Promises and pitfalls of AI for legal applications

SAYASH KAPOOR[1], PETER HENDERSON, ARVIND NARAYANAN
NOVEMBER 11, 2023

Are AI tools set to redefine the landscape of the legal profession? We argue that the current state of evaluations of AI does not allow us to answer this question. We dive into the increasingly prevalent roles of three distinct types of AI used in legal settings: generative AI, AI for automating legal judgment, and predictive AI. While generative AI could help with routine legal tasks, concerns surrounding contamination, construct validity, and prompt sensitivity warrant attention. On the other hand, applications of AI for automating legal judgment range widely in their usefulness. Some helpful interventions include finding common trademark or patent filing errors. Others are inaccurate, hard to evaluate, and suffer from common machine learning errors, such as predicting the outcome of court decisions. Finally, predictive AI is often touted as a groundbreaking tool, but it encounters serious limitations in research and real-world applications. These limitations call into question the validity of such research and applications. Diving into a series of case studies, we highlight potential pitfalls and outline necessary guardrails for evaluating AI in legal contexts.

1 INTRODUCTION

DoNotPay, a U.S.-based AI startup, claimed to sell the services of a "robot lawyer" to help customers prepare legal documents, contest parking tickets, and cancel subscriptions [1]. On January 8, 2023, CEO Joshua Browder claimed that the company would pay USD 1 million to any lawyer who used DoNotPay's robot lawyer to argue a U.S. Supreme Court case, by using an earpiece to repeat the arguments made by the company's software [2]. Even setting aside the fact that the Supreme Court prohibits electronics in the courtroom, the U.S. has several laws prohibiting unlawful legal practice—the unauthorized practice of law by individuals who are not licensed attorneys. Soon after the announcement, the CEO backed down [3], and the term "robot lawyer" was changed to "AI consumer champion" on the company's website [4]. Still, the company is facing multiple class-action lawsuits for unlawful legal practice [5].

This was far from the first time when technology was claimed to replace a lawyer, and it won't be the last. After all, the company had been claiming to sell the services of a robot lawyer for more than four years. Claims about lawyers being replaced by digital technology predate the company. A 2011 New York Times headline read: "Armies of Expensive Lawyers, Replaced by Cheaper Software." Since the article was published, the number of lawyers in the U.S. has actually *increased* by eight percent [6]. How do we separate true advances from hype?

In this position paper, we argue that the kinds of legal applications we can legitimately use AI for should be determined by the evaluations that reflect said use in the real world. In other words, when OpenAI claims that GPT-4 can "pass the bar exam," that isn't evidence that GPT-4 is becoming as capable as lawyers: after all, it's not a lawyer's job to answer bar exam questions all day. Evaluations should be used to identify not just how well AI does on a given task, but also *which types of tasks* AI can be useful for.

[1]Author emails: sayashk@princeton.edu, phend@stanford.edu, arvindn@cs.princeton.edu. Parts of this paper are based on online blog posts by two authors (https://aisnakeoil.com) and a Senate testimony by one of the authors (https://www.cs.princeton.edu/~arvindn/talks/insight_forum_statement.pdf).

In our analysis, we look at the challenges that arise in meaningful evaluations of AI in legal settings and offer recommendations for overcoming them. Legal applications of AI vary in how difficult they are to evaluate. For some applications, evaluation is relatively easy. For example, a tool that categorizes a request for legal advice into particular areas of law can be evaluated by comparing against corresponding labels from lawyers performing the same task.[1] In contrast, for other types of AI, there is no clear "correct" answer. For instance, if generative AI is used to prepare a legal filing, there is no single correct answer on how the document should be written—reasonable people can disagree on what strategies to take. Tasks that are harder to evaluate also tend to be those that would lead to the most significant changes to the legal profession. If AI could be useful for consequential legal tasks like preparing legal filings, that would have much broader implications for the future of legal professionals compared to labeling text for different areas of law. Unfortunately, evaluations of AI for consequential legal applications fall short of providing evidence about their usefulness and trustworthiness in real-world settings.

Our analysis revolves around case studies from three types of AI that have seen increasing adoption in legal and judicial settings: generative AI, AI for automating legal judgment, and predictive AI. Large language models (LLMs) such as GPT-4 are an example of generative AI (Section 2). Generative AI is usually trained on a vast amount of data—from the internet, as well as from private data sources [7]. One of the critical tenets of machine learning is that the data used for training an ML model should be different from the one used for evaluating it. Otherwise, it would be akin to teaching to the test—the model would perform well on the training data, but not in the real world. But the large amount of data used for training language models means that for any given evaluation, it is hard to say if the model will work as well on similar questions, or only works for the *specific* questions included in its training data. LLMs are also sensitive to the style of the input prompt. Minor changes in the phrasing of an input can lead to significant differences in the outputs. These issues make it hard to assess whether a generative AI system is actually useful for any given task, or whether it only works in narrowly defined settings or for data it is trained on.

We will then turn to AI for automating legal judgment (Section 3). In the last few years, many papers have claimed to predict the outcomes of court judgments using AI. On closer inspection, the vast majority of these efforts fall short because of well-known pitfalls in machine learning research. Still, such applications could be useful for solving narrower tasks, such as detecting common errors in trademark or patent filings.

The third type of AI we will look at is predictive AI (Section 4). Predictive AI is used to make consequential decisions about people's lives. It uses machine learning to predict how likely it is that a patient will be readmitted if charged, whether a credit applicant will pay back the loan if approved, whether a job candidate will be a good employee if hired, or whether a defendant will go on to commit another crime if released. Predictive AI is error prone because it is hard to predict the future. Unfortunately, companies have exploited public confusion around the disparate types of technologies that fall under the umbrella term "AI". While some types of AI are rapidly advancing, predictive AI is not. This fact is not widely appreciated, whether by those who buy these tools or those subject to its decisions.

In the next three sections, we will look at generative AI, AI for automating legal judgment, and predictive AI respectively. We will go over case studies that show the challenges of evaluating generative AI and AI for automating legal judgment. We will then examine the vast evidence against predictive AI, including in legal domains.

[1] As in, for example, the Learned Hands project. https://learnedhands.law.stanford.edu/

2 GENERATIVE AI

Generative AI refers to AI that can be used to create text, images, music, or other form of media. It is often trained on a vast amount of existing data. This process involves using algorithms and models to learn patterns, styles, or features of the data they have been trained on. Many recent instances of AI that have received widespread attention are examples of generative AI: Anthropic's Claude is a language model [8], OpenAI's DALL-E is a text-to-image model [9], and Meta's Make-a-video is a video generation tool [10]. In legal settings, the most common example of generative AI is language models.

2.1 Evaluating language models is a minefield

When OpenAI announced its GPT-4 language model, it claimed the model could pass a "simulated bar exam with a score around the top 10% of test takers". This led to much speculation about whether AI would soon replace lawyers. But what does a high score on the bar exam mean? Are language models like GPT-4 already capable of replacing lawyers, and are legal protections such as the U.S. rules on unlawful legal practice the only remaining hurdle? Or are there fundamental limitations to such claims? We hope to inject some reality into this conversation using three major concerns about current evaluations of generative AI.

2.1.1 Contamination. Contamination refers to including the same data in the training and evaluation data sets for a model [11]. This can lead to overoptimistic estimates of model performance since a model can simply memorize solutions in its training set instead of being able to answer new questions.

When OpenAI released its GPT-4 language model, it made a number of claims about the model's capabilities. As discussed, one prominent claim was the model's performance on the bar exam. Another claim was about the model's coding ability. To benchmark GPT-4's coding ability, OpenAI evaluated it on problems from Codeforces, a website that hosts coding competitions. The training data cutoff for the original GPT-4 model was September 2021. The model could correctly answer most Codeforces questions from before its training date cutoff, but couldn't answer questions after its training date cutoff correctly [12]. This strongly suggests that the model memorized solutions from its training set—or at least partly memorized them, enough that it could fill in what it can't recall. That is, instead of developing the capability to answer *new* coding questions, it could only answer questions it has already been trained on.

The Codeforces results in the paper were not affected by this, as OpenAI used problems from recent Codeforces competitions, which resulted in the model being evaluated on fresh problems that were not in the training set. Sure enough, GPT-4 performed very poorly [13]. But for the benchmarks other than coding, we don't know of a clean way to separate the questions by time period, so it is unlikely that OpenAI was able to avoid contamination. For the same reason, we can't experiment to test how performance varies by date. Contamination has affected AI since well before the recent wave of generative AI. We will dive deeper into its history in the section on automating legal judgment (Section 3).

2.1.2 Lack of construct validity. Construct validity refers to the extent to which an evaluation accurately represents and measures the construct it is designed to assess. For evaluating language models, this could be whether the underlying skill or capability being measured corresponds to the specific tasks or questions in the evaluation set.

For example, we can get some clarity around the goals of OpenAI's evaluations by asking what the developer is trying to measure using model performance on standardized exams. If the goal is to predict how the language model will do on real-world tasks, then the evaluations are not suited to assessing such claims. This is because, in a sense, any two bar exam or medical exam questions are more similar to each other than the tasks professionals do in the real world—something that critics of the bar exam regularly lament on, resulting in recent restructuring of the bar exam [14] and proposals for alternative pathways to certification based on real-world training [15]. And since bar exam questions are drawn from a limited pool, it is possible that including any exam or practice questions in the training corpus results in an inflated estimate of real-world usefulness.

One reason why language model evaluations suffer from the lack of construct validity is because memorization is a spectrum. Even if a language model has not seen an exact problem on a training set, it has inevitably seen examples that are pretty close, simply because of the size of the training corpus. That means it can get away with a much shallower level of reasoning. As a result, these benchmarks don't necessarily give us evidence that language models are acquiring the kind of in-depth reasoning skills that human test-takers might have. The assumption, correct or not, is that humans taking exams generalize the skills tested by the exam to a wider range of relevant tasks. While this assumption might already be somewhat limited for humans, it is unfounded for language models that might take all sorts of shortcuts [16] and memorize key information to come to the right answer without generalizing in any way.

In some real-world tasks, shallow reasoning may be sufficient—for example, it could be enough to build a chatbot to help applicants answer bar exam questions. But the world is constantly changing, so if a bot is asked to analyze the legal consequences of a new technology or a new judicial decision, it does not have much to draw upon. In short, tests designed for humans lack construct validity when applied to bots.

On top of this, professional exams, especially the bar exam, notoriously overemphasize subject-matter knowledge and underemphasize real-world skills, which are far harder to measure in a standardized, computer-administered way. In other words, not only do these exams emphasize the wrong thing, they overemphasize precisely the thing that language models are good at.

Benchmarks are already wildly overused in AI for comparing different models [17]. They have been heavily criticized for collapsing a multidimensional evaluation into a single number [18]. When used as a way to compare humans and bots, what results is a mischaracterization of how well an AI system actually performs at the task at hand.

2.1.3 Prompt sensitivity. Another issue with evaluating language models is their sensitivity to the user's prompts. Small changes to the prompt can have a significant impact on the model's outputs.

This issue is exemplified by a peer-reviewed paper that made the news in August 2023, claiming that ChatGPT has a liberal bias: Motoki et al. [19] asked the bot for its opinions on statements like "The freer the market, the freer the people." They claimed that ChatGPT gave left-leaning answers the overwhelming majority of the time. Surprisingly, in stark contrast to the original findings, a reproduction of the study found that most of the time, the GPT-3.5 and GPT-4 models refused to express an opinion [20]. This is precisely the behavior OpenAI claims the models have.

Why did Motoki et al. [19] find something very different? The authors asked ChatGPT multiple-choice questions. This would be relevant if we lived in a world where people form political opinions by asking ChatGPT multiple-choice questions. In reality, political bias is a concern because it might subtly come up in natural conversations. It's a genuine concern, and ChatGPT may have a liberal

bias, but this paper provides little evidence of this bias. Instead, it provides a limited window into when such bias might be presented: in multiple-choice, survey-like settings.

Unfortunately, we are entirely in the dark about how users use these models in the real world. Since model developers do not share information about model use, we currently have few ways to study chatbots' political bias in typical user settings, among a large set of other related and important questions. It's not just about bias. Suppose we want to evaluate inaccurate answers to legal or medical questions. In that case, the same hurdle arises because we do not know how users interact with these models in the real world. Recent large-scale evaluations of language model performance start to expand the scope of evaluations on a wider range of legal tasks [21], but even in these cases, benchmark creators pick a fixed set of prompts that are used across evaluations. It is possible that a user, particularly those who are not knowledgeable enough about either the legal domain or the limitations of language models, could see drastically different performance on the same tasks if they do not craft their prompt in the same way as the evaluation benchmark. Even the ordering of few-shot examples in a prompt can affect performance by double-digit percentage points [22].

In sum, prompt sensitivity makes it hard to test a model's capabilities without knowing how real-world users interact with the model. But it also points to an intervention that could help: more transparency by the companies that create these language models about how they are used in the real world.

2.2 Improving evaluations of generative AI

2.2.1 Transparency reports. A key limiting factor in current evaluations of language models is the lack of transparency around how users actually use these models on a day-to-day basis. Without knowing how users interact with LLMs, it is hard to understand what limitations needs to be addressed and how evaluations can best be constructed to be representative of typical use cases. Transparency reports that outline how models are used in the real world can help understand and improve the construct validity of current evaluations and avoid evaluations from falling prey to prompt sensitivity [23].

Transparency reports have been useful for previous waves of digital technology. Social media platforms, including Facebook [24], YouTube [25], and TikTok [26], provide some transparency about how their platforms are used and abused. In addition, on social media, researchers have some visibility into the spread of harmful content since much of it is public. But with generative AI, we are entirely in the dark. Transparency reporting is most critical for generative AI applications intended to be general purpose (e.g., ChatGPT) and those designed to be used in a high-risk setting (such as medicine, finance, law, or hiring).

2.2.2 Discipline-specific evaluations of LLMs. Many current evaluations of LLMs are general purpose: they measure the efficacy of language models on general tasks such as summarization, retrieval, or factuality. However, these evaluations do not tell us much about how LLMs can aid professionals in their day-to-day tasks. The involvement of domain experts in designing such evaluations is necessary to improve the status quo. Without the involvement of domain experts, benchmarks for testing language models on professional tasks are likely to suffer from the construct validity problem.

Such evaluations can be both quantitative and qualitative. An interdisciplinary group of lawyers and AI experts created the LegalBench benchmark for evaluating language models on various legal reasoning tasks [21]. This is an example of a quantitative evaluation created by professionals to measure the usefulness of generative AI in their profession.

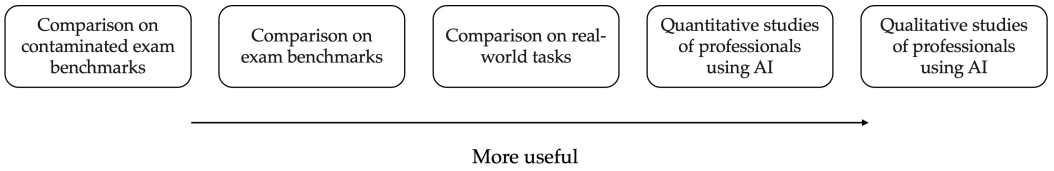

Fig. 1. **Types of evaluations of generative AI**. Current evaluations of AI are often based on exam benchmarks meant for humans, and suffer from contamination: overlaps between the training and evaluation datasets. Comparing the performance of these models on real-world tasks, especially those curated by professionals, is more likely to be useful. Since the use of generative AI is nascent, qualitative studies that observe how professionals use these tools for day-to-day professional tasks are likely to be a more useful, if expensive, way of evaluating these tools.

But there are reasons to think that qualitative studies of professionals and how they could use AI are likely to be even more useful, since these tools are so new that we still need consensus on what the right questions to ask are. To our knowledge, such qualitative studies haven't yet been conducted for legal professionals. But in other professions, several such studies have been conducted. A recent study looked at the use of ChatGPT in professional writing tasks and found that the quality of the outputs improved when people used language models compared to when they did not [27]. More importantly, two months after the experiment, writers assigned to use ChatGPT were twice as likely to continue using it, providing some evidence of its usefulness to writing professionals. Similarly, there have been several examples of qualitative studies in medicine, such as on using LLMs to create summaries [28] and respond to medical questions [29].

2.2.3 Communicating the limitations of current LLMs. Two recent cases of lawyers misusing language models have made the headlines [30, 31]. In both cases, lawyers used LLMs to generate legal citations. However, current LLMs can fabricate information even while presenting it authoritatively. When people are not aware of these limitations, it could result in massive professional damage. In both cases, the lawyers were sanctioned for fabricating information in legal filings. Even when an LLM is trained on accurate texts, such as a filtered dataset of past legal documents, it is not guaranteed to produce accurate outputs [32]. These cases highlight the need for better communication of these limitations for end users by companies providing these services [33]. Developers have added some disclaimers to language models to reduce such errors. For example, OpenAI says "ChatGPT can make mistakes. Consider checking important information" at the bottom of the ChatGPT chatbox. Anthropic goes one step further. Its disclaimer is more clear about the limitations ("Claude is in beta release and may display incorrect or harmful information,") and when the output contains URLs, there is also a disclaimer about links potentially being inaccurate. Some judges have also issued chambers' rules to clarify how lawyers should explicitly account for their use of AI [34].

3 AUTOMATING LEGAL JUDGMENT

In recent years, AI has been widely used in automating legal judgment—for example, in legal research, to predict court case outcomes based on text from the court proceedings.

Medvedeva et al. [35] systematically review papers claiming to predict judgments. They find severe shortcomings in the literature they review. Their main finding is that the vast majority of papers claiming to predict the outcomes of court judgments do not try to solve this problem at all. In many cases, the papers solve a related but ultimately less helpful problem: they use the judgment

text containing the final judgment to "predict" the verdict. Since the text of the final judgment includes the verdict, these studies do not provide real-world evidence of the usefulness of AI in judgment prediction. Medvedeva et al. point out that this error could have been caused by insufficient knowledge of the datasets being used in judgment classification and inadequate steps taken to filter out information about the verdict from the dataset. This highlights the need for both legal and AI expertise for useful applications of AI in legal settings. Moreover, for the small minority of papers that actually predict the outcomes, the accuracy of the resulting models is much lower.

The low accuracy demonstrates that automating judgments from the text of legal cases is hard. This is not surprising: legal outcomes depend on the context and specifics of cases, the available documents might not comprise the entirety of the context of the case being adjudicated, and the specific judgment might depend on a specific judge's (or set of judges') interpretation of the arguments. In addition, there is significant variability across different jurisdictions, which means the amount of data that can be used for training AI to automate judgments in any specific jurisdiction is small. Finally, the judgments made over time evolve with changes to the specific judges, the set of past cases comprising precedent, and legislation, in addition to many other factors.

Medvedeva et al.'s findings also point to the problem of contamination (Section 2.1.1). Since the text of the judgment also contains the verdict, the model essentially has access to the answers while making predictions—like teaching to the test, this vastly inflates the accuracy of the resulting models, leading to exaggerated performance estimates.

This is a well-known issue in machine learning. In traditional machine learning research, it is called *data leakage* or simply *leakage*. Legal research is far from the only field where leakage has been found: It is widespread in research that uses machine learning. Kapoor and Narayanan [36] found that it affects hundreds of papers in over a dozen scientific fields, often leading to vastly inflated performance estimates. Cases of leakage range from textbook issues such as not separating the training and test set, to as yet unsolved research problems, such as subtle dependencies between the training and test data. Surprisingly, many of these fields are independently rediscovering the issue, showing the pervasiveness of leakage and the lack of systematic solutions available to solve it. To avoid leakage, they provide model info sheets, which contain 21 questions related to the model development process that can help avoid leakage in studies that use machine learning.

This doesn't even begin to address the potential for biases, sensitivity to inputs, and other challenges for evaluating legal judgment prediction tasks. The challenges with evaluation should limit where and how judgment prediction tasks are used. A well-evaluated judgment prediction system could be used to better understand what properties of briefs could lead to poor outcomes (e.g., finding common errors). This would serve as a suggestion to attorneys that might miss common errors but not result in any binding outcome and could be ignored by the attorney. On the other hand, such systems should not be used to make final decisions or final recommendations to a decision-maker.

In a more constrained, highly issue-specific, low-stakes setting, it may be possible to get sufficient coverage to construct a thorough evaluation with confidence. For example, common errors could be identified in trademark or patent filings to reduce costs to both the filing party and the United States Patent and Trademark Office, where over 86% of patent applications received at least one non-final rejection [37]. Or, in the case of the Social Security Administration ("SSA"), a simple model is used to spot issues with judgments that might lead to a remand of the judgment on appeal [38].

These types of judgment prediction tasks are distinct from the more general case. First, they are constrained to a single or small handful of issues, which makes it possible to sample sufficient data to cover typical use cases. Second, they are (or should be) fully observable for the issue being

assessed. The model would ideally have access to the same information as the adjudicator. This is typically not true of general-purpose judgment prediction tasks, which are nearly never able to fully examine all evidence before the judge. A judge in the general case discussed above typically has access to all court transcripts and evidence (in most cases, this can span thousands of pages). And they can hold oral arguments to form their thinking on a subject matter, as well as gather more information from interacting with the people involved in the case.

On the other hand, in settings without these additional modalities—or where the issue being examined does not require examination of such external information—it is more reasonable to construct valid evaluation protocols. For example, in the SSA setting, one mistake flagged by the automated system is when the adjudicator's opinion does not address a medical claim made in a benefits claim in their denial of benefits judgment [38]. Such a mistake would almost certainly result in a remand of the decision on appeal. To make such an assessment, however, a system would not need any additional information beyond the benefits claim and the text of the decision. As such, the system operates under full observability, where more thorough evaluations can be conducted. Nonetheless, even in these cases, automated judgments should be made with extreme caution. Deployments should be structured, favoring helpful informative recommendations to both parties in a dispute, rather than being used as a binding mechanism. And a thorough appeals process should be available.

We now turn to applications of predictive AI used to make real-world decisions. These applications often suffer from exaggerated performance estimates, such as the ones we have seen in the last two sections, but they also introduce a variety of other limitations.

4 PREDICTIVE AI

Predictive AI refers to using machine learning to predict future outcomes of interest about individuals, especially focusing on making decisions using these predictions. Predictive AI stands in contrast to generative AI, AI for automating legal judgment, and other types of AI.

Questions such as whether someone will commit a crime if released on bail or pay back a loan depend on an array of factors that are unknown, and unknowable, at the time of decision making. So there is an intrinsic limit to the accuracy of predictive AI. At best, the technology can offer broad statistical generalizations. This is different from other applications of AI where there is no aleatory uncertainty about future outcomes, like extracting structured information from court opinions; in those settings, there is a consensus answer that the system can learn to generate.

Predictive AI also differs from AI for automating legal judgment. For automating legal judgment, at least in some cases we discussed, many of the facts used for making legal judgments could be available as inputs to the model. In contrast, applications of predictive AI focus on predicting outcomes that have not yet occurred. These models typically do this without sufficient observability of relevant features that would be important for such a prediction, nor do they have enough data to form a robust model of the world that would allow for such accurate predictions. Instead, these settings typically rely on extremely rough generalizations and approximations using simple linear models (when the underlying dynamics are far from linear).

Despite these issues, predictive AI has been deployed in a myriad of real-world applications for automated decision making. In criminal justice, it is used for making several decisions, including predicting if a defendant is at risk of recidivism [39, 40] and determining if a detainee should get parole or be released on probation [41]. In medicine, predictive AI determines who should be prioritized for care [42]. In finance, predictive AI determines who should be granted a loan [43]. In

education, it is used to identify students at risk of dropping out of school [44]. In hiring, employers use it to screen applications based on predictions about who would perform well at a job [45]. These applications are often accompanied by claims of high accuracy, fairness, and efficiency [46]. Yet, a closer account of how real-world deployments of predictive AI play out reveals the many flaws in applications of predictive AI.

4.1 Low accuracy of deployed applications.

Many recent studies have shown that predictive accuracy is low when AI is used to predict the future. This is true even when tens of thousands of features are collected on thousands of individuals. Salganik et al. [47] conducted a large-scale study on the predictability of life outcomes. In a prediction competition, hundreds of researchers tried to predict how well a child would do based on thousands of data points about them from the past. In total, the data consisted of over 4,000 families, with over 10,000 features collected on each family for 15 years. Despite the vast amount of data and the use of state-of-the-art machine learning methods, the results were disappointing: the best model performed only slightly better than simple regression models that only looked at four basic sociologically relevant features.

One common application of predictive tools in criminal justice is to predict recidivism. A 2016 ProPublica investigation found that COMPAS, a widely used algorithm to predict the risk of recidivism for defendants, had twice as many false positives for Black defendants as White ones [39]. Perhaps more surprisingly, the investigation found that the overall accuracy of the algorithm was only around 65%. In a follow-up study, Dressel and Farid [40] found that this accuracy was no more accurate than predictions made by people without any background in criminal justice. Moreover, while COMPAS used 137 features in its predictions, Dressel and Farid showed that an algorithm with just two features (age and number of prior arrests) performs as well as the COMPAS algorithm.

Notably, the majority of defendants predicted to be at high risk of committing violent crimes do not go on to recidivate. Fundamentally, these simple models distill into these few features a model of a person's entire future life for the next few years. They have no access to private information, like a defendant's commitment to never commit a crime. They cannot model defendants' attempts to seek help. In some cases, the only reason a defendant would fail to appear at trial is because of a lack of clarity about their court dates—and simple interventions like sending a text message to the defendant, or making the summons information easier to understand, would drastically reduce failure to appear [48]. Yet, such failure to appear is still counted as an act of recidivism. But none of this or a myriad of other possibilities could ever be considered by an algorithm that has access to superficial features about a defendant.

Similar results have been found in a number of domains for making consequential decisions. In medicine, Epic, a U.S. healthcare technology company, developed a tool for predicting which hospitalized patients are at risk of developing sepsis. It was deployed by hundreds of hospitals before an independent evaluation found that the AUC-ROC accuracy of the tool was only 63%, barely better than the flip of a coin [49]. In hiring, there are several tools used to predict how well a candidate would perform at a job. However, these tools are not accompanied by peer-reviewed validation of their performance. There have been bias audits of two leading tools [50, 51], but these were carefully scoped to exclude the more fundamental question of whether the tools even work.

These studies show that predicting the future, even with vast amounts of data and state-of-the-art machine learning methods, can be hard. In addition, in some settings, it is impossible to predict or thoroughly evaluate future predictions. Where dynamics are known and information is readily available, this might be possible—as in physical sciences, where we can build reliable approximations

of aspects of the world that we are modeling. Yet this is not the majority of cases in the law, where fundamentally, most predictions will be about people and societies.

4.2 Predictions about the wrong people.

A machine learning tool called Public Safety Assessment (PSA) is used in U.S. courts in over half the states. Like COMPAS, if the tool predicts that a defendant has a high risk of re-offending, bail could be denied. The model is trained on data from 1.5 million cases across the country. But crime patterns in specific regions differ from nationwide averages in important ways, which means that the tool fails catastrophically in some areas. Corey [52] highlights that in Cook County, Illinois, the rate of violent recidivism is *ten times* lower than the nationwide data that was used for training PSA.

This is known as the problem of *distribution shifts*: when the data used to train an ML model differs from the population on which the model is eventually deployed, models are unable to adapt well. Distribution shift is an open research problem in machine learning [16], and affects most predictive AI applications where the population of interest differs from training data [46].

4.3 The impact of leakage.

Like generative AI and AI for automating legal judgment, real-world applications of predictive AI have also suffered from leakage. Epic initially claimed that its sepsis prediction AI had an accuracy of 76–83%, far higher than its actual accuracy [49]. The hundreds of hospitals that adopted it did not challenge this claim. However, as we have seen, performance evaluation of machine learning can be notoriously tricky because of problems like data leakage. This allows vendors to get away with false or misleading claims. In Epic's case, one input to its sepsis prediction tool was whether a clinician had prescribed antibiotics. However, the prescription of antibiotics is often a sign that a clinician has already diagnosed a patient with sepsis. These cases were still counted as successful predictions, leading to vastly inflated accuracy numbers.

4.4 Diffusion of responsibility.

Vendors sell predictive AI based on the promise of full automation and elimination of jobs, but when the tools perform poorly, they retreat to the fine print, which says that the tool shouldn't be used on its own. For example, Toronto recently used an AI tool to predict when a public beach will be safe. It went horribly awry: On a majority of the days when the water was declared safe to swim in, it was actually unsafe [53]. Although the tool was not intended to be used without human oversight, it turned out that city officials responsible for oversight never questioned its recommendations. This incident illustrates the diffusion of responsibility: when the accountability for decisions and actions is spread thinly across multiple people or departments, often deliberately.

Another example comes from Optum's tools to predict patients' future healthcare costs. Hospitals used it to prioritize patients for intervention. However, it turned out that since hospitals had a history of spending less on Black patients, the tool baked in this bias and was less likely to prioritize a Black patient even if they had the same health conditions as a White patient [42]. Hospitals blamed Optum, but Optum said that the tool accurately predicted costs as designed, implying that using the tool in a way that resulted in disparate impact was the hospitals' responsibility. The individual decisions made by these systems also tend not to be contestable by decision subjects, as vendors claim that the logic of the tool is a trade secret.

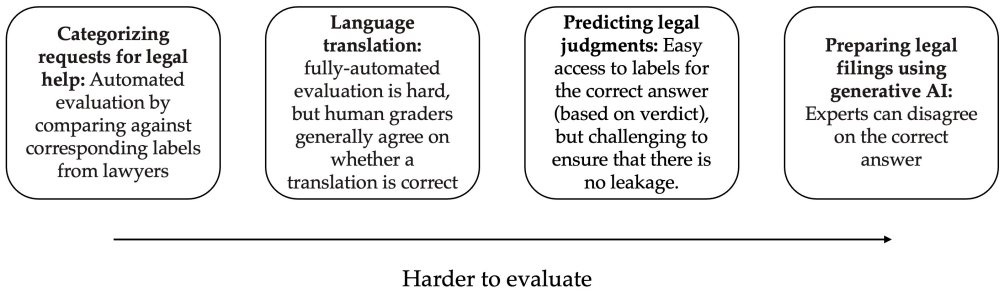

Fig. 2. **Variation in the difficulty of evaluating AI for legal tasks**. Some tasks, such as AI for categorizing requests for legal help by area of law, are easy to evaluate, whereas for other tasks, such as preparing legal filings using AI, there is no clear right answer, which makes evaluation hard.

4.5 Lack of transparency and in-house expertise.

In most cases, hospitals, court districts, or employers do not develop predictive tools in house—tools that might be tailored to their specific needs and those of the populations that they serve. Instead, they purchase or license one-size-fits-all tools from AI vendors. This exacerbates issues with evaluations since the users of the tools cannot push back against vendors' claims. The predictive AI industry is built on an inherently limited technology that has been overhyped, but avoids transparency to obscure this fact.

These issues are not specific to the examples we list above. In an analysis of eight predictive AI applications across domains, Wang et al. [46] found that these issues are widespread in domains such as finance, insurance, child welfare, and medicine, in addition to criminal justice.

Given the propensity of such applications to failure, predictive AI in the legal domain needs to be held to a much higher standard to ensure that it functions as its developers claim. This requires much stronger transparency by the developers, clear mechanisms to ensure contestability to decision subjects, and evaluations that go beyond just the technical specifications of these tools into the societal impact of these tools.

5 OUTLOOK

The effective deployment of AI in legal contexts requires shifting from technical evaluations to robust socio-technical assessments carried out in the specific context in which an AI system would be deployed. While past machine learning applications did not consider such evaluations because they were cost prohibitive, this change is necessary due to the complex nature and societal impact of AI applications in the legal field.

For generative AI, such as LLMs, incorporating domain-specific evaluations, that are representative of typical uses, is crucial. Although these evaluations could be more expensive than traditional ML evaluations, they are vital for understanding the real-world implications and limitations of AI in legal settings, especially as the cost of developing the model by far eclipses the cost of better evaluations. Engaging legal experts and closely examining the practical use of AI tools in legal work will provide insights beyond standard benchmarks, ensuring the technology is only used in settings where its use has been validated.

On the other hand, predictive AI presents a unique challenge where the evaluation process can be more complex than the model development itself. In this context, the success of predictive AI hinges not just on technical accuracy, but also on its applicability and reliability in real-world scenarios. Therefore, organizations must prioritize in-depth evaluations for social and ethical implications, such as how contestable the decisions from an AI system are. This can be accomplished by developing the capacity to perform these evaluations in house. And if organizations are developing in-house AI expertise, then it might be easier to develop the models in house too, rather than relying on off-the-shelf solutions, given performing valid evaluations can be more expensive and time consuming than model development.

To answer the question "What can I use an AI system for?", it is essential first to answer "How was this AI system evaluated?". Unfortunately, the current state of AI evaluations leaves much to be desired.

REFERENCES

[1] DoNotPay - The World's First Robot Lawyer, January 2023. URL https://web.archive.org/web/20230101170502/https://donotpay.com/.

[2] Joshua Browder [@jbrowder1]. DoNotPay will pay any lawyer or person $1,000,000 with an upcoming case in front of the United States Supreme Court to wear AirPods and let our robot lawyer argue the case by repeating exactly what it says. (1/2), January 2023. URL https://twitter.com/jbrowder1/status/1612312707398795264.

[3] Joshua Browder [@jbrowder1]. Good morning! Bad news: after receiving threats from State Bar prosecutors, it seems likely they will put me in jail for 6 months if I follow through with bringing a robot lawyer into a physical courtroom. DoNotPay is postponing our court case and sticking to consumer rights:, January 2023. URL https://twitter.com/jbrowder1/status/1618265395986857984.

[4] DoNotPay - Your AI Consumer Champion, . URL https://web.archive.org/web/20230730013643/https://donotpay.com/.

[5] ANALYSIS: DoNotPay Lawsuits: A Setback for Justice Initiatives?, . URL https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-donotpay-lawsuits-a-setback-for-justice-initiatives.

[6] Statista Research Department. U.S.: number of lawyers 2007-2022, 2023. URL https://www.statista.com/statistics/740222/number-of-lawyers-us/.

[7] Matt O'Brien. ChatGPT-maker OpenAI signs deal with AP to license news stories, July 2023. URL https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a. Section: Business.

[8] Introducing Claude, . URL https://www.anthropic.com/index/introducing-claude.

[9] DALL·E 3, . URL https://openai.com/dall-e-3.

[10] Make-A-Video, . URL https://makeavideo.studio/.

[11] Inbal Magar and Roy Schwartz. Data Contamination: From Memorization to Exploitation, March 2022. URL http://arxiv.org/abs/2203.08242. arXiv:2203.08242 [cs].

[12] Horace He [@cHHillee]. I suspect GPT-4's performance is influenced by data contamination, at least on Codeforces. Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems. This strongly points to contamination. 1/4 https://t.co/wm6yP6AmGx, March 2023. URL https://twitter.com/cHHillee/status/1635790330854526981.

[13] OpenAI. GPT-4 Technical Report, March 2023. URL http://arxiv.org/abs/2303.08774. arXiv:2303.08774 [cs].

[14] Karen Sloan. New bar exam gets lukewarm reception in previews, 2023. URL https://www.reuters.com/legal/legalindustry/new-bar-exam-gets-lukewarm-reception-previews-2023-07-19/. Accessed: 2023-11-08.

[15] Audrey Ching and Donna Hershkowitz. Report from the alternative pathway working group: Request to circulate for public comment. Board of Trustees Meeting Agenda Item, September 2023. URL https://www.courthousenews.com/wp-content/uploads/2023/09/california-bar-exam-alternative-proposal.pdf. Los Angeles Office, California State Bar.

[16] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, November 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z. Number: 11 Publisher: Nature Publishing Group.

[17] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the Everything in the Whole Wide World Benchmark. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, December 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/084b6fbb10729ed4da8c3d3f5a3ae7c9-Abstract-round2.html.

[18] Rachel L. Thomas and David Uminsky. Reliance on metrics is a fundamental challenge for AI. *Patterns*, 3(5):100476, May 2022. ISSN 2666-3899. doi: 10.1016/j.patter.2022.100476. URL https://www.sciencedirect.com/science/article/pii/S2666389922000563.

[19] Fabio Motoki, Valdemar Pinho Neto, and Victor Rodrigues. More human than human: measuring ChatGPT political bias. *Public Choice*, August 2023. ISSN 1573-7101. doi: 10.1007/s11127-023-01097-2. URL https://doi.org/10.1007/s11127-023-01097-2.

[20] Sayash Kapoor. Does ChatGPT have a liberal bias?, March 2023. URL https://www.aisnakeoil.com/p/does-chatgpt-have-a-liberal-bias.

[21] Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models, August 2023. URL http://arxiv.org/abs/2308.11462. arXiv:2308.11462 [cs].

[22] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *arXiv preprint arXiv:2104.08786*, 2021.

[23] Arvind Narayanan and Sayash Kapoor. Generative AI companies must publish transparency reports, 2023. URL http://knightcolumbia.org/blog/generative-ai-companies-must-publish-transparency-reports.

[24] Transparency reports | Transparency Center, . URL https://transparency.fb.com/reports/.

[25] YouTube Community Guidelines enforcement – Google Transparency Report, . URL https://transparencyreport.google.com/youtube-policy/removals?hl=en.

[26] Reports, . URL https://www.tiktok.com/transparency/en/reports/.

[27] Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, July 2023. doi: 10.1126/science.adh2586. URL https://www.science.org/doi/10.1126/science.adh2586. Publisher: American Association for the Advancement of Science.

[28] Ashwin Nayak, Matthew S. Alkaitis, Kristen Nayak, Margaret Nikolov, Kevin P. Weinfurt, and Kevin Schulman. Comparison of History of Present Illness Summaries Generated by a Chatbot and Senior Internal Medicine Residents. *JAMA Internal Medicine*, 183(9):1026–1027, September 2023. ISSN 2168-6106. doi: 10.1001/jamainternmed.2023.2561. URL https://doi.org/10.1001/jamainternmed.2023.2561.

[29] John W. Ayers, Adam Poliak, Mark Dredze, Eric C. Leas, Zechariah Zhu, Jessica B. Kelley, Dennis J. Faix, Aaron M. Goodman, Christopher A. Longhurst, Michael Hogarth, and Davey M. Smith. Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum. *JAMA Internal Medicine*, 183(6):589–596, June 2023. ISSN 2168-6106. doi: 10.1001/jamainternmed.2023.1838. URL https://doi.org/10.1001/jamainternmed.2023.1838.

[30] Benjamin Weiser. Here's What Happens When Your Lawyer Uses ChatGPT. *The New York Times*, May 2023. ISSN 0362-4331. URL https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html.

[31] David Wagner. This Prolific LA Eviction Law Firm Was Caught Faking Cases In Court. Did They Misuse AI?, October 2023. URL https://laist.com/news/housing-homelessness/dennis-block-chatgpt-artificial-intelligence-ai-eviction-court-los-angeles-lawyer-sanction-housing-tenant-landlord. Section: Housing and Homelessness.

[32] Michael C. Dorf. Law-Specific Large Language Model Generative AI Interim Report: Lexis+AI Versus GPT-4, November 2023. URL https://www.dorfonlaw.org/2023/11/law-specific-large-language-model.html.

[33] James Vincent. OpenAI isn't doing enough to make ChatGPT's limitations clear, May 2023. URL https://www.theverge.com/2023/5/30/23741996/openai-chatgpt-false-information-misinformation-responsibility.

[34] Hon. Bernice Bouie Donald, Hon. James C. Francis IV, Ronald J. Hedges, and Kenneth J. Withers. Generative AI and Courts: How Are They Getting Along?, February 2022. URL https://www.jamsadr.com/blog/2023/francis-james-pli-generative-ai-1023.

[35] Masha Medvedeva, Martijn Wieling, and Michel Vols. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212, March 2023. ISSN 1572-8382. doi: 10.1007/s10506-021-09306-3. URL https://doi.org/10.1007/s10506-021-09306-3.

[36] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), September 2023. ISSN 2666-3899. doi: 10.1016/j.patter.2023.100804. URL https://www.cell.com/patterns/abstract/S2666-3899(23)00159-9. Publisher: Elsevier.

[37] Michael Carley, Deepak Hedge, and Alan Marco. What is the probability of receiving a us patent. *Yale JL & Tech.*, 17: 203, 2015.

[38] Kurt Glaze, Daniel E Ho, Gerald K Ray, and Christine Tsang. Artificial intelligence for adjudication: The social security administration and ai governance. 2021.

[39] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[40] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, January 2018. ISSN 2375-2548. doi: 10.1126/sciadv.aao5580. URL https://www.science.org/doi/10.1126/sciadv.aao5580.

[41] Richard Berk. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13(2):193–216, June 2017. ISSN 1572-8315. doi: 10.1007/s11292-017-9286-2. URL https://doi.org/10.1007/s11292-017-9286-2.

[42] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, October 2019. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aax2342. URL https://www.science.org/doi/10.1126/science.aax2342.

[43] Jeff Keltner. The Importance of Variables in Assessing Credit Risk. URL https://info.upstart.com/importance-of-variables-blog.

[44] Todd Feathers. Major Universities Are Using Race as a "High Impact Predictor" of Student Success – The Markup, March 2021. URL https://themarkup.org/machine-learning/2021/03/02/major-universities-are-using-race-as-a-high-impact-predictor-of-student-success. Section: Machine Learning.

[45] Nathan Mondragon. Creating AI-driven pre-employment assessments, 2021. URL https://www.hirevue.com/blog/hiring/creating-ai-driven-pre-employment-assessments.

[46] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. Against Predictive Optimization: On the Legitimacy of Decision-Making Algorithms that Optimize Predictive Accuracy. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 626, New York, NY, USA, June 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594030. URL https://dl.acm.org/doi/10.1145/3593013.3594030.

[47] Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin Van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, April 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1915006117. URL https://pnas.org/doi/10.1073/pnas.1915006117.

[48] Alissa Fishbane, Aurelie Ouss, and Anuj K. Shah. Behavioral nudges reduce failure to appear for court. *Science*, 370(6517): eabb6591, November 2020. doi: 10.1126/science.abb6591. URL https://www.science.org/doi/10.1126/science.abb6591. Publisher: American Association for the Advancement of Science.

[49] 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. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine*, 181(8): 1065, August 2021. ISSN 2168-6106. doi: 10.1001/jamainternmed.2021.2626. URL https://jamanetwork.com/journals/jamainternalmedicine/fullarticle/2781307.

[50] Alex C. Engler. Independent auditors are struggling to hold AI companies accountable, January 2021. URL https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue.

[51] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 666–677, New York, NY, USA, March 2021. Association for Computing Machinery. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445928. URL https://dl.acm.org/doi/10.1145/3442188.3445928.

[52] Ethan Corey. How a Tool to Help Judges May Be Leading Them Astray, August 2019. URL https://theappeal.org/how-a-tool-to-help-judges-may-be-leading-them-astray/.

[53] Paris Martineau. Toronto Tapped Artificial Intelligence to Warn Swimmers. The Experiment Failed, 2022. URL https://www.theinformation.com/articles/when-artificial-intelligence-isnt-smarter.