J. R. 2023. Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.

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

Blaschke, W.; Jones, M. T.; Majnoni, G.; and Peria, S. M. 2001. Stress Testing of Financial Systems: An Overview of Issues, Methodologies, and FSAP Experiences.

Card, D.; and Smith, N. A. 2020. On consequentialism and fairness. *Frontiers in Artificial Intelligence*, 3: 34.

Castro Fernandez, R. 2023. Data-sharing markets: model, protocol, and algorithms to incentivize the formation of data-sharing consortia. *Proceedings of the ACM on Management of Data*, 1(2): 1–25.

Chen, V.; Wu, S.; Ratner, A. J.; Weng, J.; and Ré, C. 2019. Slice-based learning: A programming model for residual learning in critical data slices. *Advances in neural information processing systems*, 32.

Chohlas-Wood, A.; Coots, M.; Zhu, H.; Brunskill, E.; and Goel, S. 2021. Learning to be fair: A consequentialist approach to equitable decision-making. *arXiv preprint arXiv:2109.08792*.

Choi, B. H. 2019. Crashworthy code. *Wash. L. Rev.*, 94: 39.

Clark, L. T.; Watkins, L.; Piña, I. L.; Elmer, M.; Akinboboye, O.; Gorham, M.; Jamerson, B.; McCullough, C.; Pierre, C.; Polis, A. B.; et al. 2019. Increasing diversity in clinical trials: overcoming critical barriers. *Current problems in cardiology*, 44(5): 148–172.

Claybrook, J.; and Kildare, S. 2018. Autonomous vehicles: No driver... no regulation? *Science*, 361(6397): 36–37.

Corbett-Davies, S.; and Goel, S. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S.; Pierson, E.; Feller, A.; Goel, S.; and Huq, A. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 797–806.

Coston, A.; Rambachan, A.; and Chouldechova, A. 2021. Characterizing fairness over the set of good models under selective labels. In *International Conference on Machine Learning*, 2144–2155. PMLR.

DMV. 2022. Autonomous Vehicle Testing Permit Holders. https://www.dmv.ca.gov/portal/vehicle-industry-services/autonomous-vehicles/autonomous-vehicle-testing-permit-holders/.

D'Onfro, J. 2018. Waymo has been testing self-driving cars with 400 riders in Phoenix for a year. Here's what it's learned so far. https://www.cnbc.com/2018/06/13/alphabet-waymo-testing-early-riders-interview-with-saswat-panigrahi.html.

Donoho, D. 2024. Data science at the singularity. *Harvard Data Science Review*, 6(1).

Dustin, E.; Garrett, T.; and Gauf, B. 2009. *Implementing automated software testing: How to save time and lower costs while raising quality*. Pearson Education.

Engber, D. 2006. Baby, You Can Crash My Car. https://slate.com/news-and-politics/2006/08/how-do-automotive-crash-tests-work.html.

Federal Reserve Board. 2024. Comprehensive Capital Analysis and Review.

Ferris, R. 2022. Why companies spend millions on crash test dummies. https://www.cnbc.com/2022/03/19/why-companies-spend-millions-on-crash-test-dummies.html.

Food, U.; and Administration, D. 1998. Institutional Review Boards Frequently Asked Questions.

https://www.fda.gov/regulatory-information/search-fda-guidance-documents/institutional-review-boards-frequently-asked-questions.

Fox, L. 2023. No, Congress should not expand the use of categorical exclusions. *The Hill*.

Friedman, M. 1953. The methodology of positive economics.

Gelman, A.; Hullman, J.; and Kennedy, L. 2023. Causal quartets: Different ways to attain the same average treatment effect. *The American Statistician*, 1–6.

Ghorbani, A.; and Zou, J. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, 2242–2251. PMLR.

Goel, K.; Rajani, N.; Vig, J.; Tan, S.; Wu, J.; Zheng, S.; Xiong, C.; Bansal, M.; and Ré, C. 2021. Robustness gym: Unifying the nlp evaluation landscape. *arXiv preprint arXiv:2101.04840*.

Goel, N.; and Faltings, B. 2019. Crowdsourcing with fairness, diversity and budget constraints. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 297–304.

Google. 2023. Adversarial Testing for Generative AI. https://developers.google.com/machine-learning/resources/adv-testing.

Hall, H. 2015. GM engineer: Today's crash test dummies cost up to $500K, saving more lives. https://engineering.vanderbilt.edu/news/2015/gm-engineer-todays-crash-test-dummies-cost-up-to-500k-saving-more-lives/.

Hopkins, A.; and Booth, S. 2021. Machine learning practices outside big tech: How resource constraints challenge responsible development. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 134–145.

Hutchinson, B.; Rostamzadeh, N.; Greer, C.; Heller, K.; and Prabhakaran, V. 2022. Evaluation gaps in machine learning practice. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1859–1876.

Jiang, S.; Martin, J.; and Wilson, C. 2019. Who's the Guinea Pig? Investigating Online A/B/n Tests in-the-Wild. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 201–210.

Jo, E. S.; and Gebru, T. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 306–316.

Jobin, A.; Ienca, M.; and Vayena, E. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9): 389–399.

Karkkainen, B. C. 2002. Toward a smarter NEPA: monitoring and managing government's environmental performance. *Columbia Law Review*, 903–972.

Kramer, A. D.; Guillory, J. E.; and Hancock, J. T. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National academy of Sciences of the United States of America*, 111(24): 8788.

Krasnova, H.; Wenninger, H.; Widjaja, T.; and Buxmann, P. 2013. Envy on Facebook: a hidden threat to users' life satisfaction?

Lacoste, A.; Luccioni, A.; Schmidt, V.; and Dandres, T. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Lakkaraju, H.; Aguiar, E.; Shan, C.; Miller, D.; Bhanpuri, N.; Ghani, R.; and Addison, K. L. 2015. A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 1909–1918.

Lee, P. 2016. Learning from Tay's introduction. https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

Liao, Y.-H.; Kar, A.; and Fidler, S. 2021. Towards good practices for efficiently annotating large-scale image classification datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4350–4359.

Liu, B.; Ding, M.; Shaham, S.; Rahayu, W.; Farokhi, F.; and Lin, Z. 2021. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2): 1–36.

Liu, L. T.; Barocas, S.; Kleinberg, J.; and Levy, K. 2024. On the actionability of outcome prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 22240–22249.

Liu, L. T.; Wang, S.; Britton, T.; and Abebe, R. 2023. Reimagining the machine learning life cycle to improve educational outcomes of students. *Proceedings of the National Academy of Sciences*, 120(9): e2204781120.

Martínez-Fernández, S.; Bogner, J.; Franch, X.; Oriol, M.; Siebert, J.; Trendowicz, A.; Vollmer, A. M.; and Wagner, S. 2022. Software engineering for AI-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 31(2): 1–59.

Marx, C.; Calmon, F.; and Ustun, B. 2020. Predictive multiplicity in classification. In *International Conference on Machine Learning*, 6765–6774. PMLR.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6): 1–35.

Meyer, M. N. 2015. Two cheers for corporate experimentation: The A/B illusion and the virtues of data-driven innovation. *Colo. Tech. LJ*, 13: 273.

Meyer, M. N.; Heck, P. R.; Holtzman, G. S.; Anderson, S. M.; Cai, W.; Watts, D. J.; and Chabris, C. F. 2019. Objecting to experiments that compare two unobjectionable policies or treatments. *Proceedings of the National Academy of Sciences*, 116(22): 10723–10728.

Meyersohn, N. 2023. How liberals unintentionall made it harder to flight climate change. *CNN*.

Miller, F. G.; and Brody, H. 2002. What makes placebo-controlled trials unethical? *The American Journal of Bioethics*, 2(2): 3–9.

Millum, J.; and Grady, C. 2013. The ethics of placebo-controlled trials: methodological justifications. *Contemporary clinical trials*, 36(2): 510–514.

Nayak, M.; Laing, K.; and Hull, D. 2022. A Trial for Tesla's Autopilot.

Nidhi Kalra, S. M. P. 2016. Driving to Safety. Technical report, RAND Corporation.

Novelli, C.; Casolari, F.; Rotolo, A.; Taddeo, M.; and Floridi, L. 2024. AI Risk Assessment: A Scenario-Based, proportional methodology for the AI act. *Digital Society*, 3(1): 13.

on Artificial Intelligence, H.-L. E. G. 2019. ETHICS GUIDELINES FOR TRUSTWORTHY AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf.

on Environmental Quality Executive Office of the President, C. 2021. A Citizen's Guide to NEPA. https://ceq.doe.gov/docs/get-involved/citizens-guide-to-nepa-2021.pdf.

Perdomo, J. C. 2023. The Relative Value of Prediction in Algorithmic Decision Making. *arXiv preprint arXiv:2312.08511*.

Polonioli, A.; Ghioni, R.; Greco, C.; Juneja, P.; Tagliabue, J.; Watson, D.; and Floridi, L. 2023. The Ethics of Online Controlled Experiments (A/B Testing). *Minds and Machines*, 1–27.

Pujol, D.; and Machanavajjhala, A. 2021. Equity and privacy: More than just a tradeoff. *IEEE Security & Privacy*, 19(6): 93–97.

Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, E. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 145–151.

Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 557–571.

Riccio, V.; Jahangirova, G.; Stocco, A.; Humbatova, N.; Weiss, M.; and Tonella, P. 2020. Testing machine learning based systems: a systematic mapping. *Empirical Software Engineering*, 25: 5193–5254.

Samuelson, P. A.; Samuelson, P. A.; Samuelson, P. A.; Economiste, E.-U.; and Samuelson, P. A. 1983. *Foundations of economic analysis*, volume 197. Harvard university press Cambridge, MA.

Savage, L. J. 1972. *The foundations of statistics*. Courier Corporation.

Shevlane, T.; Farquhar, S.; Garfinkel, B.; Phuong, M.; Whittlestone, J.; Leung, J.; Kokotajlo, D.; Marchal, N.; Anderljung, M.; Kolt, N.; et al. 2023. Model evaluation for extreme risks. *arXiv preprint arXiv:2305.15324*.

Siddiqui, F.; Lerman, R.; and Merrill, J. B. 2022. Teslas running Autopilot involved in 273 crashes reported since last year. https://www.washingtonpost.com/technology/2022/06/15/tesla-autopilot-crashes/.

Sim, R. H. L.; Zhang, Y.; Chan, M. C.; and Low, B. K. H. 2020. Collaborative machine learning with incentive-aware model rewards. In *International conference on machine learning*, 8927–8936. PMLR.

Song, Q.; Borg, M.; Engström, E.; Ardö, H.; and Rico, S. 2022. Exploring ML testing in practice: Lessons learned from an interactive rapid review with axis communications. In *Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI*, 10–21.

Steele, K.; and Stefánsson, H. O. 2015. Decision theory.

Tjeng, V.; Xiao, K.; and Tedrake, R. 2017. Evaluating robustness of neural networks with mixed integer programming. *arXiv preprint arXiv:1711.07356*.

U.S. Nuclear Regulatory Commission. 2024. Probabilistic Risk Assessment (PRA). NRC Fact Sheet. Available at https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/probabilistic-risk-asses.html.

Veale, M.; and Zuiderveen Borgesius, F. 2021. Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, 22(4): 97–112.

Von Neumann, J.; and Morgenstern, O. 2007. *Theory of games and economic behavior (60th Anniversary Commemorative Edition)*. Princeton university press.

Wang, J.; Zhang, L.; Huang, Y.; and Zhao, J. 2020. Safety of autonomous vehicles. *Journal of advanced transportation*, 2020.

Wolf, M. J.; Miller, K.; and Grodzinsky, F. S. 2017. Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, 47(3): 54–64.

Wu, E.; Wu, K.; Daneshjou, R.; Ouyang, D.; Ho, D. E.; and Zou, J. 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine*, 27(4): 582–584.

Zhang, J. M.; Harman, M.; Ma, L.; and Liu, Y. 2020. Machine learning testing: Survey, landscapes and horizons. *IEEE Transactions on Software Engineering*.