WISP 2023 Proceedings

Pre-ICIS Workshop on Information Security and Privacy (SIGSEC)

Winter 12-10-2023

# Effectiveness of Organizational Mitigations for Cybersecurity, Privacy, and IT Failure Risks of Artificial Intelligence

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

John-Patrick O. Akinyemi
*University of Texas at Austin*

# Effectiveness of Organizational Mitigations for Cybersecurity, Privacy, and IT Failure Risks of Artificial Intelligence

**Hüseyin Tanriverdi**[1]
Red McCombs School of Business
University of Texas at Austin,
Austin, Texas, U.S.A

**John-Patrick O. Akinyemi**
Red McCombs School of Business
University of Texas at Austin,
Austin, Texas, U.S.A

## ABSTRACT

Emerging cybersecurity, privacy, and IT failure risks of Artificial intelligence (AI) threaten AI's business value potential and performance of organizations that develop and use AI. Current research on mitigations for these AI risks is limited to technical and data science level mitigations. There is limited research on organizational mitigations for AI risks. We address this gap by framing organizational mitigations for AI's cybersecurity, privacy, and IT failure risks and test their effectiveness in a sample of 498 AI algorithms. Developer organizations, which design AI, and user organizations which use AI, are able to reduce the likelihood and the impact of AI's cybersecurity breach, privacy breach, and IT failure risks if they collaborate to jointly institute organizational mitigations over AI's risks.

**Keywords:** AI, cybersecurity, privacy, IT failure, organizational mitigations

## INTRODUCTION

Despite their many promises, artificial intelligence (AI) algorithms also face emerging risks that threaten their business value potential (Dasgupta et al. 2021; Goodfellow et al. 2015; Szegedy et al. 2014). Adversarial attacks on AI target confidentiality, integrity, and availability of AI models. Some AI algorithms breach the privacy of their stakeholders. Some AI algorithms malfunction due to failures in their sensory IT systems that collect their big data inputs.

[1] Corresponding author. huseyin.tanriverdi@mccombs.utexas.edu +1 512 232 9164

In confidentiality attack, attacker interacts with the AI's black box as a regular user to provide inputs, observe the AI model's outputs, and train a "shadow model" with the same input/output pairs to steal the confidential statistical model of the AI. Attacker can also steal confidential training data of the AI by simply interacting with the algorithm (Wiggers 2021).

In integrity attack, attacker uses "data poisoning" to manipulate the integrity of the AI's training data and inputs. Poisoning refers to imperceptible changes to inputs. Attackers use poisoning attacks to take control of an AI model by imperceptibly modifying a training set, injecting fake data into it, or tampering with the algorithm itself (Duca 2021) and shifting the decision boundary of the AI model in their favor (Khan 2018). Poisoning attacks are prevalent in online learning models that update their learning dynamically as new data emerge.

In availability attack, also known as a "sponge attack," the attacker crafts complex inputs that would maximize the energy consumption and latency of the targeted AI. The algorithm uses all its computational resources to solve the complex problem and becomes unavailable to provide services to legitimate users. The availability attacks delay decisions where real-time performance is necessary, such as in autonomous driving algorithms' perception detection and object classification tasks (Shumailov et al. 2020).

A privacy breach happens if AI algorithm collects and uses personal identifying information (PII) without users' informed consent; or if the algorithm goes beyond the scope of the consent and uses the PII for purposes other than the original purpose for which users gave consent; or if the algorithm compromises users' ability to control their PII (e.g., users cannot opt-out) (Bélanger and Crossler 2011).

An IT failure happens if there is a glitch, outage, or malfunction in the algorithm's sensory IT ecosystem: e.g., sensors, accelerometers, microphones, cameras, LIDAR, telecommunications, etc. (Triche and Walden 2018). The IT failure systematically changes the algorithm's input data and causes it to malfunction or produce unreliable performance.

Despite the mounting evidence on these AI risks, there is limited research on how organizations can defend against then. Most current research on the defense mechanisms is at the level of data science methods: e.g., adversarial deep learning frameworks in academia (Chivukula et al. 2023) and MITRE ATT&CK® framework in practice[2]. There is a shortage of research on organizational-level mitigation mechanisms for these AI risks. We address this gap by theorizing and testing the effectiveness of organizational-level mitigation mechanisms for AI algorithms' cybersecurity, privacy, and IT failure risks.

## THEORETICAL BACKGROUND

We build on organizational control theory (Tanriverdi and Du 2020) in proposing organizational mitigations for AI risks. An organizational control is a process effected by an organization's board of directors, management, and other personnel, designed to provide reasonable assurance regarding the achievement of the organization's objectives.

Cybersecurity is an enterprise-wide risk management issue, not just an IT or algorithm issue (E&Y 2018). Some developer organizations relegate their AI portfolio's cybersecurity risks to data science teams. Relative to such organizations, developer organizations, which institute board-level oversight of their AI portfolio's cybersecurity risks (NACD 2023), disclose the AI portfolio's cybersecurity risks, and have independent audits of their governance and controls over the AI portfolio's cybersecurity risks (Schoenfeld 2022), are more likely to mitigate the AI's cybersecurity risks. However, developer organization's mitigations alone may not suffice to secure an AI in use. User organizations of the AI should also institute similar governance and

[2] https://atlas.mitre.org/

controls over AI algorithms they acquire from developer organizations and train an algorithm's stakeholders about cybersecurity risks and protections.

AI privacy risk management also entails an enterprise-wide approach covering management structures and policies around PII collected and processed by AI algorithms: e.g., policies around notice, collection, use, sharing, retention, and disposal of PII throughout its lifecycle. Compared to developer organizations that do not systematically manage privacy risks of AI, developer organizations that commit to "privacy by design" principles (Cavoukian 2009); view privacy as a core organizational value (van de Poel 2020); and use organizational privacy policies to govern the privacy rights, expectations, and concerns of their AI algorithms' stakeholders, are more likely to mitigate AI's privacy risks. User organizations that acquire AI from the developer organizations should also institute similar privacy governance and controls to complement those of the developer organizations.

AI algorithms run over developer and user organizations' digital foundations, which collect sensory data inputs for AI and provide computational resources for AI. Developer and user organizations of AI should govern the operational risks of those IT foundations to ensure that the AI's big data inputs are collected and processed accurately. Developer and user organizations that have board-level oversight of operational IT risks (Benaroch and Chernobai 2017), disclose IT risks, and use IT Governance Frameworks (e.g., COBIT, ITIL, ISO27000 standards) and IT controls are more likely to mitigate operational IT risks of the AI algorithms such as errors, glitches, and outages in sensory IT components of AI.

## METHODS

Our sampling frame was a repository of problematic algorithms maintained by "AI Algorithmic and Automation Incidents and Controversies" (AIAAIC), an initiative that supports responsible AI use and development. We downloaded the repository in September 2022. We also supplemented it by systematically searching for problematic algorithms in Google, Factiva, EBSCOhost, Web of Science, and Google Scholar. We then analyzed all allegedly problematic algorithms to select the ones that satisfy the following inclusion criteria: **(i) Algorithm:** Problematic algorithm met the definition of an intelligent algorithm. It learns from patterns in big data inputs and alters its behavior based on input changes. **(ii) Problem type:** The algorithm had a cybersecurity breach, or a privacy breach, or an IT failure. The repository also contained algorithms with biased outcomes or model failures. These were included only if the algorithm had one of the other problems central to our study. **(iii) Realized or Potential problem:** The problematic algorithm had a realized problem. We excluded entries that discussed concerns that have not been realized 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. **(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 problem emerged was known. **(vii) Location of user organization:** The user organization of the problematic algorithm had to be incorporated in the US.

We found a matching, problem-free algorithm to create a matched pair for each problematic algorithm using the following matching criteria:

**(i) Timing:** The matching algorithm had to be in use as of the year of the problematic algorithm's problem emergence. All matching criteria had to be satisfied as of that year. **(ii) Problem status:** The matching algorithm had to be free of any reports of IT Failure, Privacy breach, Cybersecurity breach, Bias, and Model failure in the matching year. **(iii) Application**

**domain**: The matching algorithm had to be in the same application domain as the problematic algorithm. **(iv) Function:** The matching algorithm had to have the same function as the problematic algorithm. **(v) Platform status:** The matching algorithm had the same on-platform/off-platform status as the problematic algorithm**. (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/private sector status:** The matching algorithm's user organization had to be in the same sector. **(vii) Industry:** The pair's developer organization had to have similar NAICS industry and SIC sector codes.

The final sample had 249 pairs of problematic and problem-free algorithms, i.e., 498 algorithms, from 16 industry segments (e.g., HRTech, FinTech, Criminal Justice, Education, etc.) and 121 functional categories (e.g., content moderation, text-to-speech, search-matching, price prediction, etc.) being used in the U.S. between 2007 and 2022. There were 88 pairs with a cybersecurity breach, 120 with a privacy breach, and 73 with IT failure. Appendix Table 1 explains the sample construction process.

### Source Documents and Coding Instrument

A combination of about 40 undergraduate and 35 graduate IS students collected source documents needed for coding the study variables. They did systematic keyword searches in the Factiva database, SEC filings, company websites, and Google to find sources discussing the characteristics of an algorithm and its developer and user organizations (e.g., peer-reviewed academic publications, mainstream news articles, investigative journalism articles, 10K and DEF14A filings to the SEC, company websites, etc.). After students found the relevant source documents, two expert coders did the coding. We developed and validated a guideline for coding 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. A Ph.D. student and a Master's student with degrees and professional experiences in IS and training and research experience in Data Science served as two independent expert coders who read the source documents to code the variables by following the validated coding guideline. We established the reliability of the coding guidelines by following an iterative process across several rounds of coding. In the first round, the two independent coders used the guidelines to code a small sample of five algorithms. We assessed the inter-coder agreement rate after each coding round. After the first round, the agreement rate was 68%. The coders discussed coding discrepancies to find that some variables were not tightly defined. Hence, we revised the definitions. After three iterations, the inter-coder agreement rate increased above the 90% threshold for establishing the reliability of the coding instrument.

## Dependent Variables

Risk is conceptualized in terms of the chance of loss (i.e., the likelihood of an occurrence) and the magnitude of loss (i.e., damages caused by an occurrence) (Tanriverdi & Ruefli, 2004). Inspired by this principle, we use two sets of dependent measures of AI risk.

### *Algorithmic Problem*

Independent coders measured an algorithm’s problems by assessing if it had (i) an IT Failure, (ii) a Cybersecurity Breach, or (iii) a Privacy Breach.

**(i) IT Failure:** An algorithm was marked as [1: IT Failure] if the independent coders observed a breakdown or malfunction in any component in the algorithm's IT ecosystem that rendered the IT ecosystem incapable of performing its intended tasks (Triche & Walden, 2018), or [0: No IT Failure] if there was no evidence of such a failure. **(ii) Cybersecurity Breach:** The coders selected [1: Cybersecurity Breach] if there was evidence of a malfunction in the algorithm or algorithm ecosystem due to malicious, unauthorized access that compromised the algorithm's or its data's confidentiality, integrity, or availability (Samonas & Coss, 2014); or evidence of an adversarial attack that fooled the algorithm. If no such evidence was found, coders selected [0: No Cybersecurity Breach]. **(iii) Privacy Breach:** [1: Privacy Breach] was selected if the algorithm collected and used PII without users' informed consent or if it went beyond the scope of the consent and used the PII for purposes other than the original purpose for which users gave the consent; or if the algorithm compromised users' ability to control their PII (Belanger et al., 2002); (Clarke, 1999). If no such evidence was found, the coder selected [0: No Privacy Breach].

***Damages caused by Algorithmic Problem***

Independent coders used four items to measure if a user organization of an algorithm suffered damages due to an IT Failure, Cybersecurity Breach, or Privacy Breach problem in the algorithm. The coders read every source on the problematic algorithm to code if it: (1) harmed customers or employees of the user organization [1] or not [0]; (2) caused financial loss (e.g., regulatory fine, compensation to victims) to the user organization [1], or not [0]; (3) harmed the user organization's reputation (e.g., bad press and pressure on the user organization to meet socially accepted standards) [1], or not [0]); and (4) led to a lawsuit [1] on the user organization, or not [0]. Cronbach's Alpha of the items was 0.847, indicating sufficient reliability.

## Independent Variables

The independent coders reviewed the source documents code to determine if there was evidence indicating that the organization had governance and controls to mitigate risks related to cybersecurity, privacy, or IT failure to its portfolio of algorithms. For each mitigation, the following scale was used: [0]: no evidence of the mitigation; [1]: symbolic evidence of the mitigation; and [2]: substantive evidence of the mitigation. As a result, we created three multi-item constructs. Each construct consisted of six measurement items, three from the developer organization and three from user organization. Cronbach's Alpha values of all three constructs (0.826, 0.854, and 0.894) demonstrated sufficient reliability. **(i) Organizations' Cybersecurity Risk Disclosures and Mitigations:** We measured if developer and user organizations made Cybersecurity Risk Disclosures and had Board-level Oversight of Cybersecurity Risks. For developer organization, we also measured if the developer had a System and Organization Controls (SOC) Report prepared by an independent auditing firm. For user organizations, we measured if the user organization had a Cybersecurity Risk Training program. **(ii) Organizations' Privacy Risk Disclosures and Mitigations:** We measured if developer and user organizations had Privacy Policies, adopted Privacy by Design principles, and viewed Privacy as a Core Value. **(iii) Organizations' IT Failure Risk Disclosures and Mitigations:** We measured if developer and user organizations made IT Failure Risk Disclosures, had Board-level Oversight of IT Failure Risks, and IT Failure Risk Mitigations. We added fifteen controls for alternative explanations and potential endogeneity concerns, as shown in Appendix Table 2.

Appendix Table 3 provides illustrative evidence of our coding for cybersecurity mitigations. Two key assumptions were made during coding. First, we sought the specific business unit that used the algorithm to code user organization-related variables for organizations in which the user and developer org were alike. Second, for algorithmic businesses, which use the terms algorithm and technology interchangeably, we looked for mention of technology risk mitigations; we did not necessarily require specific mention of algorithm risk mitigations.

## RESULTS

Tables 1 and 2 present the results on the likelihood and impact (magnitude of damages) of the AI Risks. The first result column in these tables uses a pooled sample of all pairs of algorithms. In contrast, the second, third, and fourth results columns use the subsamples of cybersecurity, privacy, and IT failure pairs.

The first results column in Table 1 shows that the combined cybersecurity, privacy, and IT failure mitigations of the developer and the user organizations are ineffective in reducing the likelihood of a cybersecurity, privacy, or IT failure problem in AI. However, Table 2 shows that the combined mitigations are effective in significantly reducing the magnitude of the damages caused by a cybersecurity, privacy, or IT failure problem in AI. The second results column in Table 1 shows that the combined cybersecurity mitigations of developer and user organizations significantly reduce the likelihood of a cybersecurity breach in AI. Table 2 shows further that the combined cybersecurity mitigations also significantly reduce the magnitude of damages caused by cybersecurity breaches in AI. The third results column in Table 1 shows that the combined privacy mitigations of developer and user organizations are ineffective in reducing the likelihood of a privacy breach in AI. However, Table 2 shows that the combined privacy mitigations effectively reduce the magnitude of damages caused by a privacy breach in AI. The fourth results column in Table 1 shows that the combined IT failure mitigations of developer and user organizations significantly reduce the likelihood of an IT failure in AI's IT ecosystem. Table 2 shows further that the combined IT failure mitigations also significantly reduce the magnitude of damages caused by an IT failure in AI's IT ecosystem.

The results on the control variables also generate exploratory insights. Specifically, AI's fairness goal, optimization approach, decision support mode, size of target audience served, and stakeholder management quality significantly affect the likelihood of one or more of the three AI risks. As for the magnitude of damages caused, AI's ground truth status, optimization approach, decision support mode, size of target audience served, number of stakeholders, for-profit status, industry similarity (between AI's developer and user), and stakeholder management quality significantly affect the impact of one or more of the three AI risks.

**Table 1.** Likelihood of Problem Emergence in Algorithm

| | Problem | | | | | CyberSecurityBreach | | | | | PrivacyBreach | | | | | IT Failure | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B | S.E. | Exp(B) | Sig. | | B | S.E. | Exp(B) | Sig. | | B | S.E. | Exp(B) | Sig. | | B | S.E. | Exp(B) | Sig. | |
| GT_Well_Established | -0.316 | 0.235 | 0.729 | 0.178 | | -0.021 | 0.483 | 0.979 | 0.965 | | -0.660 | 0.374 | 0.517 | 0.078 | + | -0.058 | 0.486 | 0.943 | 0.904 | |
| GT_Multiple | -0.163 | 0.400 | 0.850 | 0.685 | | 0.659 | 0.678 | 1.933 | 0.331 | | -0.143 | 0.666 | 0.867 | 0.830 | | -0.329 | 1.008 | 0.720 | 0.744 | |
| SupervisedLearning | -0.106 | 0.313 | 0.899 | 0.734 | | 0.862 | 0.632 | 2.369 | 0.173 | | -0.443 | 0.463 | 0.642 | 0.340 | | -0.760 | 0.883 | 0.468 | 0.389 | |
| HybridLearning | 0.032 | 0.291 | 1.032 | 0.914 | | 0.934 | 0.592 | 2.545 | 0.114 | | -0.140 | 0.421 | 0.869 | 0.740 | | -0.084 | 0.827 | 0.920 | 0.919 | |
| Fairness | -1.264 | 0.367 | 0.283 | 0.001 | *** | -21.901 | 12187.011 | 0.000 | 0.999 | | -0.979 | 0.437 | 0.376 | 0.025 | * | -2.492 | 1.130 | 0.083 | 0.027 | * |
| Humanized_Moderate | 0.470 | 0.315 | 1.600 | 0.135 | | 0.562 | 0.577 | 1.754 | 0.330 | | -0.009 | 0.609 | 0.991 | 0.989 | | 0.089 | 0.599 | 1.093 | 0.882 | |
| Humanized_High | 0.548 | 0.449 | 1.730 | 0.223 | | 0.788 | 0.716 | 2.199 | 0.271 | | 1.054 | 0.776 | 2.870 | 0.174 | | 0.394 | 1.072 | 1.483 | 0.713 | |
| InteractionCapabilities | 0.067 | 0.131 | 1.069 | 0.611 | | 0.089 | 0.259 | 1.093 | 0.731 | | 0.175 | 0.237 | 1.192 | 0.460 | | 0.029 | 0.257 | 1.029 | 0.911 | |
| UnilateralOptimization | 1.537 | 0.268 | 4.651 | 0.000 | *** | 1.325 | 0.570 | 3.762 | 0.020 | * | 2.137 | 0.387 | 8.471 | 0.000 | *** | 1.116 | 0.638 | 3.051 | 0.080 | + |
| MultilateralOptimization | -0.693 | 0.282 | 0.500 | 0.014 | * | -1.014 | 0.500 | 0.363 | 0.043 | * | -0.130 | 0.467 | 0.878 | 0.780 | | -0.689 | 0.562 | 0.502 | 0.220 | |
| UserAcquisitionMode_Collaboration | -0.204 | 0.376 | 0.815 | 0.586 | | -0.936 | 0.737 | 0.392 | 0.204 | | -0.096 | 0.524 | 0.908 | 0.854 | | -1.346 | 1.090 | 0.260 | 0.217 | |
| UserAcquisitionMode_InHouse | 0.207 | 0.311 | 1.230 | 0.506 | | 0.517 | 0.554 | 1.677 | 0.351 | | -0.021 | 0.477 | 0.979 | 0.965 | | 0.535 | 0.742 | 1.708 | 0.471 | |
| NumberStakeholder | 0.052 | 0.072 | 1.054 | 0.469 | | 0.089 | 0.135 | 1.093 | 0.508 | | -0.066 | 0.127 | 0.937 | 0.607 | | 0.043 | 0.141 | 1.044 | 0.758 | |
| StakeholderManagementQuality | 0.174 | 0.311 | 1.190 | 0.577 | | 0.489 | 0.625 | 1.631 | 0.434 | | 0.581 | 0.545 | 1.788 | 0.286 | | 0.141 | 0.593 | 1.151 | 0.812 | |
| TargetAudienceQuantity | 0.326 | 0.189 | 1.386 | 0.084 | + | 0.842 | 0.351 | 2.320 | 0.016 | * | 0.112 | 0.295 | 1.118 | 0.706 | | 0.733 | 0.409 | 2.081 | 0.073 | + |
| OnPlatformOrOffPlatform | 0.059 | 0.268 | 1.060 | 0.827 | | 0.134 | 0.504 | 1.143 | 0.791 | | 0.382 | 0.432 | 1.466 | 0.376 | | -0.354 | 0.689 | 0.702 | 0.607 | |
| ForProfitorNonProfit | 0.034 | 0.298 | 1.034 | 0.910 | | 0.182 | 0.599 | 1.199 | 0.761 | | -0.039 | 0.420 | 0.962 | 0.926 | | 0.193 | 0.781 | 1.213 | 0.805 | |
| MachineMakesFinalDec | 0.013 | 0.392 | 1.013 | 0.974 | | 0.356 | 0.639 | 1.427 | 0.578 | | -0.113 | 0.742 | 0.893 | 0.879 | | 0.259 | 0.877 | 1.296 | 0.768 | |
| HumanMakesFinalDec | 0.276 | 0.302 | 1.318 | 0.361 | | 0.952 | 0.550 | 2.590 | 0.084 | + | 0.349 | 0.457 | 1.418 | 0.445 | | -0.038 | 0.836 | 0.963 | 0.964 | |
| HumanMachineCollaboration | -0.434 | 0.343 | 0.648 | 0.206 | | 0.230 | 0.705 | 1.259 | 0.744 | | -1.027 | 0.722 | 0.358 | 0.154 | | -0.114 | 0.527 | 0.893 | 0.829 | |
| AlgorithmRepurposed | 0.533 | 0.340 | 1.704 | 0.117 | | 0.630 | 0.651 | 1.877 | 0.334 | | 0.750 | 0.471 | 2.117 | 0.112 | | 0.630 | 0.788 | 1.877 | 0.424 | |
| IndustrySimilarity | 0.234 | 0.305 | 1.263 | 0.444 | | -0.217 | 0.581 | 0.805 | 0.709 | | 0.794 | 0.479 | 2.213 | 0.097 | + | -0.478 | 0.731 | 0.620 | 0.513 | |
| **CombinedMitigations** | **-0.328** | **0.253** | **0.720** | **0.194** | | | | | | | | | | | | | | | | |
| **CombinedCyberMitigations** | | | | | | **-1.180** | **0.420** | **0.307** | **0.005** | ** | | | | | | | | | | |
| **CombinedPrivacyMitigations** | | | | | | | | | | | **-0.050** | **0.324** | **0.951** | **0.878** | | | | | | |
| **CombinedITFailureMitigations** | | | | | | | | | | | | | | | | **-1.651** | **0.468** | **0.192** | **0.000** | *** |
| Constant | -1.482 | 0.686 | 0.227 | 0.031 | | -3.570 | 1.426 | 0.028 | 0.012 | | -1.236 | 1.134 | 0.290 | 0.276 | | -0.564 | 1.591 | 0.569 | 0.723 | |

**Table 2.** Impact of Algorithmic Problems (Damages Caused)

| | Damages (Overall Sample) | | | | Damages (CybersecurityPairs) | | | | Damages (PrivacyPairs) | | | | Damages (ITFailurePairs) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B | Std. Error | Sig. | | B | Std. Error | Sig. | | B | Std. Error | Sig. | | B | Std. Error | Sig. | |
| CSecurityBreach | 0.410 | 0.027 | 0.000 | *** | 0.416 | 0.027 | 0.000 | *** | 0.414 | 0.027 | 0.000 | *** | 0.030 | 0.069 | 0.665 | |
| PrivacyBreach | 0.452 | 0.027 | 0.000 | *** | 0.450 | 0.027 | 0.000 | *** | 0.456 | 0.027 | 0.000 | *** | 0.065 | 0.068 | 0.340 | |
| ITFailure | 0.420 | 0.031 | 0.000 | *** | 0.421 | 0.031 | 0.000 | *** | 0.421 | 0.031 | 0.000 | *** | 0.639 | 0.042 | 0.000 | *** |
| ModelFailure | 0.129 | 0.036 | 0.000 | *** | 0.125 | 0.036 | 0.000 | *** | 0.128 | 0.036 | 0.000 | *** | 0.112 | 0.052 | 0.034 | * |
| BiasPresent | -0.012 | 0.035 | 0.727 | | -0.009 | 0.036 | 0.809 | | -0.014 | 0.036 | 0.690 | | -0.132 | 0.057 | 0.022 | * |
| GT_Well_Established | 0.048 | 0.023 | 0.037 | * | 0.046 | 0.023 | 0.046 | * | 0.047 | 0.023 | 0.042 | * | 0.039 | 0.037 | 0.293 | |
| GT_Multiple | -0.011 | 0.039 | 0.780 | | -0.016 | 0.039 | 0.687 | | -0.008 | 0.039 | 0.846 | | 0.016 | 0.077 | 0.831 | |
| SupervisedLearning | -0.017 | 0.030 | 0.574 | | -0.013 | 0.031 | 0.660 | | -0.013 | 0.030 | 0.680 | | -0.040 | 0.066 | 0.543 | |
| HybridLearning | -0.037 | 0.028 | 0.191 | | -0.036 | 0.028 | 0.198 | | -0.038 | 0.028 | 0.182 | | -0.041 | 0.063 | 0.521 | |
| Fairness | -0.030 | 0.033 | 0.360 | | -0.032 | 0.033 | 0.339 | | -0.027 | 0.033 | 0.423 | | -0.020 | 0.076 | 0.790 | |
| Humanized_Moderate | 0.016 | 0.031 | 0.602 | | 0.018 | 0.031 | 0.557 | | 0.014 | 0.031 | 0.661 | | -0.051 | 0.050 | 0.306 | |
| Humanized_High | 0.020 | 0.046 | 0.661 | | 0.021 | 0.047 | 0.648 | | 0.021 | 0.047 | 0.656 | | -0.035 | 0.084 | 0.679 | |
| InteractionCapabilities | 0.012 | 0.013 | 0.348 | | 0.014 | 0.013 | 0.284 | | 0.012 | 0.013 | 0.334 | | -0.023 | 0.020 | 0.250 | |
| UnilateralOptimization | 0.070 | 0.027 | 0.008 | ** | 0.072 | 0.027 | 0.007 | ** | 0.070 | 0.027 | 0.009 | ** | -0.041 | 0.048 | 0.391 | |
| MultilateralOptimization | -0.041 | 0.027 | 0.126 | | -0.042 | 0.027 | 0.122 | | -0.041 | 0.027 | 0.130 | | -0.021 | 0.045 | 0.644 | |
| UserAcquisitionMode_Collaboration | 0.002 | 0.036 | 0.963 | | -0.007 | 0.036 | 0.846 | | -0.009 | 0.036 | 0.793 | | 0.007 | 0.075 | 0.930 | |
| UserAcquisitionMode_InHouse | 0.004 | 0.030 | 0.884 | | -0.001 | 0.030 | 0.979 | | -0.002 | 0.030 | 0.938 | | 0.050 | 0.058 | 0.391 | |
| NumberStakeholder | 0.020 | 0.007 | 0.004 | ** | 0.020 | 0.007 | 0.006 | ** | 0.018 | 0.007 | 0.011 | * | 0.021 | 0.011 | 0.060 | + |
| StakeholderManagementQuality | -0.022 | 0.031 | 0.477 | | -0.025 | 0.031 | 0.408 | | -0.026 | 0.031 | 0.395 | | -0.122 | 0.048 | 0.012 | * |
| TargetAudienceQuantity | 0.033 | 0.018 | 0.070 | + | 0.026 | 0.018 | 0.150 | | 0.027 | 0.018 | 0.133 | | 0.014 | 0.031 | 0.645 | |
| OnPlatformOrOffPlatform | 0.022 | 0.026 | 0.408 | | 0.016 | 0.026 | 0.549 | | 0.027 | 0.026 | 0.300 | | -0.059 | 0.051 | 0.254 | |
| ForProfitorNonProfit | 0.054 | 0.029 | 0.063 | + | 0.052 | 0.029 | 0.071 | + | 0.048 | 0.029 | 0.099 | + | 0.003 | 0.063 | 0.962 | |
| MachineMakesFinalDec | -0.024 | 0.038 | 0.539 | | -0.025 | 0.038 | 0.522 | | -0.028 | 0.038 | 0.471 | | 0.043 | 0.069 | 0.536 | |
| HumanMakesFinalDec | 0.044 | 0.029 | 0.133 | | 0.042 | 0.029 | 0.151 | | 0.046 | 0.029 | 0.115 | | -0.035 | 0.061 | 0.566 | |
| HumanMachineCollaboration | 0.062 | 0.034 | 0.071 | + | 0.053 | 0.034 | 0.121 | | 0.062 | 0.034 | 0.073 | + | 0.116 | 0.042 | 0.007 | ** |
| AlgorithmRepurposed | -0.026 | 0.032 | 0.417 | | -0.024 | 0.032 | 0.455 | | -0.024 | 0.032 | 0.464 | | 0.021 | 0.059 | 0.724 | |
| IndustrySimilarity | -0.073 | 0.030 | 0.014 | * | -0.076 | 0.030 | 0.011 | * | -0.073 | 0.030 | 0.015 | * | -0.118 | 0.057 | 0.041 | * |
| **CombinedMitigations** | **-0.082** | **0.025** | **0.001** | *** | | | | | | | | | | | | |
| **CombinedCyberMitigations** | | | | | **-0.049** | **0.021** | **0.018** | * | | | | | | | | |
| **CombinedPrivacyMitigations** | | | | | | | | | **-0.049** | **0.019** | **0.011** | * | | | | |
| **CombinedITFailureMitigations** | | | | | | | | | | | | | **-0.041** | **0.036** | **0.255** | |
| (Constant) | -0.109 | 0.068 | 0.109 | | -0.111 | 0.069 | 0.107 | | -0.089 | 0.068 | 0.195 | | 0.056 | 0.125 | 0.657 | |

## DISCUSSION

**Contributions to research.** The study advances the literature on AI risk mitigations. Conceptually, the study goes beyond the extant data scientific mitigations focusing on adversarial learning frameworks to address AI risks. It complements them with organizational mitigation mechanisms. Factor analysis results reveal an interesting insight. Measurement items of developer and user organizations' mitigations for a given AI risk load on the same factor. For instance, the three AI cybersecurity risk mitigation items of the developer organization and the three AI cybersecurity risk mitigation items of the user organization load onto the same factor. Likewise, developer and user organizations' three AI privacy risk mitigation items load onto the same factor. The developer and user organizations' three IT failure risk mitigation items load onto the same factor. These patterns point to dependencies between the developer organization's and the user organizations' AI risk mitigation mechanisms. These dependencies require the developer and user organizations to jointly institute mitigations for AI risks.

Empirically, to our knowledge, this is the first study to conduct a large sample empirical test of the effectiveness of organizational-level mitigations for AI's cybersecurity, privacy, and IT failure risks. The findings indicate that organizational mitigations are generally effective in mitigating AI's risks. However, they also have limitations. For instance, privacy mitigations are unable to reduce the likelihood of privacy breaches in AI, but if privacy breaches emerge, they reduce the magnitude of damages. Similarly, IT failure mitigations reduce the likelihood of an IT failure in AI's IT ecosystem, but if an IT failure emerges, they do not reduce the magnitude of damages.

**Contributions to practice.** The results alert executives of developer and user organizations of AI that they should not independently institute organizational mitigations over AI risks. Rather, developer and user organizations should collaborate to jointly institute organizational mitigations that complement each other.

**Limitations and future work.** A limitation of the study was its lack of access to technical mitigations of algorithms. This limitation inhibited our ability to measure data scientific mitigations for cybersecurity, privacy, and IT failure risks. We do not know if the organizational mitigations we were able to measure might serve as proxies for technical mitigations implemented in the algorithms. Future research designs can aim to study both types of mitigations simultaneously to understand their respective roles and relative effectiveness in mitigating AI risks. Another limitation was that we had to create new data sources from scratch. There is currently no systematic database that contains data on the characteristics of a large sample of algorithms and their developers' and users' AI risk mitigation mechanisms. Our theory needs further testing and verifying as alternative data sources emerge. A third limitation is that we could not measure our variables for all years an algorithm existed. As longitudinal datasets emerge on algorithms, we can conduct longitudinal analyses on how time-varying characteristics of algorithms might affect the emergence of AI risks. Finally, this study focused on the defense side of the equation. Future research can also focus on the attack side to understand which methods attackers use to breach or fool AI.

## REFERENCES

AIAAIC. 2023. "Aiaaic Repository Governance." Retrieved June 14, 2023, from https://www.aiaaic.org/aiaaic-repository/governance

Anley, C. 2020. "Practical Attacks on Machine Learning Systems." Manchester, UK: NCC Group, Plc., pp. 1-26.

Bélanger, F., and Crossler, R. E. 2011. "Privacy in the Digital Age: A Review of Information Privacy Research in Information Systems," *MIS Quarterly*).

Benaroch, M., and Chernobai, A. 2017. "Operational IT Failures, IT Value Destruction, and Board-Level IT Governance Changes," *MIS Quarterly* (41:3), pp. 729-+.

Blut, M., Wang, C., Wnderlich, N. V., and 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), pp. 632-658.

C3.ai. 2022. "Glossary - Ground Truth." Retrieved April 23, 2022, from https://tinyurl.com/54usztwt

Cavoukian, A. 2009. "Privacy by Design: The 7 Foundational Principles," *Information and privacy commissioner of Ontario, Canada* (5), p. 12.

Chivukula, A. S., Yang, X., Liu, B., Liu, W., and Zhou, W. 2023. *Adversarial Machine Learning: Attack Surfaces, Defence Mechanisms, Learning Theories in Artificial Intelligence*. Springer International Publishing.

Cinà, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B., Pelillo, M., and Roli, F. 2022. "Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning," *ACM Computing Surveys*).

Cohen, V. 2019. "Opengpt-2: We Replicated Gpt-2 Because You Can Too." Retrieved May 3, 2022, from https://tinyurl.com/jjhpk9yz

Cordoba, P. A., and Gonzalez, J. R. 2022. "Google to Pay Arizona $85m in Privacy Suit That Alleged 'Deceptive' Location Tracking." Retrieved June 15, 2023, from https://tinyurl.com/y3b527st

Dasgupta, P., Collis, J. B., and Mittu, R. 2021. *Adversary-Aware Learning Techniques and Trends in Cybersecurity*. Cham, Switzerland: Springer.

Duca, A. L. 2021. "Adversarial Machine Learning: Attacks and Possible Defense Strategies." Retrieved May 3, 2022, from https://tinyurl.com/2p9fbrz3

E&Y. 2018. "Cybersecurity Disclosure Benchmarking," EY Conter for Board Matters, pp. 1-6.

Eitel-Porter, R. 2020. "Beyond the Promise: Implementing Ethical AI," *AI and Ethics* (1).

Evans, D. S., Hagiu, A., and Schmalensee, R. 2008. *Invisible Engines: How Software Platforms Drive Innovation and Transform Industries*. The MIT Press.

Freeman, R. E., Kujala, J., and Sachs, S. 2017. *Stakeholder Engagement: Clinical Research Cases*. New York: Springer.

Goodfellow, I. J., Shlens, J., and Szegedy, C. 2015. "Explaining and Harnessing Adversarial Examples," *CoRR* (abs/1412.6572).

Hailu, R. 2019. "Fitbits and Other Wearables May Not Accurately Track Heart Rates in People of Color." Retrieved June 15, 2023, from https://tinyurl.com/2nbh6snn

Herbert-Voss, A. 2020. "Practical Defenses against Adversarial Machine Learning." Retrieved April 3, 2022, from https://www.youtube.com/watch?v=RdHYZJ2S_Zk

Humphries, M. 2022. "Iphone 14 Crash Detection Calls 911 on Roller Coasters." Retrieved June 15, 2023, from https://tinyurl.com/ycxcyuja

Iansiti, M., and Lakhani, K. R. 2020. *Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World*. Boston, MA: Harvard Business Review Press.

Khan, M. 2018. "How Can Companies Defend against Adversarial Machine Learning Attacks in the Age of AI?" Retrieved MAy 3, 2022, from https://tinyurl.com/2p8a2ez3

Kurakin, A., Goodfellow, I. J., and Bengio, S. 2017. "Adversarial Examples in the Physical World," *ArXiv* (abs/1607.02533).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., and Gershman, S. J. 2017. "Building Machines That Learn and Think Like People," *Behavioral and Brain Sciences* (40).

Langer, M., and Landers, R. 2021. "The Future of Artificial Intelligence at Work: A Review on Effects of Decision Automation and Augmentation on Workers Targeted by Algorithms and Third-Party Observers," *Computers in Human Behavior* (123).

Lee, M. K., and 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*).

MBFC. 2023. "Media Bias/Fact Check." from https://mediabiasfactcheck.com/

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. 2021. "A Survey on Bias and Fairness in Machine Learning," *ACM Computing Surveys* (54:6).

Mitchell, R. K., Agle, B. R., and Wood, D. J. 1997. "Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts," *Academy of Management Review* (22:4), pp. 853-886.

Muller, M., Wolf, C. T., Andres, J., Desmond, M., Joshi, N. N., Ashktorab, Z., Sharma, A., Brimijoin, K., Pan, Q., and Duesterwald, E. 2021. "Designing Ground Truth and the Social Life of Labels," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1-16.

NACD. 2023. "Director's Handbook on Cyber-Risk Oversight," The National Association of Corporate Directors and the Internet Security Alliance, pp. 1-100.

Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*:5), pp. 206-215.

Samonas, S., and Coss, D. 2014. "The CIA Strikes Back: Redefining Confidentiality, Integrity and Availability in Security," *Journal of Information System Security* (10:3).

Schoenfeld, J. 2022. "Cyber Risk and Voluntary Service Organization Control (Soc) Audits," *Review of Accounting Studies*).

Shumailov, I., Zhao, Y., Bates, D., Papernot, N., Mullins, R. D., and Anderson, R. 2020. "Sponge Examples: Energy-Latency Attacks on Neural Networks," *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*), pp. 212-231.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. 2014. "Intriguing Properties of Neural Networks," *CoRR* (abs/1312.6199).

Tanriverdi, H., and Du, K. 2020. "Corporate Strategy and Information Technology Control Effectiveness," *MIS Quarterly* (44:4).

Teodorescu, M. H., Morse, L., Awwad, Y., and Kane, G. C. 2021. "Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation," *MIS Quarterly* (45:3).

Thompson, E. 2021. "U.S. Technology Company Clearview AI Violated Canadian Privacy Law." Retrieved June 15, 2023, from https://tinyurl.com/3u4n6vbe

Triche, J. H., and Walden, E. 2018. "The Use of Impression Management Strategies to Manage Stock Market Reactions to IT Failures," *Journal of the Association for Information Systems*).

van de Poel, I. 2020. "Core Values and Value Conflicts in Cybersecurity: Beyond Privacy Versus Security," in *The Ethics of Cybersecurity,* M. Christen, B. Gordijn and M. Loi (eds.). Cham: Springer International Publishing, pp. 45-71.

Wiggers, K. 2021. "Adversarial Attacks in Machine Learning: What They Are and How to Stop Them." Retrieved June 15, 2023, from https://tinyurl.com/2vv9f5hc

Zuo, Z. M., Watson, M., Budgen, D., Hall, R., Kennelly, C., and Al Moubayed, N. 2021. "Data Anonymization for Pervasive Health Care: Systematic Literature Mapping Study," *Jmir Medical Informatics* (9:10).

# APPENDIX A – ILLUSTRATIVE TABLES OF METHOD AND CODING EVIDENCE

**Appendix Table 1.** Sample Construction Process

| Step | Description of action taken | Size |
|---|---|---|
| 1 | Download problematic algorithms which have IT Failure, Privacy, and Cybersecurity problems reported in the AIAAIC repository as of 07/21/2022 | 878 |
| 2 | Complement the AIAAIC sample with additional problematic algorithms found through keyword searches in Google, Factiva, EBSCOhost, and Web of Science | 232 |
| **Subtotal of problematic algorithms before applying inclusion criteria** | | **1110** |
| 3 | Drop algorithms whose user organizations are not incorporated in the US | 707 |
| 4 | Drop algorithms in the ideation phase that are not yet used with actual data and users | 666 |
| 5 | Drop algorithms failing to satisfy the definition of an intelligent algorithm | 574 |
| 6 | Drop algorithms that do not have: (i) IT failure, (ii) privacy, or (iii) cybersecurity breach | 302 |
| 7 | Drop algorithms that: (i) were developed by an individual rather than an organization; (ii) whose developer organizations were not specified | 275 |
| 8 | Drop problematic algorithms which no matching problem-free algorithms were found | 249 |
| **Subsample of problematic algorithms** | | **249** |
| 9 | For each problematic algorithm, go to the year of problem emergence and find a problem-free algorithm that satisfies criteria listed in method | 249 |
| **Subsample of problem-free algorithms** | | **249** |
| **Final sample:** Pairs of problematic (n1=249) and problem-free algorithms (n2=249) | | **498** |

**Appendix Table 2.** Control Variables and Measurements

| Control Variable | Control Variable Description and Definition | Measurements |
|---|---|---|
| (i) Ground Truth Status | Ground truth is the sum of all the data collected, checked, and labeled in the context of a specific decision task (C3.ai 2022; Muller et al. 2021). Coders evaluated ground truth status to train and calibrate algorithm decision-making rules. | [0] No well-established ground truth (GT); [1] Well-established GT; [2] Multiple conflicting GTs |
| (ii) Learning Method | The methods used to train an algorithm that learns from data on the world (Rudin 2019). It can learn to map features X to a label Y, such that Y is a measure of the object of interest. Or, an algorithm learns latent concepts found within data. | [0] Unsupervised Learning; [1] Strict Supervised Learning; [2] Hybrid Learning |
| (iii) Fairness Goal | The organization aims to develop algorithms to avoid prejudice toward a group based on their inherent characteristics. (Mehrabi et al. 2021) | [0] Algorithm Fairness, not a stated goal<br>[1] Algorithm Fairness is a stated goal |
| (iv) Anthropomorphism | Refers to any non-human entity, such as an algorithm, with humanized characteristics (Blut et al. 2021). Frequently, algorithms have been designed with humanized features to encourage users to perceive algorithmic messages delivered by a human. | [0] No Humanized features; [1] Moderate degree of Humanized; [2] High degree of Humanized features |
| (v) Algorithm Interaction Capabilities | *Vision, speech, emotion, cognitive, and sensory algorithmic capabilities:* the process algorithms inspect and analyze images, analyze human language, recognize emotions in human text, sense surroundings, or understand the meaning of sensory inputs, explaining what they are doing, intend, or have done (Lake et al. 2017). | [1] vision (or not [0]); [1] speech (or not [0]); [1] emotion (or not [0]); [1] cognition (or not [0]); senses (or not [0]) |
| (vi) Stakeholder Utility Optimization | User organizations use algorithms to support complex decisions that impact multiple stakeholders. Different user organizations can choose to prioritize the utilities of different stakeholders or only their organization's priorities (Lee and Baykal 2017). | [0] No Utility Optimization; [1] Unilateral Utility Optimization; [2] Multilateral Utility Optimization |
| (vii) Acquisition Mode | From the governance modes literature, a user organization can acquire a new algorithm in three ways (Zuo et al. 2021). The user organization's choice of algorithm acquisition mode can affect how much it can influence the algorithm's design choices and mitigations. | [0] User org purchased the algorithm off-the-shelf; [1] User org collaborated with the developer org; [2] User org developed this algorithm on its own |
| (viii) Number of Stakeholders | Any group or individual that has an influence over the algorithm or is influenced by its objectives in the form of power, legitimacy, or urgency relationship with the algorithm's developers (Mitchell et al. 1997). | *Count of algorithm's stakeholders identified in algorithm source documents.* |
| (ix) Stakeholder Management quality | A multi-item construct with a Cronbach's Alpha of 0.754, based on four measurements related to a user organization algorithm's stakeholder: (i) relations, (ii) communication, (iii) learning with and from, and (iv) integrative engagement (Freeman et al. 2017). The construct looks at user organizations' actions taken to understand better how value is created with and for algorithm stakeholders, interacting with them to understand the role of social and political surroundings to answer concerns, engaging with them to learn about complex activities, and focusing on the power of stakeholder relationships | [1] created value with and for at least 50% of stakeholders (or not [0]); [1] sought to understand social surroundings (or not [0]); [1] learned with and from stakeholders to create value (or not [0]); [1] integrative view on interconnections of stakeholders (or not [0]) |
| (x) Target Audience Quantity | The number of people whose lives, work, decisions, and opportunities are directly affected by the decision outputs of the algorithm (Langer and Landers 2021). | [0] Few people; [1] Hundreds; [2] Thousands; [3] Millions; [4] Billions |
| (xi) Platform status | The algorithm runs on a multi-sided digital platform that has (a) two or more user groups; (b) those who need each other; (c) but who cannot capture value by themselves (Evans et al. 2008). | [0] Not on a multi-sided platform<br>[1] Runs on a multi-sided platform |
| (xii) For-Profit Status | The organization distributes profits to owners, as opposed to not seeking to produce profits. | [0] Not-for-profit; [1] For-profit |
| (xiii) Algorithm Decision-Making Support | An algorithm either fully automates a task, augments with either 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<br>[2] Augmentation – Human Final DM; [3] Hybrid |
| (xiv) Algorithm Repurposed | A repurposed algorithm was developed for another purpose in another context and is being used for a new purpose (Eitel-Porter 2020) | [0] Not Repurposed - Original Context<br>[1] Repurposed: Used for New Purpose |
| (xv) Industry Similarity | Whether based on unique 2-digit sector code (SIC) of the (NAICS), user and developer organizations are in the same industry. | [0] Same NAICS; [1] Different NAICS |

**Appendix Table 3.** Illustrative Coding of Some Study Variables

| Dev or User Org, Algorithm, Year | IV Coding | Problem Coding and Evidence URL | Problem Description | Damages by Problem Coding | Evidence for IV Coding |
|---|---|---|---|---|---|
| ***1.1 Developer Org Cyber Risk Disclosure:*** One year before the problem emergence, was there any evidence that the Developer Organization **disclosed** any cybersecurity risks of its algorithms? | | | | | |
| **Boston Dynamics, Autonomous Machine Algorithm, 2021** | [0]: User org had no evidence of disclosing cybersecurity risks of this algorithm in the year before the problem emergence | [1] Cybersecurity Breach; [0] No Privacy Breach; [0] No IT Failure https://tinyurl.com/38xk6ce8 | Algorithm's integrity violated due to misuse | [0] Harm caused to Stakeholders; [0] No Financial Loss; [1] Reputational Damage; [0] No Lawsuit | Boston Dynamics had no official information about cybersecurity management policies in 2020. No unofficial documents found in 2020 of the company disclosing algorithm cybersecurity risks. Keywords: "Boston Dynamics cybersecurity", "Boston Dynamics disclosure." |
| **Proofpoint, Email Protection Algorithm, 2019** | [1]: User org symbolically disclosed cybersecurity risks of algorithm with generic, boilerplate language in year before the problem emergence | [1] Cybersecurity Breach; [0] No Privacy Breach; [0] No IT Failure https://tinyurl.com/bdz8vmt4 | Proofpoint algorithm's IP stolen by ML researchers | [0] No harm caused to Stakeholders; [0] No Financial Loss; [1] Reputational Damage; [0] No Lawsuit | Proofpoint lists results that could occur due to different cybersecurity threats or other vulnerabilities in their AI system. https://tinyurl.com/muaaubtt |
| **GM, SuperCruise Self-Driving Algorithm, 2019** | [2]: User org substantively disclosed cybersecurity risks of this algorithm using organization and algorithm specific language prior to emergence | [0] No Cybersecurity Breach; [0] No Privacy Breach; [0] No IT Failure | No problem related to dependent variables | [0] No harm caused to Stakeholders; [0] No Financial Loss; [0] No Reputational Damage; [0] No Lawsuit | The Risk Factors section of GM's 10-K filing of 2019 provides a detailed discussion of the cybersecurity risks of autonomous vehicle technologies. Noting that these technologies are subject to various "cybersecurity and data privacy risks," https://tinyurl.com/6pb5a9hp |
| 1.2 ***Developer Org Cyber Risk Board Oversight:*** One year before the problem emergence, was there any evidence that the Developer Organization had **board-level oversight** of cybersecurity risks of its algorithms? | | | | | |
| **Olive AI, Admin Task Automation Algorithm, 2020** | [0]: Developer org had no evidence of board-level oversight of cybersecurity risks of algorithms in the year before problem emergence | [0] No Cybersecurity Breach; [0] No Privacy Breach; [0] No IT Failure | No problem related to dependent variables | [0] No harm caused to Stakeholders; [0] No Financial Loss; [0] No Reputational Damage; [0] No Lawsuit | Searches: 10K, and DEF14A, were unavailable. Google, Company's website, privacy policy (2019), Wayback machine. Keywords: "Olive AI Inc Board of Directors," "Olive AI Official Filing," "Olive AI Report," cybersecurity", "risks", "information," "security." |
| **OpenAI, GPT-3 Offensive Speech Filter Algorithm, 2021** | [1]: Developer org had symbolic evidence of board-level oversight of cybersecurity risks of algorithms in year before the problem emergence | [1] Cybersecurity Breach; [1] Privacy Breach; [0] No IT Failure https://tinyurl.com/yckz4m6w | Children's data use without consent and integrity break | [1] Harm caused to Stakeholders; [0] No Financial Loss; [1] Reputational Damage; [0] No Lawsuit | No explicit source for board-level oversight of cybersecurity risk was found. But OpenAIs "Coordinated Vulnerable Disclosure" hints at risks and guidelines for good hackers for their efforts to detect it. https://tinyurl.com/ydrpx6fz |
| **Twitter, Bot Detection Algorithm, 2021** | [2]: Developer org had substantive evidence of board-level oversight of cybersecurity risks of its algorithms in year before problem emergence | [1] Cybersecurity Breach; [0] No Privacy Breach; [0] No IT Failure https://tinyurl.com/38bppwnh | Data poisoning leading to lack of algorithm data integrity | [0] No harm caused to Stakeholders; [1] Financial Loss; [1] Reputational Damage; [0] No Lawsuit | 2021 DEF14A mentions cybersecurity knowledge a skill searched for in board nominees. Additionally, "cybersecurity is a critical part of risk management at Twitter". Cybersecurity mentioned numerous in risk oversight for board and audit committee. https://tinyurl.com/4de6uah5 |
| ***1.3 Developer Org SOC Reporting:*** One year before the problem, was there evidence that Dev Org had **Service Organization Control (SOC) reports that evaluated** cybersecurity controls of its algorithms? | | | | | |
| **Poshmark, Password Hashing Algorithm, 2019** | [0]: Developer org had no SOC report evaluating cybersecurity controls of its algorithms in the year before the problem emergence | [1] Cybersecurity Breach; [0] No Privacy Breach; [0] No IT Failure https://tinyurl.com/yc6s4wcf | Confidentiality breach due to bypassing of hash algorithm | [1] Harm caused to Stakeholders; [1] Financial Loss; [1] Reputational Damage; [0] No Lawsuit | Keywords for wayback machine "poshmark.com", website from 2018 but no mention of SOC report. Keywords for FACTIVA: "Poshmark", "Poshmark SOC" yielded 44 results from 01/01/2018 to 01/01/2019. None mentioned SOC. Also checked the AICPA website. |
| **HikVision, Body Thermal Scanner Algorithm, 2020** | [1]: Developer org had SOC 1® reports evaluating cybersecurity controls of its algorithms in the year before the problem emergence | [0] Cybersecurity Breach; [0] No Privacy Breach; [1] No IT Failure https://tinyurl.com/ekac22ud | Sensor data incorrectly detect and fed temperature | [1] Harm caused to Stakeholders; [1] Financial Loss; [1] Reputational Damage; [0] No Lawsuit | Obtained the information system security-level protection registration certificate (Class 3), as a technical requirement in line with Cyber Security Law; Through ISO / IEC29151: 2017 certification, the standard covers the requirements. https://tinyurl.com/bddjt7fz |
| **Gaggle, Behavioural Monitoring Algorithm, 2021** | [2]: Developer org had SOC 2® or SOC 3® reports evaluating cybersecurity controls of algorithms in year before problem emergence | [0] No Cybersecurity Breach; [1] Privacy Breach; [0] No IT Failure https://tinyurl.com/bdf2auf5 | Consent for data taken by algorithm of students lacked | [1] Harm caused to Stakeholders; [0] No Financial Loss; [1] Reputational Damage; [0] No Lawsuit | "The main purpose of the SOC 2 Type 1 report is to show our customers that an independent third party has evaluated our systems and controls and our adherence to those systems and controls." https://tinyurl.com/yantfkay |
| ***2.3 User Org Cyber Risk Training:*** One year before the problem emergence, was there any evidence the User Org **trained the algorithm's stakeholders** on cybersecurity risks and protections of the algorithm? | | | | | |
| **Uber, Tracking Algorithm, 2014** | [0]: User org had no evidence of training stakeholders about cybersecurity risks and protections of this algorithm prior to problem | [1] Cybersecurity Breach; [1] No Privacy Breach; [0] No IT Failure https://tinyurl.com/yzxe899y | PII overreach and, customer confidentiality data breach | [1] Harm caused to Stakeholders; [1] Financial Loss; [1] Reputational Damage; [1] No Lawsuit | "Until at least September 2014, failed to implement reasonable security training and guidance". The keywords for search: "training uber 2013", "cyber training by uber", and "stakeholder training by uber". https://tinyurl.com/4272tc3k |
| **GoodrX, Price Comparison Algorithm, 2020** | [1]: User org symbolically recognized training stakeholders on cybersecurity risks algorithm, but no substantive training prior to problem | [0] No Cybersecurity Breach; [1] Privacy Breach; [0] No IT Failure https://tinyurl.com/22te5c4r | Medical data sharing to third parties without consent | [1] Harm caused to Stakeholders; [0] No Financial Loss; [1] Reputational Damage; [0] No Lawsuit | As stated in 2020 Annual report: "We have prepared a remediation plan for each of the material weaknesses and begun training process owners, developing new controls, and monitoring results." https://tinyurl.com/dffunhp6 |
| **Tinder, Facial Recognition Algorithm, 2020** | [2]: User org had substantive evidence training the stakeholders on cybersecurity risks and protections of algorithms before the problem | [1] Cybersecurity Breach; [0] No Privacy Breach; [0] No IT Failure https://tinyurl.com/4ttk4fbv | Failure to retain the confidentiality of users | [1] Harm caused to Stakeholders; [1] Financial Loss; [1] Reputational Damage; [0] No Lawsuit | "At Tinder, security awareness begins on day one and it is a continuous process thereafter. All employees undergo security and privacy training the moment they start as annually." https://tinyurl.com/5bhkk24r |