Digital Society (2022) 1:14
https://doi.org/10.1007/s44206-022-00017-z

ORIGINAL PAPER

Algorithmic Bias and Risk Assessments: Lessons from Practice

Ali Hasan[1,2] · **Shea Brown**[2,3] · **Jovana Davidovic**[1,2] · **Benjamin Lange**[2,4] · **Mitt Regan**[2,5]

Received: 2 March 2022 / Accepted: 1 August 2022 / Published online: 19 August 2022

Abstract

In this paper, we distinguish between different sorts of assessments of algorithmic systems, describe our process of assessing such systems for ethical risk, and share some key challenges and lessons for future algorithm assessments and audits. Given the distinctive nature and function of a third-party audit, and the uncertain and shifting regulatory landscape, we suggest that second-party assessments are currently the primary mechanisms for analyzing the social impacts of systems that incorporate artificial intelligence. We then discuss two kinds of assessments: an ethical risk assessment and a narrower, technical algorithmic bias assessment. We explain how the two assessments depend on each other, highlight the importance of situating the algorithm within its particular socio-technical context, and discuss a number of lessons and challenges for algorithm assessments and, potentially, for algorithm audits. The discussion builds on our team's experience of advising and conducting ethical risk assessments for clients across different industries in the last 4 years. Our main goal is to reflect on the key factors that are potentially ethically relevant in the use of algorithms and draw lessons for the nascent algorithm assessment and audit industry, in the hope of helping all parties minimize the risk of harm from their use.

Keywords AI ethics · Algorithmic bias · Responsible AI · Algorithm audit · Ethical risk assessment

1 Introduction

In this paper, we distinguish between different sorts of assessments of algorithmic systems, describe our process of assessing such systems for ethical risk, and share some key challenges and lessons for future algorithm assessments and audits. We draw from our team's experience of advising and conducting ethical risk assessments

✉ Ali Hasan
Ali-hasan@uiowa.edu

Extended author information available on the last page of the article

for clients across different industries in the last four years. Our main goal is to reflect on the key factors that are potentially ethically relevant to the use of algorithms, and draw lessons for the nascent algorithm assessment and audit industry, in the hope of helping all parties minimize the risk of harm from their use.

What do we mean by an "algorithm"? In a very broad sense, an algorithm is any decision procedure, any set of rules or instructions, for solving a problem or performing a task. By an algorithm, we mean, more specifically, such a decision procedure or set of instructions as carried out by a computer. While all computers involve algorithms (programs), humans are increasingly using computers or computer systems that involve machine learning (ML). Roughly, these are computer systems that can improve their performance over time, and in this sense "learn," with the help of large amounts of data. Though there are various forms of ML, at a high level of abstraction, they all involve algorithms that come up with and revise "models"—other algorithms—to solve a problem or perform a task more successfully. These computer systems can be significantly more powerful and are being used increasingly to make, or help us make, important decisions in education, transportation, hiring and performance review, finance, social media, criminal justice, the military, medicine, etc.

In a related point of terminology, we use the term "Artificial Intelligence" or "AI" to stand roughly for any machine or computer system capable of doing things that normally require intelligence and reflection—thinking, learning, or problem solving—when done by humans.[1] Algorithms that employ ML thus count as a form of AI. Strictly speaking, a machine or computer that does not employ ML can achieve the level of performance (in a limited domain) that matches or surpasses that which normally requires intelligent reflection on the part of humans—in playing chess, for example. But many systems that rely on ML outperform programs written by humans. Since the systems we have assessed employed or relied on ML to some extent, they count as forms of AI. However, non-AI features of the technology or service were examined and considered as well; the assessments we conducted paid special attention to, but extended beyond, the AI features of the system. When we discuss algorithms and algorithmic systems below, we generally mean those that are or are embedded in an AI system, though we recognize that, strictly speaking, not all algorithms rise to the level of AI.

Our main goal, to repeat, is to draw lessons from our experience of conducting assessments of such systems, lessons for the algorithm assessment and audit industry, in the hope of helping all parties minimize the risk of harm from the use of algorithms. The term "use of algorithms" underscores that the focus of such an assessment is the broader socio-technical context of the algorithm, how the algorithm is employed to serve certain purposes of an organization, and how it affects the rights and interests of stake-holders—including whether it is unfair or biased in some way (Selbst et al., 2019; Brown et al., 2021). Such assessments go beyond, but guide and in turn rely upon, the more technical assessment or testing of the algorithm itself for such things as accuracy, bias, and interpretability. We use the term "ethical risk assessment" or simply "ethical

[1] See, e.g., Liao (2020, 3), and Russell and Norvig (2010, 2) for similar definitions.

assessment" in the remainder of this article to refer to this broader focus.[2] As testing for algorithmic bias is typically the most central task in the more technical assessments and audits, it is that aspect of them that we focus on in our discussion below of "algorithmic bias assessments."

Different frameworks, at various levels of abstraction, have been proposed for such assessments (e.g., Mökander & Floridi, 2021; Moss et al., 2021; Selbst, 2021), and some organizations have begun the hard work of adapting and implementing them for particular contexts.[3] However, there is little consensus on what an ethical assessment of algorithms should look like, and no accepted set of standards for conducting them. Moreover, holistic ethical assessments of the use of algorithms—i.e., assessments that are broad in scope and attentive to the complexities of the sociotechnical context—remain rare, and case studies of such assessments are rarer still.[4] And it is difficult to find any public or academic documentation and discussion of them by those who have conducted such assessments. Reflections on algorithm assessments guided by the experience of actually performing them is critical to the development of improved ethical risk management of algorithmic systems. To our knowledge, our article is one of the first in the current literature to do this.

Our discussion of our approach is a novel contribution to the literature in that (a) it separates the identification of potential harms and their ranking or prioritization as distinct stages in the ethical risk assessment, (b) it distinguishes between ethical risk assessments and algorithmic bias assessments and highlights the interplay and feedback between them, and (c) to repeat, it is one of the first discussions of lessons and recommendations for ethical risk assessments guided by actual practice. We are not claiming here that no one who has actually conducted ethical risk assessments has used a similar approach; indeed, we think that structuring assessments in this way makes a great deal of sense, and so, it would not surprise us if some others have used at least roughly similar approaches. But it is difficult to find any discussion in the literature that shares or recommends the approach captured by (a) and (b).

A note on our discussion of examples: Due to our being bound by client confidentiality agreements, the examples we use below for illustration are not from actual assessments conducted. However, central features of the examples, and the related lessons and recommendations, are based on assessments that were actually conducted, and on our collective reflection on these assessments.

We start by distinguishing audits, assurances, and assessments (Sect. 2). We then present our assessment process, for both the broader ethical risk assessment and the

[2] We use "ethical risk assessment" rather than "ethical impact assessment" because the latter naturally suggests the actual impact or consequences of the use of an algorithm, while the former covers all ethical risks, whether they come to be realized or not. We recognize, however, that "impact" is often used in the broader sense that covers risk.

[3] See, for example, the Canadian government's (2021) algorithmic impact assessment tool: https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

[4] An exception is a case study by Mökander and Floridi (2022). They provide a detailed case study based on an observation of an "ethics-based audit" of AztraZeneca conducted by a third party (not by the authors).

more technical bias assessment (Sect. 3). In the final sections, we explain how these parts of the assessment depend on each other in important ways (Sect. 4) and draw some other lessons for risk assessments and, potentially, audits (Sect. 5).

2 Taxonomy of Audits, Assurances, and Assessments

It is important to first clarify the key differences among the notions of audit, assurance, and assessment, as we use these terms. These terms are used in different ways in the literature. For example, "algorithm audit" is sometimes used in a broad sense that covers any assessment of an algorithm, but we use it here in a sense that is closer to the use of "audit" in finance where it is understood as an official examination of an organization's accounts by an independent body. The taxonomy provided here follows the work of Carrier and Brown (2021).

An audit is an independent assessment or evaluation of an organization's algorithm, using transparent rules or laws, and is intended to serve society, the public, users, or some other body independent of the evaluated organization. They thus involve three parties—the auditor, the organization being assessed, and those on whose behalf the audit or assurance is conducted.[5] An assurance is the same, except that audit criteria are constructed to yield a binary output, i.e., output indicating that the system is compliant/non-compliant with respect to rules or laws, while assurances can involve the application of criteria to yield non-binary output (e.g., a score or grade, and/or a qualitative evaluation). Assurances can be guided by "soft-law"—principles, rules, or standards that are not legally binding or not codified in law. The latter standards, while important and informative, can be more difficult to apply objectively and consistently. Examples include OECD's AI Principles[6] and the US Department of Health and Human Services' (2021) Trustworthy AI Playbook.[7] Examples of audits include upcoming regulation such as the New York City Council's (2021) law requiring "bias audits" for companies using hiring algorithms and the EU Artificial Intelligence act which will come into effect in 2023, as well as the recently revived Algorithmic Accountability Act (Wyden et al., 2022), which explicitly requires companies using "automated decision systems" to perform and disclose impact assessments.

By contrast to audits and assurances, assessments are non-independent, internal, or second-party evaluations of an organization's algorithm and intended as a service to the organization. Like assurances, they are non-binary in nature. They are aimed at providing feedback and usually at building recommendations and advising clients on how to perform better with respect to some legal and/or ethical standard. These assessments can have both technical and non-technical components (Mökander &

[5] Carrier and Brown's (2021) taxonomy also includes an "internal audit" which is carried out by a group or unit working independently within and in service of the organization rather than society or some other party but otherwise has the characteristics of an audit.

[6] https://oecd.ai/en/ai-principles

[7] https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

Floridi, 2021); the most common examples are technical assessments of bias[8] and ethical risk or impact assessments (Brown et al., 2021).

Ethical or ethics-based assessments involve the application of ethical values and principles that extend beyond legal compliance. As we discuss further below, unlike audits and assurances, assessments allow for extended consulting and advice, and a more collaborative engagement tailored to the client's needs and context. Ethical assessments tend to focus on the negative impacts, and so tend to take the form of "ethical risk assessments," because those are the impacts that organizations and regulators are primarily interested in identifying and limiting. However, positive impacts should not be completely ignored. Positive benefits such as the expansion of education and job opportunities or improved health and well-being to users and to society are important to recognize and sustain. Moreover, even in the context of ethical risk management, these sorts of positive impacts can be important. For example, significant benefits to improvement of patient care or the safety of travelers could justify some modest invasions of privacy in medicine and public transport respectively.

There are few laws or settled criteria specific to the use of algorithms, and the regulatory landscape is likely to change in uneven and complex ways. Assessments are therefore currently the primary mechanisms for analyzing the social impacts of existing algorithmic systems; they have a broader purview and can be conducted in ways that incorporate anticipated requirements of nascent and future audits or assurances while also being tailored to the institution. But there is more to be said. We return to the comparison of second-party risk assessments and third-party audits and assurances in discussing lessons learned in the final section.

3 Ethical Risk Assessment and Algorithmic Bias Assessment

3.1 Ethical Risk Assessment

By an ethical risk assessment of an algorithm, we mean an assessment of the risk that the use of the algorithm negatively impacts the rights and interests of stakeholders, with a corresponding identification of situations of the context and/or features of the algorithm which give rise or contribute to these negative impacts. Possible negative impacts include harm to one's physical body or psychology, damage or loss to one's property, and infringements or undermining of moral rights (whether or not they are enshrined in law) such as rights to privacy, autonomy, freedom of speech and expression, and the right to fair and non-discriminatory treatment. It might include impacts to other central interests of stakeholders, such as relations of trust between stakeholders. In talking of risk of harm below, we use the term "harm"

[8] See, for example, the Ada Lovelace Institute (2020) report: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

broadly (more broadly than some philosophers do) to cover any such impacts to one's interests or rights, including unfair or discriminatory treatment.[9]

A stakeholder relative to the use of some technology is, quite simply, any individual, organization, or group (including society at large) whose interests or rights could be affected, positively or negatively, by the technology—any who might have something at stake when it comes to the use of technology in a particular context. The assessment is an *ethical* or *ethics-based* risk assessment in the sense that it is primarily concerned with the impact on the central interests, well-being, and moral rights of any individuals, groups, or institutions that might be affected. Some of these risks may overlap with or have implications for compliance and legal risks, and the assessment may take some of this into consideration. However, ethical assessments are not primarily concerned with these risks and will tend to go beyond them.

The ethical risk assessment should survey, organize, and assess the significance of a range of potential harms that could occur from use of the algorithm or AI technology. We separate the ethical risk assessment into two main stages: the identification stage, which is concerned with identifying the possible or potential harms as broadly or comprehensively as possible, followed by a prioritization stage, which evaluates these potential harms to determine which are most significant. The separation of these stages aids in (a) ensuring that nothing of potential significance is simply missed or overlooked, while (b) guiding the assessors' decisions regarding what potential harms to assess and advise on. This is important, as constraints of time and resources often preclude that all potential harms and their sources can be assessed and studied to the same extent.

3.1.1 Identification

The first step in an ethical risk assessment is to identify key stakeholders and the potential harms that use of the product could cause to them. Our assessments typically start with a careful examination of a range of academic and media sources to generate a preliminary list of potential harms. This is complemented by conversations with key stakeholders, including customers or end-users, designers, developers, support teams, and others in the organization, to identify the broadest possible range of harms that could occur from use of the product. To that end, our desk research typically contained both academic research that discussed technology similar to that being assessed and media sources reporting various experiences or concerns from use of the product. The aim of this first stage is *not to evaluate potential harms*, assessing their significance, but to ensure that we identify possible or potential harms as broadly or comprehensively as possible. Many of these may, at the prioritization stage which we discuss next, be set aside as being relatively insignificant or extremely unlikely.

[9] In some cases, one might be unfairly treated or discriminated against without being worse off than one otherwise would have been, and at least in this sense one might not be "harmed" by unfairness. We use 'harm' in a broader sense that includes such cases of unfair or discriminatory treatment.

The identification of potential harms for end users, or members of society on whom algorithms are deployed, is paramount, since the risk of harm is typically highest and most direct for them (e.g., job applicants, housing or credit-card applicants, passengers of autonomous vehicles, students and test-takers, defendants in the justice system, consumers of online media and news), with risks to other stakeholders (algorithm developers and vendors, public or private organizations using the algorithms, regulatory bodies, society at large) depending largely on risks to this group. That said, it is important in the identification stage to not assume that this is true, and so to be on the lookout for independent risks to other stakeholders.

3.1.2 Prioritization

The next step in our process is to categorize all the potential ethical risks according to the underlying features of the product or technology that is the primary driver for that risk, while also determining how essential or important these underlying features are to the ultimate purpose of the algorithm. The latter can help us identify the features, if any, that can be dispensed with or altered to remove or reduce sources of risks without compromising the integrity or purpose of the system. We then assess the magnitude of the potential harm and the likelihood (and frequency) of it occurring that would result from use of the algorithm and take the product of these values to estimate the expected risk for each type of harm. We do the same for the use of the most relevant alternative, if any, for achieving the algorithm's goals. In other words, we not only assess the significance of the risks of harm of the AI system but also of the risks that the closest relevant AI-free alternative might have. This helps us identify risks that are unique to use of the product or service and determine which should be prioritized over others.

For example, and at a risk of oversimplifying, the risk of an autonomous vehicle hitting pedestrians in a certain time frame might be significantly lower than the risk of human-operated vehicles hitting pedestrians. This would suggest that when it comes to at least one type of harm, physical harm, the autonomous vehicles might perform better than the alternative, and that the risk of such harm is comparatively less significant. Nonetheless, if the distribution of the physical harm is biased towards, for example, people of color (due to bias in the computer vision systems) or biased towards particular neighborhoods, as compared to similar harms in non-autonomous vehicles, then, such harms ought to be counted as significant. To further illustrate, we might consider comparing the sorts of harms that arise from using a hiring algorithm to similar harms in hiring environments that do not use such algorithms in their hiring decisions. To prioritize harms then, we do not simply measure their likelihood and magnitude, but we also compare them to relevant alternatives.

Our ethical risk assessments complement and are complemented by algorithmic bias assessments. More will be said about the interplay between the two, but first, we turn to a quick sketch of the bias assessment.

3.2 Algorithmic Bias Assessment

The presence of bias in AI systems, especially those that employ ML, has been a significant cause of harm in recent years (Buolamwini & Gebru, 2018; Mehrabi et al., 2022). Detecting and mitigating bias in AI systems are often the first need that motivates companies to seek external assessments, and usually the main focus of technical assessments and audits of algorithms. Indeed, many seem to treat bias assessments as synonymous with "Responsible AI."

Due to the prominence of such worries, bias assessments have been a key part of our risk assessment framework, and we focus on them here. However, it is worth noting that other technical assessments extending beyond algorithmic bias can be ethically relevant, and crucial in some contexts (e.g., assessments of transparency of architecture, explainability of outputs, and vulnerability to hacking). Moreover, as we discuss further in Sect. 4, conducting bias assessments without first understanding the salient ethical risks runs the risk of missing the critical perspective needed to truly detect and mitigate harm. Structuring the assessment process as we do respects the value-ladenness of bias assessments and allows for a feedback loop between the ethical risk assessment and the bias assessment.

While the nuances of bias testing may be different for every new algorithmic system, our approach follows a three-step process:

1. **Determining the targets of evaluation:** We work with the client to define the scope of the assessment, including which algorithms and datasets will be used, and how much of the client's current testing can be incorporated into the report. We also agree upon which elements of the evaluation will be made public, if any.
2. **Conduct the ethical risk assessment:** Once the target of evaluation and the report outputs have been defined, we conduct the ethical risk assessment as described in Sect. 3.1. Based on the results of the risk assessment, we may recommend new testing or reporting due to new risks that were uncovered. This will be discussed and agreed upon with the client. Once this agreement has been made, the scope of the bias assessment and the report will not be changed, which is meant to maintain some level of independence.
3. **Conduct the bias assessment:** We conduct the agreed assessments for bias or disparate impact through a combination of direct testing and verification, and documentation of the client's own testing.

Most of the technical aspects of the bias assessments are beyond the scope of this article and vary widely depending on the type of algorithm, e.g., computer vision, natural language processing, and recommendation systems. However, two common features are the need for adequate testing data and clear definitions of testing metrics:

Data collection: In order to test for bias, we need data that meet a number of requirements. The testing data has to (1) mimic, to the extent possible, the conditions under which the algorithm is deployed, (2) be labeled by self-identified race, gender, and

other protected attributes of interest in order to construct intersectional groups, and (3) contain sufficient number of datapoints in the regions of the parameter space that are ethically salient, to obtain statistically significant test results. One important issue relevant to data collection has to do with what some call the problem of "ground truth" in ML. The problem generally is how to ensure that one's testing data (and often the training data too, in supervised and semi-supervised learning) is labeled or categorized correctly—that it provides an objective test and is anchored in reality or in the "ground truth." For example, we cannot test an algorithm's reliability in detecting a disease without having a good, independent way of determining whether the disease is present. Moreover, we cannot tell how good the algorithm is at detecting the disease *across different racial groups* without having an independent way of determining what racial group the people in the testing data belong to. Race, gender, and other such attributes are, in a sense that is difficult to make precise, "socially constructed." But they obviously are not merely arbitrary or purely subjective, and as such, even here, there is a "ground truth" so to speak that one needs to respect, and risk of bias and misclassification that one will want to avoid. It is thus important to take care that we have a reliable way of applying these labels to the testing data. Labeling based on *self-identified* social attributes seems ideal here; other things being equal, individuals are in the best position to make these sorts of determinations for themselves. At the same time, we need to be cautious of the risks to privacy and use of data beyond subject expectations. Fair and effective demographic data collection is thus no simple matter.[10]

Testing metrics: We must also define appropriate metrics for our testing. These metrics are determined in step 1 of our process above based on client needs and our best guess for the relevant harms that may occur, then reexamined after the ethical risk assessment. Examples range from simple measurements of differential false-positive and false-negative rates between groups, to more nuanced group fairness metrics that track particular notions of discrimination.[11] We will now discuss how this process is affected by, and in turn affects, the ethical risk assessment.

4 Interplay between Ethical Risk Assessment and Algorithmic Bias Assessment

4.1 How the Ethical Risk Assessment Informs the Bias Assessment

Our experience has impressed upon us an important point that applies to algorithm assessments in general: the choices that must be made when deciding to test an algorithm—what specifically to test for and how to go about it—must be guided by an initial assessment of ethical risk. This should not be that surprising, as it

[10] For more on this, see Andrus and Villenueve (2022) and Benjamin (2019).

[11] For example, see Watcher et al.'s (2021) "conditional demographic disparity," and IBM Research's list of metrics on the AI fairness 360 site: https://aif360.mybluemix.net

reflects the fact that the design of an algorithm for a specific purpose in a particular socio-technical context is itself value-laden, and the unique features of this system that give rise to risks are not obvious absent the detailed ethical analysis described above. While the value-ladenness of algorithm design and development has been discussed in the literature,[12] its bearing on approaches to structuring and conducting assessments has received much less attention.

Understanding the context and purpose of the algorithm, the different normative or value-laden assumptions being made, and the potential risks of harm or bias from the use of the algorithm in this context, are crucial for determining what the technical bias assessment should be testing for, and how best to go about it.

Let us take as an example software used to help make hiring decisions by screening and ranking candidates that are qualified and fit for a particular job. It might use natural language processing to parse candidate resumés (and perhaps other application materials, online profiles, etc.) that vary in style and structure and feed the information into a ranking algorithm. Knowledge of the purpose, and specifically of the required or desired qualifications for the job, is relevant to determining which features of the data are likely to be good or reliable indicators of these qualifications—and, importantly, whether they are appropriate and fair. Knowledge of the stakeholders and context, including the ways in which the technology and various users interact, can help identify ways that the algorithm might perform poorly with harmful consequences, and invite us to ask: can we test for that? One cannot test for everything, and the initial ethical risk assessment can help one decide what sorts of questions to ask about the algorithm, and what possible errors and biases to test for. All this will in turn inform one's selection of testing data and testing metrics.

Various disparities in the training data could lead to a biased algorithm, one that relies, for example, on race, gender, or socioeconomic class—or rather, proxies for them—in screening and ranking applications. So, it is important to be on the lookout for unobvious proxies of race, gender, or socioeconomic status that ML algorithms may be relying on without our knowledge—e.g., school names, hobbies and interests--within the resumés. The testing data should therefore include resumés of applicants that vary in self-identified race, gender, etc., and the assessment should test the ranking of applicants across different groups, including testing that screens off disparities in other features that are directly related to qualifications for the job—e.g., comparing how the algorithm ranks applicants of different race or gender who are equally qualified, or whose applications are otherwise identical. In an ideal scenario, the testing data should include real resumés of individuals that are likely to apply to such jobs, and not fabricated or artificial resumés that could, in subtle ways, introduce artificial elements or yield an unrepresentative data set, and it should be large and varied enough to allow for testing of the algorithm's sensitivity to specific features. However, in many cases, it is not possible to do all of this for a variety of reasons, including limited access to real resumés that are appropriately labeled with demographic data, something that is especially true for small vendors or startups. It

[12] For some discussion of this point, see Dotan (2021) and Fazelpour and Danks (2021).

is thus critical in these situations to narrow down the possible failures in the algorithm and tie them to potential risks, and to conduct testing and create synthetic training data that targets those failures.

There are various metrics one might focus on, including false positives, where the algorithm gives a high ranking or an "interview" recommendation for low-quality applications, and false negatives, where the algorithm gives a low ranking or "don't interview" recommendation for high-quality applications. (One will obviously need some independent, unbiased way of assessing or ranking applications in the testing data, which is no trivial matter!) An overall high false-negative rate might be concerning given that the purpose is to find highly qualified candidates and not miss them, and given that such results would unfairly block many qualified individuals from consideration. On the other hand, a high false-positive rate for high-stakes jobs in society (e.g., in engineering or medicine) should be a serious concern. But it is important—and more work—to also compare these rates across different relevant social groups: a comparatively high false-negative rate for persons of color, for example, would be particularly concerning.

We thus see how the choice of groups relative to which the algorithm should be assessed, the choice of metrics to compare, and other choices having to do with the testing data and its quality are all guided by the ethical risk assessment.

4.2 How the Bias Assessment Informs the Ethical Risk Assessment

The bias assessment in turn informs the ultimate assessment of the main ethical risks and helps guide proposed recommendations to the client. The bias testing enables us to assess whether and to what extent varying specific features of the data—e.g., in the hiring algorithm example, varying features of the resumés—affects the performance of the algorithm, and we can then incorporate these results into the overall assessment and comparison of types of harm. And we can also compare the results with reports of harm or unfairness of the algorithm from users, through interviews and/or reports in the news and social media, and use this to suggest strategies for transparent and clear communication with users and other stakeholders. Specifically, in some cases, the anxiety and lack of transparency or understanding around the use of algorithms might also create potential risks, but the remedies to such risks do not come from technological solutions or better algorithms, but transparency and clarity around the use of the algorithm. We elaborate a bit more on transparency in the next section.

5 Lessons Learned and Problems to Anticipate

In this section, we share some of the key lessons and recommendations. We begin with a brief discussion of some general lessons for developers, providers, or deployers of AI. This is followed by lessons about the assessment process itself, which apply to the assessment and audit industry.

5.1 Lessons for Development and Use of AI

One of the main lessons when it comes to the use of AI technology is that misunderstanding and ignorance about an AI system can be a major source of potential harm and therefore ethical risk. The lack of knowledge can make users uneasy and distrustful, and easily lead to misunderstanding how the technology works, and to falsely ascribing a broad range of problematic features to the product. The plethora of negative press and social-media attention to the use of AI in various contexts to track, monitor, profile, evaluate, and influence encourages users to (quite reasonably!) distrust the use of AI in the particular setting in which they encounter it. This ignorance and misunderstanding can lead to an increase in user anxiety and distrust, and a reluctance to make use of products even when they are beneficial, safe, and fair. The lack of relevant knowledge can also lead to misinterpretation and misuse of the technology, not only by end users and society but by companies and employees that develop and deploy them.

The dangers of ignorance and misunderstanding point to three related recommendations that are likely to apply quite broadly to use of AI.

First, it is crucial that there be a high degree of transparency about the technology (and key elements of the socio-technical system) and its intended and actual use. Exactly what to make transparent, and to which stakeholders, will naturally vary from context to context. But it helps to distinguish between different kinds of transparency[13]:

(i) Transparency of architecture: Availability of knowledge about the algorithm itself, its range of inputs, and outputs, whether it involves ML, is a neural net, what is the code, or the weights of the network, etc.
(ii) Explainability and interpretability: Ability to explain why and how the algorithm generates the outputs that it does.
(iii) Transparency of use: Knowledge of the fact that an algorithm is being used, and for such-and-such a purpose; this includes some relatively non-technical knowledge of points of contact with the system (e.g., that it uses one's visual input from a camera, textual input typed into a computer).
(iv) Transparency of data use and collection: Knowledge of what sort of data is collected, how long it is stored, and for what purpose.

(i) and (ii) tend to be important and useful goals for those involved in developing, improving, and deploying algorithms, while (iii) and (iv) pertain more to end-users and society.

There are serious limits to the achievement of some forms of transparency. Consider explainability of outputs (ii). It is epistemically relevant, i.e., relevant to assessments of accuracy. And in some contexts, a robust level of explainability may be needed to satisfy an important non-epistemic demand, such as to help set up

[13] See Brown et al. (2021) for a similar list. For more on transparency and its ethical relevance, see Basl et al. (2021).

chains of accountability, filling in "responsibility gaps," or to play normative roles in the legal system (Mathias, 2004; Baum et al., 2022; Brennan-Marquez, 2017). But the ability to explain specific outputs is a highly demanding requirement for some algorithms; in many cases, we can arrive at idealized, approximate, partial explanations *at best* (Mittlestaudt, 2019). And explainability is arguably not required in some cases; an algorithm might be highly reliable and verified to be so, and this could justify its use in some contexts (e.g., in diagnosing a disease) despite its being unexplainable (Zerilli et al., 2018). Matters can thus get very complicated. However, organizations using AI technology should err on the side of more transparency of kind (i) and (ii) at least on the part of experts involved in developing, improving, and deploying such systems. Any reliance on systems that are opaque and unexplainable even to experts should be examined carefully and honestly, to make sure that they are well justified.

Such transparency will need to avoid compromising the central mission or purpose of the system in which AI is used (e.g., by exposing vulnerabilities or ways to game or abuse the system), or divulging the AI vendor or institution's proprietary assets. This could help justify restricting or limiting transparency of architecture and explainability from public view. Moreover, a deep or detailed level of transparency of these forms is not useful to most end users and might cause more confusion. Indeed, some have argued that users are rarely equipped to understand disclosures even of sorts (iii) and (iv), often interpret them in biased or fallacious ways, and typically ignore them even when they might be useful.[14]

This leads to the second point regarding ignorance and transparency: AI developers and providers should share *clear and effective communication* to the public and end users, both directly and through client institutions (if applicable) and their staff, about the system, what it does and does not do, its benefits and risks, and the resources and technical support available. This is no easy task, as simply sharing complex disclosures will not do. But given that (in our experience) harms to users were often due to misunderstandings that could have been avoided by clear communication at critical points, it is a task worth the effort.

This clear and effective communication will often require the coordination of different units and stakeholders, at different stages in the product journey, in order to determine what is helpful to share when, and in what form. This leads to our third point: *the need for training*. Teams involved with developing, maintaining, deploying, or providing support for an AI-driven product or service should appreciate and build a thorough understanding of the sources of risk, including likely misunderstandings of the technology, and receive training that is tailored to their particular function or role in the organization. Staff should be provided with clear red lines and best practices, and be trained to look for, identify, and manage potential ethical risks, and in turn train client institutions to do the same. Moreover, it is important that appropriate training be given to

[14] See, for example, Ben-Shahar and Schneider (2014). Thanks to an anonymous referee for raising this important point.

end users as well, about how to use a product or service, and to not merely rely on disclosures and assume that the users will understand and take the guidance seriously. In some cases, for example, a short and user-friendly training module or onboarding process may be required before being allowed to use the product or service.

Another key lesson has to do with how AI is fielded. Organizations should consider carefully their reasons for using AI for augmenting or replacing human decision-making and should have clear frameworks for deciding when, whether and how to incorporate AI into their decision-making. Rushed fielding of AI that might not be carefully considered, or fit for purpose, could cause significant harm both directly as well as indirectly by causing the kinds of misunderstanding and anxiety discussed above. Such frameworks for decision-making ought to be modular, easy to use, and reflect the purpose of the organization.

The reference to a need for organizational frameworks for the use and deployment of AI points to another key lesson, namely, that organizations ought to develop key functions whose primary role is to oversee and consider AI with respect to its ethical, legal, and reputational risks. Whether these key functions are housed under a Chief AI Ethics Officer, an internal committee, or an oversight committee will depend on the particular systems of that organization. But having AI ethics be a functional responsibility of someone in the organization is a necessary step for minimizing ethical harms of the use of AI. Those that occupy such roles must have an ability to understand harms beyond simple compliance risks and must be able to engage with key stakeholders, including developers—i.e., they ought to be able to understand the basic technical language.

It may help to say a bit more about why a governance framework that includes a Chief AI Ethics Officer, internal committee, or some similar body, is so important. This is something that has been motivated in detail by others (e.g., Sandler & Basl, 2019). Particularly in large organizations, there is often a compartmentalized structure or culture, with complex relations between units. The ethical risks that could arise from the use of AI are, as we have seen, varying in nature and source, and identifying and addressing them often requires the cooperation of more than one unit. Moreover, the field of AI is rapidly changing as is the regulatory landscape and best practices are quickly evolving. Having a Chief AI Ethics Officer or ethics committee ensures that someone in the organization has the authority and responsibility, and some degree of independence, to stay informed about ethical standards and best practices; help draft internal and public-facing principles and handbooks; ensure the cooperation of different units in investigating problems, overseeing assessments, and addressing risks; and provide or arrange for appropriate training. Institutions using AI at scale need a forward-looking internal AI-governance structure to operationalize processes for evaluation and review of AI development and use.

Holistic assessments of ethical risk of algorithms should not be treated as optional; they are absolutely critical. (Some of the lessons in the next section also support this.) Accepting this, and organizing institutions using AI to facilitate such assessments, is crucial to mitigating ethical, compliance, and reputational risk.

5.2 Lessons about the Assessment or Auditing Process

We turn now to lessons about the assessment process, starting with a comparison of assessments and audits.

Recall that (second party) assessments need not focus on or be limited to established criteria as in the case of an audit, and typically go beyond an audit in serving an advisory role for the client. In contrast to an auditor, the assessment team and client can work together in a more transparent and cooperative fashion and are better able to tailor the assessment to the specific product in its socio-technical context. They can guide the assessment process with an eye towards providing concrete recommendations for improvement and for addressing ethical risks, including risks that go beyond an audit's purview. While it is important to have proper audits guided by independently established criteria, the development and evolution of these criteria, and guidelines for their proper application, can benefit from an examination of second-party assessments. Auditing teams may benefit from learning about broader or more comprehensive assessments and their findings, especially at the early stages of auditing and as AI technology itself continues to change. A second-party assessment team can in turn build on a prior audit in providing broader assessment and advice to its client and work closely with the client in preparing for and complying with audit requests. The advisory role and more expansive nature of assessments make it likely that, even after regulations are in place and audits of algorithmic systems become a regular affair, assessments will have a very important role to play.

As already discussed, no assessment of the use of an algorithm is complete without a robust understanding of how the algorithm works, and technical tests of the algorithm for accuracy and bias. Beyond this, however, it is important to focus not simply on the technology, but on the agents and organizational processes involved in its use. A clear understanding of the product journey from vendors to client institutions to end users can help assessors identify points at which it is possible to reduce various types of risk.

Another general lesson is the importance of identifying stakeholders in an expansive way, so as to include all those who are affected by and may in turn affect the technological system, and distinguishing their different roles, responsibilities, and any power disparities or vulnerabilities at play. Stakeholders are an essential part of the socio-technical system being assessed. Identifying stakeholders will guide one's discovery of potential harms. For example, learning about the demographic distribution of job applicants can highlight specific biases to test for, and interviewing them about their experience in the application process can reveal perceived harms and concerns that might be unobvious to others. Learning about the individuals in the hiring firm who are responsible for narrowing the selections based on the algorithm's guidance, how they interface with the technology, and what sort of training (if any) they have received can suggest ways they might misunderstand or misinterpret the algorithm's outputs and places where human bias can enter. And interviewing them can reveal important concerns and vulnerabilities with the system and the ways that they interface with it.

Less obviously perhaps, identifying potential sources of risk in the algorithmic system can help identify, or further specify, the stakeholders who might be affected.

For example, suppose that a facialrecognition algorithm used by law enforcement sometimes mistakes innocent persons for suspects or persons of interest. Learning that the system relies heavily on a database of mugshots populated primarily by persons of color suggests that such persons may be particularly vulnerable to unfair treatment. Or, to give another example, discovering that an algorithm used by an autonomous vehicle is poor at categorizing smaller bodies, or poor at tracking their behavior once they move behind and are occluded by vehicles or other objects, could raise risks to the safety of bicyclists, pedestrians, and children.

A knowledge of the structural or institutional features of the context, including the different stakeholders and their roles, is essential to uncovering impediments to and resources for ethical risk mitigation. As such, timely and effective access to and cooperation of various stakeholders, including key decision makers, developers, and users, is needed for the assessment work to proceed efficiently and effectively. An overly compartmentalized organization structure or culture, or lack of cooperation between various units in assisting the assessment process, can lead to incomplete assessments, overlooking serious risks, and failures to affect positive change. Unlike audits (and assurances), second-party advisory assessments allow for a more open and cooperative relationship with the potential to affect deep, positive change. On the other hand, these assessments, even when the reports are made public, do not provide the level of assurance of independence and objectivity of a third-party audit or assurance.

There is a related lesson here, about the value of the commitment of the company or organization to the assessment and to taking its output seriously. It is important that the organization and/or relevant parties take ownership of or responsibility for the product and its effects and are committed to the implementation of recommendations for risk mitigation. As this is a newly developing industry, with a very complex type of technology at its core, organizations are not automatically equipped or structured to implement even modest recommendations. In order to bring about impactful change, the client must have a clear plan and be able to situate action items and recommendations in relevant functional areas, and in collaboration determine how each can be taken forward in the organization. Without serious commitment from the client, the natural challenges and impediments to change will be difficult to overcome, and the assessment and advice may come to nothing. A company that is committed to the assessment process and to the mitigation of ethical risks should ensure that the assessment team is assisted by a responsive and connected project manager, and that it has access to and the assistance of high-level officers (CEO, Chief AI Officer or Chief Ethics Officer if any, etc.) who can discuss and execute the required directives, and ensure the cooperation of all key players.

Some of the most important lessons concern the connection between the bias assessment and the ethical risk assessment. As already discussed, a bias assessment cannot be undertaken without the risk assessment; choices that must be made when deciding to evaluate an algorithm for bias—what specifically to test for and how to go about it—should be guided by an initial assessment of ethical risk. The relevance of the algorithmic inputs to the intended purpose of the algorithm, the choice of groups relative to which the algorithm should be assessed, the choice of metrics to compare, and other choices having to do with the testing data and its quality is all guided by the ethical risk assessment.

One particular issue regarding testing data is worth highlighting. It is essential to conduct the bias assessment in a fair and just manner. Given that assessments are not as independent as audits, and that there is no canon of best practices and standard test-data sets, teams must use forethought in procuring and selecting data sets, and labeling them correctly and consistently. Importantly, there needs to be a level of independence between the curation of the testing dataset and bias testing itself. Especially in cases where the data sets are small, or the effect you are trying to detect is small, the addition or removal of data can have a significant effect on the testing results. Given our small team, it is not possible to have different people curating the data and testing the algorithm, which we recommend is the preferred structure, so we set up procedural limitations. This involves deliberating, approving, and fixing the dataset prior to testing, and accepting a formal policy to not change the dataset after testing has begun without repeating and documenting the deliberation process.

The bias assessment in turn informs the ultimate assessment of the main ethical risks and influences our recommendations. The prioritizations in the risk assessment—the comparison and ranking of potential harms—can only be completed in light of the bias assessment. Inaccuracy and bias of algorithms are, after all, primary sources of potential harm and unfairness.

A final lesson is the value of interdisciplinarity for the assessment process. Our team consists of academics and consultants with overlapping backgrounds in philosophy, law, human rights, organizational ethics, statistics, and machine learning. The diversity of specializations reflects the diversity of abilities needed to perform a comprehensive assessment of algorithms in their socio-technical context. The assessments and preparation of recommendations were highly collaborative. Having multiple perspectives in conversation was crucial to the whole process, guiding its next steps, and allowing the team to see the forest for the trees.

6 Conclusion

This concludes our discussion of our team's process of conducting assessments of AI systems for ethical risk, and the key lessons for use of AI and for algorithm assessments and audits. A central overarching theme is the importance of situating the algorithm or AI being assessed within its particular socio-technical context—it is that socio-technical system and not merely the algorithm itself that is the proper object of assessment. This is also what drives the relationship between bias assessment and risk assessment; the interplay between them is one key feature of the way our team does algorithmic assessments overall, and how we think audits should proceed in the future. Just as algorithms should not be assessed independently of their context and use, so assessments should not focus on the code, but on the interplay between the code, its outputs, and the users. Acknowledging that assessment methodology is, or should be, driven by realities of socio-technical aspects of the algorithm is not new. Showing what that acknowledgement looks like in practice is. We hope this work provides a general structure and some specific suggestions for

teams conducting internal or self-assessments, external second-party assessments, and potentially third-party audits and assurances as well, and that the lessons help guide all parties interested in mitigating ethical risk and harnessing the power of AI technology for good.

Data Availability Data sharing not applicable to this article as no datasets are directly relevant to the content of this article.

Declarations

Conflict of Interest The article is based on the authors' work for BABL AI, a consultancy that focuses on responsible AI governance, algorithmic risk and bias assessments, and corporate training on responsible AI.

References

Ada Lovelace Institute. (2020). Examining the Black Box: Tools for assessing algorithmic systems. Retrieved February 20, 2022. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/

Andrus, M., & Villeneuve, S. (2022). Demographic-reliant algorithmic fairness: Characterizing the risks of demographic data collection in the pursuit of fairness. In 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea. ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3531146.3533226

Basl J., Sandler, R., and Tiel, S. (2021). Getting from commitment to content in AI and data ethics: Justice and explainability. Atlantic Council. Retrieved May 30, 2022. https://www.atlanticcouncil.org/in-depth-research-reports/report/specifying-normative-content/?mkt_tok=NjU5LVdaWC0wNzUAAAF_slunuNBmXLNnheGh0w-KgEPaF8uewmUN3T7b1fFhbKHlDLa-V9Hw7UxOQVcPMrTBbngaUICIzBLDNXD7S30ZcxaKgKSvyTD6BF69Z2MH

Baum, K., Mantel, S., Speith, T., & Schmidt, E. (2022). From responsibility to reason-giving explainable artificial intelligence. *Philosophy and Technology, 35*(1), 1–30.

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code.* Polity.

Ben-Shahar, O., & Schneider, C. E. (2014). *More than you wanted to know: The failure of mandated disclosure.* Princeton University Press.

Brennan-Marquez, K. (2017). Plausible cause: Explanatory standards in the age of powerful machines. *Vanderbilt Law Review*, Vol. 70.

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society*. https://doi.org/10.1177/2F2053951720983865

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1–15.

Canadian Government. (2021). Algorithmic impact assessment tool. Retrieved February 10, 2022. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html

Carrier, R., & Brown, S. (2021). Taxonomy: AI audit, assurance, and assessment. Retrieved February 20, 2022. https://forhumanity.center/blog/taxonomy-ai-audit-assurance-assessment/

Dotan, R. (2021). Theory choice, non-epistemic values, and machine learning.*Synthese* 198, 11081–11101. https://doi.org/10.1007/s11229-020-02773-2

Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass, 16*(8), e12760. https://doi.org/10.1111/phc3.12760

IBM Research. AI Fairness 360. Retrieved May 30, 2022. https://aif360.mybluemix.net

Liao, M. (2020). *Ethics of artificial intelligence*. Oxford University Press.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology, 6*(3), 175–183.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, AR. (2022). A survey on bias and fairness in machine learning. Retrieved February 20, 2022, from the arXiv database. https://arxiv.org/abs/1908.09635

Mittelstadt, B. (2019). Explaining explanations in AI. *FAT* 2019 Proceedings* 1. https://doi.org/10.1145/3287560.3287574

Mökander, J., & Floridi, L. (2021). Ethics-based auditing to develop trustworthy AI. *Minds & Machines, 31*, 323–327. https://doi.org/10.1007/s11023-021-09557-8

Mökander, J., & Floridi, L. (2022). Operationalising AI governance through ethics-based auditing: an industry case study. *AI Ethics*. https://doi.org/10.1007/s43681-022-00171-7

Moss, E., Watkins, E.A., Singh, R., Elish, M.C., & Metcalf, J. (2021). Assembling accountability: Algorithmic impact assessment for the public interest. Retrieved February 20, 2020. https://datasociety.net/library/assembling-accountability-algorithmic-impact-assessment-for-the-public-interest/

New York City Council. (2021). A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools. Retrieved February 20, 2022. https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9

OECD. AI principles overview. Retrieved May 30, 2022. https://oecd.ai/en/ai-principles

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). Prentice Hall.

Sandler, R., & Basl, J. (2019). Building Data and AI Ethics Committees. https://www.accenture.com/_acnmedia/PDF-107/Accenture-AI-And-Data-Ethics-Committee-Report-11.pdf#zoom=50

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *FAT* '19: Proceedings of the Conference on Fairness, Accountability, and Transparency,* 59–69. https://doi.org/10.1145/3287560.3287598

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. *35 Harvard Journal of Law & Technology* 117, *UCLA School of Law, Public Law Research Paper No. 21–25*. Available at SSRN: https://ssrn.com/abstract=3867634

US Department of Health and Human Services. (2021). Trustworthy AI playbook. Retrieved May 30, 2022. https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf

Wachter, S., Mittelstadt, B., & Russell, C. (2021). Why fairness cannot be automated: Bridging the gap between EU Non-Discrimination Law and AI. *Computer Law & Security Review* 41 (2021): 105567. https://doi.org/10.2139/ssrn.3547922

Wyden, R., Booker, C., & Clarke, Y. (2022). Algorithmic Accountability Act. Retrieved February 20, 2022. https://www.wyden.senate.gov/imo/media/doc/Algorithmic%20Accountability%20Act%20of%202022%20Bill%20Text.pdf

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2018). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology, 32*(4), 661–683.

Authors and Affiliations

Ali Hasan[1,2] · Shea Brown[2,3] · Jovana Davidovic[1,2] · Benjamin Lange[2,4] · Mitt Regan[2,5]

Shea Brown
Shea-brown@uiowa.edu

Jovana Davidovic
Jovana-davidovic@uiowa.edu

Benjamin Lange
benjamin@bablai.com

Mitt Regan
regan@georgetown.edu

[1] Department of Philosophy, University of Iowa, Iowa City, IA, USA

[2] BABL AI, Iowa City, IA, USA

[3] Department of Physics and Astronomy, University of Iowa, Iowa City, IA, USA

[4] Ludwig-Maximilians-Universität München, Munich, Germany

[5] Georgetown Law Center, Georgetown University, Washington, DC, USA