Mitigating Bias in Organizational Development and Use of Artificial Intelligence

Hüseyin Tanriverdi
University of Texas at Austin, huseyin.tanriverdi@mccombs.utexas.edu

John-Patrick Akinyemi
University of Texas at Austin, johnpatrick.akinyemi@mccombs.utexas.edu

Neumann, Terrence
University of Texas at Austin, terrence.neumann@mccombs.utexas.edu

Mitigating Bias in Organizational Development and Use of Artificial Intelligence

Completed Research Paper

Huseyin Tanriverdi
University of Texas at Austin
Huseyin.Tanriverdi@mccombs.utexas.edu

John-Patrick Olatunji Akinyemi
University of Texas at Austin
Johnpatrick.Akinyemi@mccombs.utexas.edu

Terrence Neumann
University of Texas at Austin
Terrence.Neumann@mccombs.utexas.edu

Abstract

We theorize why some artificial intelligence (AI) algorithms unexpectedly treat protected classes unfairly. We hypothesize that mechanisms by which AI assumes agencies, rights, and responsibilities of its stakeholders can affect AI bias by increasing complexity and irreducible uncertainties: e.g., AI's learning method, anthropomorphism level, stakeholder utility optimization approach, and acquisition mode (make, buy, collaborate). In a sample of 726 agentic AI, we find that unsupervised and hybrid learning methods increase the likelihood of AI bias, whereas "strict" supervised learning reduces it. Highly anthropomorphic AI increases the likelihood of AI bias. Using AI to optimize one stakeholder's utility increases AI bias risk, whereas jointly optimizing the utilities of multiple stakeholders reduces it. User organizations that co-create AI with developer organizations instead of developing it in-house or acquiring it off-the-shelf reduce AI bias risk. The proposed theory and the findings advance our understanding of responsible development and use of agentic AI.

Keywords: Agentic, responsible AI, bias, learning, anthropomorphism, optimization

Introduction

Organizations are increasingly turning to machine learning (ML) and artificial intelligence (AI) algorithms (AI hereafter) with a clear goal: to make accurate decisions devoid of human prejudices and biases (Purdy, 2020). Yet, not all AI rise to the occasion. Some inadvertently echo and intensify the very societal inequalities and biases we aim to overcome (Clarke, 2022; Eubanks, 2018). At its core, AI bias is the unjust treatment of individuals or groups by AI (Friedman & Nissenbaum, 1996). These biases often sideline specific groups and deny them access to essential resources and opportunities. Tragically, these biases have adversely impacted marginalized communities across domains, from healthcare and finance to housing and education (Clarke, 2022; Eubanks, 2018; O'Neil, 2016).

These injustices underscore the need for accountability and regulation in organizations' design and use of intelligent AI, which has the capacity to learn and dynamically update their learning based on changing patterns in their big input data. This urgency hasn't gone unnoticed by policymakers around the world. For instance, in America, policymakers have proposed the Algorithmic Accountability Acts of 2019 and 2022, the White House rolled out "A Blueprint for AI Bill of Rights" in October 2022, and the National Institute of Standards and Technology (NIST) launched its "Artificial Intelligence Risk Management Framework (AI RMF 1.0)" in January 2023. However, developing a regulatory framework to prevent algorithmic bias

presents a unique set of challenges, primarily due to the intricate nature of AI design and the diverse applications of AI. Accordingly, the wide variety of applications and designs of AI has led to tensions in the academic literature regarding the fundamental properties of human-AI interaction. For example, some empirical studies show "appreciation" of AI advice (Logg et al., 2019), while others show "aversion" (Dietvorst et al., 2015). These tensions point to the need for a theory to anticipate the consequences of certain AI design and usage choices.

In this paper, we define AI as agentic information systems (IS) that have the capacity to learn, adapt, and act autonomously (Baird & Maruping, 2021). Recent theoretical work suggests that understanding how agency, rights, and responsibilities are delegated between humans and AI is pivotal for anticipating the outcomes of these systems (Baird & Maruping, 2021). Relatedly, Kim (2020) highlighted "principal-agent problems" in the context of AI as potential sources of AI bias. We hypothesize that the following agency transfer mechanisms between AI's stakeholders (principals) and AI (agent) would affect the likelihood and impact of AI bias: AI's learning method, anthropomorphism level, stakeholder utility optimization approach, and acquisition mode (make, buy, collaborate).

We test these ideas in a novel sample of n=363 matched pairs of intelligent AI algorithms from 18 industry segments and 150 functional categories in the U.S. In each pair, there is (i) a "problematic algorithm" whose stakeholders reported a bias, a model failure, or an IT failure problem, and (ii) a "problem-free algorithm," which was very similar in characteristics to the problematic algorithm, but whose stakeholders did not report any problems as of the year in which the problematic algorithm was observed. The results support the thesis of the study. Unsupervised learning increases the likelihood of AI bias, whereas strictly supervised learning reduces it. Highly anthropomorphic AI increases the likelihood but not the impact of AI bias, whereas moderately anthropomorphic AI increases the impact but not the likelihood of AI bias. User organizations that co-create AI with developer organizations instead of developing it in-house or acquiring it off-the-shelf reduce AI bias risk. Using AI to optimize one stakeholder's utility increases AI bias risk, whereas jointly optimizing the utilities of multiple stakeholders reduces it.

Background

Operationalizing AI Bias and Fairness

Interpreting AI bias requires understanding what it means for an AI to be fair (Friedler et al., 2016). Within the IS discipline, Kordzadeh and Ghasemaghaei (2022) provided an interpretation of AI bias grounded on unfairness: *"the outputs of an algorithm benefit or disadvantage certain individuals or groups more than others without a justified reason for such unequal impacts"* (pg. 1). A significant body of research has formed around statistical notions of fairness, in which a particular metric, a quantified benefit or harm, must be equal amongst groups or individuals (Srivastava et al., 2019). Mehrabi et al. (2021) find that the definitions can be described as (i) giving similar predictions to similar individuals, (ii) treating different groups equally by partitioning individuals by intersections of protected attributes, or (iii) combining the best properties of groups' and individuals' fairness notions. Demographic parity, equalized odds, equal opportunity, differential fairness, and non-parity unfairness are measurements proposed to assess group fairness based on various ethical principles that may be relevant depending on the decision context (Hardt et al., 2016; Teodorescu et al., 2021). However, such statistical fairness notions can be challenging for researchers to operationalize for deployed AI, as AI is often an organization's intellectual property. Additionally, statistical fairness may not be relevant to all learning algorithms and decision contexts. Given these challenges, we assess AI bias by examining reports from AI stakeholders that are published in credible journalistic sources. Specifically, we operationalize AI bias as an evaluation by the stakeholders, asserting that the AI has systematically and unfairly discriminated against certain stakeholders or groups of stakeholders in favor of others.

Transfer of Principals' Agencies to AI

Recent research applied agency theory to algorithmic governance (Kim, 2020) and human-AI interaction (Baird & Maruping, 2021). Principals entrust agents to act towards a goal on their behalf. The principal-agent problem (PAP) occurs when the agent's goal does not align with the principal's goal (Kim, 2020); this

is called a *malfeasance* by the agent. In the context of human-AI interaction, AI bias is an example of malfeasance. AI has two characteristics that increase the likelihood and the impact of malfeasance.

First, unlike in other principal-agent relationships, where there might be one principal, AI has a complex ecosystem of principals, each expecting the AI to learn and serve its own agency, values, and interests: e.g., user organizations, developer organizations, targets of AI's decisions, etc. In this paper, we focus on three of these stakeholders. A "user organization" is an organization that acquires and uses AI to support its business decisions. A "developer organization" is an organization that produces the AI.[1] "Targets" of algorithmic decisions are actors directly impacted by the AI's decisions. Targets may be the user organization itself or external actors. The competing and conflicting agencies of multiple principals could make it infeasible for the agent (AI) to align its goals with the goals of a particular principal. Thus, principals are likely to perceive AI malfeasance.

Second, an intelligent AI is capable of learning from dynamically changing patterns in big data inputs and dynamically updating its learning and decision rules over time. As the agencies, values, and interests of the principals evolve over time, so do the agencies, values, and interests learned by the AI. This dynamic agency transfer process makes it highly challenging to govern and control whose agency, rights, and responsibilities AI might be transferring and prioritizing. As the AI's principals compete with each other implicitly or explicitly to get the AI to learn their own agency and serve their own interests, the AI is likely to face complexity and thus irreducible uncertainties in its agency transfer processes and, ultimately, its decision rules (Cilliers, 1998). Irreducible uncertainty arises from intense, reciprocal interactions of smart, connected, and mutually dependent principles (Cilliers, 1998). Thus, AI might make emergent, unpredictable, and unexpected decisions that conflict with the principals' goals, leading to AI bias.

Hypotheses on Developer Organization's AI Design Choices

Developer Organization's Choice of Learning Method

In AI development, the learning method refers to the approach chosen by a developer organization to model the underlying structure of the data. The choice of AI's learning method can determine the extent of control the developer might have over whose agencies and interests the AI learns.

The "supervised learning" method refers to a statistical model that predicts an output based on one or more inputs (James et al., 2013). "Strict supervised learning" is a subset of supervised learning, which maps a set of expert-chosen, human-interpretable input variables to an output. These are "hand-crafted" features (Georgescu et al., 2019) and are chosen based on domain expertise and data familiarity within a given context. Much of the academic research on AI bias has focused on statistical notions of fairness relevant to supervised learning models, especially strict supervised learning models (Hardt et al., 2016). Accordingly, developer organizations that use strict supervised learning have access to a more advanced and generalized toolkit for assessing statistical fairness of the learning algorithm. For example, IBM has developed AIF360, a data science toolkit that detects and removes bias from strict supervised classifiers that contain protected attributes (Bellamy et al., 2019). Therefore, developers have more control over the agency learned by the algorithm and are more likely to create a fair model when choosing strict supervised learning.

"Unsupervised learning" method refers to a statistical model used to uncover patterns in data, given only inputs (James et al., 2013). Applications are wide-ranging but mainly fall into clustering, dimensionality reduction, and graphical techniques. Relatedly, we define "hybrid learning" methods as bespoke combinations of unsupervised, supervised, and reinforcement learning for model development. One notable example of hybrid learning is deep learning (LeCun et al., 2015), which is often operationalized as a supervised learning method in which no explicit predictor variables are provided to an AI to guide the learning process; instead, the network architecture is designed such that the AI can automatically discern predictive features that map to outputs from raw data, often without AI developers understanding why the AI behaves as it does. Therefore, in unsupervised and hybrid learning, the relevant feature set is often determined entirely by the AI, increasing the risk that the agentic AI acts towards different goals or

[1] Developer and user organizations can be the same (i.e. an "in-house" development team).

otherwise learns a different agency than the developer intended at training time (Kim, 2020). For instance, deep learning models are sensitive to erroneous cues that humans would ignore, such as the presence of a ruler in a photo indicating malignant carcinoma (Akhila et al., 2018). Additionally, there are relatively few statistical fairness metrics and development toolkits relevant to unsupervised and hybrid learning approaches. This makes it challenging for developers to ensure fairness in the agency learned by the AI when they employ unsupervised or hybrid learning approaches.

Hypothesis 1 (H1): The developer organization's choice of learning method affects AI bias risk. Strict supervised learning reduces AI bias risk relative to unsupervised and hybrid learning methods.

Developer Organization's Choice of Anthropomorphism Level

The anthropomorphism level of AI is defined as the extent to which the developer organization designs the AI's user interface to have human-like characteristics: e.g., name, gender, limbs, emotive expressions (Nowak et al., 2015), or appear to be "cognitive" or "animated." Machines perceived to be intelligent, such as agentic AI, are particularly easy to anthropomorphize (Novak & Hoffman, 2019).

The anthropomorphism level of AI can significantly affect whether users view the algorithm as a machine devoid of human agency and biases or as a human entity imbued with agency and biases. How the users view AI can, in turn, affect the level of trust or distrust they place in the AI system. Consequently, users may either blindly transfer their agency and decision-making rights to the AI or engage in vigilant information-seeking and processing to exert control over the AI's decisions. Interestingly, there are two distinct streams of literature that offer seemingly contradictory perspectives on these questions.

The automation bias literature found that users trusted decisions recommended by machines more than decisions recommended by human beings because they assumed that machines did not have agency and thus that they made rational decisions free of human biases. Unwarranted trust in machines causes an "automation bias," i.e., the "*tendency to use automated cues as a heuristic replacement for vigilant information seeking and processing*" (Mosier & Skitka, 1999, p. 344). These findings imply that if developers design AI to have a high anthropomorphism level, users might perceive the AI as a human with agency and biases, and hence, they *would not trust* the AI and would not blindly transfer their agency and decision rights to the AI. Instead, users would vigilantly seek information to check AI's decisions and reduce AI bias.

The recent literature on anthropomorphic AI makes the opposite argument. It assumes that high anthropomorphism levels can increase AI's social presence, user engagement, decision persuasiveness (Sundar et al., 2008), user trust in AI, and user willingness to adopt AI's recommendations (Blut et al., 2021). Glikson and Woolley (2020) argue that a high anthropomorphism level can increase emotional and cognitive trust towards AI. Additionally, trust in highly anthropomorphic AI persists even when the accuracy of the information provided by the agent degraded over time (De Visser et al., 2016). While these assumptions and arguments have yet to link anthropomorphism to bias, if taken at face value, they imply that a high level of anthropomorphism *would increase trust* in AI, engender the users to transfer agency and decision rights to AI, and increase the likelihood of AI bias due to lack of vigilant information seeking and processing by the users.

According to Granovetter (1985), malfeasance in principal-agent relationships is more likely to occur in states of trust rather than distrust. When developers imbue an algorithm with human-like qualities, both the user organization and the targets of the algorithmic decision may be more inclined to trust the algorithmic decisions. Unwarranted trust in the algorithm can lead to blind transfer of agency and decision rights from users to AI and allow the AI to make unchecked decisions, which could increase the likelihood of AI bias. However, this risk could potentially be mitigated by designing the AI to have moderate rather than high levels of anthropomorphism. At moderate anthropomorphism levels, users can be aware that AI is neither fully machine nor fully human. Thus, they are unlikely to blindly transfer their agency and decision rights to AI. Rather, users would more likely engage in vigilant information seeking and processing to check the AI's decisions and reduce the likelihood of AI bias.

Hypothesis 2 (H2): Relative to no anthropomorphism, a high anthropomorphism level is likely to increase AI bias, whereas a moderate anthropomorphism level is likely to reduce AI bias.

Hypotheses on User Organization's AI Usage Choices

User Organization's Choice of AI's Stakeholder Utility Optimization

User organizations use AI to optimize how scarce resources and opportunities are allocated among stakeholders. AI serves a wide variety of stakeholders (e.g., shareholders, customers, suppliers, employees, etc.) who often have competing utilities and expectations from the AI. This creates irreducible uncertainty for AI's resource and opportunity allocation decisions as it is infeasible to maximize the competing utilities of all stakeholders. When implementing AI, user organizations, in coordination with developer organizations, choose which stakeholders' agencies and utilities AI should prioritize. Unilateral utility optimization prioritizes the agency and utility of only one stakeholder. Joint utility optimization seeks to reach a compromise or a tradeoff among the competing agencies and utilities of multiple stakeholders. The choice of the optimization approach can affect AI bias.

Under unilateral utility optimization, except for the prioritized stakeholder, all other stakeholders' agencies and utilities are disregarded. The excluded stakeholders can perceive the resource and opportunity allocation decisions of the AI to be unfair to them (Lee & Baykal, 2017). To take into account the agencies and utilities of a wider variety of stakeholders in algorithmic decision-making, scholars have suggested technical (Andrew & Qi, 2022) and design-based approaches based on ethical notions of "procedural" (Lee et al., 2019) and "distributive" (Gajane & Pechenizkiy, 2017) justice. When user organizations choose a multilateral utility optimization approach and seek a compromise or a tradeoff among the conflicting utilities of the stakeholders, they may achieve procedural and distributive justice. While no single stakeholder might be perfectly happy with AI's decisions, they would not perceive any unfair treatment either and would be less likely to report AI biases.

Kleinberg et al. (2018) demonstrate how organizations can use improved accuracy from AI to optimize utility for different stakeholders in the context of judicial bail decisions. The bail system was designed to ensure that people return to court for their court date and, in certain cases, for the general public's safety. There are numerous stakeholders in this context, such as law enforcement, the accused/arrested, the accuser, the public, and relevant governing bodies. These stakeholders have different agencies and utilities from an algorithm that supports bail decisions. Some stakeholders advocate for *more stringent* pre-trial detention to enhance public safety. Others advocate for *less stringent* pre-trial detentions to reduce the significant personal costs for those detained and the costs to the public to hold people in jails (e.g., losing a job). The authors suggest that the user organization can use this algorithm to optimize different goals for different stakeholders. For instance, they experiment with policy outcomes by changing the threshold probability for the binary "release or detain" judgment, which effectively varies the *stringency* of the algorithmic decision. They find that if judges decided to keep the algorithm as stringent as human judges historically have been, the increased accuracy from algorithmic predictions significantly lowers the crime rate, an outcome that serves the public's utility. However, if other stakeholders demand a reduction in the prison population (and thus *less stringency* in bail decisions), judges could increase the release rate without any associated increase in crime due, again, to the increased accuracy of the algorithmic predictions (Kleinberg et al., 2018, p. 27). The user organization can choose which stakeholders' utilities are optimized.

If AI is unilaterally optimized to serve the user organization's agency and interests, targets of the algorithmic decisions or other impacted stakeholders may feel that their agencies are not represented in the AI. Likewise, those that control the algorithm – the user organization – will likely have increased agency, as they will be able "to do more with less," a key promise of AI. So long as the welfare of the user organization and targets of algorithmic decisions are not in direct conflict, multilateral utility optimization amongst a broader set of stakeholders should produce algorithmic decisions that are mutually agreeable and, therefore, invite fewer perceptions of algorithmic bias.

Hypothesis 3 (H3): User organization's choice of multilateral utility optimization reduces the likelihood of AI bias relative to unilateral stakeholder utility optimization.

User Organization's Choice of AI Acquisition Mode

We define the acquisition mode of AI as a user organization's choice among the buy, make, or collaboration options for AI development (Beulen et al., 2022). The literature on governance modes informs us that a user organization can acquire AI either (i) off-the-shelf, (ii) develop the AI in-house, or (iii) collaborate with an external developer organization to jointly develop the AI (Beulen et al., 2022; Rubenstein, 2021).

The choice of AI acquisition mode can affect a user organization's ability to tame irreducible uncertainties in AI design. Opting for off-the-shelf AI solutions means missing the chance to tailor the system to the organization's specific values and fairness metrics. Contrary to the myth that off-the-shelf algorithms can be seamlessly integrated (Kottler, 2020), this approach often leads to adopting the developer's values and policies. As Rubenstein (2021) notes, this can create a misalignment between the developer's and user's objectives. Such discrepancies can make the AI serve the developer's interests over those of the user organization and its target audience. If the AI exhibits bias, the user organization may lack the control to rectify it. As a result, off-the-shelf solutions are more likely to neglect local stakeholder needs, increasing the risk of AI bias (Abbasi et al., 2019).

User organizations may also have some algorithm development capability and choose to make an algorithm in-house to ensure their agency of specific domain knowledge is reflected in the algorithm decision-making. However, in-house development would mean that the user organization is on its own in facing all the irreducible uncertainties of the design phase. Unlike external vendors specializing in responsible AI development, user organizations may not have sufficient expertise in adopting fairness metrics that reflect their own agency and preferences (Hopkins & Booth, 2021). For example, a developer organization may be conscious of using strict supervised learning in algorithm development. As we argued in H1, the learning method choice determines the type of agency learned in the algorithm and, thereby, the extent of bias learned from the environment. Thus, when user organizations develop an algorithm in-house, in all but a few cases, they may be less likely to properly maintain the rights and responsibilities necessary to safeguard from algorithmic bias (Rubenstein, 2021).

User organizations that co-develop AI with developer organizations are able to tame the irreducible uncertainties of the design phase and reduce AI bias. Large external AI developers have often established responsible AI standards of practice across their full portfolio of AI development projects. For instance, Accenture and Boston Consulting Group have made their practices and values toward AI development publicly available. External developer organizations can provide controls and safeguards learned from work with clients across multiple sectors and guide user organizations on best practices for mitigating algorithm bias. Based on this, if a user organization collaborates with a developer organization on the design and development of an algorithm, it would have opportunities to discuss and choose the best practices related to agency and safeguarding against AI bias.

Hypothesis 4 (H4): User organizations that develop algorithms in collaboration with developer organizations reduce AI bias risk compared to user organizations that acquire AI off-the-shelf or develop them on their own in-house.

Methods

Sample and Matching Criteria

Our sampling frame was a repository of problematic algorithms maintained by "AI Algorithmic and Automation Incidents and Controversies" (AIAAIC), a non-partisan independent public interest initiative. Proponents of ethical, responsible AI use and development submit evidence of problematic algorithms to the AIAAIC repository. For each problematic algorithm, the AIAAIC repository points to news articles, blogs, and other data sources that discuss the alleged problem in the algorithm. We downloaded the repository in July 2022. We also supplemented the AIAAIC repository by systematically searching for problematic algorithms in Google, Factiva news database, and academic databases such as EBSCOhost, Web of Science, and Google Scholar. At this stage, an average of five unique URLs were gathered that

provided information on the algorithm's characteristics and problematic nature. We then analyzed all **allegedly problematic algorithms** to select the ones that satisfy the following **inclusion criteria**:

(i) Algorithm: A problematic algorithm met the definition of an intelligent algorithm. It learns from patterns in big data inputs and alters its behavior over time based on changes in big data inputs.
(ii) Problem type: The algorithm had to have bias, model failure, or IT failure. The repository also had malicious IT failures (i.e., algorithm privacy). These were also included.
(iii) Realized or potential problem: The problematic algorithm had a realized problem. We excluded entries that discussed concerns about potential algorithmic problems that have not emerged yet.
(iv) Usage status: When the problem emerged, the algorithm was used with actual data and users during at least a pilot study, if not in full production. Entries that discussed problems detected during the ideation phases of algorithms were excluded.
(v) Developer Organization: The developer of the problematic algorithm was an organization. If an individual developed an algorithm, it was excluded.
(vi) User organization: The user organization of the algorithm in which the algorithmic problem emerged was known. Algorithmic problems whose user organizations were unknown were excluded. If user and developer organizations were the same, subunits that developed and used the algorithm were distinguished, and their characteristics were coded separately.
(vii) Location of user organization: The user organization of the problematic algorithm had to be incorporated in the US. We excluded user organizations from international locations.

We found a **problem-free matching algorithm** for each problematic algorithm to create a matched pair of problematic and problem-free algorithms. We used the following **matching criteria**:

(i) Timing of match: The matching algorithm had to be in existence and used as of the year of the problematic algorithm's problem emergence. All other matching criteria had to be satisfied as of that year.
(ii) Problem status: The matching algorithm had to be free of bias, IT failure, model failure, privacy breach, and cybersecurity breach problems in the year of matching.
(iii) Application domain: The matching algorithm had to be in the same application domain as the problematic algorithm (e.g., insurance, healthcare, HR, etc.).
(iv) Function: The matching algorithm had to have the same function as the problematic algorithm (e.g., voice assistant, recommender, search-matching, etc.).
(v) Platform status: The matching algorithm had the same on-platform/off-platform status as the problematic algorithm. On-platform algorithms worked on multi-sided platforms (MSP); off-platform algorithms were used by organizations that did not use MSP.
(vi) For-profit status: The matching algorithm's user organization had to have the same not-for-profit/for-profit status as the problematic algorithm.
(vii) Public status: The matching algorithm's user organization was in the same public/private sector.
(viii) Industry: The matching algorithm's developer organization had to have the same NAICS industry and SIC sector code as the developer of the problematic algorithm.

The final sample had 363 pairs of problematic and problem-free algorithms, i.e., 726 algorithms, from 18 industry segments (e.g., Retail, Manufacturing, Information, etc.) and 150 functional categories (e.g., cancer prevention, predictive policing, credit scoring, content moderation, autonomous driving, etc.) in the U.S. between 2006 and 2022. Table 1 summarizes the sample construction process.

Step	Description of action taken	Size
1	Download problematic algorithms reported in the "AI Algorithmic and Automation Incidents and Controversies" (AIAAIC) repository as of 07/21/2022[1]	878
2	Complement the AIAAIC sample with additional problematic algorithms found through keyword searches in Google, Factiva, EBSCOhost, and Web of Science	131
Subtotal of problematic algorithms before applying sample selection criteria		1009
3	Drop algorithms whose user organizations are not incorporated in the US	611
4	Drop algorithms in the ideation phase that are not yet used with actual data and users	573
5	Drop algorithms failing to satisfy the definition of an intelligent algorithm	483

Step	Description of action taken	Size
6	Drop algorithms that do not have any of the following problems: (i) bias, (ii) IT failure, (iii) model failure, (iv) privacy breach	413
7	Drop algorithms that (i) were developed by an individual rather than an organization; (ii) whose developer organizations were not specified	396
8	Drop problematic algorithms which no matching problem-free algorithms were found	363
Subsample of problematic algorithms		363
9	For each problematic algorithm, go to the year of problem emergence and find a problem-free algorithm that satisfies (1) criteria in steps 3-8 above; same (2) function, application domain, private status, for-profit status, and on-platform status as the problematic algorithm; and (3) has not yet developed any bias, IT or model failure, and privacy breach.	363
Subsample of problem-free algorithms		363
Final sample: Pairs of problematic (n1=363) and problem-free algorithms (n2=363)		726
[1]https://www.aiaaic.org/aiaaic-repository/ai-and-algorithmic-incidents-and-controversies		

Table 1. Construction of matched sample of problematic and problem-free algorithms

Source Documents

About 140 students collected the source documents needed for coding the study variables. We guided students through systematic keyword searches in the Factiva database, company websites, and Google to find sources discussing the characteristics of an algorithm and its developer and user organizations. We collected six unique source document types: (i) peer-reviewed academic publications, (ii) investigative journalism articles (in-depth journalist techniques used to expose matters concealed behind conditions that confuse people's understanding (Kaplan, 2013)), (iii) court cases, (iv) mainstream news articles, (v) developer organization documents and websites, and (vi) user organization documents and websites. However, students did not code the variables. Instead, we reserved the coding tasks for two expert coders.

We considered peer-reviewed academic publications, investigative journalism, and court cases as credible since they go through a rigorous process by publishers and also by AIAAIC experts. For news articles, we used an independent media fact-checker (MBFC, 2023) to analyze the credibility and factual reporting of the 192 unique news outlets in our sample. Two outlets had low factual reporting and credibility ratings, and three had low credibility. Therefore, a total of 12 articles associated with these outlets were dropped. In addition, algorithms with unreliable or insufficient source documents were also dropped. Table 2 lists the 9834 sources (by type) collected and used for coding the study variables.

	Academic Articles	Investigative Journalism	Court Cases	News Articles	User Org Sources	Dev Org Sources	Total
Problematic Algorithms	121	1214	44	972	794	2149	5294
Problem-Free Algorithms	93	649	13	825	501	2459	4540
Total URLs	214	1853	57	1797	1295	4608	9834

Table 2. Source Documents Used to Code Study Variables

User and Developer organization sources include the following nine types of evidence: (i) Annual statements, (ii) ESG and CSR reports, (iii) Policy documents (e.g., privacy or ethics statements, compliance pages, data handling procedures), (iv) Business Standards documents (e.g., controls and methods document, corporate governance guidelines), (v) HR pages and LinkedIn company profiles, (vi) Investor documents (e.g., Audit Committee Charters, Oversight board structures, Shareholder meeting notes), (vii) Corporate Blog pages (e.g., Internal news pages, product, and service description pages), (viii) Corporate Internal Research (e.g., technical publications, whitepapers), and (ix) General organization webpages (e.g., About us page, Vision and Mission statement page). Furthermore, for investigative website sources (e.g., ProPublica, The Verge, Wired, MIT Technology Review, New York Times, etc.), we used (i) investigations

on organizations, (ii) investigations specifically on algorithms, (iii) interviews of prominent figures of the organization, and (iv) third-party analysis of corporate documents.

Table 3 illustrates the measurements of each independent variable and an illustrative example for each with applicable evidence used to support coding. Table 4 provides an example of two algorithms, their user and developer organizations, and their associated evidence used to code the dependent variables.

Coding Instrument and Intercoder Reliability

We developed and validated a coding guideline to code the study variables from the source documents. Definitions of variables were adapted from the published literature or practitioner articles where no academic articles were available. Two Ph.D. students with degrees and professional experiences in IS and strong training and research experience in Data Science served as two independent expert coders who read the source documents to code the study variables by following the validated coding guidelines.

We established the reliability of the coding guidelines as follows. In each round of an iterative process, the two independent coders used the guidelines to code a small sample of algorithms (e.g., 20). We assessed the inter-coder agreement rate after each episode of coding. After the first round, the agreement rate was 68%. The two coders discussed the sources of coding discrepancies to find that some variables were not tightly defined. As a result, we revised and tightened the definitions. After three rounds of iterations, the inter-coder agreement rate increased above the 90% threshold for establishing the reliability of the coding instrument. In the coding of n=726 algorithms, agreement rates were 95% for the dependent and 94% for the independent variables. The remaining discrepancies were discussed and resolved by the two coders.

Variables and Multi-Item Scales

Independent and control variables of the study were coded within the year prior to the emergence of the algorithmic problem. In Table 5, we list all dependent and control variables with their definitions and measurements. Further, we created six dummy variables: Choice of Learning Method, Anthropomorphism Level, AI's Stakeholder Utility Optimization, AI Acquisition Mode, Algorithm Decision-Making Support, and Organization Industry. Table 6 provides correlation values among all the study variables. We had three multi-item scale constructs. Factor analysis indicated that measurement items of each of the multi-item constructs load onto their respective factors and have very low loadings on other factors.

Damages caused by algorithmic bias consisted of the following four measures: (a) harm to stakeholders; (b) financial loss to user organization; (c) reputational harm to user organization; and (d) lawsuit on user organization. Cronbach's Alpha was 0.85, indicating sufficient reliability.

Developer Organization Mitigations measurement was taken from eight developer organization mitigations, each supported by literature (e.g., Clarke (2022), RDS (2022), Guszcza et al. (2018), and Leslie (2019)): (a) board-level oversight of algorithmic risks, (b) commitment to ethical development and use of A.I, (c) algorithmic transparency, (d) algorithmic accountability, (e) algorithmic audits, (f) the use of F.A.C.T (Fairness, Accuracy, Confidentiality, and Transparency) principles by data science teams, (g) diversity of data science team members' backgrounds and perspectives, and (h) governance and controls of algorithms' inputs, logic, and outputs. Cronbach's Alpha was 0.80, demonstrating sufficient reliability.

User Organization Mitigations measurement was taken from a total of three items, focused on whether the user organization for a given algorithm: (a) committed to the ethical use of the algorithm (Leslie, 2019); (b) accepted accountability for the algorithm's decisions (Koene et al., 2019); and (c) exhibited transparency about variables the algorithm used to make decisions (Koene et al., 2019). Cronbach's Alpha for the three items was 0.71, indicating sufficient reliability.

Furthermore, though developer and user organizations can exist in the same organization (e.g., Google), they are still different organizational units. For example, Google's Data Science unit may have developed the algorithm, but the user organization could be YouTube, Google Maps, or another organizational unit of Google. We distinguished the units with the same organization and coded their characteristics accordingly.

Coding	Developer Org – Algorithm, Year	Evidence	Evidence URL
Choice of Learning Method. What method was used by the developer organization to train an algorithm before the emergence of the algorithm problem?			
[0] Unsupervised Learning	OpenAI – Image Recognition Algorithm, 2021	"CLIP is intended to explore how A.I. systems might **learn to identify objects without close supervision** by **training on huge image and text pairs databases**. OpenAI used some **400 million image-text pairs scraped from the internet to train CLIP**."	https://tinyurl.com/y7m2uur
[1] Strict Supervised Learning	Microsoft – Content Moderation Algorithm, 2020	"LinkedIn to adopt a machine learning approach **trained** on public member profile content…**accounts labeled as either "inappropriate" or "appropriate"**… LinkedIn identified problematic words and sampled appropriate accounts from the member base containing these words. **The accounts were then manually labeled and added to the training set**, after which the model was trained."	https://tinyurl.com/2sn7e9tx
[2] Hybrid Learning	Amazon – Speech Recognition Algorithm, 2021	"Using semi-supervised learning, Amazon scientists were able to train a model and reduce speech recognition error rates by 10-22% compared to methods on supervised learning…The model was **trained with 7,000 hours of labeled data, then 1 million hours of unannotated or unlabeled data."**	https://tinyurl.com/2s3rb94m
Anthropomorphism Level. Refers to the level of any non-human entity with humanized characteristics such as talking, singing, etc. Some algorithms have humanized features to encourage users to perceive algorithmic messages delivered by a human. To what extent is this algorithm anthropomorphic?			
[0] No Humanized features at all	Instacart – Payment Algorithm, 2019	"I**t's a learning algorithm that takes into account all kinds of different factors, including things like distance, time of day, the market, the items being shopped,** and whether they're difficult in some way."	https://tinyurl.com/7eubpkyh
[1] Moderate degree of Humanized features	Mya System – Conversational Chatbot Algorithm, 2015	"Hiring **chatbo**t Mya guides candidates through the entire hiring process, starting from the job search and up to the onboarding. To allow **natural conversation with the candidates**, Mya leverages state-of-the-art approaches from natural language processing and understanding…She keeps existing databases warm and **engaged**, refreshing profile contents and attracting best-fit candidates to open roles."	https://tinyurl.com/bddwhc38
[2] High degree of Humanized features	Sense.ly – Virtual Nurse Assistant Algorithm, 2016	"It's human-like. It **talks to patients naturally**, and they talk to the nurse as they would a real nurse or doctor. W**e have an avatar that responds like a real person, with empathy**, who we hope can illicit long-term use and honesty… patients can interact with one of many Sense.ly nurses, which vary in gender, ethnicity, and accents. "	https://tinyurl.com/3e5ypxpz
AI's Stakeholder Utility Optimization. Whose objectives did the algorithm try to optimize?			
[0] No Utility Optimization	Microsoft Azure – Facial Recognition Algorithm, 2020	Stakeholder groups: Azure Data Science Team, Third-Party Azure App Developers, Microsoft Azure Users "With face verification, two face templates are compared to see if they are a match. On a practical level, the purpose of a facial recognition algorithm is to evaluate whether two faces belong to the same person."	https://tinyurl.com/247t7965
[1] Unilateral Utility Optimization	Amazon – Search-Ranking Algorithm, 2016	Stakeholder groups: Marketplace Sellers, Amazon Retail Unit, Amazon "Fulfilled by" Vendors, Customers "An investigation by The Markup found that **Amazon places products from its house brands and products exclusive to the site ahead of those from competitors**…The company appears to be using **a proprietary algorithm to advantage itself** at the expense of sellers and many customers."	https://tinyurl.com/mvwxhcn3
[2] Multilateral Utility Optimization	DoorDash – On-Demand Matching Algorithm, 2020	Stakeholder groups: Delivery Persons (Dashers), Consumers, Restaurants, DoorDash Business Unit "Through **optimal matching**, we **ensure dashers get more done in less time, consumers receive their orders quickly, and merchants have a reliable partner** to help them grow their businesses."	https://tinyurl.com/4radjn29

Coding	Developer Org – Algorithm, Year	Evidence	Evidence URL
AI Acquisition Mode. How did the user organization acquire this algorithm?			
[0] User organization purchased the algorithm off-the-shelf	IBM Watson for Oncology – Treatment Recommendation Algorithm, 2017	"Jupiter Medical Center **will adopt** Watson for Oncology trained by Memorial Sloan Kettering, a cognitive computing platform to provide insights to physicians to help them deliver personalized, evidence-based cancer treatment. Jupiter is the first U.S. community hospital to adopt Watson for Oncology, which will go live at the facility in the beginning of March." Jupiter Medical Center's President states, "**Watson for Oncology is part of our significant investment** in creating a world-class cancer program."	https://tinyurl.com/43ynrhy4
[1] User organization collaborated with dev organization	CloudMedx – Predictive Healthcare Model Algorithm, 2020	"Anthem has launched a digital tool that aims to allow public health officials and other health and community leaders to track and predict the impacts of COVID-19. **The tool was built in partnership with CloudMedx**, an artificial intelligence startup." This work is part of Anthem's best-in-class data scientists and clinicians collaborating with a global alliance of leaders. **This collaborative effort** to introduce C19 Explorer and C19 Navigator is another example of Anthem's commitment to leadership."	https://tinyurl.com/rwkwrdra
[2] User organization developed this algorithm on its own	HealthTap – Healthcare Conversational Chatbot Algorithm, 2017	"Through **our user-friendly, AI-driven app and website**, we provide unparalleled personalized care to every HealthTap member." HealthTap's new Dr. A.I. considers both patient context and the clinical expertise of doctors who have helped triage hundreds of millions of patients worldwide…Over the past six years, **we've collected data** from tens of thousands of the leading U.S. doctors who have triaged millions of patients throughout their careers."	https://tinyurl.com/2p9ap659

Table 3. Illustrative Examples of Coding of Independent Variables

Bias Coding	Magnitude of Damage	Dev Org – User org of Algorithm, Year of use	Description	Evidence URL
Algorithmic Bias. **Assess if an algorithm was perceived as systematically and unfairly discriminating against certain individuals or groups of individuals in favor of others.** Federally protected classes were the target of algorithmic bias: i.e., age, sex (gender, pregnancy, sexual orientation, and gender identity), physical or mental disability, race, color, religion or creed, citizenship, national origin, or ancestry, veteran status, and socioeconomic status.				
Damages caused by algorithmic bias. **Assess if a user organization of an algorithm suffered damages due to an algorithmic bias problem**. Did the problematic algorithm (1) harm customers or employees of the user organization, (2) cause financial loss (e.g., regulatory fine) to the user organization, (3) harm the user organization's reputation, (e.g., bad press or pressure on the organization to use socially accepted norms), or (4) led to a lawsuit on the user organization.				
[1] Bias	[1] Harm to People, [0] No Financial Loss [1] Reputational Harm [0] No Lawsuit	TaskRabbit – *TaskRabbit Business Operation's use of* Ranking Algorithm, 2015	TaskRabbit was found to **systematically and unfairly treat women and minorities** by being less likely to recommend them in search results, **even if they have the same or better qualifications as their white male counterparts**. The study's lead researcher states, "What I suspect is going on with TaskRabbit's algorithm is that social feedback, such as **reviewer comments, are considered in determining the ranking**, and we know that **social feedback can be biased**."	https://tinyurl.com/3dc2k3se
[1] Bias	[1] Harm to People, [1] Financial Loss, [1] Reputational Harm, [1] Lawsuit	Checkr – *Uber's use of* Background Check Algorithm, 2019	Checkr Background check algorithm is biased against **individuals wrongly accused of crimes or who have committed minor offenses**. As evident from lawsuits, this group of people is **systematically and unfairly treated** by preventing them from obtaining gig worker jobs. "In court documents, the plaintiffs have accused Checkr of including ***criminal activity that is too old to report under the law***."	https://tinyurl.com/5crhpmst

Table 4. Illustrative Examples of Coding of Dependent Variables

Org	Variable Name	Variable Definition	Measurements
Dependent Variables			
User	*Algorithmic Bias*	An algorithm systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others (Friedman & Nissenbaum, 1996)	[0] No Perceived Algorithmic Bias [1] Perceived Algorithmic Bias
	Damages caused by algorithmic bias	An algorithm causes negative physical or psychological effects on the organization's stakeholders, resulting in the mistreatment of the stakeholders. Algorithmic bias causes financial loss, bad press, or legal violations for an organization (Agrafiotis et al., 2018)	[1] Harm people/[0] None; [1] Financial loss/[0] None; [1] Reputational Damage/[0] None; [1] Lawsuit/[0] None
Control Variables			
Dev	*UX Group*	A group focused on user research, usability testing, and user experience design.	[0] No UX Group; [1] UX Group
	Fairness Goal	The organization's objective for developing algorithms is to avoid prejudice toward a group based on their inherent characteristics. (Mehrabi et al., 2021)	[0] Algorithm Fairness, not a stated goal [1] Algorithm Fairness is a stated goal
	Developer Organization Mitigations	The developer organization of an algorithm had governance and controls to mitigate the risks of its portfolio of algorithms *(8 items listed in the text)*.	[0] No evidence of mitigation [1] Symbolic evidence of mitigation [2] Substantive evidence of mitigation
	Model Failure	An algorithm's analytical model fails to deliver acceptable output accuracy due to missing, wrong, incorrect proxies or unexpected interactions amongst variables.	[0] No model failure; [1] Model failure
	Organization Industry	The developer organization's unique 2-digit sector code (SIC) is based on the North American Industry Classification System (NAICS).	[0] Dev in Government; [1] Dev in Manufacturing; [2] Dev in Services
	Interaction Capabilities	The count of an algorithm's human-like interaction abilities is based on having one or more capabilities of vision, speech, emotion, cognition, and touch.	[1] Vision, [1] Speech, [1] Emotion, [1] Cognition, [1] Touch / [0] None
User	*For-Profit Status*	The organization distributes profits to owners.	[1] For-profit; [0] Not-for-profit
	Accuracy Problem	An algorithm's undesirable ratio of correct predicted outcomes to all observations.	[1] Accuracy; [0] No Accuracy discussed
	User Organization Mitigations	The user organization of an algorithm had mitigations at the level of the specific algorithm in question (3 items listed in the text).	[0] No evidence of mitigation [1] Symbolic evidence of mitigation [2] Substantive evidence of mitigation
	Target Audience Quantity	The number of people whose lives, work, decisions, and opportunities are directly affected by the decision outputs of the algorithm.	[0] Few people; [1] Hundreds; [2] Thousands; [3] Millions; [4] Billions
	Algorithm Runs on a Multi-sided Platform	The algorithm runs on a multi-sided digital platform that has (a) two or more user groups, (b) who need each other, and (c) who cannot capture value by themselves.	[0] Not on a multi-sided platform [1] Runs on a multi-sided platform
	Algorithm Decision-Making Support	The degree to which an algorithm supports human decision-making - based on an algorithm that either fully automates a task, augments with a human or machine as the final decision maker (DM), or a hybrid of automation and augmentation (Teodorescu et al., 2021)	[0] Automation [1] Augmentation – Algorithm Final DM [2] Augmentation – Human Final DM [3] Hybrid
User & Dev	*Industry Similarity*	If based on NAICS SIC, user and developer organizations are in the same industry.	[0] Same NAICS; [1] Different NAICS
	Privacy Breach	An algorithm violates the interest an individual has in influencing the handling of their data by secondary use or collection of data from an unaware individual.	[0] No privacy breach [1] Privacy breach
	IT Failure	A breakdown or malfunction in any component in the algorithm's IT ecosystem	[0] No IT Failure; [1] IT failure

Table 5. Overview of Dependent and Control Study Variables

	Variables	Mean	SD	N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
1	Algorithmic Bias	0.30	0.46	726	1.00																											
2	Damages caused by Algorithmic Bias	0.33	0.37	726	.564**	1.00																										
3	FairnessGoal	0.14	0.34	726	-0.03	-.110**	1.00																									
4	StrictSupervisedLearning	0.20	0.40	726	-0.05	-0.05	0.04	1.00																								
5	HybridLearning	0.45	0.50	726	-.087*	0.00	0.04	-.448**	1.00																							
6	UnilateralUtilityOptimization	0.35	0.48	726	.482**	.478**	-0.04	-0.01	-.085*	1.00																						
7	MultilateralUtilityOptimization	0.29	0.45	726	-.304**	-.363**	.075*	0.05	-0.06	-.468**	1.00																					
8	InteractionCapabilities	1.76	0.80	726	-.157**	-0.01	-0.02	0.03	.341**	-.169**	-.074*	1.00																				
9	AnthropomorphicFeatures_Moderate	0.13	0.33	726	-.141**	0.04	-.081*	-.099**	.261**	-.172**	-0.03	.341**	1.00																			
10	AnthropomorphicFeatures_High	0.04	0.19	726	-0.01	-0.03	-0.06	-0.01	0.04	-.118**	-0.03	.150**	-.077*	1.00																		
11	DeveloperUXGroup	0.75	0.44	726	0.07	.128**	-0.05	-.129**	0.01	0.04	-0.07	-0.01	0.05	-.081*	1.00																	
12	UserCollaborationDev	0.10	0.30	726	-.090*	-.090**	.103**	0.05	0.00	-0.07	0.06	0.03	-0.05	0.05	-0.01	1.00																
13	UserDevelopedOnItsOwn	0.65	0.48	726	.079*	.129**	-.122**	-.163**	-.094*	-0.01	-.100**	-0.06	.146**	0.00	.251**	-.455**	1.00															
14	MachineMakesFinalDecision	0.10	0.29	726	0.00	0.03	0.02	.073*	-0.02	0.03	0.02	-0.04	-0.05	-0.04	-0.02	0.03	-.084*	1.00														
15	HumanMakesFinalDecision	0.22	0.42	726	0.06	0.01	.133**	.156**	0.06	.083*	0.06	-0.01	-.155**	-.107**	-.177**	.116**	-.319**	-.173**	1.00													
16	HumanMachineCollaborates	0.13	0.34	726	-.112**	0.01	-0.06	-.073*	.247**	-.093*	-0.05	.262**	.349**	0.03	.097**	-0.05	0.03	-.127**	-.208**	1.00												
17	AccuracyProblem	0.21	0.41	726	.416**	.414**	0.06	.084*	-0.02	.259**	-.196**	0.03	-0.01	-0.03	-0.04	0.04	-0.07	0.04	.159**	-.100**	1.00											
18	PrivacyBreach	0.14	0.35	726	.214**	.339**	0.00	0.00	.092*	.281**	-.168**	.105**	-0.05	-0.02	-0.01	0.03	-.099**	0.02	.118**	-.088*	.310**	1.00										
19	ModelFailure	0.24	0.43	726	.424**	.483**	-0.06	0.04	-0.05	.224**	-.216**	-0.04	0.05	-0.03	.124**	-0.05	.108**	0.03	-0.03	-0.05	.313**	0.04	1.00									
20	ITFailure	0.08	0.27	726	-0.04	.273**	-.103**	-0.01	.093*	-0.01	-.096**	.179**	.130**	-0.01	0.07	-0.07	.080*	-0.03	-0.05	.170**	0.06	-0.03	0.00	1.00								
21	DeveloperOrgMitigations	0.75	0.44	726	-.105**	-.096**	.145**	-0.02	0.01	-.073*	.134**	-0.02	-0.01	-.067*	.263**	.078*	0.03	-0.03	-0.01	-0.02	-0.06	-0.06	0.00	-0.05	1.00							
22	UserOrgMitigations	0.88	0.61	726	-.159**	-.189**	.140**	-0.04	0.04	-.203**	.150**	0.04	.086**	-0.03	.162**	0.01	.163**	0.01	-.115**	0.03	-.105**	-.122**	-.071*	-0.04	.281**	1.00						
23	TargetAudienceQuantity	2.60	0.70	726	0.05	.112**	0.01	-.174**	-.088*	.075*	-0.04	-.103**	-.088*	-0.04	.245**	-0.05	.274**	-0.01	-.209**	-0.05	-0.07	-0.05	0.07	-0.03	.125**	.087**	1.00					
24	OrgIndustry_DevInManuIndustry	0.10	0.30	726	-.194**	0.04	-.091*	-0.07	.255**	-.137**	-0.03	.311**	.429**	0.03	0.05	-.096*	.118**	-.075*	-.098**	.364**	-.112**	-0.07	-0.03	.228**	-0.05	0.05	-.075*	1.00				
25	OrgIndustry_DevInServiceIndustry	0.86	0.35	726	.137**	-0.04	0.05	0.00	-.194**	.105**	0.03	-.246**	-.337**	0.00	.132**	0.05	-.130**	0.05	-0.02	-.281**	0.05	0.05	0.03	-.184**	.067*	-0.03	.140**	-.805**	1.00		
26	ForProfitStatus	0.75	0.43	726	-.075*	0.00	-.077*	-.178**	-.097**	-0.06	-0.01	-0.02	.124**	.081*	.387**	-0.07	.445**	-.076*	-.349**	.082*	-.187**	-.146**	0.00	0.06	.110**	.096**	.247**	.113**	.122**	1.00	
27	AIgorithmRunsOnPlatform	0.35	0.48	726	.106**	0.03	-0.03	-.297**	-.233**	0.07	0.04	-.317**	-.171**	-0.04	.331**	-.106**	.359**	-0.01	-.269**	-.143**	-.151**	-.157**	0.04	-.122**	.086**	.066*	.434**	-.233**	.294**	.409**	1.00	
28	IndustrySimilarity	0.65	0.48	726	0.05	.097**	-0.06	-.120**	-.089*	-0.02	-.088*	-0.06	.086*	0.03	.145**	-0.05	.548**	-0.06	-.236**	0.06	-0.05	-0.06	0.07	0.03	0.05	.139**	.217**	0.02	-0.06	.432**	.311**	1.00

Two-tailed test of significance; * p<0.05; ** p<0.01; *** p<0.001

Kendall's tau_b for variables that are measured on an ordinal scale.

Table 6. Correlation Table of Study Variables

Estimation Models

The first dependent variable (DV-1), algorithmic bias [1]/ [0], is dichotomous. Thus, we use a binary logistic model to estimate the likelihood of algorithmic bias. The second dependent variable (DV-2), damages caused by algorithmic bias, is a multi-item scale. Thus, we use OLS in estimating the damages.

One identification concern in our empirical modeling is organization-level unobserved heterogeneity. Some developer organizations (e.g., Google, Facebook, etc.) have multiple problematic algorithms in our sample. Thus, we need to account for possible correlations between algorithms developed by the same developer organization. Certain heterogeneity may exist at the developer organization level. We cluster standard errors by unique developer organizations to account for potential unobserved heterogeneity.

Another concern is potential endogeneity. It is possible that both developer and user organizations self-select into implementing governance and control mechanisms due to an omitted variable, and their decisions subsequently influence the likelihood of the emergence of algorithmic bias. This means that both variables are endogenous and can bias the estimation results. We used the instrumental variable estimation technique to address the potential endogeneity problem. A valid instrumental variable must be correlated with the endogenous variables but not directly affect the dependent variable. We used the following instrument: the first proposal of the Algorithmic Accountability Act (AAA) in the U.S. in 2019. Though this proposal was not signed into law, it may have affected the developer and user organizations' adoption of governance and controls by signaling that the government would hold companies accountable for algorithmic biases. Therefore, the proposal was not likely to have a direct effect on the likelihood of the emergence of algorithmic bias. However, it could have an indirect effect through its effects on the developer and user organization's mitigations. Hence, we classified matching pair algorithms into two groups based on the year of emergence of algorithmic bias. If the year of emergence was on or before (vs. after) 2019, the year of the AAA, the developer and user organizations of an algorithm had relatively low (vs. high) awareness of algorithmic bias risks and were less (vs. more) likely to implement controls.

Results

Table 7 illustrates the results. After accounting for the effects of the control variables, we find that relative to unsupervised learning, strictly supervised learning significantly reduced the likelihood of perceived algorithmic bias (beta=-0.77, p<0.05). In contrast, hybrid learning did not have a significant impact (beta=-0.21,p>0.10). Algorithms with high levels of anthropomorphism in their user interfaces increased the likelihood of perceived algorithmic bias (beta=1.36, p<0.05), and moderate anthropomorphism levels increased the magnitude of damages caused by algorithmic bias (beta=0.07, p<0.05). Unilateral stakeholder utility optimization increased the likelihood of algorithmic bias (beta=1.62, p<0.001), while joint stakeholder utility optimization reduced it marginally (beta= -0.65, p<0.10). Collaborating with developer organizations to co-develop algorithms reduced the likelihood of algorithmic bias (beta=-0.95, p<0.10). While developing algorithms in-house did not have a significant impact relative to acquiring algorithms off-the-shelf. These results support H1, H2, H3, and H4. We also ran two-stage least squares (2SLS) with the Algorithmic Accountability Act of 2019 as the instrument variable to account for potential self-selection endogeneity. The results remained qualitatively the same.

	Perceived Algorithmic Bias					Damages caused by Algorithmic Bias			
	Binary Logistic					OLS			
Variables	B	Std. Error	Exp (B)	Sig.		B	Std. Error	Sig.	
Constant	-0.81	0.87	0.44	0.35		0.06	0.07	0.37	
StrictSupervisedLearning	-0.77	0.37	0.46	0.04	*	-0.04	0.03	0.18	
HybridLearning	-0.21	0.29	0.81	0.47		-0.01	0.02	0.61	
AnthropomorphicFeatures_Moderate	-0.46	0.46	0.63	0.31		0.07	0.03	0.03	*
AnthropomorphicFeatures_High	1.36	0.57	3.90	0.02	*	0.07	0.05	0.15	
UnilateralUtilityOptimization	1.62	0.27	5.07	0.00	***	0.22	0.02	0.00	***
MultilateralUtilityOptimization	-0.65	0.37	0.52	0.08	+	-0.03	0.02	0.23	
UserCollaborationDev	-0.95	0.52	0.39	0.07	+	-0.04	0.04	0.32	
UserDevelopedOnItsOwn	0.22	0.39	1.25	0.57		0.01	0.03	0.73	
DeveloperUXGroup	0.84	0.35	2.32	0.02	*	0.06	0.03	0.03	*
FairnessGoal	0.13	0.36	1.14	0.71		-0.02	0.03	0.55	
DeveloperOrgMitigations	-0.83	0.31	0.44	0.01	**	-0.05	0.02	0.05	+
ModelFailure	1.78	0.26	5.94	0.00	***	0.29	0.02	0.00	***
OrgIndustry_DevInManuIndustry	-2.40	1.04	0.09	0.02	*	0.02	0.06	0.77	
OrgIndustry_DevInServiceIndustry	-0.50	0.68	0.61	0.46		-0.08	0.05	0.13	
InteractionCapabilities	-0.35	0.19	0.71	0.07	+	-0.03	0.01	0.06	+
ForProfitStatus	-0.62	0.37	0.54	0.09	+	-0.02	0.03	0.60	
AccuracyProblem	1.64	0.28	5.13	0.00	***	0.14	0.03	0.00	***
UserOrgMitigations	-0.22	0.22	0.80	0.31		-0.05	0.02	0.00	**
TargetAudienceQuantity	-0.08	0.19	0.92	0.65		0.05	0.01	0.00	***
AlgorithmRunsOnPlatform	0.40	0.34	1.49	0.25		0.02	0.03	0.38	
MachineMakesFinalDecision	-0.05	0.38	0.95	0.89		0.05	0.03	0.09	+
HumanMakesFinalDecision	0.22	0.34	1.25	0.51		0.03	0.03	0.19	
HumanMachineCollaborates	0.16	0.43	1.17	0.72		0.05	0.03	0.12	
IndustrySimilarity	0.30	0.32	1.35	0.35		0.04	0.02	0.09	+
PrivacyBreach	0.40	0.32	1.49	0.20		0.23	0.03	0.00	***
ITFailure	-0.22	0.42	0.81	0.61		0.35	0.03	0.00	***
-2 Log likelihood	508.03				Multiple R	0.77			
Cox & Snell R^2	0.40				R^2	0.60			
Nagelkerke R^2	0.57				Adjusted R	0.58			

Two-tailed test of significance; * p<0.05; ** p<0.01; *** p<0.001

Table 7. Likelihood of Perceived Algorithmic bias and damages by Algorithmic Bias

Discussion and Conclusion

This study challenges two commonly held beliefs in the data science community. The first is that issues in AI design and usage can lead to inaccuracies, which subsequently result in AI bias. The second is that these issues can be mitigated through responsible AI design and usage methods. While these beliefs may hold true for certain algorithms, they fail to consider the complexities and irreducible uncertainties in intelligent, agentic AI systems. These agentic AI systems have multiple stakeholders and principals, each with their own objectives or agencies to transfer to the AI. This transfer process is complicated by two main factors: conflicting objectives among stakeholders and the AI's continual updating of its objectives as stakeholders' objectives evolve. According to complexity science, such complexities create irreducible uncertainties, which cannot be fully resolved but could be tamed (Cilliers, 1998).

The learning method of AI serves as a mechanism for transferring agency. Developer organizations have the option to use strict supervised learning methods to manage the irreducible uncertainties regarding which agencies the AI is learning from the data. Involving domain experts early in the training process can enhance the developer organization's ability to make sense of the AI's learning phase, as well as define and measure relevant constructs and fairness metrics, thereby taming the irreducible uncertainties and minimizing AI bias. Although there are costs associated with involving human experts, our findings indicate that strict supervised learning methods are effective in reducing both the likelihood and impact of AI bias. In contrast, in unsupervised or hybrid learning methods where the algorithms may define the constructs themselves, developer organizations lack the ability to govern and control which agencies the AI may adopt and whether those agencies prioritize fairness. As a result, AI bias is more likely to emerge in unsupervised or hybrid learning methods.

The level of anthropomorphism in AI significantly influences how users perceive the AI's agency. A high level of anthropomorphism makes the AI appear human-like, introducing a form of irreducible uncertainty for users. On one side, users who perceive the AI as human may expect it to exhibit human agency, which is often fraught with biases. On the other side, users may place undue trust in a highly anthropomorphic AI, transferring their agency and decision-making rights to it without scrutinizing the AI's decisions. Our findings indicate that stakeholders report more instances of bias in AI systems with high levels of anthropomorphism. In contrast, AI with moderate levels of anthropomorphism is less likely to be reported as biased. Contrary to popular belief, making AI appear more human-like is not always advantageous. Developer organizations should carefully consider both the advantages and disadvantages of increasing the anthropomorphism levels in intelligent AI systems.

The optimization approach of AI has a significant impact on which stakeholders' agencies the AI prioritizes in its resource and opportunity allocation decisions. AI systems serve a diverse range of stakeholders, who often have conflicting utilities and expectations. This introduces an element of irreducible uncertainty into the AI's decision-making process, as it is impractical to maximize the competing utilities of all stakeholders simultaneously. Our research shows that when user organizations deploy AI systems that focus solely on maximizing the utility of a single stakeholder group, other stakeholders are more likely to report instances of AI bias. Conversely, when user organizations opt for a multilateral utility optimization approach, they manage to tame this irreducible uncertainty by seeking tradeoffs among the conflicting utilities of various stakeholders. Although no stakeholder group may be entirely satisfied with the AI's decisions under this approach, they are less likely to report instances of AI bias.

The mode of AI acquisition has a substantial impact on a user organization's ability to address the irreducible uncertainties encountered during the AI's design phase. Organizations that purchase off-the-shelf AI solutions miss the chance to influence the AI's design choices, such as its learning methods, which can align the AI with their own objectives or "agencies." On the other hand, organizations that opt for in-house AI development must confront all the design-phase uncertainties but lack the benefit of external expertise to tame them effectively. Organizations that choose to co-develop AI with specialized developer organizations are better positioned to manage these irreducible uncertainties and, as a result, reduce the likelihood of AI bias.

A boundary condition and a limitation of this study is that it had to create new data sources from scratch as there is currently no systematic database that contains data on the characteristics of a large sample of algorithms and their developers' and users' characteristics. Our theory needs further testing and verifying as alternative data sources emerge. Another limitation is that we could not measure our variables for all years an algorithm existed. As longitudinal datasets emerge on algorithms, we can analyze how time-varying characteristics of algorithms may affect AI bias risks.

References

Agrafiotis, I., Nurse, J. R. C., Goldsmith, M., Creese, S., & Upton, D. (2018). A Taxonomy of Cyber-Harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity*, *4*(1).

Akhila, N., Brett, K., Kavita, S., Roberto, N., & Justin, K. (2018). Automated Classification of Skin Lesions: From Pixels to Practice. *Journal of Investigative Dermatology*, *138*(10), 2108-2110.

Andrew, A., & Qi, Z. (2022). Distributed fairness-guided optimization for coordinated demand response in multi-stakeholder process networks. *Computers & Chemical Engineering*, *161*.

Baird, A., & Maruping, L. M. (2021). The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts. *MIS Quarterly*, *45*(1).

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., & Mojsilovi. (2019). AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4--1.

Beulen, E., Plugge, A., & Van Hillegersberg, J. (2022). Formal and relational governance of artificial intelligence outsourcing. *Information Systems and e-Business Management*, *20*(4), 719-748.

Blut, M., Wang, C., Wnderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: a meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*, *49*, 632-658.

Cilliers, P. (1998). *Complexity and postmodernism: Understanding complex systems*. Routledge.

Clarke, Y. D. (2022). *H.R.6580 - 117th Congress (2021-2022): Algorithmic Accountability Act of 2022*. Retrieved May 28 from https://tinyurl.com/yru354rw

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3).

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of experimental psychology. General*, 114-126.

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Picador.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). On the (im) possibility of fairness. *arXiv preprint*.

Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, *14*(3), 330–347.

Gajane, P., & Pechenizkiy, M. (2017). On formalizing fairness in prediction with machine learning. *arXiv preprint*

Georgescu, M.-I., Ionescu, R. T., & Popescu, M. (2019). Local learning with deep and handcrafted features for facial expression recognition. *IEEE Access*, *7*, 64827-64836.

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627-660.

Guszcza, J., Rahwan, I., Bible, W., Cebrian, M., & Katyal, V. (2018). *Why We Need to Audit Algorithms*. Harvard Business Review. Retrieved May 7 from https://tinyurl.com/326f9nb3

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Hopkins, A., & Booth, S. (2021, 2021). Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.

Kaplan, D. E. (2013). Global investigative journalism: Strategies for support.

Kim, E.-S. (2020). Deep learning and principal–agent problems of algorithmic governance: The new materialism perspective. *Technology in Society*, *63*, 101378.
Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237--293.
Koene, A., Clifton, C., Hatada, Y., Webb, H., Patel, M., & Machado, C. (2019). A governance framework for algorithmic accountability and transparency. *European Parliamentary Research*.
Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388-409.
LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436--444.
Lee, M. K., & Baykal, S. (2017). Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*.
Lee, M. K., Kusbit, D., Kahng, A., Kim, J. T., Yuan, X., Chan, A., See, D., Noothigattu, R., Lee, S., Psomas, A., & Procaccia, A. D. (2019). WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.*, *3*(CSCW).
Leslie, D. (2019). *Understanding Artificial Intelligence Ethics and Safety: A guide for the responsible design and implementation of AI systems in the public sector* (The Alan Turing Institute, Issue.
Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*.
MBFC. (2023). *Media Bias/Fact Check*. https://mediabiasfactcheck.com/
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, *54*(6).
Mosier, K. L., & Skitka, L. J. (1999). Automation use and automation bias. *Human Factors and Ergonomics Society*, *43*(3), 344-348.
Novak, T. P., & Hoffman, D. L. (2019). Relationship journeys in the internet of things: a new framework for understanding interactions between consumers and smart objects. *Journal of the Academy of Marketing Science*, *47*, 216-237.
Nowak, K. L., Fox, J., & Ranjit, Y. S. (2015). Inferences about avatars: Sexism, appropriateness, anthropomorphism, and the objectification of female virtual representations. *Journal of Computer-Mediated Communication*, *20*(5), 554-569.
O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books.
Purdy, M. (2020). *Unlocking AI's Potential for Social Good*. Harvard Business Review. Retrieved April 29 from https://tinyurl.com/yx9tkvr9
RDS. (2022). *Fact (Fairness, Accuracy, Confidentiality, and Transparency)*. Retrieved April 29 from https://redasci.org/
Rubenstein, D. S. (2021). Acquiring ethical AI. *Fla. L. Rev.*, *73*, 747.
Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. Proceedings of the 25th ACM SIGKDD,
Sundar, S. S., Oeldorf-Hirsch, A., & Garga, A. (2008). A cognitive-heuristics approach to understanding presence in virtual environments. International Workshop on Presence,
Teodorescu, M. H., Morse, L., Awwad, Y., & Kane, G. C. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Quarterly*, *45*(3).