How Does AI Fail Us? A Typological Theorization of AI Failures

Xinhui Zhan
University of Oklahoma, xzhan@ou.edu

Heshan Sun
University of Oklahoma, sunh@ou.edu

Shaila M. Miranda
University of Oklahoma, smiranda@walton.uark.edu

How Does AI Fail Us? A Typological Theorization of AI Failures

Completed Research Paper

Xinhui Zhan
University of Oklahoma
307 W. Brooks, Norman, OK
xzhan@ou.edu

Heshan Sun
University of Oklahoma
307 W. Brooks, Norman, OK
sunh@ou.edu

Shaila Miranda
University of Arkansas
220 N. Mcllroy Avenue, Fayetteville, AR
smiranda@walton.uark.edu

Abstract

AI incidents, often resulting from the complex interplay of algorithms, human agents, and situations, violate norms and can cause minor or catastrophic errors. This study systematically examines these incidents by developing a typology of AI failure and linking these modes to AI task types. Using a computationally intensive grounded theory approach, we analyzed 466 unique reported real-world AI incidents from 2013 to 2023. Our findings reveal an AI failure typology with six modes, including artifact malfunction, artifact misuse, algorithmic bias, agency oversight, situational unresponsiveness, and value misalignment. Furthermore, we explore the relationship between these failure modes and the tasks performed by AI, uncovering four propositions that provide a framework for future research. Our study contributes to the literature by offering a more holistic perspective on the challenges faced by AI-powered systems, beyond the critical challenges of fairness, transparency, and responsibility noted by the literature.

Keywords: AI failure modes, artifact malfunction, artifact misuse, algorithmic bias, agency oversight, situational unresponsiveness, value misalignment

Introduction

In late March 2023, an open letter endorsed by nearly 30,000 individuals, including AI experts and industry leaders, called for a temporary halt to the development of powerful AI systems. The letter underscored the potential risks that advancements in AI could pose to society and humanity, emphasizing the need for robust regulations to mitigate unintended consequences of AI-powered systems. Numerous instances of AI failures have been reported, ranging from the amusing, such as an AI-powered camera mistaking a bald head for a football during a live stream[1], to the more concerning, such as biased decision-making in AI-powered hiring tools (Kordzadeh and Ghasemaghaei 2022), to tragic, such as fatal accidents caused by autonomous vehicles (Grigorescu et al. 2020). These incidents result from the complex interplay of an algorithm, a user, and a situation, rather than being solely attributable to the AI artifact itself.

Indeed, AI is not infallible and the public often perceives these incidents as AI failures[2] (Yampolskiy & Spellchecker, 2016), resulting in widespread concern about the uncertainty and risks tied to AI. Specifically,

---

[1] https://thenextweb.com/news/ai-mistakes-referees-bald-head-for-football-hilarity-ensued
[2] Please note that in this paper, "AI incident" refers to AI-related negative events that has been perceived by public as an "AI failure".

AI incidents significantly impact the public's impression of AI and may become an inhibitor to AI diffusion. For individuals, a single unpleasant AI incident can undermine the reputation of AI systems as a whole, deterring their adoption (Goot et al. 2020; Janssen et al. 2021). For organizations, AI incidents may increase the uncertainty of deploying AI and thus impact organizational strategic use of AI (Li et al. 2021). In light of these concerns, it is crucial to gain a comprehensive understanding of AI failure modes, which refer to *the manner in which a breakdown is perceived to occur at the intersection of the technical components, human agents, and situations surrounding an AI artifact.*

The current information system (IS) literature scrutinizes AI from various perspectives, including fairness (Akter et al. 2021; Jussupow et al. 2021a; Kordzadeh and Ghasemaghaei 2022; Teodorescu et al. 2021), transparency (Arrieta et al. 2020; Fernández-Loría et al. 2022; Lebovitz et al. 2022; Lu et al. 2019), and responsibility (Mikalef et al. 2022; Rai et al. 2019). However, those metrics are still fragmented and at a granular level. We have only a limited understanding of the real-world AI failures, phenomena at the intersections of AI algorithms, human stakeholders, and the situation.

Although past literature on IS failure (Rezazade Mehrizi et al. 2022; Salo et al. 2020; Tan et al. 2016) offers valuable insights, those insights do not translate directly to the study of AI failures. First, AI represents a fundamentally different type of technology system with higher complexity (Russell and Norvig 2002), making it more difficult to diagnose when failures occur. Unlike traditional code-based systems designed to automate specific tasks, AI takes on a variety of tasks that often have unclear specifications and that operate in uncertain situations (Baird and Maruping 2021; Kane et al. 2021; Mikalef et al. 2022). For example, a traditional IS in e-commerce would handle straightforward tasks like inventory management or payment processing, confined to clear rules and procedures. Conversely, an AI system might dynamically adjust pricing based on real-time supply and demand data, all of which operate under far less well-defined conditions. Moreover, the human-technology interaction is evolving. Scholars suggest that the term "use," traditionally employed in IS research, is increasingly being replaced by the concept of "delegation" to better describe how people interact with AI systems (Baird and Maruping 2021). This evolving relationship becomes especially complex as AI is becoming an integral component in diverse domains, such as medical diagnosis (Jussupow et al. 2021b; Lebovitz et al. 2021; Lebovitz et al. 2022), human resources knowledge generation (van den Broek et al. 2021), online labor market management (Möhlmann et al. 2021), image classification decision-making (Fügener et al. 2022), as well as financial investment advising (Ge et al. 2021).

To bridge the aforementioned knowledge gaps, our objective in conducting this research is to develop a typological theorization of AI failures that account for not only the technical components, but also human agents, and situations in relation to different types of tasks. Specifically, we focus on the following research questions:

1. *How is AI perceived to fail us and what are the failure modes?*
2. *How do the failure modes of AI incidents relate to the tasks performed by AI?*

Given AI failures are new phenomena that are yet to be theorized, we adopted the computationally intensive grounded theory approach (Berente et al. 2019; Miranda et al. 2022a). We started by compiling a comprehensive list of AI incidents from two open data resources: AI Incident Database (AIID) and AI, Algorithmic and Automation Incidents and Controversies (AIAAIC)[3]. We then eliminated duplicate entries using text similarity spatial analysis, resulting in a dataset of 466 unique AI incidents from 2013 to 2023. Then, we employed a combination of machine learning algorithms and manual coding to analyze the text (i.e., incident description) of our incident dataset. Through this analysis, we identified six failure modes: *artifact malfunction, artifact misuse, algorithmic bias, agency oversight, situational unresponsiveness,* and *value misalignment.* To investigate the relationship between these failure modes and tasks performed by AI, we tagged each incident with the AI system's dominant task type and performed contingency analysis to understand the association between failure modes and task types (i.e., generative, intellective, judgmental, psychomotor, and emotional labor tasks).

---

[3] AI Incident Database https://incidentdatabase.ai/ and AIAAIC Repository https://www.aiaaic.org/aiaaic-repository

Our study makes two theoretical contributions to the literature. First, this research contributes a typology of six AI failure modes. By incorporating the human-AI relationship and AI-context relationship into this consideration, we enrich the existing AI literature. Second, we tackled the complexity of AI failures by examining the failure modes in relation to AI tasks. We formulated four propositions, providing a framework for future research to consider AI task types when evaluating AI.

Literature Review

IS Failure and Characteristics of AI

Prior research has studied failures in traditional IS contexts (Rezazade Mehrizi et al. 2022; Salo et al. 2020; Tan et al. 2016). For example, Tan et al. (2016) classified e-commerce service failures into three categories: information, functional, and system, each with its own set of characteristics. IS failures are also characterized by their size, complexity (Loch et al. 1992; Lyytinen and Robey 1999), intentionality (Ekelhart et al. 2015), and urgency (Spears et al. 2013). Moreover, IS failures can harm developers and deployers by inducing dissatisfaction, discontinued use, switching, and negative word-of-mouth (Salo et al. 2020).

Transitioning from this traditional understanding of IS failures, it's important to consider that AI failures present their own unique challenges. AI is often described as a machine's capability to carry out tasks typically associated with human cognition, including "perception, reasoning, learning, interacting with the environment, problem-solving, decision-making, and even demonstrating creativity" (Rai et al. 2019, p. iii). Indeed, there is wide agreement that AI is a novel, broad, and multi-dimensional concept (Brachman 2006; Russell and Norvig 2002). Wirth (2018) classified AI based on its capabilities as narrow, strong, or hybrid AI. Building on these insights, we identify two distinct aspects of AI that differentiate AI from traditional IS. First, unlike traditional technology programmed to perform a task, AI is programmed to learn to complete a task and function on its own (Nilsson 2009). Second, AI systems often rely on real-time information that is gathered from the environment to perform tasks (Hamm and Klesel 2021) and make judgments (Crowston and Bolici 2019). Third, Lee et al. (2015) defined an 'IS artifact' as a system composed of three integral subsystems: a technology artifact, an information artifact, and a social artifact. We propose that an AI system is composed of three subsystems: an algorithmic artifact, a data artifact, and a social artifact. Specifically, the algorithmic artifact refers to the machine learning or computational algorithms that drive the AI's decision-making and action. The data artifact is the data set on which the AI is trained and which it uses for ongoing learning. The social artifact refers to the human interactions and societal contexts that shape how AI is deployed and used. These three components together make AI systems far more dynamic compared to traditional IS.

Due to these distinctions, AI failures are not merely the result of straightforward bugs or user errors. Instead, they often emerge from interactions among algorithms, humans, and the specific situations in which they are deployed. This multifaceted nature of AI failure demands a comprehensive approach that considers both the technical and situational complexities involved.

Task Literature and AI Task

AI has the capability to perform a wide range of tasks, ranging from intellective medical diagnoses (Jussupow et al. 2021b; Lebovitz et al. 2021; Lebovitz et al. 2022), making judgments (Fernández-Loría et al. 2022; Fügener et al. 2022), and automated robotic motions (Esterwood and Robert Jr 2023). From this agentic point of view, AI is designed to exhibit agency or autonomy and is increasingly proficient in handling more challenging tasks (Baird and Maruping 2021). For example, AI-powered language models, such as GPT-4, can generate human-like text based on a given prompt; companion chatbots, such as Replika, use natural language processing and machine learning to have personalized conversations and improve users' mental well-being.

In order to identify and abstract the tasks performed by AI, we adapt prior literature on human group tasks. McGrath (1984) introduced a task circumplex model, which offers a systematic framework to classify a wide range of tasks performed by human groups and thus can be adapted to classify the wide range of AI tasks. This model organizes tasks into eight categories based on two dimensions: cooperation-competition (i.e.,

the degree to which people work together versus work against each other) and cognitive-behavioral requirements (i.e., whether the task requires cognitive versus behavioral efforts). The task groups consist of planning, intellective, judgment, performance, creativity, negotiation, mixed-motive, and cognitive conflict tasks. To adapt this framework to the AI context, we selected four of the eight task types from McGrath's (1984) model: creative, intellective, judgmental, and psychomotor tasks. We excluded task types focused primarily on human groups, such as planning, cognitive conflict, mixed-motive, and contest tasks as these were currently absent from our incidents dataset. However, as AI has been used for emotional support (Meng and Dai 2021), we expanded the task domains by including emotional labor as an additional type (Wharton 2009). Table 1 summarizes these task types, with definitions and examples of AI tasks.

Task Type	Definition	Example of AI Task
Generative	generate novel ideas, solutions, or products	OpenAI's GPT models generate human-like text.
Intellective	cognitive processing of information, logical reasoning, and analysis for problem-solving	Google Translate uses AI to provide translations between various languages, often with impressive accuracy.
Judgmental	evaluating options, weighing pros and cons, and selecting the most appropriate decision.	IBM's Watson is being used to support doctors in diagnosing diseases, analyzing medical images, and predicting patient outcomes.
Psychomotor	perform specific actions or tasks to achieve a goal	Autopilot vehicles perform both cognitive (sensing, processing, decision-making) and physical (steering, accelerating, braking).
Emotional labor	display appropriate emotion through interaction with human	Companion chatbots, such as Replika, use natural language processing and machine learning to have personalized conversations.

Table 1. Task Typology Used in Our Study (adapted from McGrath (1984) and Wharton (2009))

Diverse Criteria for Evaluating AI

One issue that has received significant attention in the literature is AI fairness. Researchers and activists have raised concerns about the potential biases embedded in AI algorithms, which can perpetuate and even amplify existing societal biases and discrimination (Akter et al. 2021; Cavazos et al. 2020). Unfairness in AI can stem from various sources and impact different stages of the AI pipeline, such as data collection, preprocessing, modeling, testing, and deployment (Barocas et al. 2019). First, the data used to train an AI system may be incomplete, unrepresentative, or contain historical biases that may propagate or amplify unfairness in the system's outputs (Kane et al. 2021). For example, if an AI system is trained to analyze job applicants' resumes and decide which ones to consider for interviews, it may inadvertently discriminate against certain groups of people, such as women or people of color, if the training data is biased toward those groups. Second, the algorithms used to process the data or make decisions may introduce or exacerbate unfairness due to choices made during their design or their optimization objectives (Bolukbasi et al. 2016). In fact, different stakeholders may have different preferences or expectations about what constitutes fair outcomes or processes (Binns et al. 2018). Furthermore, the context in which an AI system is deployed may influence its performance and impact different groups or individuals (Eubanks 2018). In order to mitigate biases, researchers attempted to address fairness in AI systems and develop methods to detect biases (Cavazos et al. 2020; Kordzadeh and Ghasemaghaei 2022; Parikh et al. 2019) and demonstrate the relationship between unjust outcomes and biased training data. However, despite progress in using advanced computational methods to improve fairness, AI tools are still incapable of automating fairness and require human argumentation (Teodorescu et al. 2021).

A second issue is transparency, also referred to as explainability, which is a critical aspect of evaluating AI systems. Transparency refers to the ability of AI systems to explain their decisions, actions, and behaviors to human users, stakeholders, and regulators. They are particularly important in fields such as healthcare, where the consequences of AI decisions can be significant (Fernández-Loría et al. 2022). Jussupow et al. (2021a) also discussed the importance of explainability in machine learning models, arguing that it is essential for users to understand the rationale behind AI decisions. Arrieta et al. (2020) argued that
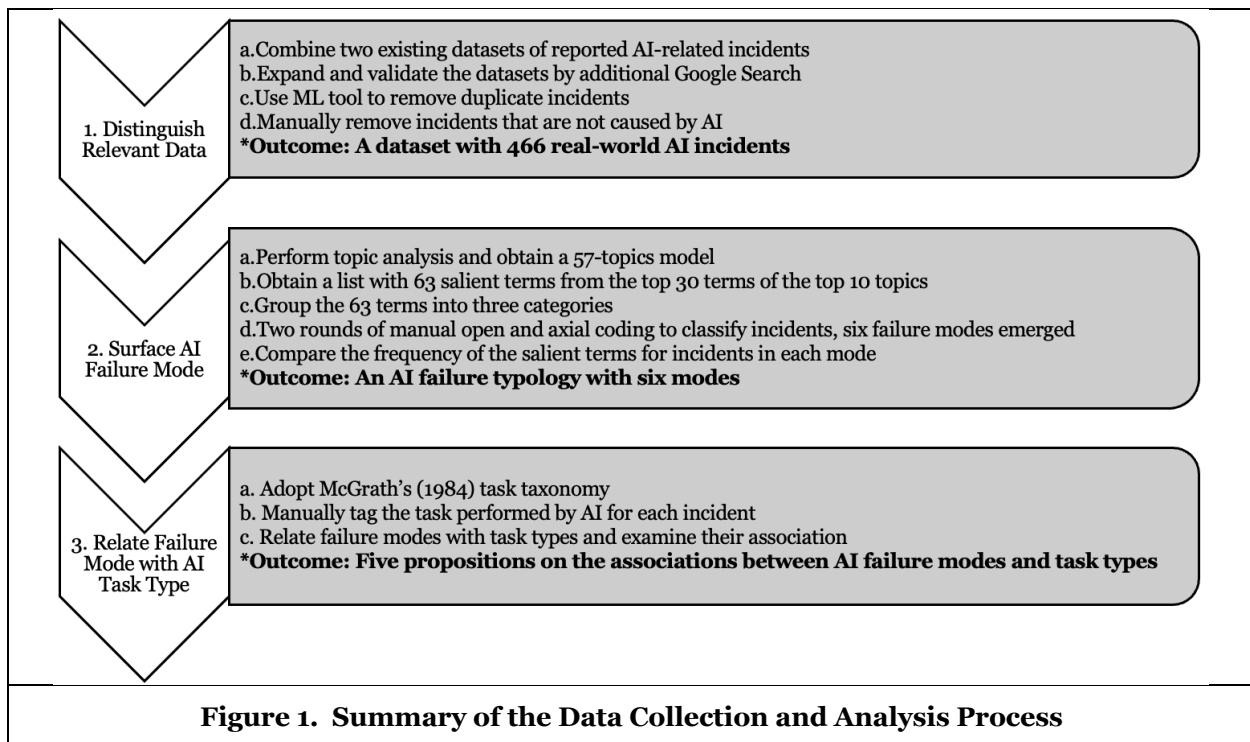
understanding how AI systems make decisions is essential for users to trust and effectively collaborate with them. Lu et al. (2019) advocated that good explanations should enable users to advance their goals or objectives. However, achieving AI transparency is not a straightforward task, as it involves technical, ethical, and social challenges. For example, some AI algorithms, such as neural networks, may be too complex or dynamic to be fully explained (Lu et al. 2019). Moreover, different users may have different expectations and preferences for how AI systems should explain themselves, depending on their background, context, and objectives (Arya et al. 2019). Therefore, AI transparency requires a careful balance between providing sufficient information and respecting various constraints and trade-offs.

A third issue is accountability. Busuioc (2021) argued that AI system developers and deployers have the responsibility to ensure that AI systems are designed and deployed in a manner that upholds the standards of fairness and integrity. Nushi et al. (2018) examined the role of developers in assuring accountability in AI systems, arguing that accountability should be built into the design of AI systems from the outset. Similarly, Rai et al. (2019) discussed the need for accountability in AI, arguing that it is essential to ensure that AI systems are designed and deployed in an ethical and responsible manner.

In sum, the literature on AI with the fairness, transparency, and accountability perspectives has grown substantially in recent years. It highlights the potential problems associated with AI and, at the same time, shows the complexity of understanding AI-related phenomena. Nevertheless, there is a scarcity of holistic and systematic investigations into AI failures that focus on the multifaceted interplay among algorithm design, human stakeholders, contextual factors, as well as social norms and values.

Method

Since existing theories are insufficient to address our research questions, we adopted a grounded theory approach that employs computational tools to process and analyze data (Lindberg et al. 2016; Miranda et al. 2022b; Miranda et al. 2022c). This approach is suitable for studying AI incidents for two reasons. First, the vast amount of digital traces left by these phenomena requires the researcher to be computationally assisted in order to make sense of it in its entirety (Lazer et al. 2009). Second, a grounded theory approach is well suited to the investigation of novel phenomena for which existing theoretical lenses may be ill-matched (Edmondson and McManus 2007). Figure 1 summarizes the data collection and analysis process.

1. Distinguish Relevant Data
a. Combine two existing datasets of reported AI-related incidents
b. Expand and validate the datasets by additional Google Search
c. Use ML tool to remove duplicate incidents
d. Manually remove incidents that are not caused by AI
*Outcome: A dataset with 466 real-world AI incidents

2. Surface AI Failure Mode
a. Perform topic analysis and obtain a 57-topics model
b. Obtain a list with 63 salient terms from the top 30 terms of the top 10 topics
c. Group the 63 terms into three categories
d. Two rounds of manual open and axial coding to classify incidents, six failure modes emerged
e. Compare the frequency of the salient terms for incidents in each mode
*Outcome: An AI failure typology with six modes

3. Relate Failure Mode with AI Task Type
a. Adopt McGrath's (1984) task taxonomy
b. Manually tag the task performed by AI for each incident
c. Relate failure modes with task types and examine their association
*Outcome: Five propositions on the associations between AI failure modes and task types

Figure 1. Summary of the Data Collection and Analysis Process

Sampling and Data Collection

We sampled reported AI incidents in three stages. First, we obtained the AI Incident Database (AIID), a database dedicated to indexing the harms or near harms realized in the real world by the deployment of artificial intelligence systems. AIID has 2,523 incident reports (as of March 29, 2023). Second, we obtained the AI, Algorithmic and Automation Incidents and Controversies (AIAAIC) database, which includes 972 AI incident reports (as of April 20, 2023). Combining the two datasets resulted in a dataset with 3,495 reports. Each report includes the year, name of the AI system, developer of AI, headline, as well as a short description – 15 to 100 words – of the incident. Third, we performed a Google search using search terms such as "AI accident", "AI failure", "AI error", and "AI glitch." We also refined the search by adding specific industries or technologies, such as "self-driving car accidents," "AI healthcare errors," and "robotics malfunctions." These searches yielded no new cases, convincing us of the comprehensiveness of the compiled AIID and AIAAIC databases.

Data Processing

In the dataset obtained from the previous step, we noticed numerous duplicates (i.e., multiple reports discussing the same incident). To address this issue, we cleaned the data and identified unique AI incidents. Following the methodology outlined in earlier studies (Bail 2012), we conducted a text similarity spatial analysis to detect exact and near matches between pairs of reports. We employed *SentenceTransformer*, a Python package that facilitates text embeddings. Then, these embeddings were compared using cosine-similarity to identify sentences with semantically similar meanings (Reimers and Gurevych 2019). For instance, the description of a report "*Local firefighters were only able to stop a Cruise AV from driving over fire hoses that were in use in an active fire scene when they shattered its front window*" was determined to be similar to another report "*A Cruise AV ran over a fire hose that was being used in an active firefighting area*." Next, we also limited the scope of the included reports to those occurring between 2013 and 2023. Through these steps, we compiled a dataset of 466 unique AI incidents. We proceeded to tag and address the missing values for various features of the identified incidents, including the year of occurrence, country, developer, technology, number of reports, description, and headline.

To construct an inductive typology of AI failures, we utilized content analysis techniques in accordance with grounded theory principles (Croidieu and Kim 2018; Gioia et al. 2013). Initially, one of the authors engaged in open and axial coding, examining the text of each AI incident in the dataset on a sentence-by-sentence basis to identify basic elements. For instance, an incident involving ChatGPT in grant applications was coded with labels such as "User" and "Misuse." Similarly, an incident where an autopilot car crashed into a stopped truck received codes like "Malfunction" and "Careless Driving." During the first round of open coding, we relied on the 'name of the AI system' column in the dataset and tagged each incident with the AI's dominant task type. For example, the dominant task type for generative AI, such as ChatGPT and Midjourney, is a generative task, while for AI drones and robotic arms, it is a psychomotor task. After the open coding phase, we moved on to axial coding, where we identified relationships among these various codes, grouping them as multiple incidents exhibited shared characteristics. After two iterations, we surfaced six failure modes, including *artifact malfunction, artifact misuse, algorithmic bias, agency oversight, situational unresponsiveness,* and *value misalignment.*

It's important to note that some AI systems may encompass multiple task types. For instance, an AI-enabled translation service may primarily perform an intellective task (e.g., translating language) but is also required to be sensitive to the socio-cultural context of its outputs. In this sense, it performs an emotional labor task as well. To ensure the reliability of our coding, we employed investigator triangulation (Denzin 2017). All three authors independently coded a sample of the data, and then discussed and reconciled the differences. We chose to focus on the most dominant task type for the AI in each incident to maintain clarity and simplicity in our analysis.

Next, we employed machine learning algorithms to analyze the data and validate our initial round of manual coding. We began by applying Latent Dirichlet Allocation (LDA), a powerful algorithm for discovering underlying topics within large text corpora (Blei et al. 2003), to analyze the content of the "description" of the incidents. Utilizing the *pyLDAvis* Python package, we assessed the relative coherence of alternative

models by examining the frequency of co-occurrence among the most probable terms within a topic (Miranda et al. 2022c). This allowed us to determine an optimal topic number and derive a 57-topic model. Then, we extracted the top 30 prominent terms from the top 10 prevalence topic, since the density value dropped considerably since the 11th topic. We further examined the 300 terms and obtained a list with 63 words after removing duplicate and meaningless terms. The 63 terms provided valuable insights into the subject matter and helped us establish coherent labels within the data. Thus, we organized and classified the 63 terms into three broad categories, each with its own sub-categories: (1) **technological components**, including data, algorithm, and system, (2) **agents**, including users, developers/deployers, and inanimate, and (3) **actions**, including behavior and misbehavior. Table 2 summarizes the categorization.

Theme		Salient Terms
Technical Components	Data	information, content
	Algorithm	facial recognition, NLP, algorithm
	System	tool, program, vehicle, robot, autopilot, software, recommendation, echo, chatgpt, bot
Agents	User	user, driver, worker, women, men, patient, student, employee, member
	Developer/deployer	tesla, google, uber, company, facebook, amazon, twitter, tiktok
	Inanimate	highway, fire, feature, voice, privacy, face, language, image
Actions	Behavior	use, drive, automate, flag, cruise, operate, produce, alert, deliver, proctor, detect, transfer, create
	Misbehavior	kill, misidentify, offense, bias, crash, violate, injury, racist, disproportionate
Table 2. Themes and Their Salient Terms		

In order to further examine and validate the differences among the six modes, we divided the corpus into groups of texts corresponding to the six failure modes. For each group of text, we counted the number of words and obtained the aggregated frequency of each term (as detailed in Table 2). Next, we compared the frequencies across the six failure modes, as shown in Table 3. The text within each mode differs from the text across the other failure modes, and we will discuss these differences separately for each mode in the following sections.

Failure Mode	Technical Components			Agents			Actions		Word Count	Incident Count
	Data	Algorithm	Artifact	User	Developer	Inanimate	Behavior	Misbehavior	8,041	466
Artifact Malfunction									2,158	130
Artifact Misuse									1,152	70
Algorithmic Bias									1,682	95
Agency Oversight									900	46
Situational Unresponsiveness									958	60
Value Misalignment									1,192	65
Table 3. Frequency of the Salient Terms across Six Failure Modes										

Theorizing the Bases for Perceived AI Failures

As mentioned earlier, we focused our analysis on AI incidents reported in the past 10 years, which led to the identification of six failure modes. We then tagged the type of task performed by AI for each incident. In the following sections, to address our first research question regarding the various ways AI fails, we describe and differentiate the six failure modes that emerged from our dataset. To answer the second

research question, we illustrate the extent to which each failure mode is associated with the task types. Finally, we also discuss some observations obtained from the data.

Six AI Failure Modes

We define an AI failure mode as **the manner in which a breakdown is perceived to occur at the intersection of the technical components, human agents, and actions surrounding an AI artifact**.

Failure mode 1: Artifact Malfunction

In this mode of failure, the **AI system malfunctions, producing unreliable or harmful outputs**. This happens because the system was not adequately tested, verified, or validated. In our dataset, 130 incidents were classified under this category as the AI system failed to function as intended. For example, incident #22 reports that *a worker was crushed to death by a robot at a Volkswagen plant when the robot pinned him to a metal plate*. This malfunction occurred due to the robot's failure to operate as intended, resulting in tragic consequences. Moreover*, an automated maneuvering system malfunctioned due to faulty sensor data, leading to a tragic loss of life* (incident #3). This incident highlights the importance of ensuring that AI systems receive accurate and reliable input data to function properly. Overall, Table 3 shows that terms related to the artifact, such as "system", "tool" and "vehicle", and to the developer are prominent in this mode. This mode of failure emphasizes the need for comprehensive testing and evaluation before deploying such systems.

Failure Mode 2: Artifact Misuse

In this mode of failure, the AI enables **adversarial inputs or malicious interference from external agents.** This potentially results in harmful outcomes. This mode of failure represents an "unfaithful appropriation," in which the system's features are used in a manner that is not consistent with the spirit and structural design (Markus and Silver 2008).

In our dataset, 70 incidents were grouped into this category as they all involve the misuse of AI systems by external agents. These incidents include the spread of misinformation, compromising personal privacy, promoting offensive content, and weaponizing AI against individuals. For example, incident #6 reports that *Microsoft's Tay, which was deployed on Twitter in 2016, with the goal of simulating the personality of a teenage girl through interactions with Twitter users. However, some Twitter users left all sorts of hateful conversations to Tay, and unfortunately, Tay quickly learned to respond in offensive ways. In less than 24 hours, Microsoft removed Tay from Twitter*. This incident revealed the susceptibility of AI to manipulation and the importance of implementing safeguards and monitoring user interactions. Offensive content was generated as a result of the AI's vulnerability to manipulation by malicious users. Similarly, incident #180 involved *a quickly-debunked video of Ukrainian President Volodymyr Zelenskyy encouraging Ukrainians to surrender to Russian forces during the Russia-Ukraine war*. In this case, AI allowed for the creation of realistic, fake content with the potential to spread misinformation and manipulate public opinion. Moreover, AI-driven language models, such as OpenAI's ChatGPT (# 411), have also been reported to be abused by cybercriminals with minimal coding skills to develop malware, ransomware, and other malicious software. The accessibility of such technology makes it easier for unethical actors to 'wreak havoc' in the digital world. Table 3 shows that terms related to behavior are prominent in this mode, specifically related to the usage of the AI, i.e., "use", "transfer" and "create". It raises questions about the responsibility of AI developers to prevent such misuse. Promoting ethical and responsible AI development (Mikalef et al. 2022) as well as establishing clear policies and regulations regarding AI usage are crucial to alleviating this mode of failure.

Failure Mode 3: Algorithmic Bias

In this failure mode, the **AI provides recommendations or classifications that differentially disadvantage or disparage a category of people**. This is the classic problem of AI bias and unfairness

((Kordzadeh and Ghasemaghaei 2022; Teodorescu et al. 2021)) and occurs because biased or unrepresentative training data were used, causing the algorithm to learn and perpetuate existing societal prejudices.

In our dataset, 95 incidents were grouped into this category because the AI-powered system has been shown to produce biased, discriminatory, or harmful results, often reflecting societal prejudices. For example, incident #32 described that *Amazon's AI recruiting tool was found to down-rank female applicants*, perpetuating gender discrimination in the hiring process. Similarly, incidents #17 and #14 demonstrate how AI can inadvertently amplify sexist and racist biases, with *Google AdSense producing sexist and racist results and Google Photos mislabeling a black couple as "gorillas."* AI systems have also exhibited biased behaviors that could lead to dangerous consequences. For instance, incidents involving predictive policing algorithms (#45) and the Intelligence-Led Policing model (#177) demonstrate how biased data can lead to the unjust targeting of certain communities. Table 3 shows that terms related to human are prominent in this mode, with approximately 2% of the text specifically related to gender, represented by terms such as "women" or "men." Moreover, terms related to misbehavior are also noticeable, particularly those related to unfairness, such as "bias" and "racist".

In fact, an AI system often relies heavily on both training data and incoming data (Ntoutsi et al. 2020). Therefore, the issues from biased, unexpected, or abnormal historical data may be inherited in working with incoming data (Kane et al. 2021). In addition to biased data, biases can also be introduced into an AI system if the training data is not representative of the population it is meant to serve, or if the algorithms or models used in the system are not designed to mitigate bias. Consequently, it is crucial to ensure that AI development processes incorporate diverse perspectives, utilize unbiased and representative data, and employ rigorous testing and monitoring to mitigate the failures of bias.

Failure Mode 4: Agency Oversight

In this mode of failure, the AI was **unable to function correctly with available human supervision**. Additional human oversight could have mitigated or avoided its harmful consequences. In our dataset, 46 incidents were grouped into this category as these incidents highlight a lack of human oversight or intervention. One typical example would be the case of employees laid off by an AI personnel system (#30, #98). It highlights the risks of over-reliance on automation in decision-making processes. The lack of human oversight in crucial or sensitive decision-making tasks align with the Human-in-the-Loop literature, emphasizing the need for human intervention, complementary, and augmentation (Fügener et al. 2021; Teodorescu et al. 2021). Moreover, Amazon Flex's dismissal of contract delivery drivers (incident #98) based on automated performance evaluations emphasizes the risks of relying solely on AI-generated metrics for decision-making. Human intervention could have ensured that drivers had the opportunity to defend themselves and appeal decisions. Table 3 shows that terms related to data and developer are prominent in this mode, with approximately 1.2% of the text specifically mentioning "data". Moreover, terms related to developers are noticeable, specifically several companies that have been blamed for a lack of supervision and guidance in using AI.

This mode highlights the crucial role of human actors in compensating for technology errors (Jussupow et al. 2021b). Both designers and users have the important role of mitigating the failures of AI-based systems and deciding whether AI advice should be transformed into concrete action. Making the right balance between automation and human involvement is key to ensuring responsibility in AI systems.

Failure Mode 5: Situational Unresponsiveness

In this failure mode, the AI was **unable to adapt to unexpected situations or changing circumstances**. This led to accidents, misinterpretations, and other unintended consequences. In our dataset, 60 incidents were grouped into this category as they demonstrate that AI systems are still limited in their ability to adapt to real-world complexities and unexpected situations. For example, the Tesla Model S crash in 2016 (#43), Uber AV pedestrian fatality (#4), and Google autonomous test vehicle collision (#62, #8) illustrate the limitations of current autonomous driving systems when faced with unusual or unexpected events, i.e., a pedestrian or unexpected object in the environment. Similarly, the Facebook translation error (#63) and Gmail Smart Reply tool incidents (#15) show that AI systems can struggle with

context and nuance in language processing. This can lead to misunderstandings and inappropriate recommendations, as demonstrated in both cases. In fact, AI system often needs to deal with loosely defined or formatted data that comes in at high speed and is hard to process (O'Leary 2013). Thus, even if an AI system functions well in a well-defined environment, it may still cause unexpected incidents when put into real use because of environmental complexity. As shown in Table 3, the text in this mode is not distinctively associated with the terms considered. This highlights the inherent complexity of this failure mode.

Failure mode 6: Value Misalignment

In this failure mode, the **values manifested by the AI were inconsistent with those of the humans affected by its outputs**. In our dataset, incidents were grouped into this category as those AI systems were used to automate processes, but the algorithms had not been fully optimized to account for ethical considerations or were not designed to align with human values, leading to negative consequences. For example, *Kronos's scheduling algorithm and its use by Starbucks managers allegedly negatively impacted financial and scheduling stability for Starbucks employees, which disadvantaged wage workers* (#9). In this case, the AI algorithm generated employees' work schedules based on factors such as sales projections, weather, and traffic patterns. However, the algorithm allegedly did not take into account employee preferences or availability, leading to significant scheduling conflicts and financial instability for employees. On the other hand, AI algorithms can also violate user privacy, as in the case of Clearview AI (#244), which scraped images from social media sites without user consent. Similarly, the Blue Wolf surveillance program in Palestine raises ethical concerns about the use of AI for military purposes, which can lead to human rights violations (#339). This failure mode underlines the need for AI systems to be developed in a manner that prioritizes the needs and well-being of humans. As shown in Table 3, terms associated with AI algorithms are prominent in this mode, with around 1.6% of the text explicitly mentioning "algorithm." This highlights the importance of incorporating ethical considerations and ensuring alignment with human values when designing AI algorithms.

In summary, the six AI failure modes represent various ways in which AI systems can fail to meet expectations. We discuss two perspectives to examine our typology, as summarized in Table 4. First, the typology underscores the challenges and pitfalls in the design, deployment, and use of AI systems. (1) Algorithmic bias and value misalignment often originate during the design stage. Algorithmic bias results from unrepresentative or unjust training data, leading to unfair and discriminatory outcomes. Value misalignment between AI algorithms and human values emphasizes the importance of prioritizing ethical considerations and human well-being during AI development. (2) Artifact misuse, which highlights the vulnerability of these systems to adversarial inputs and malicious interference, stems from the usage stage of AI systems. (3) Agency oversight, artifact malfunction, and situational unresponsiveness may arise during either the deployment or usage stage of AI systems. Agency oversight mode emphasizes the need for human involvement and oversight in AI decision-making processes, while situational unresponsiveness mode underlines the limitations of AI systems in adapting to real-world complexities and unexpected situations. Artifact malfunction mode may result from inadequate testing, verification, or validation, leading to unreliable and potentially harmful outcomes.

Second, the typology emphasizes the complexity of understanding AI failure, as it involves not only AI itself but also the stakeholders and the broader environment. The artifact malfunction mode highlights the interplay between the AI artifact and the developer who bears the responsibility for adequately testing or validating the system. Artifact Misuse mode highlights the interplay between AI and human users. Algorithmic bias mode emphasizes the interplay between data and social justice. Agency oversight highlights the interplay between agents, which include both human stakeholders and AI. Situational unresponsiveness highlights the interplay between AI and the environment. Lastly, value misalignment emphasizes the interplay between social value systems, ethical considerations, and AI.

Failure Mode	Characteristics of the Failure	Connection with Current Literature
Artifact Malfunction	The AI system malfunctions, producing unreliable or harmful outputs.	Ties into issues of accountability and transparency. Inadequate design or deployment practices could result in malfunction.
Artifact Misuse	The AI enables adversarial inputs or malicious interference from external agents.	Ties into accountability and Human-in-the-Loop literature. Developers need to foresee and prevent possible misuses.
Algorithmic Bias	The AI provides recommendations or classifications that differentially disadvantage or disparage a category of people.	Ties into the literature on fairness.
Agency Oversight	The AI was unable to function correctly with available human supervision.	Ties into the issues of accountability and Human-in-the-Loop literature.
Situational Unresponsiveness	The AI was unable to adapt to unexpected situations or changing circumstances.	Ties into the literature on transparency.
Value Misalignment	The values manifested by the AI were inconsistent with those of the humans affected by its outputs.	Ties into literature on fairness. Indicates that AI systems must align with human values to be considered fair.
Table 4. Modes of Perceived AI Failure		

Failure Modes in Relation to AI Tasks

Our second research question addressed the relationship between failure modes and task types. Figure 2 depicts the 466 incidents in our dataset in relation to their reporting year, failure modes, and task types. Unsurprisingly, the first portion of the chart highlights the increasing frequency of reported incidents over time. The next portion depicts the flows of incidents to the six failure modes over time, with the widths of the flows indicating the proportion of incidents associated with each mode. The final portion of the chart depicts the flows from the failure modes to task types, highlighting the correspondence between failure mode and task type, with the widths of the flows indicating the proportion of each failure mode associated with each task type.

Figure 2. AI incidents from 2013 to 2023 across Failure Modes and Task Types
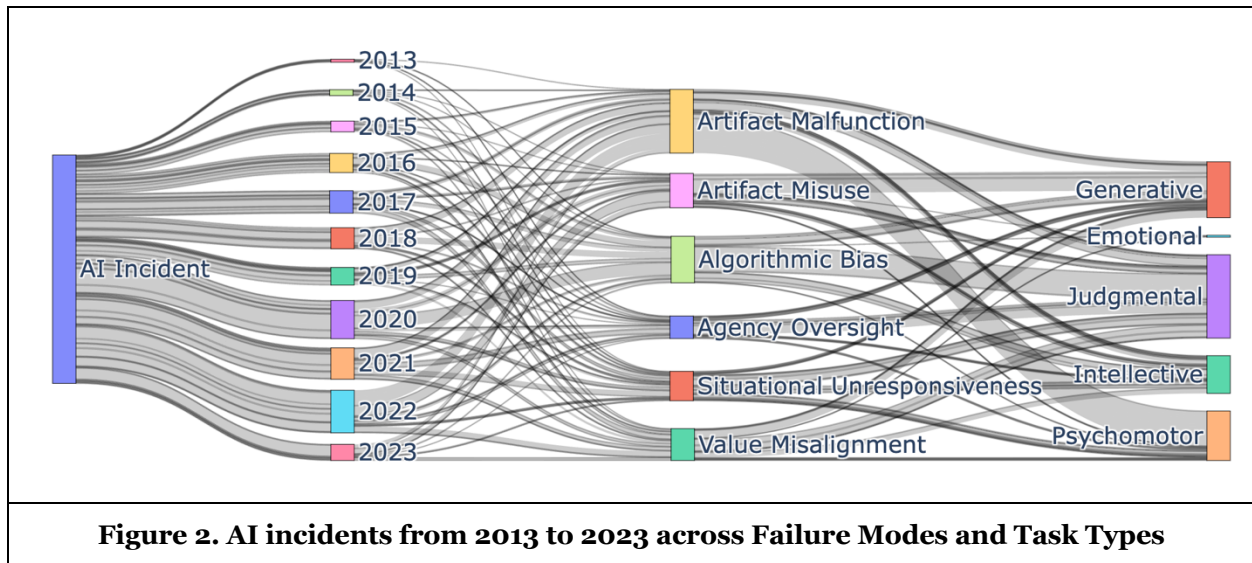
To more systematically examine the patterning of failure modes and task types noted in Figure 2, the contingency analysis test was conducted to examine the association with the failure modes and task types. The results show a significant association between the two variables, with a Pearson chi-square value χ^2 (20, 466) = 181.42, $p < 0.001$. Overall, there is a significant relationship between task types and failure modes, suggesting that different types of tasks are more likely to result in certain types of failures. We then examine the statistical power for each cell (i.e., each combination of task and failure mode). Table 5 details a cross-tabulation of task types and failure modes, along with counts and percentages of each combination.

	Agency Oversight		Algorithmic Bias		Artifact Malfunction		Artifact Misuse		Situational Unresponsiveness		Value Misalignment		Total Count
	Count	AR	Count	AR	Count	AR	Count	AR	Count	AR	Count	AR	
Emotional	0	⇨ -0.7	1	⇨ 0.2	0	⇨ -1.2	1	⇨ 0.6	0	⇨ -0.8	2	⬆ 2.1	4
Generative	9	⇨ -0.8	20	⇨ -0.9	19	⬇ -3.1	40	⬆ 6.9	7	⬇ -2.5	19	⇨ 1	114
Intellective	5	⇨ -1.1	20	⇨ 1.3	17	⇨ -1.2	10	⇨ -0.5	14	⇨ 1.5	11	⇨ 0.1	77
Judgmental	28	⬆ 3.6	52	⬆ 4.1	23	⬇ -5.2	16	⬇ -2.6	24	⇨ 0.6	27	⇨ 0.9	170
Psychomotor	4	⬇ -2.3	2	⬇ -5.2	71	⬆ 10.7	3	⬇ -3.8	15	⇨ 0.7	6	⬇ -2.6	101
Total Count	46		95		130		70		60		65		466

Table 5. Failure Modes and Task Types

Note: AR refers to adjusted residuals, which indicate the degree of deviation from expected values. According to Agresti (2002)), an adjusted residual that is more than 1.96 indicates that the number of cases in that cell is significantly larger than would be expected (marked by ⬆), while an adjusted residual that is less than -1.96 indicates that the number of cases in that cell is significantly smaller than would be expected (marked by ⬇).

Table 5 shows that the most common task type is judgmental (170 cases, 36.5% of the total number of incidents), while the most common failure mode is artifact malfunctions (129 cases, 27.7% of the total). The combination of psychomotor task type and artifact malfunctions failure mode has the highest count (71 cases, 15.24% of the total). The contingency analysis results indicate that the task being performed may influence the likelihood of certain types of failures. Specifically, the contingency analysis further indicates that the combination of emotional tasks and value misalignment failures is more common than expected (i.e., adjusted residual = 2.1), as shown in Table 5. Figure 2 also visualizes this pattern. Thus we propose that:

Proposition 1: AI used for emotional tasks is more susceptible to value misalignment.

The contingency analysis indicates that the combination of generative tasks and artifact misuse is more common than expected (i.e., adjusted residual = 6.9) and less susceptible to artifact malfunction and situational unresponsiveness (i.e., adjusted residual = -3.1, -2.5, respectively), as shown in Table 5. Thus we propose that:

Proposition 2: AI used for generative tasks is more susceptible to artifact misuse and less susceptible to artifact malfunction and situational unresponsiveness.

The contingency analysis indicates that the combination of judgmental tasks and agency oversight, and algorithmic bias are more common than expected (i.e., adjusted residual = 3.6, 4.1, respectively). Moreover, the combination of judgmental tasks and artifact malfunction, and artifact misuse are less common than expected (i.e., adjusted residual = -5.2, -2.6, respectively), as shown in Table 5. Thus we propose that:

Proposition 3: AI used for judgmental tasks is more susceptible to agency oversight and algorithmic bias and less susceptible to artifact malfunction and artifact misuse.

The contingency analysis indicates that the combination of psychomotor tasks and artifact malfunction failures is more common than expected (i.e., adjusted residual = 10.7) and less common than expected for agency oversight, algorithmic bias, artifact misuse, and value misalignment (i.e., adjusted residual =2.3, -5.2, -3.8, and -2.6, respectively), as shown in Table 5. Thus, we propose that:

Proposition 4: AI used for psychomotor tasks is more susceptible to artifact malfunction and less susceptible to agency oversight, algorithmic bias, artifact misuse, and value misalignment.

We also observed that AI used for intellective tasks may not be particularly susceptible to any of the six failure modes identified in our typology. Additionally, we made several observations regarding changes in the modes over time. First, corresponding with the growing emphasis on AI ethics, we see a marked improvement in AI ethics. This is evident from the decreasing instances of algorithmic bias failures in recent years, as illustrated in Table 6. Conversely, we also note an upward trend in the prevalence of artifact misuse in recent years. As AI continues to become more integrated into personal use, we emphasize the need for future studies to concentrate on addressing and understanding this phenomenon of AI misuse.

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	Total
Artifact Malfunction	1	2	5	8	11	16	10	17	20	31	9	130
Artifact Misuse	0	1	1	2	6	2	6	13	5	22	12	70
Algorithmic Bias	2	1	7	13	9	13	8	26	6	8	2	95
Agency Oversight	2	4	2	2	3	4	1	7	12	8	1	46
Situational Unresponsiveness	1	3	3	7	7	5	6	8	12	5	3	60
Value Misalignment	0	1	4	7	10	3	5	7	9	13	6	65
Total	6	12	22	39	46	43	36	78	64	87	33	466

Table 6. Failure Modes over Time

Note: We report the count of incidents across six failure modes and mark the adjusted residuals in this table (the rules are similar to Table 5). The data in 2023 is limited to the first quarter of the year.

Discussion

The objective of our study was to offer insights into AI incidents, which occur at the intersection of the AI artifact, humans, and situations. Through our data-grounded research, we offer a typology of six failure modes of AI-powered systems. These failure modes highlight the complexity and nuances associated with AI systems. We also related these failure modes to the tasks delegated to the AI artifact. In this section, we discuss the implications of our findings for the existing literature and practitioners of AI systems.

Contributions and Research Implications

Our study makes several key contributions to the literature. First, we extend the understanding of AI incidents by offering a typology of failure modes that accounts for the intersectionality of the AI artifact, humans, and situations. The six failure modes identified a wide range of AI-related incidents, providing a more holistic perspective on the challenges faced by AI-powered systems, beyond the critical challenges of fairness, transparency, and responsibility noted by the current literature. Second, we examine these AI failure modes in relation to AI tasks and formulate propositions. These propositions provide a framework to guide future research that accounts for task types when evaluating AI. Third, we attempt to bridge the gap between theoretical understanding and practical application, offering a more comprehensive perspective on the complexities and challenges associated with AI systems in real-life situations.

Our findings suggest that AI incidents are indeed different from IS failures studied in prior research. AI systems are more prone to failure due to their complex interplay among AI, user, developer, the unpredictable environments in which they operate, and the societal value system. Our study emphasizes

the need for further investigation into the unique challenges associated with AI systems to ensure their effective and safe deployment in various industries.

Practical Implications

Our findings have important practical implications for the development, deployment, and management of AI systems. First, by identifying the six failure modes, practitioners can better understand the potential risks and vulnerabilities associated with AI systems, leading to more informed decision-making when designing, implementing, and managing AI-powered solutions. In particular, our study can guide practitioners in identifying task scenarios where additional attention and resources may be necessary to mitigate potential failure. Second, our study serves as a valuable resource for policymakers tasked with creating robust regulatory frameworks for AI. We underscore the importance of developing guidelines and regulations that specifically address the unique challenges associated with each failure mode. Lastly, for end-users, our work illuminates the real-world challenges and uncertainties tied to AI systems. This heightened awareness can lead to more informed decisions in the everyday use of AI-enabled systems.

Limitations and Future Research

Our study has several limitations that could be addressed in the future. First, our study utilized two datasets that only included AI incidents that attracted sufficient public attention (i.e., reported by the media). While this approach allowed us to capture a wide range of AI incidents, future research could explore the types of AI incidents that are less visible to the public but still have significant consequences for individuals, organizations, or society. Second, our study did not explore the underlying causes of the identified failure modes. As such, the possible relationships among failure modes have yet to be considered. We intend to address this limitation through a root cause analysis method (Tucker et al. 2001). Future research may also explore the reasons behind each failure mode and investigate how they can be mitigated or prevented. Lastly, these failure modes are meant to be a foundation, not an exhaustive or exclusive list. As AI technology develops, possibly new ideal types may be added to the framework. Third, this research does not consider the interconnection and interdependency among the six AI failure modes. For example, agency oversight failure is likely to cause failure in value misalignment mode. Our next step is to focus on root cause analysis to understand the potential cascading effects among the six failure modes.

Conclusion

We proposed a typology for understanding the complexity of AI incidents. These failures are not just about AI itself but also involve human, and situations, which are all crucial factors for fostering more robust and reliable AI systems. By addressing these aspects holistically, developers, users, and policymakers can work together to create AI solutions that benefit humanity.

References

Agresti, A. 2002. *Categorical Data Analysis*, (2nd ed.). John Wiley & Sons, Inc., New York.

Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y. K., D'Ambra, J., and Shen, K. N. 2021. "Algorithmic Bias in Data-Driven Innovation in the Age of AI," *International Journal of Information Management* (60), p. 102387.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., and Benjamins, R. 2020. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," *Information Fusion* (58), pp. 82-115.

Arya, V., Bellamy, R. K., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., and Mojsilović, A. 2019. "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques," *arXiv preprint arXiv:1909.03012*.

Bail, C. A. 2012. "The Fringe Effect: Civil Society Organizations and the Evolution of Media Discourse About Islam since the September 11th Attacks," *American Sociological Review* (77:6), pp. 855-879.

Baird, A., and 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), pp. 315-341.

Barocas, S., Hardt, M., and Narayanan, A. 2019. "Fairness and Machine Learning: Limitations and Opportunities," *Fairness and Machine Learning: Limitation and Oppotunities*.

Berente, N., Seidel, S., and Safadi, H. 2019. "Research Commentary—Data-Driven Computationally Intensive Theory Development," *Information Systems Research* (30:1), pp. 50-64.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. 2018. "'It's Reducing a Human Being to a Percentage' Perceptions of Justice in Algorithmic Decisions," *Proceedings of the 2018 Chi conference on human factors in computing systems*, pp. 1-14.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of machine Learning research* (3:Jan), pp. 993-1022.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. 2016. "Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings," *Advances in neural information processing systems* (29).

Brachman, R. J. 2006. "AI More Than the Sum of Its Parts," *AI Magazine* (27:4), pp. 19-19.

Busuioc, M. 2021. "Accountable Artificial Intelligence: Holding Algorithms to Account," *Public Administration Review* (81:5), pp. 825-836.

Cavazos, J. G., Phillips, P. J., Castillo, C. D., and O'Toole, A. J. 2020. "Accuracy Comparison across Face Recognition Algorithms: Where Are We on Measuring Race Bias?," *IEEE transactions on biometrics, behavior, and identity science* (3:1), pp. 101-111.

Croidieu, G., and Kim, P. H. 2018. "Labor of Love: Amateurs and Lay-Expertise Legitimation in the Early Us Radio Field," *Administrative Science Quarterly* (63:1), pp. 1-42.

Crowston, K., and Bolici, F. 2019. "Impacts of Machine Learning on Work," *Proceedings of the 52nd Hawaii International Conference on System Sciences (HICSS), Hawaii, USA*.

Denzin, N. K. 2017. *The Research Act: A Theoretical Introduction to Sociological Methods*. Transaction publishers.

Edmondson, A. C., and McManus, S. E. 2007. "Methodological Fit in Management Field Research," *Academy of management review* (32:4), pp. 1246-1264.

Ekelhart, A., Kiesling, E., Grill, B., Strauss, C., and Stummer, C. 2015. "Integrating Attacker Behavior in It Security Analysis: A Discrete-Event Simulation Approach," *Information Technology and Management* (16:3), pp. 221-233.

Esterwood, C., and Robert Jr, L. P. 2023. "Three Strikes and You Are Out!: The Impacts of Multiple Human–Robot Trust Violations and Repairs on Robot Trustworthiness," *Computers in Human Behavior* (142), p. 107658.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.

Fernández-Loría, C., Provost, F., and Han, X. 2022. "Explaining Data-Driven Decisions Made by AI Systems: The Counterfactual Approach," *MIS Quarterly* (46:3), pp. 1635-1660.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. 2021. "Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI," *MIS Quarterly* (45:3), pp. 1527-1556.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. 2022. "Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path toward Productive Delegation," *Information Systems Research* (33:2), pp. 678-696.

Ge, R., Zheng, Z., Tian, X., and Liao, L. 2021. "Human–Robot Interaction: When Investors Adjust the Usage of Robo-Advisors in Peer-to-Peer Lending," *Information Systems Research* (32:3), pp. 774-785.

Gioia, D. A., Corley, K. G., and Hamilton, A. L. 2013. "Seeking Qualitative Rigor in Inductive Research: Notes on the Gioia Methodology," *Organizational research methods* (16:1), pp. 15-31.

Goot, M. J., Hafkamp, L., and Dankfort, Z. 2020. "Customer Service Chatbots: A Qualitative Interview Study into the Communication Journey of Customers," *International Workshop on Chatbot Research and Design*: Springer, pp. 190-204.

Grigorescu, S., Trasnea, B., Cocias, T., and Macesanu, G. 2020. "A Survey of Deep Learning Techniques for Autonomous Driving," *Journal of Field Robotics* (37:3), pp. 362-386.

Hamm, P., and Klesel, M. 2021. "Success Factors for the Adoption of Artificial Intelligence in Organizations: A Literature Review," *Proceedings of the 27th Americas Conference on Information Systems (AMCIS), Montreal, Canada.*

Janssen, A., Grützner, L., and Breitner, M. H. 2021. "Why Do Chatbots Fail? A Critical Success Factors Analysis," *Proceedings of the 42nd International Conference on Information Systems (ICIS), Austin, USA.*

Jussupow, E., Meza Martínez, M. A., Mädche, A., and Heinzl, A. 2021a. "Is This System Biased?–How Users React to Gender Bias in an Explainable AI System," *Proceedings of the 42nd International Conference on Information Systems (ICIS), Austin, USA.*

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. 2021b. "Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence," *Information Systems Research* (32:3), pp. 713-735.

Kane, G. C., Young, A. G., Majchrzak, A., and Ransbotham, S. 2021. "Avoiding an Oppressive Future of Machine Learning: A Design Theory for Emancipatory Assistants," *MIS Quarterly* (45:1), pp. 371-396.

Kordzadeh, N., and Ghasemaghaei, M. 2022. "Algorithmic Bias: Review, Synthesis, and Future Research Directions," *European Journal of Information Systems* (31:3), pp. 388-409.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A.-L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., and Gutmann, M. 2009. "Social Science. Computational Social Science," *Science (New York, NY)* (323:5915), pp. 721-723.

Lebovitz, S., Levina, N., and Lifshitz-Assa, H. 2021. "Is AI Ground Truth Really True? The Dangers of Training and Evaluating AI Tools Based on Experts' Know-What," *MIS Quarterly* (45:3), pp. 1501-1526.

Lebovitz, S., Lifshitz-Assaf, H., and Levina, N. 2022. "To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using Ai for Medical Diagnosis," *Organization Science* (33:1), pp. 126-148.

Lee, A. S., Thomas, M., and Baskerville, R. L. 2015. "Going Back to Basics in Design Science: From the Information Technology Artifact to the Information Systems Artifact," *Information Systems Journal* (25:1), pp. 5-21.

Li, J., Li, M., Wang, X., and Bennett Thatcher, J. 2021. "Strategic Directions for AI: The Role of Cios and Boards of Directors," *MIS Quarterly* (45:3), pp. 1603-1644.

Lindberg, A., Berente, N., Gaskin, J., and Lyytinen, K. 2016. "Coordinating Interdependencies in Online Communities: A Study of an Open Source Software Project," *Information Systems Research* (27:4), pp. 751-772.

Loch, K. D., Carr, H. H., and Warkentin, M. E. 1992. "Threats to Information Systems: Today's Reality, Yesterday's Understanding," *MIS Quarterly* (16:2), pp. 173-186.

Lu, J., Lee, D., Kim, T. W., and Danks, D. 2019. "Good Explanation for Algorithmic Transparency," *Available at SSRN 3503603*).

Lyytinen, K., and Robey, D. 1999. "Learning Failure in Information Systems Development," *Information Systems Journal* (9:2), pp. 85-101.

Markus, M. L., and Silver, M. S. 2008. "A Foundation for the Study of It Effects: A New Look at Desanctis and Poole's Concepts of Structural Features and Spirit," *Journal of the Association for Information systems* (9:10), p. 5.

McGrath, J. E. 1984. *Groups: Interaction and Performance*. Prentice-Hall Englewood Cliffs, NJ.

Meng, J., and Dai, Y. 2021. "Emotional Support from AI Chatbots: Should a Supportive Partner Self-Disclose or Not?," *Journal of Computer-Mediated Communication* (26:4), pp. 207-222.

Mikalef, P., Conboy, K., Lundström, J. E., and Popovič, A. 2022. "Thinking Responsibly About Responsible AI and 'the Dark Side'of AI." Taylor & Francis, pp. 257-268.

Miranda, S., Berente, N., Seidel, S., Safadi, H., and Burton-Jones, A. 2022a. "Editor's Comments: Computationally Intensive Theory Construction: A Primer for Authors and Reviewers," *MIS Quarterly* (46:2), pp. iii-xviii.

Miranda, S. M., Wang, D. D., and Tian, C. A. 2022b. "Discursive Fields and the Diversity-Coherence Paradox: An Ecological Perspective on the Blockchain Community Discourse," *MIS Quarterly* (46:3), pp. 1421-1451.

Miranda, S. M., Xing, Q., and Zhai, S. 2022c. "Creating Opportunity Amid Geographic Constraint on Digital Innovation Discourses," *Proceedings of the 43nd International Conference on Information Systems (ICIS), Copenhagen, Denmark.*

Möhlmann, M., Zalmanson, L., Henfridsson, O., and Gregory, R. W. 2021. "Algorithmic Management of Work on Online Labor Platforms: When Matching Meets Control," *MIS Quarterly* (45:4), pp. 1999-2022.

Nilsson, N. J. 2009. *The Quest for Artificial Intelligence*. Cambridge University Press.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., Ruggieri, S., Turini, F., Papadopoulos, S., and Krasanakis, E. 2020. "Bias in Data-Driven Artificial Intelligence Systems—an Introductory Survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* (10:3), p. e1356.

Nushi, B., Kamar, E., and Horvitz, E. 2018. "Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pp. 126-135.

O'Leary, D. E. 2013. "Artificial Intelligence and Big Data," *IEEE intelligent systems* (28:2), pp. 96-99.

Parikh, R. B., Teeple, S., and Navathe, A. S. 2019. "Addressing Bias in Artificial Intelligence in Health Care," *JAMA* (322:24), pp. 2377-2378.

Rai, A., Constantinides, P., and Sarker, S. 2019. "Next Generation Digital Platforms:: Toward Human-AI Hybrids," *MIS Quarterly* (43:1), pp. iii-ix.

Reimers, N., and Gurevych, I. 2019. "Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks," *arXiv preprint arXiv:1908.10084*).

Rezazade Mehrizi, M., NICOLINI, D., and Modol, J. R. 2022. "How Do Organizations Learn from Information System Incidents? A Synthesis of the Past, Present, and Future," *MIS Quarterly* (46:1), pp. 531-590.

Russell, S., and Norvig, P. 2002. "Artificial Intelligence: A Modern Approach," Pearson Education, Inc.

Salo, M., Makkonen, M., and Hekkala, R. 2020. "The Interplay of It Users' Coping Strategies: Uncovering Momentary Emotional Load, Routes, and Sequences," *MIS Quarterly* (44:3), pp. 1143-1176.

Spears, J. L., Barki, H., and Barton, R. R. 2013. "Theorizing the Concept and Role of Assurance in Information Systems Security," *Information & Management* (50:7), pp. 598-605.

Tan, C.-W., Benbasat, I., and Cenfetelli, R. T. 2016. "An Exploratory Study of the Formation and Impact of Electronic Service Failures," *MIS Quarterly* (40:1), pp. 1-30.

Teodorescu, M., Morse, L., Awwad, Y., and Kane, G. 2021. "Failures of Fairness in Automation Require a Deeper Understanding of Human-Ml Augmentation," *MIS Quarterly* (45:3), pp. 1483-1500.

Tucker, A. L., Edmondson, A. C., and Spear, S. 2001. "Front-Line Problem Solving: The Responses of Hospital Nurses to Work System Failures," *Academy of Management Proceedings*: Academy of Management Briarcliff Manor, NY 10510, pp. C1-C6.

van den Broek, E., Sergeeva, A., and Huysman Vrije, M. 2021. "When the Machine Meets the Expert: An Ethnography of Developing AI for Hiring," *MIS Quarterly* (45:3), pp. 1557-1580.

Wharton, A. S. 2009. "The Sociology of Emotional Labor," *Annual review of sociology* (35), pp. 147-165.

Wirth, N. 2018. "Hello Marketing, What Can Artificial Intelligence Help You With?," *International Journal of Market Research* (60:5), pp. 435-438.