The Fallacy of AI Functionality

Inioluwa Deborah Raji*
University of California, Berkeley
Berkeley, CA, USA
rajiinio@berkeley.edu

I. Elizabeth Kumar*
Brown University
Providence, RI, USA
iekumar@brown.edu

Aaron Horowitz
American Civil Liberties Union
New York City, NY, USA
ahorowitz@aclu.org

Andrew D. Selbst
University of California, Los Angeles
Los Angeles, CA, USA
aselbst@law.ucla.edu

ABSTRACT

Deployed AI systems often do not work. They can be constructed haphazardly, deployed indiscriminately, and promoted deceptively. However, despite this reality, scholars, the press, and policymakers pay too little attention to functionality. This leads to technical and policy solutions focused on "ethical" or value-aligned deployments, often skipping over the prior question of whether a given system functions, or provides any benefits at all. To describe the harms of various types of functionality failures, we analyze a set of case studies to create a taxonomy of known AI functionality issues. We then point to policy and organizational responses that are often overlooked and become more readily available once functionality is drawn into focus. We argue that functionality is a meaningful AI policy challenge, operating as a necessary first step towards protecting affected communities from algorithmic harm.

CCS CONCEPTS

• **Computing methodologies → Machine learning**; • **Applied computing → Law, social and behavioral sciences**.

ACM Reference Format:
Inioluwa Deborah Raji*, I. Elizabeth Kumar*, Aaron Horowitz, and Andrew D. Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3531146.3533158

1 INTRODUCTION

As one of over 20,000 cases falsely flagged for unemployment benefit fraud by Michigan's MIDAS algorithm [34], Brian Russell had to file for bankruptcy, undermining his ability to provide for his two young children. The state finally cleared him of the false charges two years later [51]. RealPage, one of several automated tenant screening tools producing "cheap and fast—but not necessarily accurate—reports for an estimated nine out of 10 landlords across the country", flagged Davone Jackson with a false arrest record, pushing him out of low income housing and into a small motel room with his 9-year-old daughter for nearly a year [107, 108]. Josiah Elleston-Burrell had his post-secondary admissions potentially revoked [106, 113], Robert Williams was wrongfully arrested for a false facial recognition match [90], Tammy Dobbs lost critical access to healthcare benefits [116]. The repercussions of AI-related functionality failures in high stakes scenarios cannot be overstated, and the impact reverberates in real lives for weeks, months and even years.

Despite the current public fervor over the great potential of AI, many deployed algorithmic products do not work. AI-enabled moderation tools regularly flag safe content [80, 109, 139], teacher assessment tools mark star instructors to be fired [140, 159], hospital bed assignment algorithms prioritize healthy over sick patients [133], and medical insurance service distribution and pricing systems gatekeep necessary care-taking resources [116, 159]. Deployed AI-enabled clinical support tools misallocate prescriptions [182], misread medical images [66, 132], and misdiagnose [180, 203]. The New York MTA's pilot of facial recognition had a reported 100% error rate, yet the program moved forward anyway [21]. Some of these failures have already proven to disproportionately impact some more than others: moderation tool glitches target minoritized groups [45]; facial recognition tools fail on darker skinned female faces [31]; a hospital resource allocation algorithm's misjudgements will mostly impact Black and lower income patients [133]. However, all failures in sum reveal a broader pattern of a market saturated with dysfunctional, deployed AI products.

Importantly, the hype is not limited to AI's boosters in corporations and the technology press; scholars and policymakers often assume functionality while discussing the dangers of algorithmic systems as well. In fact, many of the current critiques, policy positions and interventions in algorithmic accountability implicitly begin from the premise that such deployed algorithmic systems work, echoing narratives of super-human ability [62], broad applicability [149], and consistency [145], espoused in corporate marketing materials, academic research papers and in mainstream media. These proposals thus often fall short of acknowledging the functionality issues in AI deployments and the role of the lack of functional safety in contributing to the harm perpetuated by these systems.

If a product works, we can weigh its costs and benefits. But if the product does *not* work, the judgment is no longer a matter of pros and cons, but a much simpler calculation, exposing that this product does not deserve its spot on the market. Although notions of accuracy and product expectations are stakeholder-dependent

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

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533158

and can be contested, the assessment of such claims are often easier to empirically measure, grounding the discussion of harm in a way that is challenging to repudiate.

As an overlooked aspect of AI policy, functionality is often presented as a consideration secondary to other ethical challenges. In this paper, we argue that it is a primary concern that often precedes such problems. We start by calling out what we perceive to be a functionality assumption, prevalent in much of the discourse on AI risks. We then argue that this assumption does not hold in a large set of cases. Drawing on the AI, Algorithmic and Automation Incident and Controversy Repository (AAAIRC), we offer a taxonomy of the ways in which such failures can take form and the harms they cause, which differ from the more commonly cited critiques of AI. We then discuss the existing accountability tools to address functionality issues, that are often overlooked in AI policy literature and in practice, due in large part to this assumption of functionality.

2 RELATED WORK

A review of past work demonstrates that although there is some acknowledgement that AI has a functionality problem, little has been done to systematically discuss the range of problems specifically associated with functionality.

Recent work details that the AI research field suffers from scientific validity and evaluation problems [48, 79]. Kapoor and Narayanan [105] have demonstrated reproducibility failures in published work on predicting civil wars. Liao et al. [118] found that advances in machine learning often "evaporate under closer scrutiny or turn out to be less widely applicable than originally hoped."

There is also some work demonstrating that AI products are challenging to engineer correctly in practice. In a survey of practitioners, Wan et al. [194] describe how developers often modify traditional software engineering practices due to unique challenges presented by ML, such as the increased effort required for testing and defining requirements. They also found that ML practitioners "tend to communicate less frequently with clients" and struggle to make accurate plans for the tasks required in the development process. Sculley et al. [166] have additionally argued that ML systems "have a special capacity for incurring technical debt."

Other papers discuss how the AI label lends itself to inflated claims of functionality that the systems cannot meet. Kaltheuner et al. [102] and Broussard [28] critique hyped narratives pushed in the AI industry, joined by many similar domain-specific critiques [18, 19, 148, 173, 179, 184]. Narayanan [130] recently popularized the metaphor of "snake oil" as a description of such AI products, raising concerns about the hyperbolic claims now common on the market today. Richardson [157] has noted that despite the "intelligent" label, many deployed AI systems used by public agencies involve simple models defined by manually crafted heuristics. Similarly, Raji et al. [149] argue that AI makes claims to generality while modeling behaviour that is determined by highly constrained and context-specific data. In a study of actual AI policy discussions, Krafft et al. [110] found that policymakers often define AI with respect to how human-like a system is, and concluded that this could lead to deprioritizing issues more grounded in reality.

Finally, Vinsel [191] has argued that even critics of technology often hype the very technologies that they critique, as a way of inflating the perception of their dangers. He refers to this phenomenon as "criti-hype"—criticism which both needs and feeds on hype. As an example, he points to disinformation researchers, who embrace corporate talking points of a recommendation model that can meaningfully influence consumer behavior to the point of controlling their purchases or voting activity—when in actuality, these algorithms have little ability to do either [22, 75, 88, 95, 162]. Even the infamous Cambridge Analytica product was revealed to be "barely better than chance at applying the right [personality] scores to individuals", and the company accused explicitly of "selling snake oil" [88].

3 THE FUNCTIONALITY ASSUMPTION

It is unsurprising that promoters of AI do not tend to question its functionality. More surprising is the prevalence of criti-hype in the scholarship and political narratives around automation and machine learning—even amidst discussion of valid concerns such as trustworthiness, democratization, fairness, interpretability, and safety. These fears, though legitimate, are often premature "wishful worries"—fears that can only be realized once the technology works, or works "too well", rather than being grounded in a reality where these systems do not always function as expected [191]. In this section, we discuss how criti-hype in AI manifests as an unspoken assumption of functionality.

The functionality of AI systems is rarely explicitly mentioned in AI principle statements, policy proposals and AI ethics guidelines. In a recent review of the landscape of AI ethics guidelines, Jobin et al. [101] found that few acknowledge the possibility of AI not working as advertised. In guidelines about preventing malfeasance, the primary concern is malicious use of supposedly functional AI products by nefarious actors. Guidelines around "trust" are geared towards eliciting trust in AI systems from users or the public, implying that trusting these AI products would be to the benefit of these stakeholders and allow AI to "fulfill its world changing potential" [101]. Just one guideline of the hundreds reviewed in the survey "explicitly suggests that, instead of demanding understandability, it should be ensured that AI fulfills public expectations" [101]. Similarly, the U.S. National Institute of Standards and Technology (NIST) seeks to define "trustworthiness" based primarily on how much people are willing to use the AI systems they are interacting with [178]. This framing puts the onus on people to trust in systems, and not on institutions to make their systems reliably operational, in order to earn that trust [6, 30]. NIST's concept of trust is also limited, citing the "dependability" section of ISO/IEEE/IEC standards [96], but leaving out other critical concepts in these dependability engineering standards that represent basic functionality requirements, including assurance, claim veracity, integrity level, systematic failure, or dangerous condition. Similarly, the international trade group, the Organisation for Economic Co-operation and Development (OECD), mentions "robustness" and "trustworthy AI" in their AI principles but makes no explicit mention of expectations around basic functionality or performance assessment [207].

The ideal of “democratizing” AI systems, and the resulting AI innovation policy, is another effort premised on the assumed functionality of AI. This is the argument that access to AI tooling and AI skills should be expanded [14, 70, 83, 181]—with the corollary claim that it is problematic that only certain institutions, nations, or individuals have access to the ability to build these systems [8]. A recent example of democratization efforts was the global push for the relaxation of oversight in data sharing in order to allow for more innovation in AI tool development in the wake of the COVID-19 pandemic [11, 49, 122, 137, 196]. The goal of such efforts was to empower a wider range of non-AI domain experts to participate in AI tool development. This policy impact was long lasting and informed later efforts such as the AI National Resource (AINR) effort in the US [43] and the National Medical Imaging Platform (NMIP) executed by National Health Services (NHS) in the UK [112]. In this flurry of expedited activity, some parallel concerns were also raised about how the new COVID-19 AI tools would adequately address cybersecurity, privacy, and anti-discrimination challenges [46, 111], but the functionality and utility of the systems remained untested for some time [85, 97, 161, 206].

An extremely premature set of concerns are those of an autonomous agent becoming so intelligent that humans lose control of the system. While it is not controversial to claim that such concerns are far from being realized [13, 42, 146], this fear of misspecified objectives, runaway feedback loops, and AI alignment presumes the existence of an industry that can get AI systems to execute on any clearly declared objectives, and that the main challenge is to choose and design an appropriate goal. Needless to say, if one thinks the danger of AI is that it will work too well [168], it is a necessary precondition that it works at all.

The fear of hyper-competent AI systems also drives discussions on potential misuse [29]. For example, expressed concerns around large language models centers on hyped narratives of the models' ability to generate hyper-realistic online content, which could theoretically be used by malicious actors to facilitate harmful misinformation campaigns [176, 195]. While these are credible threats, concerns around large language models tend to dismiss the practical limitations of what these models can achieve [18], neglecting to address more mundane hazards tied to the premature deployment of a system that does not work [55, 189]. This pattern is evident in the EU draft AI regulation [9], where, even as the legislation does concern functionality to a degree, the primary concerns—questions of “manipulative systems,” “social scoring,” and “emotional or biometric categorization”—“border on the fantastical” [190, p. 98].

A major policy focus in recent years has been addressing issues of bias and fairness in AI. Fairness research is often centered around attempting to balance some notion of accuracy with some notion of fairness [59, 63, 68]. This research question presumes that an unconstrained solution without fairness restrictions is the optimal solution to the problem. However, this intuition is only valid when certain conditions and assumptions are met [67, 124, 198], such as the measurement validity of the data and labels. Scholarship on fairness also sometimes presumes that unconstrained models will be optimal or at least useful. Barocas and Selbst [15, p. 707] argued that U.S. anti-discrimination law would have difficulty addressing algorithmic bias because the "nature of data mining" means that in many cases we can assume the decision is at least statistically valid. Similarly, as an early example of technical fairness solutions, Feldman et al. [60] created a method to remove disparate impact from a model while preserving rank, which only makes sense if the unconstrained system output is correct in the first place. Industry practitioners then carry this assumption into how they approach fairness in AI deployments. For example, audits of AI hiring tools focus primarily on ensuring an 80% selection rate for protected classes (the so-called 4/5ths rule) is satisfied, and rarely mention product validation processes, demonstrating an assumed validity of the prediction task [52, 148, 199].

Another dominant theme in AI policy developments is that of explainability or interpretability. The purpose of making models explainable or interpretable differs depending on who is seen as needing to understand them. From the engineering side, interpretability is usually desired for debugging purposes [23], so it is focused on functionality. But on the legal or ethical side, things look different. There has been much discussion about whether the GDPR includes a "right to explanation" and what such a right entails [50, 103, 167, 193]. Those rights would serve different purposes. To the extent the purpose of explanation is to enable contestation [104], then functionality is likely included as an aspect of the system subject to challenge. To the extent explanation is desired to educate consumers about how to improve their chances in the future [16], such rights are only useful when the underlying model is functional. Similarly, to the extent regulators are looking into functionality, explanations aimed at regulators can assist oversight, but typically explanations are desired to check the basis for decisions, while assuming the systems work as intended.

Not all recent policy developments hold the functionality assumption strongly. The Food and Drug Administration (FDA) guidelines for AI systems integrated into software as a medical device (SaMD) has a strong emphasis on functional performance, clearly not taking product performance as a given [64]. The draft AI Act in the EU includes requirements for pre-marketing controls to establish products' safety and performance, as well as quality management for high risk systems [190]. These mentions suggest that functionality is not always ignored outright. Sometimes, it is considered in policy, but in many cases, that consideration lacks the emphasis of the other concerns presented.

4 THE MANY DIMENSIONS OF AI DYSFUNCTION

Functionality can be difficult to define precisely. The dictionary definition of “fitness for a product's intended use” [134] is useful, but incomplete, as some intended uses are impossible. Functionality could also be seen as a statement that a product lives up to the vendor's performance claims, but this, too, is incomplete; specifications chosen by the vendor could be insufficient to solve the problem at hand. Another possible definition is "meeting stakeholder expectations" more generally, but this is too broad as it sweeps in wider AI ethics concerns with those of performance or operation.

Lacking a perfectly precise definition of functionality, in this section we invert the question by creating a taxonomy that brings together disparate notions of product failure. Our taxonomy serves several other purposes, as well. Firstly, the sheer number of points of failure we were able to identify illustrates the scope of the problem.

Secondly, we offer language in which to ground future discussions of functionality in research and policy. Finally, we hope that future proposals for interventions can use this framework to concretely illustrate the way any proposed interventions might work to prevent different kinds of failure.

4.1 Methodology

To challenge the functionality assumption and demonstrate the various ways in which AI doesn't work, we developed a taxonomy of known AI failures through the systematic review of case studies. To do this, we partly relied on the AI, Algorithmic and Automation Incident and Controversy Repository (AIAAIC) spreadsheet crowdsourced from journalism professionals [35]. Out of a database of over 800 cases, we filtered the cases down to a spreadsheet of 283 cases from 2012 to 2021 based on whether the technology involved claimed to be AI, ML or data-driven, and whether the harm reported was due to a failure of the technology. In particular, we focused on describing the ways in which the artifact itself was connected to the failure, as opposed to infrastructural or environmental "meta" failures which caused harm through the artifact. We split up the rows in the resulting set and used an iterative tagging procedure to come up with categories that associate each example with a different element or cause of failure. We updated, merged, and grouped our tags in meetings between tagging sessions, resulting in the following taxonomy. We then chose known case studies from the media and academic literature to illustrate and best characterize these failure modes.

4.2 Failure Taxonomy

Here, we present a taxonomy of AI system failures and provide examples of known instances of harm. Many of these cases are direct refutations of the specific instances of the functionality assumptions in Section 3.

Table 1: Failure Taxonomy

Impossible Tasks	Conceptually Impossible Practically Impossible
Engineering Failures	Design Failures Implementation Failures Missing Safety Features
Post-Deployment Failures	Robustness Issues Failure under Adversarial Attacks Unanticipated Interactions
Communication Failures	Falsified or Overstated Capabilities Misrepresented Capabilities

4.2.1 *Impossible Tasks.*

In some situations, a system is not just "broken" in the sense that it needs to be fixed. Researchers across many fields have shown that certain prediction tasks cannot be solved with machine learning. These are settings in which no specific AI developed for the task can ever possibly work, and a functionality-centered critique can be made with respect to the task more generally. Since these general critiques sometimes rely on philosophical, controversial, or morally contested grounds, the arguments can be difficult to leverage practically and may imply the need for further evidence of failure modes along the lines of our other categories.

Conceptually Impossible. Certain classes of tasks have been scientifically or philosophically "debunked" by extensive literature. In these cases, there is no plausible connection between observable data and the proposed target of the prediction task. This includes what Stark and Hutson call "physiognomic artificial intelligence," which attempts to infer or create hierarchies about personal characteristics from data about their physical appearance [179]. Criticizing the EU Act's failure to address this inconvenient truth, Veale and Borgesius [190] pointed out that "those claiming to detect emotion use oversimplified, questionable taxonomies; incorrectly assume universality across cultures and contexts; and risk '[taking] us back to the phrenological past' of analysing character traits from facial structures."

A notorious example of technology broken by definition are attempts to infer "criminality" from a person's physical appearance. A paper claiming to do this "with no racial bias" was announced by researchers at Harrisburg University in 2020, prompting widespread criticism from the machine learning community [69]. In an open letter, the Coalition for Critical Technology note that the only plausible relationship between a person's appearance and their propensity to commit a crime is via the biased nature of the category of "criminality" itself [65]. In this setting, there is no logical basis with which to claim functionality.

Practically Impossible. There can be other, more practical reasons for why a machine learning model or algorithm cannot perform a certain task. For example, in the absence of any reasonable observable characteristics or accessible data to measure the model goals in question, attempts to represent these objectives end up being inappropriate proxies. As a construct validity issue, the constructs of the built model could not possibly meaningfully represent those relevant to the task at hand [98, 99].

Many predictive policing tools are arguably practically impossible AI systems. Predictive policing attempts to predict crime at either the granularity of location or at an individual level [61]. The data that would be required to do the task properly—accurate data about when and where crimes occur—does not and will never exist. While crime is a concept with a fairly fixed definition, it is practically impossible to predict because of structural problems in its collection. The problems with crime data are well-documented—whether in differential victim crime reporting rates [10], selection bias based on policing activities [54, 120], dirty data from periods of recorded unlawful policing [158], and more.

Due to upstream policy, data or societal choices, AI tasks can be practically impossible for one set of developers and not for another, or for different reasons in different contexts. The fragmentation, billing focus, and competing incentives of the US healthcare system have made multiple healthcare-related AI tasks practically impossible [7]. US EHR data is often erroneous, miscoded, fragmented, and incomplete [91, 92], creating a mismatch between available data and intended use [74]. Many of these challenges appeared when IBM attempted to support cancer diagnoses. In one instance, this meant using synthetic as opposed to real patients for oncology prediction data, leading to "unsafe and incorrect" recommendations for cancer treatments [164]. In another, IBM worked with MD Anderson to work on leukemia patient records, poorly extracting reliable insights from time-dependent information like therapy timelines—the components of care most likely to be mixed up in fragmented doctors' notes [171, 180].

4.2.2 Engineering Failures.

4.2.2 Engineering Failures. Algorithm developers maintain enormous discretion over a host of decisions, and make choices throughout the model development lifecycle. These engineering choices include defining problem formulation [141], setting up evaluation criteria [118, 143], and determining a variety of other details [126, 142]. Failures in AI systems can often be traced to these specific policies or decisions in the development process of the system.

Model Design Failures. Sometimes, the design specifications of a model are inappropriate for the task it is being developed for. For instance, in a classification model, choices such as which input and target variables to use, whether to prioritize accepting true positives or rejecting false negatives, and how to process the training data all factor into determining model outcomes. These choices are normative and may prioritize values such as efficiency over preventing harmful failures [47, 117].

In 2014, BBC Panorama uncovered evidence of international students systematically cheating on English language exams run by the UK's Educational Testing Service by having others take the exam for them. The Home Office began an investigation and campaign to cancel the visas of anyone who was found to have cheated. In 2015, ETS used voice recognition technology to identify this type of cheating. According to the National Audit Office [135],

> ETS identified 97% of all UK tests as "suspicious". It classified 58% of 58,459 UK tests as "invalid" and 39% as "questionable". The Home Office did not have the expertise to validate the results nor did it, at this stage, get an expert opinion on the quality of the voice recognition evidence. ... but the Home Office started cancelling visas of those individuals given an "invalid" test.

The staggering number of accusations obviously included a number of false positives. The accuracy of ETS's method was disputed between experts sought by the National Union of Students and the Home Office; the resulting estimates of error rates ranged from 1% to 30%. Yet out of 12,500 people who appealed their immigration decisions, only 3,600 won their cases—and only a fraction of these were won through actually disproving the allegations of cheating. This highly opaque system was thus notable for the disproportionate amount of emphasis that was put into finding cheaters rather than protecting those who were falsely accused. Although we cannot be sure the voice recognition model was trained to optimize for sensitivity rather than specificity, as the head of the NAO aptly put, "When the Home Office acted vigorously to exclude individuals and shut down colleges involved in the English language test cheating scandal, we think they should have taken an equally vigorous approach to protecting those who did not cheat but who were still caught up in the process, however small a proportion they might be" [135]. This is an example of a system that was not designed to prevent a particular type of harmful failure.

Model Implementation Failures. Even if a model was conceptualized in a reasonable way, some component of the system downstream from the original plan can be executed badly, lazily, or wrong. In 2011, the state of Idaho attempted to build an algorithm to set Medicaid assistance limits for individuals with developmental and intellectual disabilities. When individuals reported sudden drastic cuts to their allowances, the ACLU of Idaho tried to find out how the allowances were being calculated, only to be told it was a trade secret. The class action lawsuit that followed resulted in a court-ordered disclosure of the algorithm, which was revealed to have critical flaws. According to Richard Eppink, Legal Director of the ACLU of Idaho [177],

> There were a lot of things wrong with it. First of all, the data they used to come up with their formula for setting people's assistance limits was corrupt. They were using historical data to predict what was going to happen in the future. But they had to throw out two-thirds of the records they had before they came up with the formula because of data entry errors and data that didn't make sense.

Data validation is a critical step in the construction of a ML system, and the team that built the benefit system chose to use a highly problematic dataset to train their model. For this reason, we consider this to be an implementation failure.

Another way that failures can be attributed to poor implementation is when a testing framework was not appropriately implemented. One area in which a lack of sufficient testing has been observed in the development of AI is in the area of clinical medicine. Nagendran et al. [129] systematically examined the methods and claims of studies which compared the performance of diagnostic deep learning computer vision algorithms against that of expert clinicians. In their literature review, they identified 10 randomized clinical trials and 81 non-randomized clinical trials. Of the 81 non-randomized studies, they found the median number of clinical experts compared to the AI was 4, full access to datasets and code were unavailable in over 90% of studies, the overall risk of bias was high, and adherence to reporting standards were suboptimal, and therefore poorly substantiate their claims. Similarly, the Epic sepsis prediction model, a product actually implemented at hundreds of hospitals, was recently externally validated by Wong et al. [203], who found that the model had poor calibration to other hospital settings and discriminated against under-represented demographics. These results suggest that the model's testing prior to deployment may have been insufficient to estimate its real-world performance. Notably, the COVID-19 technology which resulted from innovation policy and democratization efforts mentioned in section 3 was later shown to be completely unsuitable for clinical deployment after the fact [85, 97, 161, 206].

Missing Safety Features. Sometimes model failures are anticipated yet difficult to prevent; in this case, engineers can sometimes take steps to ensure these points of failure will not cause harm. In 2014, a Nest Labs smoke and carbon monoxide detector was recalled [200]. The detector had a feature which allowed the user to turn it off with a "wave" gesture. However, the company discovered in testing that under certain circumstances, the sensor could be unintentionally deactivated. Detecting a wave gesture with complete accuracy is impossible, and Google acknowledges factors that contribute to the possibility of accidental wave triggering for its other home products [1]. However, the lack of a failsafe to make sure the carbon monoxide detector could not be turned off accidentally made the product dangerous.

In the same way, the National Transportation Safety Board (NTSB) cited a lack of adequate safety measures—such as "a warning/alert when the driver's hands are off the steering wheel", "remote monitoring of vehicle operators" and even the companies' "inadequate safety culture"—as the probable causes in at least two highly publicized fatal crashes of Uber [27, 197] and Tesla [25, 26] self-driving cars. As products in public beta-testing, this lack of functional safeguards was considered to be an even more serious operational hazard than any of the engineering failures involved (such as the vehicle's inability to detect an incoming pedestrian [27] or truck [25]).

This category also encompasses algorithmic decision systems in critical settings that lack a functional appeals process. This has been a recurring feature in algorithms which allocate benefits on behalf of the government [56]. Not all of these automated systems rely on machine learning, but many have been plagued by bugs and faulty data, resulting in the denial of critical resources owed to citizens. In the case of the Idaho data-driven benefit allocation system, even the people responsible for reviewing appeals were unable to act as a failsafe for the algorithm's mistakes: "They would look at the system and say, 'It's beyond my authority and my expertise to question the quality of this result' " [115].

4.2.3 Deployment Failures. Sometimes, despite attempts to anticipate failure modes during the design phase, the model does not "fail" until it is exposed to certain external factors and dynamics that arise after it is deployed.

Robustness Issues. A well-documented source of failure is a lack of robustness to changing external conditions. Liao et al. [118] have observed that the benchmarking methods used for evaluation in machine learning can suffer from both internal and external validity problems, where "internal validity refers to issues that arise within the context of a single benchmark" and "external validity asks whether progress on a benchmark transfers to other problems." If a model is developed in a certain context without strong evaluation methods for external validity, it may perform poorly when exposed to real-world conditions that were not captured by the original context. For instance, while many computer vision models developed on ImageNet are tested on synthetic image perturbations in an attempt to measure and improve robustness, but Taori et al. [183] have found that these models are not robust to real-world distribution shifts such as a change in lighting or pose.

Robustness issues are also of dangerous consequence in language models. For example, when large language models are used to process the queries of AI-powered web search [131], the models' fragility to misspellings [125, 147], or trivial changes to format [19] and context [18] can lead to unexpected results. In one case, a large language model used in Google search could not adequately handle cases of negation [55] – and so when queried with "what to do when having a seizure", the model alarmingly sourced the information for what *not* to do, unable to differentiate between the two cases [189].

Failure under Adversarial Attacks. Failures can also be induced by the actions of an adversary—an actor deliberately trying to make the model fail. Real-world examples of this often appear in the context of facial recognition, in which adversaries have some evidence that they can fool face-detection systems with, such as 3d-printed masks [144] or software-generated makeup [78]. Machine learning researchers have studied what they call "adversarial examples," or inputs that are designed to make a machine learning model fail [76]. However, some of this research has been criticized by its lack of a believable threat model— in other words, not focusing on what real-world "adversaries" are actually likely to do [136].

Unanticipated Interactions. A model can also fail to account for uses or interactions that it was not initially conceived to handle. Even if an external actor or user is not deliberately trying to break a model, their actions may induce failure if they interact with the model in a way that was not planned for by the model's designers. For instance, there is evidence that this happened at the Las Vegas Police Department:

> As new records about one popular police facial recognition system show, the quality of the probe image dramatically affects the likelihood that the system will return probable matches. But that doesn't mean police don't use bad pictures anyway. According to documents obtained by Motherboard, the Las Vegas Metropolitan Police Department (LVMPD) used "non-suitable" probe images in almost half of all the facial recognition searches it made last year, greatly increasing the chances the system would falsely identify suspects, facial recognition researchers said. [57]

This aligns with reports from Garvie [71] about other police departments inappropriately uploading sketch and celebrity photos to facial recognition tools. It is possible for designers to preempt misuse by implementing instructions, warnings, or error conditions, and failure to do so creates a system that does not function properly.

4.2.4 Communication Failures. As with other areas of software development, roles in AI development and deployment are becoming more specialized. Some roles focus on managing the data that feeds into models, others specialize in modeling, and others optimally engineer models for speed and scale [44]. There are even those in "analytics translator" roles – managers dedicated to acting as communicators between data science work and non-technical business leaders [86]. And, of course, there are salespeople. Throughout this chain of actors, potential miscommunications or outright lies can happen about the performance, functional safety or other aspects of deployed AI/ML systems. Communication failures often co-occur with other functional safety problems, and the lack of accountability for false claims – intentional or otherwise – makes these particularly pernicious and likely to occur as AI hype continues absent effective regulation.

Falsified or Overstated Capabilities. To pursue commercial or reputational interests, companies and researchers may explicitly make claims about models which are provably untrue. A common form of this are claims that a product is "AI", when in fact it mainly involves humans making decisions behind the scenes. While this in and of itself may not create unsafe products, expectations based on unreasonable claims can create unearned trust, and a potential overreliance that hurts parties who purchase the product. As an example, investors poured money into ScaleFactor, a startup that claimed to have AI that could replace accountants for small businesses, with the exciting (for accountants) tagline "Because evenings are for families, not finance" [100]. Under the hood, however,

> Instead of software producing financial statements, dozens of accountants did most of it manually from ScaleFactor's Austin headquarters or from an outsourcing office in the Philippines, according to former employees. Some customers say they received books filled with errors, and were forced to re-hire accountants, or clean up the mess themselves. [100]

Even large well-funded entities misrepresent the capabilities of their AI products. Deceptively constructed evaluation schemes allow AI product creators to make false claims. In 2018, Microsoft created machine translation with "equal accuracy to humans in Chinese to English translations" [186]. However, the study used to make this claim (still prominently displayed in press release materials) was quickly debunked by a series of outside researchers who found that at the document-level, when provided with context from nearby sentences, and/or compared to human experts, the machine translation model did not indeed achieve equal accuracy to human translators [114, 185]. This follows a pattern seen with machine learning products in general, where the advertised performance on a simple and static data benchmark, is much lower than the performance on the often more complex and diverse data encountered in practice.

Misrepresented Capabilities. A simple way to deceive customers into using prediction services is to sell the product for a purpose you know it can't reliably be used for. In 2018, the ACLU of Northern California revealed that Amazon effectively misrepresented capabilities to police departments in selling their facial recognition product, Rekognition. Building on previous work [31], the ACLU ran Rekognition with a database of mugshots against members of U.S. Congress using the default setting and found 28 members falsely matched within the database, with people of color shown as a disproportionate share of these errors [175]. This result was echoed by Raji and Buolamwini [150] months later. Amazon responded by claiming that for police use cases, the threshold for the service should be set at either 95% or 99% confidence [204]. However, based on a detailed timeline of events [5], it is clear that in selling the service through blog posts and other campaigns that thresholds were set at 80% or 85% confidence, as the ACLU had used in its investigation. In fact, suggestions to shift that threshold were buried in manuals end-users did not read or use – even when working in partnership with Amazon. At least one of Amazon's police clients also claimed being unaware of needing to modify the default threshold [123].

The hype surrounding IBM's Watson in healthcare represents another example where a product that may have been fully capable of performing *specific* helpful tasks was sold as a panacea to health care's ills. As discussed earlier, this is partially the result of functional failures like practical impossibility – but these failures were coupled with deceptively exaggerated claims. The backlash to this hype has been swift in recent years, with one venture capitalist claiming "I think what IBM is excellent at is using their sales and marketing infrastructure to convince people who have asymmetrically less knowledge to pay for something" [202]. At Memorial-Sloan Kettering, after $62 million dollars spent and may years of effort, MD Anderson famously cancelled IBM Watson contracts with no results to show for it [89].

This is particularly a problem in the context of algorithms developed by public agencies – where the AI systems can be adopted as symbols for progress, or smokescreens for undesirable policy outcomes, and thus liable to inflated narratives of performance. Green [77] discusses how the celebrated success of "self-driving shuttles" in Columbus, Ohio omits its marked failure in the lower-income Linden neighborhood, where residents were now locked out of the transportation apps due to a lack of access to a bank account, credit cards, a data plan or Wi-Fi. Similarly, Eubanks [56] demonstrates how a $1.4 billion contract with a coalition of high-tech companies led an Indiana governor to stubbornly continue a welfare automation algorithm that resulted in a 54% increase in the denials of welfare applications.

5 DEALING WITH DYSFUNCTION: OPPORTUNITIES FOR INTERVENTION ON FUNCTIONAL SAFETY

The challenge of dealing with an influx of fraudulent or dysfunctional products is one that has plagued many industries, including food safety [24], medicine [12, 17], financial modeling [170], civil aviation [87] and the automobile industry [128, 192]. In many cases, it required the active advocacy of concerned citizens to lead to the policy interventions that would effectively change the tide of these industries. The AI field seems to now be facing this same challenge.

Thankfully, as AI operates as a general purpose technology prevalent in many of these industries, there already exists a plethora of governance infrastructure to address this issue in related fields of application. In fact, healthcare is the field where AI product failures appear to be the most visible, in part due to the rigor of pre-established evaluation processes [20, 119, 160, 205]. Similarly, the transportation industry has a rich history of thorough accident reports and investigations, through organizations such as the National Transportation and Safety Board (NTSB), who have already been responsible for assessing the damage from the few known cases of self-driving car crashes from Uber and Tesla [81].

In this section, we specifically outline the legal and organizational interventions necessary to address functionality issues in general context in which AI is developed and deployed into the market. In broader terms, the concept of *functional safety* in engineering design literature [163, 174] well encapsulates the concerns articulated in this paper—namely that a system can be deployed without working very well, and that such performance issues can cause harm worth preventing.

5.1 Legal/Policy Interventions

The law has several tools at its disposal to address product failures to work correctly. They mostly fall in the category of consumer protection law. This discussion will be U.S.-based, but analogues exist in most jurisdictions.

5.1.1 *Consumer Protection.*

The Federal Trade Commission is the federal consumer protection agency within the United States with the broadest subject matter jurisdiction. Under Section 5 of the FTC Act, it has the authority to regulate "unfair and deceptive acts or practices" in commerce [58]. This is a broad grant authority to regulate practices that injure consumers. The authority to regulate deceptive practices applies to any material misleading claims relating to a consumer product. The FTC need not show intent to deceive or that deception actually occurred, only that claims are misleading. Deceptive claims can be expressed explicitly—for example, representation in the sales materials that is inaccurate—or implied, such as an aspect of the design that suggests a functionality the product lacks [82, 93]. Many of the different failures, especially impossibility, can trigger a deceptive practices claim.

The FTC's ability to address unfair practices is wider-ranging but more controversial. The FTC can reach any practice "likely to cause substantial injury to consumers[,] not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers" [58]. Thus, where dysfunctional AI is being sold and its failures causes substantial harm to consumers, the FTC could step in. Based on the FTC's approach to data security, in which the Commission has sued companies for failing to adequately secure consumer data in their possession against unknown third-party attackers [121], even post-deployment failures—if foreseeable and harmful—can be included among unfair practices, though they partially attributable to external actors.

The FTC can use this authority to seek an injunction, requiring companies to cease the practice. Formally, the FTC does not have the power to issue fines under its Section 5 authority, but the Commission frequently enters into long-term consent decrees with companies that it sues, permitting continuing jurisdiction, monitoring, and fines for future violations [3, 40]. The Commission does not have general rulemaking authority, so most of its actions to date have taken the form of public education and enforcement. The Commission does, however, have authority to make rules regarding unfair or deceptive practices under the Magnuson-Moss Warranty Act. Though it has created no new rules since 1980, in July 2021, the FTC voted to change internal agency policies to make it easier to do so [41].

Other federal agencies also have the ability to regulate faulty AI systems, depending on their subject matter. The Consumer Product Safety Commission governs the risks of physical injury due to consumer products. They can create mandatory standards for products, can require certifications of adherence to those rules, and can investigate products that have caused harm, leading to bans or mandatory recalls [39]. The National Highway Safety Administration offers similar oversight for automobiles specifically. The Consumer Finance Protection Bureau can regulate harms from products dealing with loans, banking, or other consumer finance issues [4].

In addition to various federal agencies, all states have consumer protection statutes that bar deceptive practices and many bar unfair practices as well, like the FTC Act [33]. False advertising laws are related and also common. State attorneys general often take active roles as enforcers of those laws [38]. Of course, the efficacy of such laws varies from state to state, but in principle, they become another source of law and enforcement to look to for the same reasons that the FTC can regulate under Section 5. One particular state law worth noting is California's Unfair Competition Law, which allows individuals to sue for injunctive relief to halt conduct that violates other laws, even if individuals could not otherwise sue under that law [2].

It is certainly no great revelation that federal and state regulatory apparatuses exist. Rather, our point is that while concerns about discrimination and due process can lead to difficult questions about the operation of existing law and proposals for legal reform, thinking about the ways that AI is *not working* makes it look like other product failures that we know how to address. Where AI doesn't work, suddenly regulatory authority is easy to find.

5.1.2 *Products Liability Law.*

Another avenue for legal accountability may come from the tort of products liability, though there are some potential hurdles. In general, if a person is injured by a defective product, they can sue the producer or seller in products liability. The plaintiff need not have purchased or used the product; it is enough that they were injured by it, and the product has a defect that rendered it unsafe.

It would stand to reason that a functionality failure in an AI system could be deemed a product defect. But surprisingly, defective software has never led to a products liability verdict. One commonly cited reason is that products liability applies most clearly to tangible things, rather than information products, and that aside from a stray comment in one appellate case [201], no court has actually ruled that software is even a "product" for these purposes [32, 53]. This would likely not be a problem for software that resides within a physical system, but for non-embodied AI, it might pose a hurdle. In a similar vein, because most software harms have typically been economic in nature, with, for example, a software crash leading to a loss of work product, courts have rejected these claims as "pure economic loss" belonging more properly in contract law than tort. But these mostly reflect courts' anxiety with intangible *injuries*, and as AI discourse has come to recognize many concrete harms, these concerns are less likely to be hurdles going forward [36].

Writing about software and tort law, Choi [36] identifies the complexity of software as a more fundamental type of hurdle. For software of nontrivial complexity, it is provably impossible to guarantee bug-free code. An important part of products liability is weighing the cost of improvements and more testing against the harms. But as no amount of testing can guarantee bug-free software, it will difficult to determine how much testing is enough to be considered reasonable or non-negligent [36, 94]. Choi analogizes this issue to car crashes: car crashes are inevitable, but courts developed the idea of crashworthiness to ask about the car's contribution to the total harm, even if the initial injury was attributable to a product defect [36]. While Choi looks to crashworthiness as a solution, the thrust of his argument is that software can cause exactly the type of injury that products liability aims to protect us from, and doctrine should reflect that.

While algorithmic systems have a similar sort of problem, the failure we describe here are more basic. Much as writing bug-free software is impossible, creating a model that handles every corner case perfectly is impossible. But the failures we address here are not about unforeseeable corner cases in models. We are concerned with easier questions of basic functionality, without which a system should never have been shipped. If a system is not functional, in the sense we describe, a court should have no problem finding that it is unreasonably defective. As discussed above, a product could be placed on the market claiming the ability to do something it cannot achieve in theory or in practice, or it can fail to be robust to unanticipated but foreseeable uses by consumers. Even where these errors might be difficult to classify in doctrinally rigid categories of defect, courts have increasingly been relying on "malfunction doctrine," which allows for circumstantial evidence to be used as proof of defect where "a product fails to perform its manifestly intended function." [155]. Courts are increasingly relying on this doctrine and it could apply here [73, 138]. Products liability could especially easily apply to engineering failures, where the error was foreseeable and an alternative, working version of the product should have been built.

5.1.3 *Warranties.*

Another area of law implicated by product failure is warranty law, which protects the purchasers of defunct AI and certain third parties who stand to benefit from the sale. Sales of goods typically come with a set of implied warranties. The implied warranty of merchantability applies to all goods and states, among other things, that the good is "fit for the ordinary purposes for which such goods are used" [187]. The implied warranty of fitness for particular purpose applies when a seller knows that the buyer has a specific purpose in mind and the buyer is relying on the seller's skill or judgment about the good's fitness, stating that the good is fit for that purpose[188]. Defunct AI will breach both these warranties. The remedy for such a breach is limited to contract damages. This area of law is concerned with ensuring that purchasers get what they pay for, so compensation will be limited roughly to value of the sale. Injuries not related to the breach of contract are meant to be worked out in tort law, as described above.

5.1.4 *Fraud.*

In extreme cases, the sale of defunct AI may constitute fraud. Fraud has many specific meanings in law, but invariably it involves a knowing or intentional misrepresentation that the victim relied on in good faith. In contract law, proving that a person was defrauded can lead to contract damages. Restitution is another possible remedy for fraud. In tort law, a claim of fraud can lead to compensation necessary to rectify any harms that come from the fraud, as well as punitive damages in egregious cases. Fraud is difficult to prove, and our examples do not clearly indicate fraud, but it is theoretically possible if someone is selling snake oil. Fraud can lead to criminal liability as well.

5.1.5 *Other Legal Avenues Already Being Explored.*

Finally, other areas of law that are already involved in the accountability discussion, such as discrimination and due process, become much easier cases to make when the AI doesn't work. Disparate impact law requires that the AI tool used be adequately predictive of the desired outcome, before even getting into the question of whether it is *too* discriminatory or not [15]. A lack of construct validity would easily subject a model's user to liability. Due process requires decisions to not be arbitrary, and AI that doesn't work loses its claim to making decisions on a sound basis [37]. Where AI doesn't work, legal cases in general become easier.

5.2 Organizational interventions

In addition to legal levers, there are many organizational interventions that can be deployed to address the range of functionality issues discussed. Due to clear conflicts of interest, the self-regulatory approaches described are far from adequate oversight for these challenges, and the presence of regulation does a lot to incentivise organizations to take these actions in the first place. However, they do provide an immediate path forward in addressing these issues.

5.2.1 *Internal Audits & Documentation.*

After similar crises of performance in fields such as aerospace, finance and medicine, such processes evolved in those industries to enforce a new level of introspection in the form of internal audits. Taking the form of anything from documentation exercises to challenge datasets as benchmarks, these processes raised the bar for deployment criteria and matured the product development pipeline in the process [152]. The AI field could certainly adopt similar techniques for increasing the scrutiny of their systems, especially given the nascent state of reflection and standardization common in ML evaluation processes [118]. For example, the "Failure modes, effects, and diagnostic analysis (FMEDA)" documentation process from the aerospace industry could support the identification of functional safety issues prior to AI deployment [152], in addition to other resources from aerospace (such as the functional hazard analyses (FHA) or Functional Design Assurance Levels (FDALS)).

Ultimately, internal audits are a self-regulatory approach—though audits conducted by independent second parties such as a consultancy firm could provide a fresh perspective on quality control and performance in reference to articulated organizational expectations [151]. The challenge with such audits, however, is that the results are rarely communicated externally and disclosure is not mandatory, nor is it incentivized. As a result, assessment outcomes are mainly for internal use only, often just to set internal quality assurance standards for deployment and prompt further engineering reflection during the evaluation process.

5.2.2 *Product Certification & Standards.*

A trickier intervention is the avenue of product certification and standards development for AI products. This concept has already made its way into AI policy discourse; CEN (European Committee for Standardisation) and CENELEC (European Committee for Electrotechnical Standardisation), two of three European Standardisation Organisations (ESOs) were heavily involved in the creation of the EU's draft AI Act [190]. On the U.S. front, industry groups IEEE and ISO regularly shape conversations, with IEEE going so far as to attempt the development of a certification program [72, 84]. In the aviation industry, much of the establishment of engineering standards happened without active government intervention, between industry peers [152]. These efforts resemble the Partnership on AI's attempt to establish norms on model documentation processes [153]. Collective industry-wide decision-making on critical issues can raise the bar for the entire industry and raise awareness within the industry of the importance of handling functionality challenges. Existing functional safety standards from the automobile (ISO 26262), aerospace (US RTCA DO-178C), defense (MIL-STD-882E) and electronics (IEEE IEC 61508 / IEC 61511) industries, amongst others, can provide a template on how to approach this challenge within the AI industry.

5.2.3 Other Interventions. There are several other organizational factors that can determine and assess the functional safety of a system. As a client making decisions on which projects to select, or permit for purchase, it can be good to set performance related requirements for procurement and leverage this procurement process in order to set expectations for functionality [127, 156, 165, 172]. Similarly, cultural expectations for safety and engineering responsibility impact the quality of the output from the product development process – setting these expectations internally and fostering a healthy safety culture can increase cooperation on other industry-wide and organizational measures [163]. Also, as functionality is a safety risk aligned with profit-oriented goals, many model logging and evaluation operations tools are available for organizations to leverage in the internal inspection of their systems – including tools for more continuous monitoring of deployed systems [154, 169].

6 CONCLUSION : THE ROAD AHEAD

We cannot take for granted that AI products work. Buying into the presented narrative of a product with at least basic utility or an industry that will soon enough "inevitably" overcome known functional issues causes us to miss important sources of harm and available legal and organizational remedies. Although functionality issues are not completely ignored in AI policy, the lack of awareness of the range in which these issues arise leads to the problems being inadequately emphasized and poorly addressed by the full scope of accountability tools available.

The fact that faulty AI products are on the market today makes this problem particularly urgent. Poorly vetted products permeate our lives, and while many readily accept the potential for harms as a tradeoff, the claims of the products' benefits go unchallenged. But addressing functionality involves more than calling out demonstrably broken products. It also means challenging those who develop AI systems to better and more honestly understand, explore, and articulate the limits of their products prior to their release into the market or public use. Adequate assessment and communication of functionality should be a minimum requirement for mass deployment of algorithmic systems. Products that do not function should not have the opportunity to affect people's lives.

ACKNOWLEDGMENTS

We thank the Mozilla Foundation and the Algorithmic Justice League for providing financial support during this project.

REFERENCES

[1] [n.d.]. Wave control - Google Nest Help. https://support.google.com/googlenest/answer/6294727?hl=en.

[2] 2013. Zhang v. Superior Ct., 304 P.3d 163 (2013).

[3] 2019. Stipulated Order for Civil Penalty, Monetary Judgment, and Injunctive Relief, No. 1:19-cv-2184, Docket 2-1 (D.D.C. July 24, 2019) (fining Facebook $5 billion for violating a prior consent decree).

[4] 12 U.S.C. § 5511 [n.d.].

[5] ACLU. 2018. ACLU Comment on New Amazon Statement Responding to Face Recognition Technology Test. https://www.aclu.org/press-releases/aclu-comment-new-amazon-statement-responding-face-recognition-technology-test. Accessed: 2022-1-12.

[6] ACLU. 2021. ACLU Comment on NIST's Proposal for Managing Bias in AI. https://www.aclu.org/letter/aclu-comment-nists-proposal-managing-bias-ai. Accessed: 2022-1-6.

[7] Raag Agrawal and Sudhakaran Prabakaran. 2020. Big data in digital healthcare: lessons learnt and recommendations for general practice. *Heredity* 124, 4 (April 2020), 525–534.

[8] Nur Ahmed and Muntasir Wahed. 2020. The De-democratization of AI: Deep Learning and the Compute Divide in Artificial Intelligence Research. *CoRR* abs/2010.15581 (2020). arXiv:2010.15581 https://arxiv.org/abs/2010.15581

[9] AI Act [n.d.]. Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts (COM(2021) 206 final).

[10] Nil-Jana Akpinar, Maria De-Arteaga, and Alexandra Chouldechova. 2021. The effect of differential victim crime reporting on predictive policing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 838–849.

[11] Zaheer Allam, Gourav Dey, and David S Jones. 2020. Artificial intelligence (AI) provided early detection of the coronavirus (COVID-19) in China and will influence future Urban health policy internationally. *AI* 1, 2 (2020), 156–165.

[12] Ann Anderson. 2015. *Snake oil, hustlers and hambones: the American medicine show.* McFarland.

[13] Robert D Atkinson. 2018. " It Is Going to Kill Us!" and Other Myths About the Future of Artificial Intelligence. *IUP Journal of Computer Sciences* 12, 4 (2018), 7–56.

[14] Pranjal Awasthi and Jordana J George. 2020. A case for Data Democratization. (2020).

[15] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[16] Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.* 80–89.

[17] R Barker Bausell. 2009. *Snake oil science: The truth about complementary and alternative medicine.* Oxford University Press.

[18] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency.* 610–623.

[19] Emily M Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* 5185–5198.

[20] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. 2020. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *NPJ digital medicine* 3, 1 (2020), 1–8.

[21] Paul Berger. 2019. MTA's Initial Foray Into Facial Recognition at High Speed Is a Bust. *The Wall Street Journal* (2019).

[22] Joseph Bernstein. 2021. Bad News. https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation/. *Harper's Magazine* (2021).

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

[24] Deborah Blum. 2018. *The Poison Squad: One Chemist's Single-minded Crusade for Food Safety at the Turn of the Twentieth Century.* Penguin.

[25] National Transportation Safety Board. 2017. Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck. https://ntsb.gov/investigations/Pages/HWY18FH010.aspx

[26] National Transportation Safety Board. 2017. Driver Errors, Overreliance on Automation, Lack of Safeguards, Led to Fatal Tesla Crash. https://www.ntsb.gov/news/press-releases/pages/pr20170912.aspx

[27] National Transportation Safety Board. 2018. Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian. https://ntsb.gov/investigations/Pages/HWY18FH010.aspx

[28] Meredith Broussard. 2018. *Artificial unintelligence: How computers misunderstand the world.* mit Press.

[29] 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 intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228* (2018).

[30] Joanna Bryson. [n.d.]. AI & Global Governance: No One Should Trust AI - United Nations University Centre for Policy Research. https://cpr.unu.edu/publications/articles/ai-global-governance-no-one-should-trust-ai.html. Accessed: 2022-1-6.

[31] Joy Buolamwini, Sorelle A Friedler, and Christo Wilson. [n.d.]. Gender shades: Intersectional accuracy disparities in commercial gender classification. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. Accessed: 2022-1-12.

[32] Ryan Calo. 2015. Robotics and the Lessons of Cyberlaw. *Calif. L. Rev.* 103 (2015), 513.

[33] Carolyn L. Carter. 2009. *Consumer Protection in the States.* Technical Report. National Consumer Law Center.

[34] Robert Charette. 2018. Michigan's MiDAS Unemployment System: Algorithm Alchemy Created Lead, Not Gold-IEEE Spectrum. *IEEE Spectrum* 18, 3 (2018), 6.

[35] Charlie Pownall. 2021. AI, Algorithmic and Automation Incident and Controversy Repository (AIAAIC). https://www.aiaaic.org/.

[36] Bryan H Choi. 2019. Crashworthy code. *Wash. L. Rev.* 94 (2019), 39.

[37] Danielle Keats Citron. 2007. Technological due process. *Wash. UL Rev.* 85 (2007), 1249.

[38] Danielle Keats Citron. 2016. The Privacy Policymaking of State Attorneys General. *Notre Dame L. Rev.* 92 (2016), 747.

[39] Consumer Product Safety Commission. [n.d.]. About Us. https://www.cpsc.gov/About-CPSC.

[40] Federal Trade Commission. 2014. In re Snapchat, Inc., File No. 132-3078, Docket No. C-4501 (consent decree).

[41] Federal Trade Commission. 2021. FTC Votes to Update Rulemaking Procedures, Sets Stage for Stronger Deterrence of Corporate Misconduct. https://www.ftc.gov/news-events/press-releases/2021/07/ftc-votes-update-rulemaking-procedures-sets-stage-stronger.

[42] Kate Crawford. 2016. Artificial intelligence's white guy problem. *The New York Times* 25, 06 (2016).

[43] Russell C. Wald Christopher Wan Daniel E. Ho, Jennifer King. 2021. Building a National AI Research Resource: A Blueprint for the National Research Cloud. https://hai.stanford.edu/sites/default/files/2022-01/HAI_NRCR_v17.pdf.

[44] Andrea De Mauro, Marco Greco, Michele Grimaldi, and Paavo Ritala. 2018. Human resources for Big Data professions: A systematic classification of job roles and required skill sets. *Inf. Process. Manag.* 54, 5 (Sept. 2018), 807–817.

[45] Ángel Díaz and Laura Hecht. 2021. Double Standards in Social Media Content Moderation. https://www.brennancenter.org/sites/default/files/2021-08/Double_Standards_Content_Moderation.pdf. *New York: Brennan Center for Justice* (2021).

[46] Digwatch. 2021. The COVID-19 crisis: A digital policy overview. https://dig.watch/trends/covid-19-crisis-digital-policy-overview/.

[47] Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. 2019. Hard Choices in Artificial Intelligence: Addressing Normative Uncertainty through Sociotechnical Commitments. (Nov. 2019). arXiv:1911.09005 [cs.AI]

[48] Will Douglas Heaven. 2020. AI is wrestling with a replication crisis. *MIT Technology Review* (Nov. 2020).

[49] Nature Editorial. 2021. Greece used AI to curb COVID: what other nations can learn. *Nature* 597, 7877 (2021), 447–448.

[50] Lilian Edwards and Michael Veale. 2017. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.* 16 (2017), 18.

[51] Paul Egan. 2019. State of Michigan's mistake led to man filing bankruptcy. https://www.freep.com/story/news/local/michigan/2019/12/22/government-artificial-intelligence-midas-computer-fraud-fiasco/4407901002/.

[52] Alex C Engler. 2021. Independent auditors are struggling to hold AI companies accountable. FastCompany.

[53] Nora Freeman Engstrom. 2013. 3-D printing and product liability: identifying the obstacles. *U. Pa. L. Rev. Online* 162 (2013), 35.

[54] Danielle Ensign, Sorelle A Friedler, Scott Neville, Carlos Scheidegger, and Suresh Venkatasubramanian. 2017. Runaway Feedback Loops in Predictive Policing. (June 2017). arXiv:1706.09847 [cs.CY]

[55] Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics* 8 (2020), 34–48.

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

[57] Todd Feathers. [n.d.]. Las Vegas Cops Used 'Unsuitable' Facial Recognition Photos To Make Arrests. *Vice* ([n. d.]). https://www.vice.com/en/article/pkyxwv/las-vegas-cops-used-unsuitable-facial-recognition-photos-to-make-arrests

[58] Federal Trade Commission Act, 15 U.S.C. § 45 [n.d.].

[59] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[60] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[61] A G Ferguson. 2016. Policing predictive policing. *Wash. UL Rev.* (2016).

[62] Chaz Firestone. 2020. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences* 117, 43 (2020), 26562–26571.

[63] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. 2016. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*. SIAM, 144–152.

[64] U.S. Food and Drug Administration. 2021. Good Machine Learning Practice for Medical Device Development: Guiding Principles. https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles.

[65] Coalition for Critical Technology. [n.d.]. Abolish the #TechToPrisonPipeline. https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16.

[66] Karoline Freeman, Julia Geppert, Chris Stinton, Daniel Todkill, Samantha Johnson, Aileen Clarke, and Sian Taylor-Phillips. 2021. Use of artificial intelligence for image analysis in breast cancer screening programmes: systematic review of test accuracy. *bmj* 374 (2021).

[67] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2021. The (Im)Possibility of Fairness: Different Value Systems Require Different Mechanisms for Fair Decision Making. *Commun. ACM* 64, 4 (mar 2021), 136–143. https://doi.org/10.1145/3433949

[68] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2019. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*. 329–338.

[69] Sidney Fussell. [n.d.]. An Algorithm That 'Predicts' Criminality Based on a Face Sparks a Furor. *Wired* ([n. d.]). https://www.wired.com/story/algorithm-predicts-criminality-based-face-sparks-furor/

[70] Colin K Garvey. 2017. On the Democratization of AI. In *Datapower Conference Proceedings*. 5–3.

[71] Clare Garvie. 2019. Garbage in, Garbage out. Face recognition on flawed data. *Georgetown Law Center on Privacy & Technology* (2019).

[72] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. 2018. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*. IEEE, 3–18.

[73] Mark A Geistfeld. 2017. A roadmap for autonomous vehicles: State tort liability, automobile insurance, and federal safety regulation. *Calif. L. Rev.* 105 (2017), 1611.

[74] Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA Intern. Med.* 178, 11 (Nov. 2018), 1544–1547.

[75] Elizabeth Gibney. 2018. The scant science behind Cambridge Analytica's controversial marketing techniques. *Nature* (2018).

[76] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[77] Ben Green. 2019. *The smart enough city: putting technology in its place to reclaim our urban future.* MIT Press.

[78] Nitzan Guetta, Asaf Shabtai, Inderjeet Singh, Satoru Momiyama, and Yuval Elovici. 2021. Dodging Attack Using Carefully Crafted Natural Makeup. *CoRR* abs/2109.06467 (2021). arXiv:2109.06467 https://arxiv.org/abs/2109.06467

[79] Benjamin Haibe-Kains, George Alexandru Adam, Ahmed Hosny, Farnoosh Khodakarami, Thakkar Shraddha, Rebecca Kusko, Susanna-Assunta Sansone, Weida Tong, Russ D. Wolfinger, Christopher E. Mason, Wendell Jones, Joaquin Dopazo, Cesare Furlanello, Levi Waldron, Bo Wang, Chris McIntosh, Anna Goldenberg, Anshul Kundaje, Casey S. Greene, Tamara Broderick, Michael M. Hoffman, Jeffrey T. Leek, Keegan Korthauer, Wolfgang Huber, Alvis Brazma, Joelle Pineau, Robert Tibshirani, Trevor Hastie, John P. A. Ioannidis, John Quackenbush, Hugo J. W. L. Aerts, and Massive Analysis Quality Control (MAQC) Society Board of Directors. 2020. Transparency and reproducibility in artificial intelligence. *Nature* 586, 7829 (2020), E14–E16. https://doi.org/10.1038/s41586-020-2766-y

[80] Isobel Asher Hamilton. 2020. Facebook's nudity-spotting AI mistook a photo of some onions for 'sexually suggestive' content. https://www.businessinsider.com/facebook-mistakes-onions-for-sexualised-content-2020-10.

[81] M Harris. 2019. NTSB investigation into deadly Uber self-driving car crash reveals lax attitude toward safety. *IEEE Spectrum* (2019).

[82] Woodrow Hartzog. 2018. *Privacy's blueprint.* Harvard University Press.

[83] Sudhir Hasbe and Ryan Lippert. [n.d.]. ([n. d.]).

[84] John C Havens and Ali Hessami. 2019. From Principles and Standards to Certification. *Computer* 52, 4 (2019), 69–72.

[85] Will Douglas Heaven. 2021. Hundreds of AI tools have been built to catch covid. None of them helped.

[86] Nicolaus Henke, Jordan Levine, and Paul McInerney. 2018. You Don't Have to Be a Data Scientist to Fill This Must-Have Analytics Role. *Harvard Business Review* (Feb. 2018).

[87] Thomas A Heppenheimer and Ta Heppenheimer. 1995. *Turbulent skies: the history of commercial aviation.* Wiley New York.

[88] Alex Hern. 2018. Cambridge Analytica: how did it turn clicks into votes. *The Guardian* 6 (2018).

[89] Matthew Herper. 2017. MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine. *Forbes Magazine* (Feb. 2017).

[90] Kashmir Hill. 2020. Wrongfully accused by an algorithm. *The New York Times* 24 (2020).

[91] Sharona Hoffman and Andy Podgurski. 2013. Big bad data: law, public health, and biomedical databases. *J. Law Med. Ethics* 41 Suppl 1 (March 2013), 56–60.

[92] Sharona Hoffman and Andy Podgurski. 2013. The use and misuse of biomedical data: is bigger really better? *Am. J. Law Med.* 39, 4 (2013), 497–538.

[93] Chris Jay Hoofnagle. 2016. *Federal Trade Commission: Privacy Law and Policy.* Cambridge University Press.

[94] F Patrick Hubbard. 2014. Sophisticated robots: balancing liability, regulation, and innovation. *Fla. L. Rev.* 66 (2014), 1803.

[95] Tim Hwang. 2020. *Subprime attention crisis: advertising and the time bomb at the heart of the Internet.* FSG originals.

[96] IEEE. 2006. IEEE Standard Dictionary of Measures of the Software Aspects of Dependability. *IEEE Std 982. 1-2005 (Revision of IEEE Std 982. 1-1988)* (May 2006), 1–41.

[97] Bilal Mateen Michael Wooldridge Inken von Borzyskowski, Anjali Mazumder. 2021. Data science and AI in the age of COVID-19. https://www.turing.ac.uk/sites/default/files/2021-06/data-science-and-ai-in-the-age-of-covid_full-report_2.pdf.

[98] Abigail Z Jacobs. 2021. Measurement as governance in and for responsible AI. (Sept. 2021). arXiv:2109.05658 [cs.CY]

[99] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event Canada). ACM, New York, NY, USA.

[100] David Jeans. 2020. ScaleFactor Raised $100 Million In A Year Then Blamed Covid-19 For Its Demise. Employees Say It Had Much Bigger Problems. *Forbes Magazine* (July 2020).

[101] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (Sept. 2019), 389–399.

[102] Frederike Kaltheuner, Abeba Birhane, Inioluwa Deborah Raji, Razvan Amironesei, Emily Denton, Alex Hanna, Hilary Nicole, Andrew Smart, Serena Dokuaa Oduro, James Vincent, Alexander Reben, Gemma Milne, Crofton Black, Adam Harvey, Andrew Strait, Tulsi Parida, Aparna Ashok, Fieke Jansen, Corinne Cath, and Aidan Peppin. 2021. *Fake AI.* Meatspace Press.

[103] Margot E Kaminski. 2019. The Right to Explanation, Explained. *Berkeley Technology Law Journal* 34 (2019), 189.

[104] Margot E Kaminski and Jennifer M Urban. 2021. The right to contest AI. *Columbia Law Review* 121, 7 (2021), 1957–2048.

[105] Sayash Kapoor and Arvind Narayanan. 2021. (Ir)Reproducible Machine Learning: A Case Study. https://reproducible.cs.princeton.edu/. , 6 pages. https://reproducible.cs.princeton.edu/

[106] Sean Kippin and Paul Cairney. 2021. The COVID-19 exams fiasco across the UK: four nations and two windows of opportunity. *British Politics* (2021), 1–23.

[107] Lauren Kirchner and Matthew Goldstein. 2020. Access Denied: Faulty Automated Background Checks Freeze Out Renters. *The Markup* (2020).

[108] Lauren Kirchner and Matthew Goldstein. 2020. How Automated Background Checks Freeze Out Renters. *The New York Times* 28 (May 2020).

[109] Kumba Kpakima. 2021. Tiktok's algorithm reportedly bans creators using terms 'Black' and 'BLM'. https://i-d.vice.com/en_uk/article/m7epya/tiktoks-algorithm-reportedly-bans-creators-using-terms-black-and-blm. *The Verge* (2021).

[110] P. M. Krafft, Meg Young, Michael Katell, Karen Huang, and Ghislain Bugingo. 2020. *Defining AI in Policy versus Practice.* Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/3375627.3375835

[111] Mark Krass, Peter Henderson, Michelle M Mello, David M Studdert, and Daniel E Ho. 2021. How US law will evaluate artificial intelligence for covid-19. *bmj* 372 (2021).

[112] NHS AI Lab. 2021. National Medical Imaging Platform (NMIP). https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/ai-in-imaging/national-medical-imaging-platform-nmip/.

[113] Tom Lamont. 2021. The student and the algorithm: how the exam results fiasco threatened one pupil's future.

[114] Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. (Aug. 2018). arXiv:1808.07048 [cs.CL]

[115] Colin Lecher. [n.d.]. What Happens When an Algorithm Cuts Your Health Care. *The Verge* ([n. d.]). https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy

[116] Colin Lecher. 2018. What happens when an algorithm cuts your health care. *The Verge* (2018).

[117] David Lehr and Paul Ohm. [n.d.]. Playing with the data: What legal scholars should learn about machine learning. https://lawreview.law.ucdavis.edu/issues/51/2/Symposium/51-2_Lehr_Ohm.pdf. Accessed: 2021-8-10.

[118] Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Pre-Proceedings).* https://openreview.net/forum?id=mPducS1MsEK

[119] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, and Alastair K Denniston. 2020. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *bmj* 370 (2020).

[120] Kristian Lum and William Isaac. 2016. To predict and serve? *Signif. (Oxf.)* 13, 5 (Oct. 2016), 14–19.

[121] William McGeveran. 2018. The Duty of Data Security. *Minn. L. Rev.* 103 (2018), 1135.

[122] Bruce Mellado, Jianhong Wu, Jude Dzevela Kong, Nicola Luigi Bragazzi, Ali Asgary, Mary Kawonga, Nalomotse Choma, Kentaro Hayasi, Benjamin Lieberman, Thuso Mathaha, et al. 2021. Leveraging Artificial Intelligence and Big Data to optimize COVID-19 clinical public health and vaccination roll-out strategies in Africa. *Available at SSRN 3787748* (2021).

[123] Brian Menegus. 2019. Defense of amazon's face recognition tool undermined by its only known police client.

[124] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annu. Rev. Stat. Appl.* 8, 1 (March 2021), 141–163.

[125] Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. *arXiv preprint arXiv:2108.12237* (2021).

[126] Michael Muller, Melanie Feinberg, Timothy George, Steven J Jackson, Bonnie E John, Mary Beth Kery, and Samir Passi. 2019. Human-Centered Study of Data Science Work Practices. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19).* Association for Computing Machinery, New York, NY, USA, 1–8.

[127] Deirdre K Mulligan and Kenneth A Bamberger. 2019. Procurement as policy: Administrative process for machine learning. *Berkeley Tech. LJ* 34 (2019), 773.

[128] Ralph Nader. 1965. Unsafe at any speed. The designed-in dangers of the American automobile. (1965).

[129] Myura Nagendran, Yang Chen, Christopher A Lovejoy, Anthony C Gordon, Matthieu Komorowski, Hugh Harvey, Eric J Topol, John P A Ioannidis, Gary S Collins, and Mahiben Maruthappu. 2020. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 368 (2020). https://doi.org/10.1136/bmj.m689 arXiv:https://www.bmj.com/content/368/bmj.m689.full.pdf

[130] Arvind Narayanan. 2019. How to recognize AI snake oil. *Arthur Miller Lecture on Science and Ethics* (2019).

[131] Pandu Nayak. 2019. Understanding searches better than ever before. *The Keyword* 295 (2019).

[132] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning.* 151–159.

[133] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[134] OED Online 2021. https://www.oed.com/view/Entry/54950742.

[135] National Audit Office. 2020. Investigation into the response to cheating in English language tests - national audit office (NAO) press release. https://www.nao.org.uk/press-release/investigation-into-the-response-to-cheating-in-english-language-tests/

[136] Catherine Olsson. 2019. Unsolved research problems vs. real-world threat models. https://medium.com/@catherio/unsolved-research-problems-vs-real-world-threat-models-e270e256bc9e. https://medium.com/@catherio/unsolved-research-problems-vs-real-world-threat-models-e270e256bc9e

[137] Steven Overly. 2020. White House seeks Silicon Valley help battling coronavirus.

[138] David G Owen. 2001. Manufacturing Defects. *SCL Rev.* 53 (2001), 851.

[139] Jesse O'Neill. 2021. Facebook cracks down on discussing 'hoes' in gardening group. https://nypost.com/2021/07/20/facebook-cracks-down-on-discussing-hoes-in-gardening-group/.

[140] Mark A Paige and Audrey Amrein-Beardsley. 2020. "Houston, We Have a Lawsuit": A Cautionary Tale for the Implementation of Value-Added Models for High-Stakes Employment Decisions. *Educational Researcher* 49, 5 (2020), 350–359.

[141] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19).* Association for Computing Machinery, New York, NY, USA, 39–48.

[142] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 1–28.

[143] Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society* 7, 2 (July 2020), 2053951720939605.

[144] Jay Peters. [n.d.]. Researchers fooled Chinese facial recognition terminals with just a mask. *The Verge* ([n. d.]). https://www.theverge.com/2019/12/13/21020575/china-facial-recognition-terminals-fooled-3d-mask-kneron-research-fallibility

[145] 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).

[146] Carina Prunkl and Jess Whittlestone. 2020. Beyond near-and long-term: Towards a clearer account of research priorities in AI ethics and society. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 138–143.

[147] Danish Pruthi, Bhuwan Dhingra, and Zachary C Lipton. 2019. Combating adversarial misspellings with robust word recognition. *arXiv preprint arXiv:1905.11268* (2019).

[148] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

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

[150] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

[151] Inioluwa Deborah Raji, Sasha Costanza-Chock, and Joy Buolamwini. 2022. Change From the Outside: Towards Credible Third-Party Audits of AI Systems. *Missing Links in AI Policy* (2022).

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

[153] Inioluwa Deborah Raji and Jingying Yang. 2019. About ml: Annotation and benchmarking on understanding and transparency of machine learning lifecycles. *arXiv preprint arXiv:1912.06166* (2019).

[154] Alexander Ratner, Dan Alistarh, Gustavo Alonso, David G Andersen, Peter Bailis, Sarah Bird, Nicholas Carlini, Bryan Catanzaro, Jennifer Chayes, Eric Chung, et al. 2019. MLSys: The new frontier of machine learning systems. *arXiv preprint arXiv:1904.03257* (2019).

[155] Restatement (Third) of Torts: Products Liability § 3 [n.d.].

[156] Rashida Richardson. 2021. Best Practices for Government Procurement of Data-Driven Technologies. *Available at SSRN 3855637* (2021).

[157] Rashida Richardson. 2021. Defining and Demystifying Automated Decision Systems. *Maryland Law Review, Forthcoming* (2021).

[158] Rashida Richardson, Jason Schultz, and Kate Crawford. 2019. Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice. (Feb. 2019).

[159] Rashida Richardson, Jason M Schultz, and Vincent M Southerland. 2019. Litigating Algorithms: 2019 US Report. *AI Now Institute, September* (2019).

[160] Samantha Cruz Rivera, Xiaoxuan Liu, An-Wen Chan, Alastair K Denniston, and Melanie J Calvert. 2020. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *bmj* 370 (2020).

[161] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. 2021. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence* 3, 3 (2021), 199–217.

[162] Ronald E Robertson, Jon Green, Damian Ruck, Katya Ognyanova, Christo Wilson, and David Lazer. 2021. Engagement Outweighs Exposure to Partisan and Unreliable News within Google Search. *arXiv preprint arXiv:2201.00074* (2021).

[163] Harold E Roland and Brian Moriarty. 1991. *System safety engineering and management*. John Wiley & Sons.

[164] Casey Ross, Ike Swetlitz, Rachel Cohrs, Ian Dillingham, STAT Staff, Nicholas Florko, and Maddie Bender. 2018. IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/?utm_source=STAT+Newsletters&utm_campaign=beb06f048d-MR_COPY_08&utm_medium=email&utm_term=0_8cab1d7961-beb06f048d-150085821. Accessed: 2022-1-13.

[165] David S Rubenstein. 2021. Acquiring ethical AI. *Florida Law Review* 73 (2021).

[166] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015), 2503–2511.

[167] Andrew D Selbst and Julia Powles. 2017. Meaningful information and the right to explanation. *International Data Privacy Law* 7, 4 (2017), 233–242.

[168] J Shane. 2019. Janelle Shane: The danger of AI is weirder than you think TED Talk, 10: 20. Katsottu 8.8, 2020.

[169] Shreya Shankar and Aditya Parameswaran. 2021. Towards Observability for Machine Learning Pipelines. *arXiv preprint arXiv:2108.13557* (2021).

[170] Nate Silver. 2012. *The signal and the noise: why so many predictions fail–but some don't*. Penguin.

[171] George Simon, Courtney D DiNardo, Koichi Takahashi, Tina Cascone, Cynthia Powers, Rick Stevens, Joshua Allen, Mara B Antonoff, Daniel Gomez, Pat Keane, Fernando Suarez Saiz, Quynh Nguyen, Emily Roarty, Sherry Pierce, Jianjun Zhang, Emily Hardeman Barnhill, Kate Lakhani, Kenna Shaw, Brett Smith, Stephen Swisher, Rob High, P Andrew Futreal, John Heymach, and Lynda Chin. 2019. Applying Artificial Intelligence to Address the Knowledge Gaps in Cancer Care. *Oncologist* 24, 6 (June 2019), 772–782.

[172] Mona Sloane, Rumman Chowdhury, John C Havens, Tomo Lazovich, and Luis Rincon Alba. 2021. AI and Procurement-A Primer. (2021).

[173] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. 2022. A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability. *Patterns* 3, 2 (2022), 100425.

[174] David Smith and Kenneth Simpson. 2004. *Functional safety*. Routledge.

[175] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28. Accessed: 2022-1-12.

[176] Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang. 2019. Release Strategies and the Social Impacts of Language Models. arXiv:1908.09203 [cs.CL]

[177] Jay Stanley. [n.d.]. Pitfalls of Artificial Intelligence Decisionmaking Highlighted In Idaho ACLU Case. *ACLU Blogs* ([n. d.]). https://www.aclu.org/blog/privacy-technology/pitfalls-artificial-intelligence-decisionmaking-highlighted-idaho-aclu-case

[178] Brian Stanton and Theodore Jensen. 2021. Trust and Artificial Intelligence. (March 2021).

[179] Luke Stark and Jevan Hutson. 2022. Physiognomic Artificial Intelligence. *forthcoming in Fordham Intellectual Property, Media & Entertainment Law Journal XXXII* (2022). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3927300

[180] Eliza Strickland. [n.d.]. IBM Watson Heal Thyself: How IBM Watson Overpromised And Underdeliverd On AI Health Care. https://spectrum.ieee.org/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care. Accessed: 2022-1-13.

[181] Andreas Sudmann. 2020. The Democratization of Artificial Intelligence. In *The Democratization of Artificial Intelligence*. transcript-Verlag, 9–32.

[182] Maia Szalavitz. 2021. The Pain Was Unbearable. So Why Did Doctors Turn Her Away? https://www.wired.com/story/opioid-drug-addiction-algorithm-chronic-pain/.

[183] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring Robustness to Natural Distribution Shifts in Image Classification. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 18583–18599. https://proceedings.neurips.cc/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf

[184] Chris Tennant and Jack Stilgoe. 2021. The attachments of 'autonomous' vehicles. *Social Studies of Science* 51, 6 (2021), 846–870.

[185] Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. 2018. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. (Aug. 2018). arXiv:1808.10432 [cs.CL]

[186] Microsoft Translator. 2018. Neural Machine Translation reaches historic milestone: human parity for Chinese to English translations. https://www.microsoft.com/en-us/translator/blog/2018/03/14/human-parity-for-chinese-to-english-translations/. Accessed: 2022-1-12.

[187] Uniform Commercial Code § 2-314 [n.d.].

[188] Uniform Commercial Code § 2-315 [n.d.].

[189] Sam Varghese. 2021. How a Google search could end up endangering a life. https://itwire.com/home-it/how-a-google-search-could-end-up-endangering-a-life.html.

[190] Michael Veale and Frederik Zuiderveen Borgesius. 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 (2021), 97–112.

[191] Lee Vinsel. [n.d.]. You're Doing It Wrong: Notes on Criticism and Technology Hype. ([n. d.]). https://sts-news.medium.com/youre-doing-it-wrong-notes-on-criticism-and-technology-hype-18b08b4307e5

[192] Lee Vinsel. 2019. *Moving Violations: Automobiles, Experts, and Regulations in the United States*. JHU Press.

[193] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law* 7, 2 (2017), 76–99.

[194] Zhiyuan Wan, Xin Xia, David Lo, and Gail C. Murphy. 2021. How does Machine Learning Change Software Development Practices? *IEEE Transactions on Software Engineering* 47, 9 (2021), 1857–1871. https://doi.org/10.1109/TSE.2019.2937083

[195] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from Language Models. *arXiv preprint arXiv:2112.04359* (2021).

[196] Emily Weinstein. 2020. China's Use of AI in its COVID-19 Response.

[197] Eric Weiss. 2019. 'Inadequate Safety Culture' Contributed to Uber Automated Test Vehicle Crash - NTSB Calls for Federal Review Process for Automated Vehicle Testing on Public Roads. https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx

[198] Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. 2019. Unlocking Fairness: a Trade-off Revisited. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/373e4c5d8edfa8b74fd4b6791d0cf6dc-Paper.pdf

[199] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and auditing fair algorithms: A case study in candidate screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 666–677.

[200] Nick Wingfield. 2014. Nest Labs Stops Selling Its Smoke Detector. *The New York Times* (Apr 2014). https://www.nytimes.com/2014/04/04/technology/nest-labs-citing-flaw-halts-smoke-detector-sales.html

[201] Winter v. G.P. Putnam's Sons, 938 F.2d 1033 (9th Cir. 1991) 1991.

[202] Natalia Wojcik. [n.d.]. IBM's Watson 'is a joke,' says Social Capital CEO Palihapitiya. https://www.cnbc.com/2017/05/08/ibms-watson-is-a-joke-says-social-capital-ceo-palihapitiya.html. Accessed: 2022-1-13.

[203] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. 2021. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine* 181, 8 (08 2021), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626 arXiv:https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2781307/jamainternal_wong_2021_oi_210027_1627674961.11707.pdf

[204] Matt Wood. [n.d.]. Thoughts On Machine Learning Accuracy. https://aws.amazon.com/blogs/aws/thoughts-on-machine-learning-accuracy/.

[205] Eric Wu, Kevin Wu, Roxana Daneshjou, David Ouyang, Daniel E Ho, and James Zou. 2021. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nature Medicine* 27, 4 (2021), 582–584.

[206] Laure Wynants, Ben Van Calster, Gary S Collins, Richard D Riley, Georg Heinze, Ewoud Schuit, Marc MJ Bonten, Darren L Dahly, Johanna A Damen, Thomas PA Debray, et al. 2020. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj* 369 (2020).

[207] Karen Yeung. 2020. Recommendation of the council on artificial intelligence (oecd). *International Legal Materials* 59, 1 (2020), 27–34.