# Failures in the Loop: Human Leadership in AI-Based Decision-Making

## I. Introduction

THE dark side of AI has been a persistent focus in discussions of popular science and academia (Appendix A), with some claiming that AI is "evil" [1]. Many commentators make compelling arguments for their concerns. Techno-elites have also contributed to the polarization of these discussions, with ultimatums that in this new era of industrialized AI, citizens will need to "[join] with the AI or risk being left behind" [2]. With such polarizing language, debates about AI adoption run the risk of being oversimplified. Discussion of technological trust frequently takes an *all-or-nothing* approach. All technologies – cognitive, social, material, or digital – introduce tradeoffs when they are adopted, and contain both 'light and dark' features [3]. But descriptions of these features can take on deceptively (or unintentionally) anthropomorphic tones, especially when stakeholders refer to the features as 'agents' [4], [5]. When used as an analogical heuristic, this can inform the design of AI, provide knowledge for AI operations, and potentially even predict its outcomes [6]. However, if AI agency is accepted at face value, we run the risk of having unrealistic expectations for the capabilities of these systems.

At present- and for the foreseeable future- AI has no will or agency of its own; it cannot autonomously enter a market to satisfy the needs of stakeholders, and, consequently, cannot take responsibility for anything it produces as an outcome (e.g., a prediction), as it does not act of its own accord [A1]. A "thing" is not a legal entity and cannot undergo trial in a court of law, nor can it be convicted of a crime [7]. As a Canadian airline learned, if a chatbot they use promises consumers a product or service, the company and its human employees remain accountable [8].

Much attention has been paid to the trustworthiness of AI in terms of its accuracy, reliability, and 'fit' relative to human performance. While important, such technical characteristics of AI neglect that stakeholders generally lack the competencies to assess systems in these terms. Instead of focusing on AI or AI-related predictive errors, or AI-based accidents or AI bias, this editorial considers assumptions underlying an AI's design, development, and deployment. This requires an assessment of (1) the motivation behind the development of an AI, (2) the responsibility for initiating, monitoring, and regulating its development, (3) the characteristics of group decision-making processes, (4) how values and risk management have been incorporated into the stages of development and implementation, (5) the responsibility for continual monitoring and regulation of AIs following implementation, and (6) how de-implementation of AI occurred when applicable.

## II. Assigning Responsibility in the AI Lifecycle

All tools are developed for a reason. Over time, those tools can be adapted by others for ends other than were initially intended (i.e., re-domaining [9]). Scientific tools were often framed as a means of controlling nature: physical or human. For instance, in the United States, the science of behaviorism was often referenced as a tool to shape society [10]. Modern software development can be understood in similar terms, such that data – or people *through* datafication – can be controlled: loans can be awarded, airline travel denied, academic assignments failed, and behaviors nudged in directions that benefit retailers and politicians [11], [12], [13].

### A. Lackus Reus: AIs Non-Volition and Corporate Accountability

AI cannot choose to do these things independently. Choice requires that there are alternatives, that these alternatives are perceived and understood, that appropriate resources are available to realize these choices, and that a decision-maker is autonomous. This last point is critical: without the freedom to decide, there is *no* choice. Here, again, we find the problem of AI anthropomorphism: despite referring to systems as 'autonomous', the systems cannot be truly autonomous as they have no agency. For example, autonomous weapons systems (AWS) or even self-driving cars, are unleashed (or leashed) by regulators, manufacturers, and operators. Humans remain accountable for choosing to develop, adopt, or shutdown a system. Any claim that humans are *out-of-* or *on-the-loop* ignores the fact that a choice was still made to initiate the loop and that humans are the sole agents who can understand the *meaning* of what occurs in the loop. Human decision-makers cannot be absolved from their obligations to users, consumers, and citizens with a rhetorical device of calling a system an autonomous agent as they are providing a product or service and are therefore the guarantors of trust.

Perhaps this problem is not especially unique to AI. Despite organizations being viewed as entities in the law, employees or leadership are held accountable, when successes or failures occur. The idea of *corporate personhood* in essential [14, p. 31]. However, as many hands are involved in making AI, allocating blame is not an easy task. Moreover, while the brand of an organization might remain relatively stable, those behind the brand, change. This is the problem of the

Digital Object Identifier 10.1109/TTS.2024.3378587

'Ship of Thesus': as a revered warfighter, Thesus' ship was preserved but, as the years wore on, planks of his ship were replaced until none of the original structure remained. Michael [15, p. 163] and Schoenherr [9] separately raise this point with the International Business Machine (IBM) corporation who sold counting machines to facilitate Nazi Germany's catalogue of the Jewish population being placed in concentration camps [16]. Here, we assume that the people involved in those decisions are to blame, not the organization. Despite the apparent continuity of applications like ChatGPT, the parts that make it are constantly changing – changes made by humans.

### B. Limitations in Management and Human Decision-Making

We must also acknowledge that human decision-making is limited. While behavioral economists might call humans rational actors, the best we can hope for is *bounded* rationality: making decisions with limited information in a limited amount of time [17]. Hindsight bias [18] might make use attributed blame when outcomes might not have been foreseeable. Indeed, consequentialist theories of ethics are hampered by a need to delimit an examination of outcomes for a given population, for a given period of time. Concurrently, one's actions can have multiple outcomes, positive and negative, i.e., the *doctrine of double-effect* [19]. Consequently, our limited ability to recognize and manage risk must be accommodated within organizational and national strategies. To make this task more manageable, we can attempt to delimit stages in the AI life cycle [20]. Using the concept of the AI life-cycle [21], Sullivan and Fosso-Wamba [20] describe 4 phases: (1) *initiate*, where ideas are explored, (2) *build and train*, where data is prepared and trained and models evaluated, (3) *deploy*, where AI models are run and tested, and (4) *manage and operate*, where the AI is monitored and insights are obtained regarding quality, performance and fairness. In this way, responsibility can be assigned to specific entities (i.e., individual people or groups) at each stage of the AI lifecycle [20].

This editorial considers the first "initiate" phase where the typical responsible parties for AI development include: business or government agency owner, data scientist, and data providers. Business or government agency owners traditionally are defined as companies or government entities that are the organizational body "who initiate the adoption of AI" [20, p. 918], but in this paper we are rather considering the human decision-maker(s) who provided the initial approval for the AI to be developed, and made it available to business and government agency owners for adoption. This is what could be identified as the root cause of the problem.

## III. AI as Socio-Technical Systems

### A. Socio-Technical Systems as Distributed Cognition

The notion of a socio-technical system provides a useful framework for discussing responsibility. This approach sees humans and nonhuman agents as components of a larger system. When viewed from an evolutionary perspective, the success of human-AI coordination will determine whether that socio-technical system is *replicable*, and the approach is adapted by others [9]. Crucially, in evolutionary theory, the process need not involve deliberation [22]. Blind copying often results without extensive consideration of the long-term implications of a system [6]. Contemporary concerns about AI use of power and its environmental impact can be understood in these terms: if users, distributors, and policymakers focus only on the benefits, they might ignore the costs. Similarly, if AI catastrophists only focus on the costs, they might ignore the legitimate reasons behind the adoption of innovation, leading to the two groups failing to come to a mutual understanding.

In psychological science, socio-technical systems can be described in terms of *distributed cognition* [23], [24]. Here, we assume that a task is performed within a social network, with each agent performing a specific role. These roles consist of identifying and defining problems as well as generating, verifying, and implementing solutions for the problems faced by the group. AI replaces individuals in this process, assuming the role of human agents. The critical feature of this approach is that AI need not have *agency* any more than a simple organism or pocket calculator does.

### B. AI and Trust Based on Transactive Memory

Despite limited agency, trust features prominently in discussions of AI. When humans choose to delegate an activity, they must base this choice on their knowledge of other people and AIs. In some cases, they might have direct experience. If they have the requisite knowledge, they might be able to assess the accuracy, reliability, and efficiency of these agents. This is referred to as *transactive memory*, the extent to which we keep track of those with whom we interact [25], [26], [27]. In most cases, humans are more likely to rely on stereotypes of AI as trustworthy or untrustworthy depending on the AI's domain of operation and the attitudes and beliefs of those in the users' social network. Trust is also conferred by the manufacturer. If a trustworthy company or country has produced an AI, other stakeholders will be more likely to accept it. If stakeholders use heuristics such as trust, this can lead to response biases. A recent example is the widespread adoption of large language models (LLM), such as ChatGPT. Media outlets and promotional materials praise the abilities of these systems. However, the claims are misleading. For instance, when provided with the same reasoning task, ChatGPT-3.5 provides the correct responses whereas ChatGPT-4.0 provides an incorrect response (Fig. 1). Naïve users are likely to hold the reasonable expectation that a newer version of software from the same company would produce responses that are as accurate if not more accurate than an earlier version. If an organization were to adopt such a system, it could lead to critical errors if left unsupervised.

A key determinant of transactive memory is our ability to understand the affective and cognitive states of others, i.e., empathy [28]. Empathy can take a variety of forms. However, what is most relevant for AI development is understanding that responsibility is intentionality [6], [29], in that philosophical frameworks concerned with justice prioritize intentions to avoid harm and maximize the good [30]. In psychological science, researchers have examined the

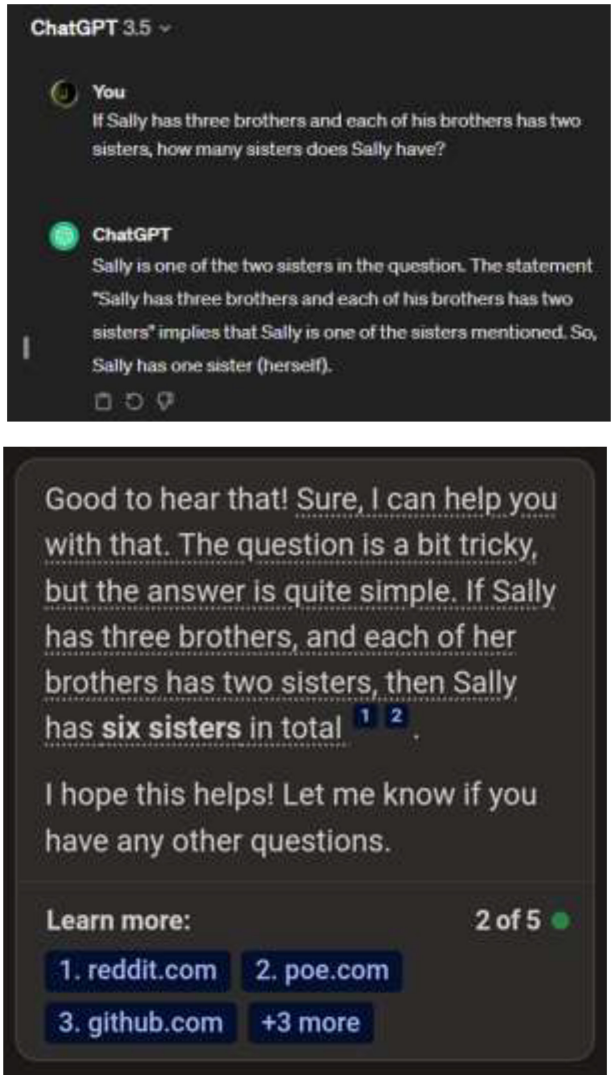


Fig. 1. Screenshot of query responses from OpenAI (ChatGPT-3.5) Bing conversational agent (ChatGPT-4) when provided with the same reasoning task (Oct 14, 2023).

development of *theory of mind*, our ability to take another person's perspective [31], [32]. Early studies suggest that humans can differentiate animate (goal-directed) and inanimate entities. Children eventually understand that other human and nonhuman agents have different mental states than they themselves do [6], with cross-cultural evidence suggesting that this is a function of the extent of exposure to different agents [33], [34]. Consequently, using a human agent analogy is generative in that it makes predictions, but we must not overlook that analogies are mental models, not reality.

## IV. Assigning Responsibility to AI

### A. Embedding Positive Intentionality in AI

In recent years, the lengthening list of the shortcomings of AI has led to the development of design and regulatory frameworks, with a noticeable shift to human-centered design [5], [9], [35], [36]. Whether designer or developer, entrepreneur or strategist, how we choose to apply AI in the name of digital transformation of organizations and societies matters. Each decision is value laden. Each decision point and context of implementation requires the consideration of multiple stakeholder values, many of which are conflicting or incommensurable. These choices are difficult for humans. They are currently impossible for AI. Despite the possibility of a 'moral governor' guiding the machine [37], AI cannot (yet) act in a morally responsible manner [38].

The possible benefits of AI are limitless. It is therefore easier for us to focus on its limitations and the vulnerabilities it creates when introduced into an existing socio-technical system. One useful distinction is that between the adverse consequences of Little and Big AI [39]. Negative outcomes from *Little AI* stem from "optimiz[ing] a function", and in so doing negatively affect a whole operation making it "more brittle or less effective." The end result reducing, rather than promoting, their agency. In contrast, *Big AI* implements "a new AI-based business model that may improve profits but manipulates customers" [39]. As we enter the industrial age of AI, we must reckon with this amorality.

While Euchner's [39] account focuses on "customers", these outcomes are equally applicable to other stakeholders such as employees, children, seniors, or citizens at large. Many of these outcomes might be isolated and incidental as many leaders are "making critical decisions, for our companies and for our society, about the proper uses of this technology, often implicitly." The speed of adaptation, combined with limited understanding of risk and uncertainty associated with largescale deployment, mean that significant consequences can result from seemingly minor decisions. For instance, there is the growing recognition that developing and sustaining AIs uses considerable natural resources. Energy and materials can have dire long-term consequences, if not given adequate consideration.

Not all of these outcomes are unintentional. The intentional use of behavioral engineering techniques to modify behaviors toward addiction [40] or nudge voters [41] or consumers (so-called, 'dark patterns of design'), can lead to both short-term individual problems and long-term societal problems, such as "fragmented work, impoverished social experiences, and a loss of privacy" [39].

### B. AI Decision-Makers Must Proactively Consider Ethics

If consequences cannot always be predicted, then we must differentiate an ethics of AI design and implementation that accounts for *reasonable* prospective and retrospective considerations. Here, the distinction between moral permissibility and moral fault [42], [43] is useful: whereas "[p]ermissibility attributions typically occur *before* an action and are focused on evaluating options to act, [wrongness and blame] are part of the evaluating fault for a violation *after* it occurs" (italics added).

Here, we assume that decision-makers must be proactive in their approach to ethical considerations, including the incorporation of principles from ethics, human factors, and the law. Critically, this must occur *prior* to development when decisions about the provisional objective of an AI can be restated, refined, rejected, or redefined, negating the need

for any moral fault post deployment and the corresponding harm that might be forthcoming. Many recent reviews have drawn attention to moral fault, including routine discrimination against minority groups and the poor in the name of operational performance and efficiency [44], [45], [46], [47]. Together, these reflect clear and present dangers to social justice and human rights. When organizations can collectively surveil, create user profiles, use these profiles to influence user decision-making, and produce consequential outcomes, we must acknowledge that users are being denied choice. They are being used by those who control the AI. For instance, Facebook Project Amplify prioritized the positive stories about Facebook in users' newsfeeds is an action directed toward shaping the perception of the organization [48].

### C. Responsibility Lies With the Human Agent(s)

Rather than placing responsibility at the feet of AI users, software developers, or organizations who adopt AI, we might instead elevate it to ask the tough questions of how fair-minded national and organizational leaders have failed to see the real opportunities that AI provides, and in the process failed themselves and all of us. Unlike the random variation that defines natural selection, *their* motives, *their* goals, and *their* commitments will shape the lifecycle of the AI as a human artifact. Individual choices to exploit users good-will and faith will concatenate into system-wide abuses. Adoption of AI ethical principles without embedding them within an organization's culture and promoting an ethical climate, represents little more than a virtue signal and fails to adhere to fundamental principles of business ethics [9]. A myopic focus on maximizing profits and minimizing operational costs can only lead to greater distrust in AI and user cynicism. In this light, the appearance of this trend for social media is unsurprising [49].

In parallel, one of the most concerning trends is changes in the trust of leaders and users. Trust can be understood as the degree of confidence one has in oneself, another, or a process. However, humans are chronically overconfident in their decisions [50]. The introduction of AI could, therefore, lead to blind spots if confidence in technologies exceeds their accuracy. A recent survey of senior executives at over 1,000 companies found that 92% of leaders reported that the use of AI had increased their confidence in decision-making [51]. Conversely, studies in the U.S. have found that the majority (62%) of people were concerned about AI, with 82% indicating that they did not trust tech executives to regulate AI [52].

AIs are ultimately created in our image (data), used as tools, and can change how we see ourselves and others. As Payne et al. [53] notes: "If AI is biased, blame the humans." To assess AI that pose serious challenges to human autonomy, we must send them through a gauntlet. If the AI can survive the "funnel" of questions" [53], then it can be released into the wild. Their first question is pertinent here: "Is the problem too risky for AI to tackle?" [53]. Certainly, we agree that AI can create a "responsibility gap" [38], once an AI is deployed, and does not perform in accordance with the hopes of the human decision-makers. But when an AI that has been deployed for nefarious purposes affects humans negatively, the responsibility without a doubt lies with the decision maker(s), and there is no "gap" to ponder; the responsibility *is* with the human decision-maker.

TABLE I
MORAL VIOLATION ATTRIBUTIONS FOR HUMAN DECISION-MAKING

| | Moral Attribution | Description |
|---|---|---|
| 1 | **Responsibility** | How responsible was the human decision-maker for the AI outcome? Did a single person make the decision, or two or more people come together to discuss the opportunity? What evidence was retained on the decision process, and reasons to proceed? |
| 2 | **Awareness** | Was the human decision-maker aware of the potential for the AI outcome and the impact on humans the deployment might have? Did they conduct adequate research into the potential risks of the AI software? Did they consult widely? Did they keep evidence of structured walkthroughs and provide what-if scenarios about how the AI software might impact business owners and end-users and others? |
| 3 | **Intention** | Did the human decision-maker intend for the outcome to occur from the outset of the idea/ conception? Was the intention to build the AI connected to supporting people or not? Was the AI application motivated by a monetary objective? Can the human decision-maker state their intent was clean and honorable using human-centered design principles, or were they attempting to satisfy corporate goals? |
| 4 | **Justified** | Was the human decision-maker justified in their role in the AI outcome? How is that justification demonstrated? What were the alternate mechanisms by which the human decision-maker could achieve the same ends without the adoption of AI? |
| 5 | **Permissible** | How permissible was it for the human decision-maker to produce this outcome stemming from the deployment of the AI? Was there evidence that there was consultation with end-users from the outset, inclusive of other stakeholders? Did the AI being considered for deployment, abide by general laws and regulations? Does the AI come into conflict with the values of the end-users? |
| 6 | **Blame** | How much do you blame the human decision-maker for the AI outcome? Are there other human agents that are to blame for the decision to proceed with the AI? Is there a hidden context that is missing for the decision to proceed with the given AI that will have negative externalities? |
| 7 | **Wrongness** | How wrong was the human decision-maker for their role in the outcome of the AI deployment? Did they listen to advice provided by others throughout the decision-making process? |

Adapted from Shank et al. [71, p. 653]

## V. MINI CASE STUDIES EXPLORING RIGHTS ABUSES

This editorial provides three mini case studies which on deeper analysis, have little to do with the capabilities of AI or AI bias, and everything to do with the moral attributions of human decision-making, including, (1) awareness, (2) responsibility, (3) intentionality, (4) justification, (5) permissibility,

TABLE II
THREE MINI CASES WHERE HUMAN DECISION-MAKERS HAVE UNLEASHED AI APPS TO THE DETRIMENT OF THE RIGHTS OF HUMANS

| Case ID | Description | Context |
|---|---|---|
| Case 1: *Past.*<br><br>Harm against workers. | AI impacting workers physical and mental health by pushing performance limits to the brink of human ability. | Worker rights; employment conditions; optimization and productivity dictated by a machine; human operators following machine suggestions without questioning the real needs of human workers for breaks at work, traffic delays in delivery, and anger management in call center work, among other contexts. |
| Case 2: *Present.*<br><br>Harm against children. | Children being used as free data annotators without their knowledge or remuneration. Privacy being impinged through the use of AI chatbots without parental supervision. | Children's rights not being respected. Entitled to carefree upbringing and full knowledge about the activities they are engaged in. They should not be exploited for work.<br><br>Children have a right to privacy. The use of a conversational agent in the form of an AI chatbot, is unacceptable when conversations cannot be supervised by an adult, and the data gathered is used for sentiment analysis and to train the AI. Especially problematic when the AI chatbot feature cannot be removed from the app the child relies on for communications with friends.<br><br>Children should not be duped because they lack the capacity to understand the intent of a new system or application. They need to be told in age appropriate language what they are engaging with. Consent should be garnered via the child's parent or guardian. |
| Case 3: *Prospective*.<br><br>Harm against those living with disability. | The potential for all humans to be screened for genetic rare diseases without their knowledge/ consent. Propensity for genetic discrimination, in gaining employment or even gaining and sustaining contractual obligations, e.g., related to health or life insurance, and other financial services like loans. | Citizen rights to privacy impinged by blanket coverage surveillance techniques. Unauthorized use of facial images for genetic analysis. Informed consent about the use of personally identifiable information unavailable. Citizens who are living with rare diseases completely unaware of the emergent technological capability, altogether. Accuracy of information, and determination may be lacking.<br>Whether the data being web scraped is related to one's face or to one's assets (e.g., vehicles), there should be an expectation of privacy in a public space. People are entitled to post images of themselves on social media without thinking that their face will end up on a repository being screened for genetic-related information. |

(6) blame, and (7) wrongness, in an attempt to explain a decision maker's behavior [42], [54], [55], [56]. In the next section we detail how AI depicted in each mini case study has been misdirected in some instances, by one or more human decision-makers. Table II presents these cases on a time continuum; first, a case that has already been heard in the courts; second, a case that has been flagged and likely to be raised in the courts in the not-too-distant future; and finally, another case that is speculative toward a future possibility. The names of the organizations are not entirely relevant here because numerous organizations are engaged in similar acts treading on worker rights, children's rights, and even disability rights through the use of AI-based apps. These rights are protected either by local laws, and/or by international conventions, such as the 11 conventions in the *International Labour Standards* [57]. Examples include C187 - Promotional Framework for Occupational Safety and Health Convention, 2006 (No. 187) [58], *The Convention on the Rights of the Child* [59], and the *United Nations Convention on the Rights of Persons with Disabilities* [60].

> *'We are pushing forward, we know what we are doing is completely new, we've looked for guidance but can't find any beyond the available legislation. We have no way of gauging if what we are doing is unlawful, but for now, we are forging forward… get in touch with us in the event we've missed something important that might shed light on our practice'* [[86], p. 70].

Box 1. Representative example of the Policy-Practice Pacing Problem inferred in [66]

### A. The Policy-Practice Pacing Problem

Regulation is often presented as a solution. However, governments and their supporting bureaucracies are neither nimble nor innovative entities. They are frequently late adopters of technologies, taking their cues from the 'will of the people' and leaders inside and outside government. The pervasive question of our time concerns what happens when techno-opportunism exceeds existing laws, regulations, and standards (refer to Box 1)] [61]. Designers and developers themselves have begun to ask for help when their solutions are sufficiently novel that there are neither legal precedents nor design principles to guide them [62, Sec. 8.4], [63, 38–45 min]. As they are working in distributed social networks, it is also unreasonable to assume that they have the time or the skills to lead such initiatives. Much like science, the imperative is typically to create rather than regulate. This pacing problem has created a unique set of circumstances that have been previously encountered in fields such as biotechnology and the life sciences where converging disciplines seem to have exacerbated the lag between innovation and regulation [64], [65].

## VI. CASE ONE: THREATS TO WORKER RIGHTS FROM AI-BASED MONITORING AND SURVEILLANCE

Workers in factories and warehouses, among many other contexts, are having their every move and work break monitored by AI-based apps, that are acting to cut pay, reduce work hours, or even fire employees. The European Parliament warned that such practices were occurring in December 2020 with the publication of a full-length report on the risks related to data subjects, digital surveillance, AI, and the

future of work [67]. Performance data in the workplace is being gathered and analyzed using a variety of technologies, providing evidence for certain decisions, a method referred to as "productivity scoring" [68]. AI-based software used by some firms has been judged to be grossly unreliable, inaccurate and discriminatory; but it has not stopped the practice. Regulators have begun to take action, imposing hefty fines in different parts of the world, signaling that such human resource practices are illegal [69]. Concurrently, researchers are now studying the effects of AI on workers from a human resource management (HRM) perspective [68], [70]. Such an approach is consistent with the historical shift from organization-centric approaches to the study of workplace (deviant) behavior, to one that considers employees' motivations in industrial-organizational psychology [71], as well as more recent work on 'insider threat' detection [72], [73].

In 2021, Deliveroo and Foodinho were both fined millions of dollars by the Italian privacy regulator "for using algorithms that discriminated against some "gig economy" workers" [74]. The regulator said: "that workers could be penalized based on how [AI] ... algorithms were being used to assess their work. But those algorithms remained secret, and workers had no way to appeal any such assessment. In addition, the regulator said, the firms could not prove that their algorithms were not being discriminatory." More recently, Amazon was fined $35 million USD for excessively intrusive practices on the collection of data related to its workers, breaking with the GDPR [75]. Human performance is being judged by AI-based algorithms, e.g., as video surveillance cameras are being used to monitor worker productivity, reviewed by other human monitors as far away as another continent.

## VII. Case Two: The Use of Children as Involuntary Data Annotators on Social Media Platforms and Data Privacy Infringements

Recently, much attention has been paid to workers in developing nations being employed to annotate data for mere dollars a day. Prominent examples of workers on low wages who are overworked, employed under exploitative conditions, and on poor labor standards, in places such as Kenya, Madagascar [76], [77], and the Philippines [78], have fostered AI 'data sweatshops' [79]. Consider that OpenAI used Kenyan workers for less than $2 per hour [80], and other companies tapping into the GlobalSouth or prisoners for 1.54 Euro per hour [81]. This is often very boring microwork, or 'ghost work' [82], where data labelers return to their homes after an initial training boot camp [83]. At times, the work is so taxing, that psychological trauma ensues [84]. This brings into question the supply chain of AI [85].

Less consideration has been given to the role that children play on social media platforms and apps. Through gamification, companies have discovered a way to engage users, making them members of a voluntary 'citizen science' style data collection effort. The act of annotation is made "less boring" and even can go undetected by the unsuspecting labeler, when it feels like just another game on a smartphone app. In this kind of setup, there are no labor costs and no liability. This is highly problematic, however, when the person being "gamed"/duped happens to be a child protected under the *Convention on the Rights of the Child*.

Additionally, organizations have embedded AI chatbots in applications for children, with the child-users inadvertently helping these organizations effortlessly acquire training data [86]. Seemingly innocent questions which become repetitive, such as "is this image a banana", encourages a response from a child with a correction of what the image actually is. The AI bot will even ask the user about their location, in order to cross-correlate where a given image/ thing is located or for more context. Try to delete the AI bot and the user is forced to delete the whole app. Beyond this, the deterioration of a child's right to privacy [87], as the child exchanges in personal conversation with a thing (i.e., a myAI), that may even direct them "to meet at the park" though no such meeting is actually occurring with any physical human being (see, e.g., the case of Australian Teagan Luketic and her daughter who was told by an AI that age was just a number, and was directed to meet at the park) [86]. These kinds of intrusions are exacerbated when certain behaviors are encouraged alongside a given platform, such as the idea of gaining extra "points" for streaking with a friend or a bot, and tracking continuous daily exchanges over a given medium between buddies [88].

Similarly, social media apps are designed to promote self-disclosure [89], [90], [91]. Studies have suggested that the more self-disclosure a child engages in on social networking sites, the more likely they are to experience cyberbullying [92], [93]. In parallel, virtual agents can also be designed to engage in simulated self-disclosure, which can promote reciprocal disclosure in children [94]. Given that the information provided by a child user could be used for the remainder of their lives, such disclosures could have dire ramifications. As it is unrealistic to expect any user – let alone, a child– to fathom the consequences of such disclosure, it is unlikely to be viewed as a valid form of consent.

## VIII. Case Three: Determining Genetic Rare Diseases Using AI and Biometrics Without Informed Consent

Facial recognition technologies have been one of the earliest uses of AI, exploiting their pattern matching capabilities. Images captured from CCTV at the scene of a crime are compared to those within a massive database [95]. The success of such usage has been variable, especially against people of color, where errors have been made in identifying a suspect accurately (e.g., [96]). The ability to Web scrape images from the Internet has meant that large databases of facial images in excess of 40 billion images [97] can be created and searched in a matter of seconds [98]. This is true in the case of ClearView AI [99], which has provided companies with the potential to amass facial images for different reasons, inclusive of missing persons, cold cases, and other crimes.

Similar pattern matching techniques have also been used to determine whether someone has genetic disorders [100], [101], [102], with their explicit consent or the consent of their parent/ guardian, and direction of their medical practitioner. The efficacy of the technology is a consequence of the association between multiple genes (genotype) and

their expression in physical characteristics (phenotype). When genes associated with genetic disorder are associated with facial features, images of a face predict the likelihood that they have a particular disorder [103]. As a noninvasive technology, the ability to rapidly detect genetic anomalies has considerable potential to promote healthcare delivery, especially in areas where genetic testing might be comparatively inaccessible.

Two issues stem from this approach. First, as with any prediction, algorithms might not be calibrated due to limited or biased datasets. Without converging evidence from other sources (e.g., direct genetic testing), the diagnostic capabilities of these systems might be questionable without the contribution of human clinicians and diagnosticians. This can lead to significant repercussions such as the denial of insurance or loans. If this data is shared and the decision-making process is not transparent, people might be unaware of the rationale for denial and cannot effectively contest these outcomes. Second, much like direct genetic testing [104], the consequences of these decisions are not restricted to a single individual. Instead, inferences might be made about the relatives of those who have been datafied, further obscuring the process used to make decisions [105]. Conceivably, the effects could also impact one's descendants, especially if third-party data aggregators or data breaches result in this information becoming widely available [106]. Even if data sets are anonymized, the risks of de-anonymization are quite high given the number of idiosyncratic features that define a person's DNA.

While there is no evidence as yet to suggest that this is happening today, without an individual's consent, the scenario is plausible and one might say inevitable in practice. Human curiosity about their nature and ancestry, health concerns (real or imagined), and the public nature of open data sets might lead to such outcomes. No matter the reasoning, this will have major privacy and security impacts on individuals and their respective families and the communities to which they belong.

## IX. Discussion: Responding to the Prospect of Unethical Practices of Humans Utilizing AI

Now that we have briefly explored these three exemplar cases, we will assess them against the moral violation attributes outlined by Shank et al. [43, p. 653]. Table III demonstrates that in each of the mini-cases of concern, the human decision to proceed with a particular type of AI should have been subject to further deliberation until the significant risks to individuals were adequately addressed. The attributes require further research and investigation on a case-by-case basis. For example, was a socio-technical risk register ever created by the human decision-maker(s), and not simply a risk register based on time, cost and specification [107]? From secondary sources of evidence, we can deduce that the human decision-maker(s) responsible for creating the AI software has not focused on the rights of the people who will be affected by the deployment of the AI. In the case of children, we see that even if the short-term consequences are minimal, their future might be adversely affected by youthful disclosures. In the case of genetic-based facial recognition technology, the negative impact might propagate out within a social network and toward future generations.

TABLE III
The Three Cases Mapped Against the Moral Violation Attributes

| | Moral Violation Attribute | Case 1: Worker Productivity | Case 2: Child Annotation & Privacy | Case 3: Rare Diseases & Disability |
|---|---|---|---|---|
| 1 | Responsibility | X | ? | ? |
| 2 | Awareness | X | ? | ? |
| 3 | Intention | X | ? | ? |
| 4 | Justified | X | ? | ? |
| 5 | Permissible | X | ? | ? |
| 6 | Blame | X | ? | ? |
| 7 | Wrongness | X | ? | ? |

Caption: Primary evidence is required to ascertain how decisions were made by humans with respect to a project's initiation. This can often take the form of a socio-technical risk register informed by initial multi-stakeholder consultation. For now, the courts determine the legality of a given AI application. The EU AI Act 2024 will shed greater light on the prohibition of certain AI practices, especially if the project is deemed "high risk" and subject to a fundamental rights impact assessment (FRIA). Thus, approaches to business ethics and legal instruments will assist in addressing the policy-practice pacing problem.

### A. Proportional Responses, Rights, and Action

A proportional response is required in each of these cases to stem long-term consequences. Regulators must respond with swift action, sending a clear message that exploitative and manipulative practices relating to these technologies are unacceptable. They must address the Policy-Practice Pacing Problem in a meaningful way, creating agile, responsive policies and regulatory guidelines [108], [109]. If this does not happen, we are destined to continue to erode the rights of workers, children, those living with disability, and human rights more generally [110]. Stakeholders might not be aware of the Faustian bargain, nor will the AI-Mephistopheles be capable of understanding that one has been made.

The rights at stake are myriad. Hickok and Maslej [68] stress the importance of (1) human dignity, among other issues such as the (2) right to privacy, (3) right to expression, (4) right to data protection, (5) right to collective action and power, (6) right to work and right to just and favorable remuneration, (7) validity and black box decisions, (8) right to due process, (9) normative judgements, (10) context and cultural specificity, (11) disability discrimination, (12) erosion of trust, (13) impact on health and safety, (14) feedback loops and (15) behavioral change. In assessing the moral violation attributes against the cases we have reviewed, it becomes clear that Hickok and Maslej [68] have captured the diverse concerns that we faced. As we have noted above, Vlad Glăveanu [111, p. 437] also emphasizes the importance of personal agency in making decisions, from possibility to action [P-ACT], reflecting on the different ways forward. Glăveanu outlines the P-ACT Model presenting the relationship "between becoming aware of and exploring possibilities." He claims that the power to pursue (or not), can come with deliberation and choice. This is

particularly pertinent for decision-makers with respect to the decision to positively impact people through the creation of AI software.

As the Policy-Practice Pacing Problem suggests, decision-makers' primary tasks when commissioning the development of an AI, should not be to consider whether they are legal but whether they are ethical. At the heart of the business ethics of AI development, we must ask whether or not a formal process has been defined for the gathering of evidence toward a decision to commission the development of an AI software, and/or whether alternate approaches to AI have been considered that includes prospective ethical and legal risks. It is important that as human decision-makers move toward the case to forge forward with an AI software/app, that they have demonstrated some depth of consultation, directly from end-users and other stakeholders that will be ultimately affected by the deployment. There should be knowledge of the AI's existence in any given context before deployment and not after-the-fact. The AIs sources, operations, and outcomes should be overt and not covert; obtrusive and not unobtrusive. Complete transparency and explainability to the end-user is a requirement to ensure that consent is valid. In fact, we would argue, in the spirit of participative design or co-design, the end-users should be brought into the process of decision-making at the initial phases of conception [5].

Ethical obligations do not end once the product is released. End-users affected by the system even if indirectly, need to have access to the data generated about them by the AI, offering corrections and deletion where required. Through preliminary testing in regulatory sandboxes, end-users might gain an awareness and provide feedback to the kinds of data that are collected about them [112]. This process is only preliminary, and can be repeated with more in-depth feedback during the full design phase.

### *B. Bridging Socio-Technical Gaps Between Actors, Their Interactions, Abstract Language, and the Process of Design*

We are not alone in these observations. Tigard et al. [113] broach this issue in what they refer to as "proper design" with a call to "multi-disciplinary attention". They note that "[w]e often speak and act as though we are blaming or praising our technological devices, despite *these practices being unfitting toward artifacts*" [113] (italics added). The researchers call for action in how we might approach the design of such systems. Likewise, Payne et al. [53], write about the need to bring together a "diverse group of people involved in making AI tools and principles that guide the use of those tools". Aizenberg and van den Hoven [110] add to these requirements by calling for a bridging of socio-technical gaps and the divide between "abstract value language" and "design requirements" in order to "facilitate nuanced, context-dependent design choices that will support moral and social values". They stress the importance of "designing for human rights in AI". They call for the design of algorithms and AI that "address stakeholder needs consistent with human rights" [110]. Selbst et al. [114, p. 60] the authors call for an emphasis on process-oriented approaches to design, away from solutionism, that "draws the boundary of abstraction to include social actors, institutions, and interactions." Fundamentally, this has to do with understanding an AI within the context of socio-technical systems [115]. AI is not solely about hardware and software, but the bigger picture perspective incorporating the social and environmental subsystems together with the technical subsystem.

Other approaches offer more concrete recommendations. They suggest identifying the affordances and how they align with ethical principles [116], [117], [118], [119]. Along with affordances that help users complete the intended tasks, the non-affordances that fail to consider essential features of a user population, and the disaffordances that limit users' interactions or dysaffordances that force users to engage in behavior or make declarations that are inaccurate in order to use technology (e.g., to select an ethnic or gender identity that is not accurate [for reviews, see [9], [120]. By adopting a human-centered approach to AI regulation, we not only accommodate human values, but also how users make value judgments, express them, and are affected by participating in this process. In organizations, this requires adopting a socio-technical systems perspective that includes understanding the power relationship between human agents and how they are mediated through AI. Even when it is claimed that AI is fully autonomous in any given system, one or more stakeholders are present.

## X. Conclusion

The introduction of any AI is supposed to transform an organization digitally, by means of automating (to varying degrees), internal and/or externally-facing business processes. There are however AIs that provide new management and operational capabilities that were not previously available using traditional software and hardware. The conception, invention, and prototyping of these emergent AIs can gather text, images, and video data that was previously beyond the reach of traditional data gathering capabilities. AI also provides a mechanism by which to scrutinize the data seeking patterns and trends. As these novel applications are being conceived and deployed, the question becomes whether decision-makers who are budgeting for the development of such AI-based applications have considered deeply their responsibility, the stakeholders they are affecting, and the potential of these applications to exacerbate existing human resource problems, or other problems.

The practice of critical reflection is invaluable. Hindsight might lead us to say 'if only…', however, some project objectives might have always retained an element of untrustworthiness or an intuitive aversiveness. It might have been simply that the objective solely considered cost optimization, rather than the health and wellbeing of users and their communities. Alternatively, it might have been focused on how to eliminate underperforming or suspect staff, rather than asking about the root cause of their failures and whether the organization can or should make changes to the culture and climate that the employees are embedded within, i.e., insider threats [121]. These are the kinds of discussions that need to be had early.

If interviews or other outreach activities had been conducted with staff of the organization that was considering deploying a performance-related AI, then the staff would have identified a different kind of AI that was necessary, e.g., how their employer might be more compliant to laws and regulations, rather than omnipresent monitoring through surveillance technologies. As human beings with personal agency, we can develop and expand our awareness, even if we do not possess it from the outset, or at the time a decision is required. But when AI is used in nefarious ways we might begin with the fundamental question of whether or not the AI itself from its very conception, could have been misapplied in a way that might cause harm to individual operators or end-users, or in some way break the law by its very existence and application. If the answer to the question: "could the AI being developed be used for harm?" is "yes", it means that the design of the AI was never intended to do good and there were no embedded guard-rails in place for its diffusion. We advocate for the development of tools and techniques, such as socio-technical risk registers that involve multi-stakeholder consultation at the outset, and would help decision-makers address complex issues early. Humans possess agency and can take responsibility in their decision-making and make the right choice to alter course where required, knowing that is the right thing to do.

## About the Special Issues

There are two special issues in this *IEEE TTS* March 2024 vol. 5. no. 1. The first is on the topic of "Public Interest Technology for Innovation in Global Development" led by Dr Roba Abbas of the University of Wollongong, and the second is on the topic of "Locating Responsibility in the Future of Human–AI Interactions" led by Dr Ehsan Nabavi of the Australian National University.

## Appendix: Related Articles

[A1] E. Nabavi, R. Nicholls, and G. Roussos, "Locating responsibility in the future of human—AI interactions," *IEEE Trans. Technol. Soc.*, submitted for publication, doi: 10.1109/TTS.2024.3386247.

## Acknowledgment

The authors would like to thank Hussein Abbass, Professor of Autonomous Systems and AI with the University of New South Wales Canberra for his initial feedback on the first draft of the paper that helped to shape its direction and contribution.

## References

[1] Ž Bjelajac, A. M. Filipović, and L. Stošić, "Can AI be evil: The criminal capacities of ANI," *Int. J. Cogn. Res. Sci., Eng. Educ. (IJCRSEE)*, vol. 11, no. 3, pp. 519–531, 2023.

[2] T. Carter. "Humans could become 'part AI' to keep up with superintelligent machines, OpenAI's chief scientist says." 2024. Accessed: Oct. 30, 2023. [Online]. Available: https://www.businessinsider.com/humans-could-merge-with-ai-in-future-openai-cofounder-2023-10?r=US&IR=T

[3] D. Grewal, A. Guha, C. B. Satornino, and E. B. Schweiger, "Artificial intelligence: The light and the darkness," *J. Busin. Res.*, vol. 136, pp. 229–236, Nov. 2021.

[4] K. Michael, R. Abbas, G. Roussos, E. Scornavacca, and S. Fosso-Wamba, "Ethics in AI and autonomous system applications design," *IEEE Trans. Technol. Soc.*, vol. 1, no. 3, pp. 114–127, Sep. 2020.

[5] J. R. Schoenherr, R. Abbas, K. Michael, P. Rivas, and T. D. Anderson, "Designing AI using a human-centered approach: Explainability and accuracy toward trustworthiness," *IEEE Trans. Technol. Soc.*, vol. 4, no. 1, pp. 9–23, Mar. 2023.

[6] D. C. Dennett, *The Intentional Stance*, Cambridge, MA, USA: MIT Press, 1989.

[7] "What is legal entity?: The law dictionary." Black's Law Dictionary. Accessed: Jul. 5, 2018. [Online]. Available: https://thelawdictionary.org/legal-entity/

[8] N. Robertson. "Air Canada must pay refund promised by AI chatbot tribunal rules." The Hill. Accessed: Feb. 18, 2024. [Online]. Available: https://thehill.com/business/4476307-air-canada-must-pay-refund-promised-by-ai-chatbot-tribunal-rules/

[9] J. R. Schoenherr, *Ethical Artificial Intelligence from Popular to Cognitive Science*. New York, NY, USA: Taylor Francis, 2022.

[10] R. M. Lemov, *World as Laboratory: Experiments With Mice, Mazes, and Men*. New York, NY, USA: Macmillan, 2005.

[11] S. Akter et al., "Algorithmic bias in machine learning-based marketing models," *J. Bus. Res.*, vol. 144, pp. 201–216, May 2022, doi: 10.1016/j.jbusres.2022.01.083.

[12] S. Akter, Y. K. Dwivedi, K. Biswas, K. Michael, R. J. Bandara, and S. Sajib, "Addressing algorithmic bias in AI-driven customer management," *J. Glob. Inf. Manag.*, vol. 29, no. 6, p. 27, 2021, doi: 10.4018/JGIM.20211101.oa3.

[13] S. Akter et al., "Algorithmic bias in data-driven innovation in the age of AI," *Int. J. Inf. Manag.*, vol. 60, Oct. 2021, Art. no. 102387, doi: 10.1016/j.ijinfomgt.2021.102387.

[14] K. Michael, "Modern indentured servitude in the gig economy: A case study on the deregulation of the taxi industry in the United States," *IEEE Technol. Soc. Mag.*, vol. 41, no. 2, pp. 30–41, Jun. 2022, doi: 10.1109/MTS.2022.3173306.

[15] K. Michael, "The technological trajectory of the automatic identification industry: the application of the systems of innovation (SI) framework for the characterization and prediction of the auto-ID industry," Ph.D. dissertation, School Inf. Technol. Comput. Sci., Univ. Wollongong, Wollongong, NSW, Australia, 2003. [Online]. Available: http://ro.uow.edu/theses/309

[16] E. Black, *IBM and the Holocaust: The Strategic Alliance between Nazi Germany and America's Most Powerful Corporation*. London, U.K.: Little, Brown, 2001.

[17] H. A. Simon, "Bounded rationality," *Utility and Probability*. London, U.K.: Palgrave Macmillan, 1990, pp. 15–18.

[18] N. J. Roese and K. D. Vohs, "Hindsight bias," *Perspect. Psychol. Sci.*, vol. 7, no. 5, pp. 411–426, 2012.

[19] A. McIntyre, "Doing away with double effect," *Ethics*, vol. 111, no. 2, pp. 219–255, 2001.

[20] Y. W. Sullivan and S. Fosso-Wamba, "Moral Judgments in the Age of artificial intelligence," *J. Busin. Ethics*, vol. 178, no. 4, pp. 917–943, 2022.

[21] K. Ishizaki (IBM, Armonk, NY, USA). *AI Model Lifecycle Management: Overview*. (Nov. 2020). [Online]. Available: https://www.ibm.com/cloud/blog/aimodel-lifecycle-management-overview

[22] K. Michael and R. Abbas, "Evolutionary economic theory: A review," in *TheoryHub Book*, S. Papagiannidis, Eds. [Online]. Available: https://open.ncl.ac.uk/

[23] E. Hutchins, "Distributed cognition," *Int. Encycl. Soc. Behav. Sci.*, vol. 138, no. 1, pp. 1–10, 2000.

[24] J. Hollan, E. Hutchins, and D. Kirsh, "Distributed cognition: Toward a new foundation for human-computer interaction research," *ACM Trans. Comput.-Hum. Interact. (TOCHI)*, vol. 7, no. 2, pp. 174–196, 2000.

[25] D. M. Wegner, "Transactive memory: A contemporary analysis of the group mind," in *Theories of Group Behavior*. New York, NY, USA: Springer, 1987, pp. 185–208.

[26] D. P. Brandon and A. B. Hollingshead, "Transactive memory systems in organizations: Matching tasks, expertise, and people," *Org. Sci.*, vol. 15, no. 6, pp. 633–644, 2004.

[27] V. Peltokorpi, "Transactive memory systems," *Rev. Gener. Psychol.*, vol. 12, no. 4, pp. 378–394, 2008.

[28] K. Stueber, *Rediscovering Empathy: Agency, Folk Psychology, and the Human Sciences*. Cambridge, MA, USA: MIT Press, 2010.

[29] N. Yuill and J. Perner, "Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction," *Develop. Psychol.*, vol. 24, no. 3, p. 358, 1988.

[30] M. Tomasello, "How children come to understand false beliefs: A shared intentionality account," *Proc. Nat. Acad. Sci.*, vol. 115, no. 34, pp. 8491–8498, 2018.

[31] K. Milligan, J. W. Astington, and L. A. Dack, "Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding," *Child Develop.*, vol. 78, no. 2, pp. 622–646, 2007.

[32] D. M. Bernstein, W. L. Thornton, and J. A. Sommerville, "Theory of mind through the ages: Older and middle-aged adults exhibit more errors than do younger adults on a continuous false belief task," *Exp. Aging Res.*, vol. 37, no. 5, pp. 481–502, 2011.

[33] H. C. Barrett et al., "Early false-belief understanding in traditional non-western societies," *Proc. Royal Soc. B, Biol. Sci.*, vol. 280, no. 1755, 2013, Art. no. 20122654.

[34] D. Liu, H. M. Wellman, T. Tardif, and M. A. Sabbagh, "Theory of mind development in Chinese children: A meta-analysis of false-belief understanding across cultures and languages," *Develop. Psychol.*, vol. 44, no. 2, p. 523, 2008.

[35] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts taxonomies opportunities and challenges toward responsible AI," *Inf. Fus.*, vol. 58, pp. 82–115, Jun. 2020.

[36] R. Chatila et al., "Trustworthy AI" in *Reflections on Artificial Intelligence for Humanity*. Cham, Switzerland: Springer, 2021, pp. 13–39. [Online]. Available: https://doi.org/10.1007/978-3-030-69128-8_2

[37] R. C. Arkin, P. Ulam, and A. R. Wagner, "Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception," *Proc. IEEE*, vol. 100, no. 3, pp. 571–589, Mar. 2012.

[38] D. W. Tigard, "Artificial moral responsibility: How we can and cannot hold machines responsible," *Cambridge Quart. Healthc. Ethics*, vol. 30, no. 3, pp. 435–447, 2021, doi: 10.1017/S0963180120000985.

[39] J. Euchner, "Little AI, big AI—Good AI, bad AI," *Res. Technol. Manag.*, vol. 62, no. 3, pp. 10–12, doi: 10.1080/08956308.2019.1587280.

[40] R. Abbas, K. Michael, M. G. Michael, C. Perakslis, and J. Pitt, "Machine learning, convergence digitalization, and the concentration of power: Enslavement by design using techno-biological behaviors," *IEEE Trans. Technol. Soc.*, vol. 3, no. 2, pp. 76–88, Jun. 2022.

[41] K. Michael, "Bots trending now: Disinformation and calculated manipulation of the masses," *IEEE Technol. Soc. Mag.*, vol. 36, no. 2, pp. 6–11, Jun. 2017, doi: 10.1109/MTS.2017.2697067.

[42] B. F. Malle, S. Guglielmo, and A. E. Monroe, "A theory of blame," *Psychol. Inquiry*, vol. 25, no. 2, pp. 147–186, 2014.

[43] D. B. Shank, A. DeSanti, and T. Maninger, "When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions," *Inf., Commun. Soc.*, vol. 22, no. 5, pp. 648–663, 2019, doi: 10.1080/1369118X.2019.1568515.

[44] V. Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish The Poor*, New York, NY, USA: St. Martin's Press, 2018.

[45] S. U. Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York, NY, USA: NYU Press, 2018.

[46] S. Wachter-Boettcher, *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech*, New York, NY, USA: WW Norton, 2017.

[47] C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, New York, NY, USA: Broadway Books, 2016.

[48] R. Mac and S. Frenkel, "No more apologies: Inside Facebook's push to defend its image." The New York Times. Sep. 21, 2021. Accessed: Oct. 9, 2021. [Online]. Available: https://www.nytimes.com/2021/09/21/technology/zuckerberg-facebook-project-amplify.html

[49] H. Kelly and E. Guskin, "Americans widely distrust Facebook, TikTok and Instagram with their data, poll finds." Washington Post. 2021. Accessed: Mar. 9, 2024. [Online]. Available: https://www.washingtonpost.com/technology/2021/12/22/tech-trust-survey/

[50] U. Hoffrage, "Overconfidence," *Cognitive Illusions*. London, U.K.: Routledge, 2022, pp. 287–306.

[51] E. Plastino. "Data modernization: Breaking the AI vicious cycle for superior decision-making." Cognizant. Aug. 2021. [Online]. Available: https://thoughtlabgroup.com/wp-content/uploads/2021/08/data-modernization-breaking-the-ai-vicious-cycle-for-superior-decision-making-codex6814.pdf

[52] R. Heath. "Exclusive poll: Americans distrust AI giants." Axios. Aug. 9, 2023. [Online]. Available: https://www.axios.com/2023/08/09/ai-voters-trust-government-regulation

[53] D. Payne, E. Peng, E. Schumaker, and R. Reader *If AI Is Biased, Blame the Humans*, POLITICO LLC., Arlington, VA, USA, 2023.

[54] J. A. Gailey and R. F. Falk, "Attribution of responsibility as a multidimensional concept," *Sociol. Spectrum*, vol. 28, no. 6, pp. 659–680, 2008.

[55] B. F. Malle and J. Knobe, "The folk concept of intentionality," *J. Exp. Soc. Psychol.*, vol. 33, no. 2, pp. 101–121, 1997.

[56] K. G. Shaver, *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. New York, NY, USA: Springer-Verlag, 1985.

[57] (Int. Labor Org., Geneva, Switzerland). *Conventions and Recommendations*. (2023). [Online]. Available: https://www.ilo.org/global/standards/introduction-to-international-labour-standards/conventions-and-recommendations

[58] (Int. Labor Org., Geneva, Switzerland). *C187—Promotional Framework for Occupational Safety and Health Convention 2006 (No. 187)*. Accessed: Feb. 4, 2024. [Online]. Available: https://www.ilo.org/dyn/normlex/en/f?p=NORMLEXPUB:12100:::NO:12100:P12100_ILO_CODE:C187:NO

[59] (UNICEF, New York, NY, USA). *The Convention on the Rights of the Child: The Children's Version*. Feb. 4, 2024. [Online]. Available: https://www.unicef.org/child-rights-convention/convention-text-childrens-version

[60] (Nat. Disabil. Author., Dublin, Ireland). *United Nations Convention on the Rights of Persons with Disabilities*. (2024). Accessed: Feb. 4, 2024. [Online]. Available: https://nda.ie/disability-policy/uncrpd

[61] G. E. Marchant, *The Growing Gap Between Emerging Technologies and Legal-Ethical Oversight The Pacing Problem* (The International Library of Ethics, Law and Technology), vol. 7, B. R. Allenby and J. R. Herkert, Eds. Dordrecht, The Netherlands: Springer, 2011. [Online]. Available: https://link.springer.com/book/10.1007/978-94-007-1356-7

[62] (Aeroranger, West Perth, WA, USA). *General Terms of Use: Section VIII-D: Legal Compliance*. (2021). [Online]. Available: https://www.aeroranger.com/general-terms-of-use

[63] R. Abbas and K. Michael, "Geospatial big data analytics: Opportunities & challenges in present & future modes of operation," in *Proc. Spatial Data Sci. Symp.*, 2021. [Online]. Available: https://www.youtube.com/watch?v=jBNOnd9jCNg

[64] H. Field (CNBC, Englewood Cliffs, NJ, USA). *AI Lobbying Spikes 1% as Calls for Regulation Surge*. Feb. 2, 2024. [Online]. Available: https://www.cnbc.com/2024/02/02/ai-lobbying-spikes-nearly-200percent-as-calls-for-regulation-surge.html

[65] K. Michael, R. Abbas, G. Roussos, E. Scornavacca and S. Fosso-Wamba, "Dealing with technological trajectories: Where we have come from and where we are going," *IEEE Trans. Technol. Soc.*, vol. 1, no. 1, pp. 2–7, Mar. 2020, doi: 10.1109/TTS.2020.2976425.

[66] K. Michael, R. Abbas, R. A. Calvo, G. Roussos, E. Scornavacca, and S. F. Wamba, "Manufacturing consent: The modern pandemic of technosolutionism," *IEEE Trans. Technol. Soc.*, vol. 1, no. 2, pp. 68–72, Jun. 2020, doi: 10.1109/TTS.2020.2994381.

[67] P. V. Moore (European Parliam., Strasbourg, France). *Data Subjects, Digital Surveillance, AI and the Future Of Work: STOA \Panel for the Future of Science and Technology*. (Dec. 2020). Accessed: Feb. 4, 2024. [Online]. Available: https://www.europarl.europa.eu/RegData/etudes/STUD/2020/656305/EPRS_STU(2020)

[68] M. Hickok and N. Maslej, "A policy primer and roadmap on AI worker surveillance and productivity scoring tools," *AI Ethic.*, vol. 3, pp. 673–687, Mar. 2023. [Online]. Available: https://doi-org.ezproxy.uow.edu.au/10.1007/s43681-023-00275-8

[69] S. Greenhouse. "Constantly monitored': The pushback against AI surveillance at work." The Guardian. Jan. 7, 2024. Accessed: Feb. 4, 2024. [Online]. Available: https://www.theguardian.com/technology/2024/jan/07/artificial-intelligence-surveillance-workers

[70] V. De Stefano and V. Doellgast, "Introduction to the Transfer special issue. Regulating AI at work: Labour relations, automation, and algorithmic management," *Transfer, Eur. Rev. Labour Res.*, vol. 29, no. 1, pp. 9–20, 2023, doi: 10.1177/10242589231157656.

[71] L. M. Andersson and C. M. Pearson, "Tit for tat? The spiraling effect of incivility in the workplace," *Acad. Manag. Rev.*, vol. 24, no. 3, pp. 452–471, 1999.

[72] J. R. Schoenherr, "Insider threats and individual differences: Intention and unintentional motivations," *IEEE Trans. Technol. Soc.*, vol. 3, no. 3, pp. 175–184, Sep. 2022.

[73] J. R. Schoenherr, K. Lilja-Lolax, and D. Gioe, "Multiple approach paths to insider threat (MAP-IT): Intentional, ambivalent and unintentional insider threats," *Counter Insider Threat Res. Pract.*, vol. 1, no. 1, pp. 1–23, 2022.

[74] M. J. Schwartz, "Food delivery services face GDPR fines over AI algorithms." BankInfo Security. Aug. 4, 2021. Accessed: Feb. 7, 2024. [Online]. Available: https://www.bankinfosecurity.eu/food-delivery-services-face-gdpr-fines-over-ai-algorithms-a-17212

[75] W. Davis, "France fines Amazon $35 million over 'excessive' worker surveillance." The Verge. Jan. 23, 2024. Accessed: Feb. 4, 2024. [Online]. Available: https://www.theverge.com/2024/1/23/24048197/amazon-fine-employee-surveillance-france-cnil-gdpr-privacy

[76] C. Le Ludec, M. Cornet, and A. A. Casilli, "The problem with annotation. Human labour and outsourcing between France and Madagascar," *Big Data Soc.*, vol. 10, no. 2, 2023, Art. no. 20539517231188723, doi: 10.1177/20539517231188723.

[77] C. Le Ludec and M. Cornet, "How low-paid workers in Madagascar power French tech's AI ambitions." The Conversation. 2023. Accessed: Mar. 30, 2023. [Online]. Available: https://theconversation.com/how-low-paid-workers-in-madagascar-power-french-techs-ai-ambitions-202421

[78] R. Tan and R. Cabato. "Behind the AI boom, an army of overseas workers in 'digital sweatshops'." The Washington Post. Aug. 28, 2023. https://www.washingtonpost.com/world/2023/08/28/scale-ai-remotasks-philippines-artificial-intelligence/

[79] P. Kant (Enabled Intell. Inc., Arlington, VA, USA). *AI Data "Sweatshops" are Bad News—And Threaten National Security*. (Aug. 29, 2023). Accessed: Feb. 4, 2024. [Online]. Available: https://enabledintelligence.net/news/opinion-ai-data-sweatshops-are-bad-news-and-threaten-national-security/

[80] B. Perrigo, "OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic." Time. Jan. 18, 2023. Accessed: Feb. 4, 2024. [Online]. Available: https://time.com/6247678/openai-chatgpt-kenya-workers/

[81] B. Lee Taylor. "Long hours and low wages: The human labour powering AI's development." The Conversation. Nov. 16, 2023. Accessed: Feb. 4, 2024. [Online]. Available: https://theconversation.com/long-hours-and-low-wages-the-human-labour-powering-ais-development-217038

[82] A. Naylor. "Underpaid workers are being forced to train biased AI on mechanical turk." VICE. Mar. 8, 2021. Accessed: Feb. 4, 2024. [Online]. Available: https://www.vice.com/en/article/88apnv/underpaid-workers-are-being-forced-to-train-biased-ai-on-mechanical-turk

[83] J. Dzieza. "AI is a lot of work." The Verge. Feb. 4, 2024. [Online]. Available: https://www.theverge.com/features/23764584/ai-artificial-intelligence-data-notation-labor-scale-surge-remotasks-openai-chatbots

[84] N. Rowe. "'It's destroyed me completely': Kenyan moderators decry toll of training of AI models." The Guardian. Aug. 3, 2024. Accessed: Feb. 4, 2024. [Online]. Available: https://www.theguardian.com/technology/2023/aug/02/ai-chatbot-training-human-toll-content-moderator-meta-openai

[85] K. Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT, USA: Yale Univ. Press, 2021.

[86] B. Palmada. "Mom slams Snapchat's 'creepy' AI bot after it poses as 25-year-old man and invites daughter, 13, to meet in park." New York Post. Sep. 2, 2023. [Online]. Available: https://nypost.com/2023/09/02/mom-says-snapchat-ai-bot-tried-to-entrap-teen-daughter/

[87] B. Lovejoy. "Snapchat My AI may be banned in UK over 'worrying' child privacy concerns." 9to5Mac. Feb. 4, 2024. [Online]. Available: https://9to5mac.com/2023/10/09/snapchat-my-ai-child-privacy/

[88] W. Antonelli, "How to start a Snapchat Streak and keep it alive to boost your Snap Score." Business Insider. Aug. 18, 2022. Accessed: Feb. 5, 2024. [Online]. Available: https://www.businessinsider.com/guides/tech/snapchat-streak

[89] N. N. Bazarova and Y. H. Choi, "Self-disclosure in social media: Extending the functional approach to disclosure motivations and characteristics on social network sites," *J. Commun.*, vol. 64, no. 4, pp. 635–657, 2014.

[90] S. Conger, J. H. Pratt, and K. D. Loch, "Personal information privacy and emerging technologies," *Inf. Syst. J.*, vol. 23, no. 5, pp. 401–417, 2013.

[91] P. Hodkinson, "Bedrooms and beyond: Youth, identity and privacy on social network sites," *New Media Soc.*, vol. 19, no. 2, pp. 272–288, 2017.

[92] B. K. Wiederhold and G. Riva, "Online social networking and the experience of cyber-bullying," in *Annual Review of Cybertherapy and Telemedicine*. Amsterdam, The Netherlands: IOS Press, 2012, pp. 212–217.

[93] D. Aizenkot, "Social networking and online self-disclosure as predictors of cyberbullying victimization among children and youth," *Child. Youth Serv. Rev.*, vol. 119, Dec. 2020, Art. no. 105695.

[94] F. Burger, J. Broekens, and M. A. Neerincx, "Fostering relatedness between children and virtual agents through reciprocal self-disclosure," in *Proc. 28th Benelux Conf. Artif. Intell.*, 2017, pp. 137–154.

[95] K. Michael, R. Abbas, P. Jayashree, R. J. Bandara, and A. Aloudat, "Biometrics and AI Bias," *IEEE Trans. Technol. Soc.*, vol. 3, no. 1, pp. 2–8, Mar. 2022, doi: 10.1109/TTS.2022.3156405.

[96] K. Hill. "Wrongfully accused by an algorithm." The New York Times. Jun. 24, 2020. [Online]. Available: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html

[97] C. Burt. "Clearview AI tops 40 billion reference images in facial recognition database." Biometric Update. Nov. 24, 2023. [Online]. Available: https://www.biometricupdate.com/202311/clearview-ai-tops-40-billion-reference-images-in-facial-recognition-database

[98] J. Clayton and B. Derico. "Clearview AI used nearly 1m times by US police, it tells the BBC." BBC News. Mar. 27, 2023. [Online]. Available: https://www.bbc.com/news/technology-65057011

[99] Staff. "Clearview AI." Accessed: April 13, 2024. [Online]. Available: https://www.clearview.ai/

[100] Y. Gurovich et al., "Identifying facial phenotypes of genetic disorders using deep learning," *Nat. Med.*, vol. 25, no. 1, pp. 60–64, 2019.

[101] H.-S. Rabia et al., "Automatic recognition of the XLHED phenotype from facial images," *Am. J. Med. Genet. Part A*, vol. 173, no. 9, pp. 2408–2414, 2017.

[102] A. J. Dingemans et al., "PhenoScore quantifies phenotypic variation for rare genetic diseases by combining facial analysis with other clinical features using a machine-learning framework," *Nat. Genet.*, vol. 55, no. 9, pp. 1598–1607, 2023.

[103] K. Michael, "Racial and genetic discrimination in automated face analysis," in *Proc. 5th Workshop Demogr. Variat. Perform. Biometr. Rel. Technol.*, Waikoloa, Hawaii, pp. 1–2, 2024. [Online]. Available: https://sites.google.com/msu.edu/dvbpa2024/keynote

[104] E. Niemiec and H. C. Howard, "Ethical issues in consumer genome sequencing: Use of consumers' samples and data," *Appl. Transl. Genom.*, vol. 8, pp. 23–30, 2016.

[105] K. Michael, "Biometric surveillance and the future of war" *Computer*, vol. 56, no. 7, pp. 21–30, 2023, doi: 10.1109/MC.2023.3249416.

[106] K. Michael, "The legal, social and ethical controversy of the collection and storage of fingerprint profiles and DNA samples in forensic science," in *Proc. IEEE Int. Symp. Technol. Soc.*, 2010, pp. 48–60, doi: 10.1109/ISTAS.2010.5514654.

[107] T. M. Willams, "Using a risk register to integrate risk management in project definition," *Int. J. Project Manag.*, vol. 12, no. 1, pp. 17–22, 1994.

[108] C. I. Gutierrez, G. E. Marchant, and K. Michael, "Effective and trustworthy implementation of AI soft law governance," *IEEE Trans. Technol. Soc.*, vol. 2, no. 4, pp. 168–170, Dec. 2021, doi: 10.1109/TTS.2021.3121959.

[109] C. I. Gutierrez, G. E. Marchant, and K. Michael, "Ideas on optimizing the future soft law governance of AI," *IEEE Technol. Soc. Mag.*, vol. 40, no. 4, pp. 10–13, Dec. 2021, doi: 10.1109/MTS.2021.3123746.

[110] E. Aizenberg and J. Van Den Hoven, "Designing for human rights in AI," *Big Data Soc.*, vol. 7, no. 2, 2020, Art. no. 2053951720949566, doi: 10.1177/2053951720949566.

[111] V. Glaveanu, "Possibility Spaces: An invitation to foster transformative experiences of the possible," *Possibil. Stud. Soc.*, vol. 1, no. 4, pp. 436–450, 2023, doi: 10.1177/27538699231214520.

[112] D. H. Guston, "Understanding 'anticipatory governance'," *Soc Stud Sci*, vol. 44, no. 2, pp. 218–42, Apr. 2014, doi: 10.1177/0306312713508669.

[113] D. W. Tigard, N. H. Conradie, and S. K. Nagel, "Socially responsive technologies: Toward a co-developmental path," *AI Soc.*, vol. 35, no. 4, pp. 885–893, 2020.

[114] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proc. Conf. Fair., Accountab. Transp. (FAT)*, 2019, pp. 59–68.

[115] R. Abbas and K. Michael, "Socio-technical theory: A review," in *TheoryHub Book*, S. Papagiannidis, Eds. [Online]. Available: https://open.ncl.ac.uk/9781739604400.2023

[116] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren, "Value sensitive design and information systems," in *The Handbook of Information and Computer Ethics*. New York, NY, USA: Wiley, 2008, pp. 69–101.

[117] P. Dalsgaard, K. Halskov, S. David, and J. Jofish Kaye, "Reflective design," in *Proc. 4th Decenn. Conf. Crit. Comput. Between Sense Sensibil.*, 2005, pp. 49–58.

[118] D. E. Wittkower, "Principles of anti-discriminatory design," presented at IEEE Int. Symp. Ethics Eng., Sci. Technol. (ETHICS), vol. 28, 2016, pp. 1–8.
[119] D. Wittkower, "Disaffordances and dysaffordances in code," in *Proc. AoIR Sel. Papers Internet Res.*, 2017. [Online]. Available: https://journals.uic.edu/ojs/index.php/spir/article/view/10174
[120] J. R. Schoenherr, "Learning engineering is ethical," in *Learning Engineering Toolkit*. London, U.K.: Routledge. 2022, pp. 201–228.
[121] J. R. Schoenherr and R. Thomson, "Insider threat detection: A solution in search of a problem," in *Proc. Int. Conf. Cyber Secur. Protect. Digit. Services (Cyber Security)*, 2020, pp. 1–7.

KATINA MICHAEL
School for the Future of Innovation in Society
Arizona State University
Tempe, AZ 85287 USA
School of Computing and Augmented Intelligence
Arizona State University
Tempe, AZ 85287 USA
E-mail: katina.michael@asu.edu

JORDAN RICHARD SCHOENHERR
Department of Psychology, Applied AI Institute
Concordia University
Montreal, QC H3G 1M8, Canada
Department of Psychology, Institute for Data Science
Carleton University
Ottawa, ON K1S 5B6, Canada
E-mail: Jordan.Schoenherr@carleton.ca

KATHLEEN M. VOGEL
School for the Future of Innovation in Society
Arizona State University
Tempe, AZ 85287 USA
E-mail: Kathleen.Vogel@asu.edu

**Katina Michael** (Senior Member, IEEE) is a Professor with Arizona State University and a Senior Global Futures Scientist with the Global Futures Laboratory and has a joint appointment with the School for the Future of Innovation in Society and the School of Computing and Augmented Intelligence. Prior to academia, she was employed with Nortel Networks, Anderson Consulting, and OTIS Elevator Company. She has been funded by the National Science Foundation, the Canadian Social Sciences and Humanities Research Council, and the Australian Research Council. She is the Director of the Society Policy Engineering Collective and the founding Editor-in-Chief of the IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY. She has been a technical editor, an editor, and the editor-in-chief since 2005 for several publishers. She is also the founding Chair of the inaugural Master of Science in Public Interest Technology.

**Jordan Richard Schoenherr** is an Assistant Professor with the Department of Psychology and a member of the Applied AI Institute, Concordia University, and an Adjunct Research Professor with the Department of Psychology and a member of the Institute for Data Science, Carleton University. He also serves as a Research Fellow with the Center for AI and Digital Policy. He is a former Postdoctoral Fellow with the University of Ottawa's Skills and Simulation Centre and a former Visiting Scholar with the U.S. Military Academy, West Point. He has acted as an ethics consultant for the Ombudsman, Integrity, and Resolution Office (Health Canada/PHAC), the Office of the Chief Scientist (Health Canada), the Canadian Border Services Agency, and the Department of National Defence. His primary areas of interest are learning and decision-making and metacognition with application in cyberpsychology (cybersecurity, disinformation, ethical AI, and XAI) and organizational behavior (incivility, insider threat, and knowledge management).

Dr. Schoenherr serves as a member of the Canadian Association of Chiefs of Police Artificial Intelligence Steering Committee.

**Kathleen M. Vogel** is a Professor with the School for the Future of Innovation in Society, Arizona State University. Her work has kept a close engagement between academia, intelligence, and policy. She is the author of *Phantom Menace or Looming Danger?: A New Framework for Assessing Bioweapons Threats* (Johns Hopkins University, 2013). Her research focuses on the production of knowledge and big data in intelligence assessments. She is also a Senior Editor of the IEEE TRANSACTIONS ON TECHNOLOGY AND SOCIETY.