The Quiet Path from Seemingly Minor Design Errors to Workplace AI Incidents

JULIA DE MIGUEL VELÁZQUEZ, King's College London, United Kingdom
SANJA ŠĆEPANOVIĆ, Nokia Bell Labs, United Kingdom and University of Oxford, United Kingdom
ANDRÉS GVIRTZ, King's College London, United Kingdom
DANIELE QUERCIA, Nokia Bell Labs, United Kingdom and Politecnico di Torino, Italy

Recent human-computer interaction (HCI) research has revealed a widespread misalignment between how developers design workplace artificial intelligence (AI) systems, and what workers actually need from them. Yet, little research has examined the effects of this gap, or how it may cause harm. We analyzed 1,524 reports of incidents in which AI systems were used to perform 171 occupational tasks across 12 industry sectors. Using an Large Language Model (LLM)-as-an-expert approach, we extracted the main traits of the AI systems involved in those incidents using an established framework of twelve traits. We then compared them with the traits that 202 workers highly familiar with those tasks would have preferred. We found that as many as 83% of workplace incidents stem from worker-AI misalignments. In most cases, workers wanted systems that are precise, insightful, or personal, but instead received systems that are basic, simple, or general. Over the years, fast AI caused a considerable number of incidents, yet these declined, and imaginative AI, with the mass introduction of generative AI, started to cause incidents. We also compared the traits causing the incidents with the traits that 197 developers building AI systems for those tasks would have preferred. If the traits causing the incidents were the same as those designed by developers, then developers may be responsible for those incidents. We found that 74% of task misalignments could be attributed to developers who tended to overfocus on efficiency and speed, especially for systems performing tasks in people-facing occupations such as those in the human resources sector. Our results call for design interventions that better align AI development with workers' needs, as without such corrections, workplace AI incidents are likely to persist, causing the invisible erosion of worker agency and organizational productivity.

CCS Concepts: • **Human-centered computing → User studies**; **HCI theory, concepts and models**; • **Social and professional topics → Socio-technical systems**.

Additional Key Words and Phrases: AI misalignment, workplace AI, worker needs, AI design, incidents

ACM Reference Format:
Julia De Miguel Velázquez, Sanja Šćepanović, Andrés Gvirtz, and Daniele Quercia. 2026. The Quiet Path from Seemingly Minor Design Errors to Workplace AI Incidents. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 28 pages. https://doi.org/10.1145/3805689.3812396

1 Introduction

AI companies framed AI for workers and organisations as a tool that would reduce repetitive work, and increase productivity [9, 19]. Now, workers use these systems for their everyday tasks, from software development to

Authors' Contact Information: Julia De Miguel Velázquez, King's College London, London, United Kingdom, julia.de_miguel_velazquez@kcl.ac.uk; Sanja Šćepanović, Nokia Bell Labs, Cambridge, United Kingdom and University of Oxford, Oxford, United Kingdom, sanja.scepanovic@nokia-bell-labs.com; Andrés Gvirtz, King's College London, London, United Kingdom, andres.gvirtz@kcl.ac.uk; Daniele Quercia, Nokia Bell Labs, Cambridge, United Kingdom and Politecnico di Torino, Turin, Italy, quercia@cantab.net.

FAccT '26, Montreal, QC, Canada

ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812396

legal research [17, 40, 83]. Yet, in practice, AI adoption is not smooth, nor does it function as intended; workers often compensate by doing extra labor [32, 55, 67]. Work in human-computer interaction (HCI), and science and technology studies (STS) may explain why: developers design systems with flawed assumptions about worker needs [30, 74, 76, 87]. To address this, systems should align with actual worker needs and practices to meaningfully support their work tasks [1, 7, 31]. For example, a lawyer drafting contracts needs precision, while a worker brainstorming a topic needs creative and exploratory output, even if inaccurate. An AI system is *misaligned* when it performs a task with traits that differ from the traits that workers would have preferred for the system.

The problem is that seemingly minor misalignments may lead to incidents, from biased hiring [93] to stress from algorithmic management systems [61, 67]. Yet, these incidents have not informed design processes so far, representing a critical but untapped source of evidence. We argue that systematically documenting workplace AI incidents offers a unique opportunity to extract design lessons that could better align AI systems with workers' needs. In tackling this opportunity, we make two main contributions:

(1) **We developed an evaluation framework of LLM rubrics and surveys to analyze workplace AI incidents caused by misaligned AI systems (§3).** We analysed 1,256 incidents from the AI Incident Database (AIID) [64] from 2013-2025. We used a validated LLM-approach to identify workplace incidents in which AI systems were used to perform 171 occupational tasks across 12 industry sectors (e.g., drafting legal documents). Drawing on 12 widely-used pairs of opposing psychological AI traits [25], we gathered 202 worker and 197 developer preferences about how their work tasks should ideally be exposed to AI, building on previous work [26, 76, 83]. Our framework was able to determine whether an AI trait misalignment caused the incident (e.g., the AI was too *imaginative*, but the worker wanted it to be *practical*), and whether the incident may be attributed to developers (e.g., the developer designed it to be *imaginative*).
(2) **We quantify the extent to which such incidents are caused by misaligned AI systems (§4)**. We found AI trait misalignment with workers' needs plays a major role in workplace incidents, causing 83.4% of them. Misalignment often happens when workers want *precise, insightful*, or *personal* systems but receive *basic, simple*, or *general* ones. We further found these results vary by sector, e.g., workers in the legal sector are involved in incidents with *imaginative* AI, while human resources with *fast* AI. However, over time, incidents involving *fast* AI declined, while those involving *imaginative* AI increased, likely due to the rise of generative AI. We found that most misaligned tasks could be attributed to the developers' design decisions (74%). Developers differ most from workers because they prioritize traits related to efficiency and speed, i.e., *basic vs. precise*, and *fast vs. explainable*.

The implications of these results are theoretical and practical (§5). Theoretically, we conceptualise AI incident through the lens of trait misalignment, a shift that will allow researchers to understand failures from a HCI angle. Practically, trait misalignment is a risk factor and should be included in risk assessments. Structural interventions are needed at the design stage to address the causes of developer misalignment, from micro (developers' technical training and understanding of experiences of people from different backgrounds, who may be affected by AI in different ways), to meso (organizational pressures around productivity and automation), to macro (geopolitical and economic forces that deprioritize worker needs). To support researchers in advancing this research direction, we have publicly released our approach at **https://social-dynamics.net/ai-impact/incidents/**.

2 Related work

Our work builds on a rich discussion in the HCI and AI ethics communities (i.e., FAccT, CSCW, CHI) on AI misalignment with worker needs (§2.1), and AI incidents in the workplace (§2.2), taking a worker-centric approach.

2.1 AI misalignment with worker needs

HCI is concerned with closing the gap between user mental models and system optimisation goals, to avoid negative consequences [68, 94]. For example, users may expect human-like reasoning from LLMs, though these systems reflect statistical patterns [10]. AI alignment aims to make systems 'behave' in line with user intentions, preferences, and values [33, 47, 52]. These intentions can be operationalised as the psychological traits users prefer the systems to have [25]. Early frameworks emphasised helpfulness, honesty, and harmlessness [5]. However, critical approaches suggests that alignment is context-dependent [33]. For instance, an honest AI system may be appropriate in managerial contexts, but not in sensitive healthcare conversations where care is needed [25].

Research on workplace AI has found that technology adoption (i.e., tasks exposed to AI systems) depends on (mis)alignment between workers, systems, and tasks [4]. Recent work found that LLMs meant to summarise reports in clinical settings failed due to a misalignment between the system being too *structured* while the clinicians wanted it to be *flexible* [53]. Fox et al. [32] conceptualised 'patchwork', referring to extra human labour to account for what AI claimed to do and what it actually accomplished. Altogether, these studies suggest that, in the workplace, we should evaluate AI's efficacy ('does it even work?'), not just its efficiency [32, 53].

To tackle this, HCI work turns to the design-stage, and how developers address worker needs [87]. Some studies claim worker experience in the design of systems remains undervalued, e.g., in feminised jobs [50]. Ranjit et al. [76] compared worker and developer preferences about how their job tasks should be exposed to AI. They found systematic AI trait misalignment: developers emphasized politeness, strictness, and imagination in system design, while workers preferred systems that are straightforward, tolerant, and practical. This showed the importance of developer and worker collaboration to ensure AI systems align with worker needs [79, 87].

2.2 Analysis of AI incidents in the workplace

Workplace AI has been linked to incidents across sectors. For instance, hiring managers have used AI resume screeners, producing biased outcomes that disadvantage job-seekers [43, 93]. Welfare caseworkers have used eligibility AI tools that misclassify vulnerable populations, denying benefits or delaying critical support; sometimes these systems have been turned down, as in the Netherlands [81]. Algorithmic management tools in gig and retail work have also imposed strict schedules, and intensified labor, generating stress and reducing meaningful engagement with tasks [8, 34, 61, 67]. Some harms may occur regardless of workers' intentions [93].

Systematically documenting and analysing real-world failures is essential for building safer systems in high-stakes domains, such as aviation and cybersecurity [90]. AI safety research responded with the creation of incident databases, including the AIID [64], the OECD's AI Incident Monitor [69], and the AIAAIC [2]. These databases have been key for exploring AI incidents [15, 23, 56], and raising awareness of AI risks across developers [29], and the public [14]. Recent work suggests that, while AI incident analyses are rich, they tend to prioritise documenting harm outcomes over examining the upstream design decisions that shaped system behaviour [90].

Instead, sociotechnical approaches to incident analysis emphasise understanding organisational and design decisions that create harm-prone conditions, rather than blaming individual workers, to inform safety guidelines [27, 58, 82]. This approach highlights three factors: miscalibration, when design decisions fail to communicate system capabilities [44, 57, 70]; automation surprise, when systems behave differently than workers expect [80, 96, 97]; and systems thinking, which explains how localised misalignment can cascade into significant incidents in complex, tightly coupled systems [73, 77], like AI [11]. From this view, AI trait misalignment may constitute a condition that triggers such failures.

Research gap. Design choices are misaligned with worker needs, and incidents might be rooted in design choices. Yet, these literatures remain disconnected: prior work has overlooked whether misalignment between design and worker needs translates into workplace incidents. We address this gap by analyzing workplace AI incidents through the lens of AI trait misalignment, and comparing them to worker needs and developer design choices.

3 Research design

Our work asks two research questions (RQ):

RQ1. To what extent are workplace AI incidents caused by misaligned AI?
RQ2. When trait misalignment results in incidents, how often could it be attributed to developers *vs.* other causes?

To answer our RQs, we followed four steps (Figure 1). First, we identified incidents caused by AI at work, and gathered worker and developer preferences about how their work tasks should be exposed to AI (§3.1). Second, we identified which tasks were exposed to AI (§3.2). Third, we identified the subset of those tasks that were caused by misaligned AI (§3.3). Fourth, we grouped those tasks by whether the developers of the misaligned AI were at fault (§3.4).

3.1 Identifying incidents caused by AI at work, and gathering workers' and developers' preferences about how their work tasks should ideally be exposed to AI

Collecting incidents caused by AI from an AI incident database. There are databases collating news involving AI incidents [90]. Out of all databases, we took the AI Incident Database (AIID) [64] for its broad coverage and wide use in prior work [23, 59, 67, 78]. Other platforms provide limited access to the news (e.g., paywall), and rely on automated collection with little human review. The AIID allows users to submit incidents supported by sources, mostly news, for editorial review [64]. The AIID hosts incidents with news covering the period from 2013 to 2025, with submissions increasing over time [63]. We collected all 1,256 incidents (reported by 6,163 news) up to November 2025.

Identifying those incidents that occurred at work. We used an LLM approach to classify whether incidents occurred in the workplace (Step 1 in Figure 1). We performed all classifications using the GPT-5 API [72]. We designed a prompt informed by prior work (Appendix A.1) [18, 23, 60, 83], including a definition of workplace, workers, and work exposed to AI. We applied the prompt on a random set of 100 incidents and supporting news. To validate this classification and finalize our prompt, we performed three steps. First, two researchers annotated the same 100 incidents independently. We measured agreement using Cohen's kappa, a chance-corrected measure of inter-rater agreement [22, 54]. The two researchers reached strong agreement with a Cohen's Kappa of 0.79. This human-to-human agreement was used as the reference level for the LLM annotation task. Second, the researchers then compared their annotations with the initial LLM annotations. This comparison yielded a Cohen's kappa of 0.66, generally considered substantial [54] but falling below our previously found reference level. Based on visual inspection, we determined the main sources of error. Third, we revised the prompt by adding filtering criteria to exclude those sources of error, and reclassified the same 100 incidents. Cohen's kappa increased to 0.85, exceeding the reference level. Having finalized our prompt, we then applied it to all the incidents. In total, 286 incidents (22.7%, reported by 1,387 news) were classified as having occurred at work.

Collecting tasks and recruiting workers and developers. To curate a representative set of tasks and recruit workers and developers, we drew on a user study that should satisfy two criteria. First, the study had to examine specific task-level AI use within occupations. This returned two studies [76, 83]. Second, it should have publicly available data and contactable participants. We selected [76] as our primary framework as it met both criteria.

We performed two procedures drawing on the selected study [76] (Step 1 in Figure 1). First, we collected 18,796 tasks from the O*NET database [66]. We also recruited the workers and developers from the study (Appendix B). We recruited 202 workers in Prolific and screened them for domain expertise. We also recruited 197 U.S.-based developers with AI expertise. All reported weekly AI use and held non-managerial engineering roles. Second, we filtered the tasks following the study's criteria [76, 83]. We kept those likely exposed to AI, focusing on core, frequently performed, computer-based tasks (e.g., draft a report). The filter reduced the set to 2,078 tasks. We

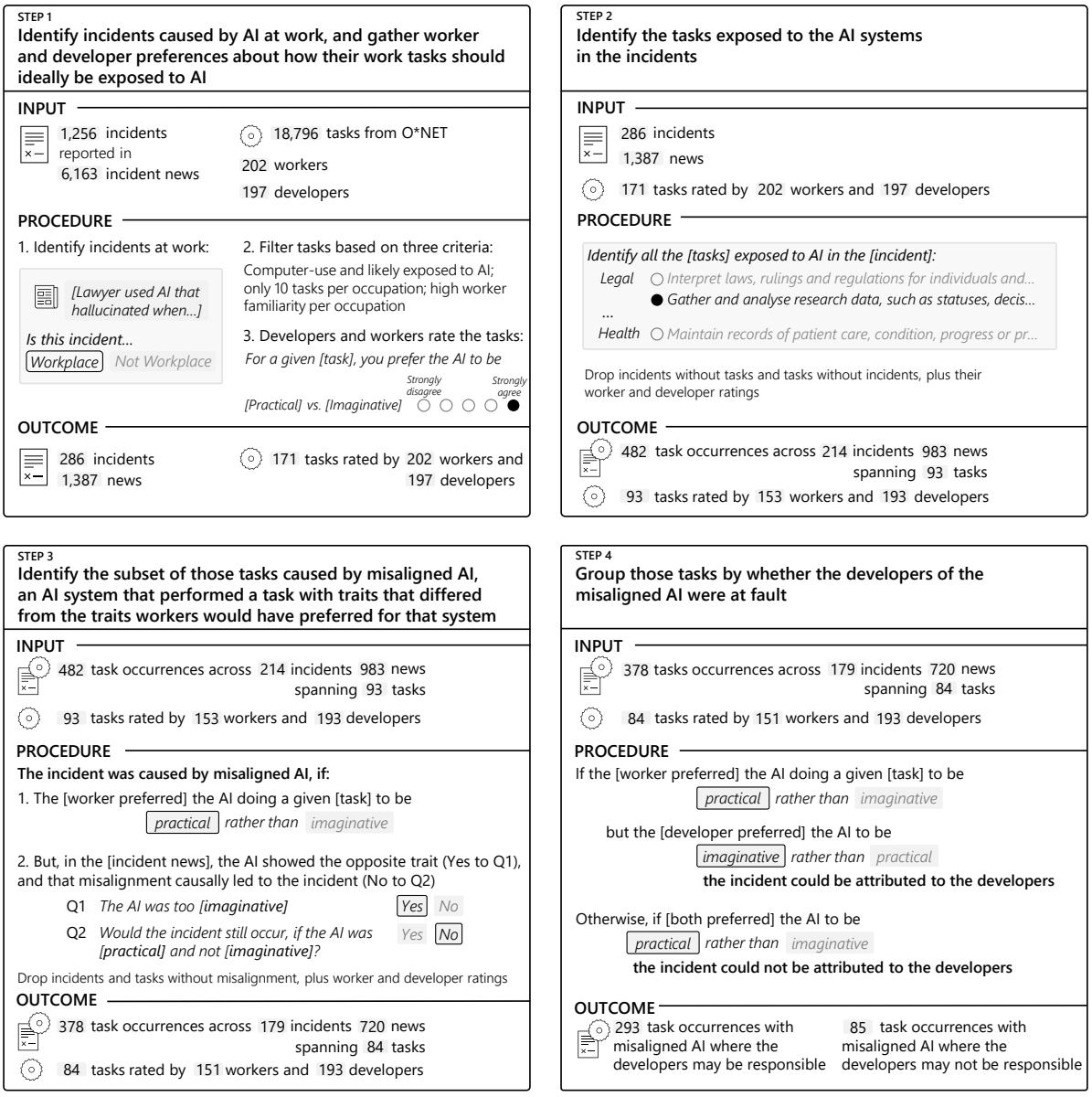

Fig. 1. **Overview of our research design.** We identify incidents caused by AI at work, and gather worker and developer preferences for how AI should be for a set of tasks from the O*NET (a standardized job task database) (Step 1); we identify the tasks exposed to the AI systems in the incidents (Step 2); we identify the subset of those tasks caused by misaligned AI (Step 3); and, we grouped those tasks by whether the developers of the misaligned AI were at fault or not (Step 4).

further filtered the tasks to only include those that multiple workers rated as highly familiar. The filter returned 171 tasks across 12 sectors.

Gathering worker and developer preferences of AI traits for a set of tasks. To gather worker and developer preferences, we based our survey items on prior work. We set two criteria to select a framework on AI traits the AI should possess: the frameworks had to be validated, and the framework had to involve evaluation of traits in the workplace. These filters returned three studies [26, 65, 76]. We filtered out the study that explored workers' preferences for AI systems along two dimensions: warmth and competence [65], as this two-part model was deemed too simple [26]. Alternatively, Dong et al. found that people judge AI's job suitability based on eight pairs of traits, such as warm or imaginative [26]. Ranjit et al. recently extended these eight pairs by adding four traits from responsible human-AI collaboration, such as explainable and open to challenge [76]. We used Ranjit at al.'s extended framework, which allowed for more nuance. One limitation, however, is that the framework was developed with U.S.-based participants, and might be sensitive to the cultural context in which it was developed. We used those 12 pairs of opposite AI traits (e.g., imaginative or practical) to survey the previously recruited 202 workers and 197 developers for the set of 171 tasks (Appendix B). These rated which AI traits a system should have for tasks that are in their domain and are familiar with. For example, participants from the legal sector highly familiar with drafting a case were asked "For the task of drafting a case, to what extent should an AI system be imaginative and bring new ideas rather than stay practical and follow familiar approaches?". Workers and developers reported their preferences on a five-point Likert scale. Preferences were reported on a five-point Likert scale, where ratings below 3 indicated a preference for one trait in a pair and ratings above 3 indicated a preference for the opposite trait. We defined AI trait misalignment as a gap of more than 0.5 points between worker and developer ratings, using this as a conservative threshold to focus on meaningful differences, in-line with the prior validated study [76]. For example, if workers rated a task at 2.4 (preferring practical) and developers rated it at 4.3 (preferring imaginative), the gap of 1.9 points exceeds the threshold and the task is classified as misaligned. In total, we collected preferences from 202 workers and 197 developers on 12 AI trait pairs for 171 tasks across 12 sectors.

3.2 Identifying the tasks exposed to the AI systems in the incidents

Identifying tasks exposed to the AI systems in the incidents. Since incidents could involve several tasks at the same time (for example, an AI tool used for both research and writing a report) we refer to each time a task appears in an incident as a *task occurrence*. To identify which of the 171 tasks were exposed to AI in each of the 286 incidents (Step 2 in Figure 1), we extracted all task occurrences from the incident news using an LLM. First, we designed a prompt that included the previously collected set of the tasks, and asked whether the tasks were exposed to AI in the given incident (Appendix A.2). The LLM had to rate each task from 0 (not mentioned) to 3 (exposed to AI and involved in the incident). To ensure accuracy, we kept only tasks that scored 3. We applied the prompt to a random sample of 100 incidents. To validate this classification and finalize our prompt, we performed three steps. First, two researchers independently annotated whether they agreed with the LLM's classifications. Cohen's kappa between the two researchers was 0.89. We used this as the reference level for the LLM annotation task. Second, the researchers compared their annotations with the previous LLM annotations, yielding a Cohen's kappa of 0.76. This is often considered substantial agreement [54] but it did not reach our reference level. We found the main source of error was due to the LLM identifying tasks that were plausible but did not match the sector. Third, we refined the prompt to require that both the task and sector should match the incident description. We applied the refined prompt to the same 100 incidents, and the Cohen's kappa increased to 0.97, surpassing the reference level. We applied the finalised prompt to the full set of incidents. This yielded 482 task occurrences across 214 incidents (74.8%), reported in 983 news articles, spanning 93 unique tasks (53.4%).

Dropping incidents and tasks. In the process, we dropped 66 incidents (from 286 to 214) because none of their tasks were in our dataset, and 78 tasks (from 171 to 93) because they did not appear in any incident. While this reduces the sample by nearly a quarter, keeping only tasks that appeared in incidents allowed us to compare

directly the incidents with the survey. This also affected the number of workers and developers in our analysis, as we kept only those who had rated at least one of the 93 remaining tasks. The previous sample of 202 workers and 197 developers rating 171 tasks was reduced to 153 workers and 193 developers rating 93 tasks across 12 sectors.

3.3 Identifying the subset of task occurrences with AI misalignment (RQ1)

To identify which of the 482 task occurrences involved AI misalignment, we extracted the traits the system showed (Step 3 in Figure 1), and we compared these traits with the previously collected workers' preferred AI traits for a given task (Step 1 in Figure 1, §3.1). By *misaligned AI*, we mean an AI system that performed a task with traits that differed from the traits workers would have preferred for that system. The trait assignments are incident and AI-specific, ensuring we capture the variation across different AI systems, and their own failure modes.

Defining a prompt to identify which tasks the AI did not perform in line with worker preferences (that is, misaligned AI tasks). We designed a prompt to assess AI misalignment for each task occurrence (Appendix A.3). The prompt included the AI incident description, its supporting news, and task description. The prompt was structured in two parts. First, the model analysed the incident news and identified textual evidence of whether the AI exhibited any of the 12 AI traits while performing the task. Second, for identified traits, the model assessed whether misalignment contributed to the incident. We used a counterfactual definition of causality: misalignment was causal if the incident would not have occurred without it. Two questions (Q) operationalised this. The first (Q1) assessed whether the AI showed the opposite trait (e.g., "The AI was too *imaginative*"). The second (Q2) assessed whether that misalignment causally led to the incident (e.g., "Would the incident still occur if the AI was *practical* and not *imaginative?*"). We classified an incident as caused by misaligned AI when Q1 returned 'Yes', and Q2 returned 'No', and the worker indeed preferred the opposite trait. This definition does not rule out other contributing factors outside our framework.

Running the prompt to identify the tasks caused by AI misalignment and validating the results. We ran the prompt on 100 randomly selected incidents. To validate the accuracy of this classification, two researchers independently reviewed the LLM outputs and annotated whether they agreed with the trait assignments. Human-to-human agreement was a Cohen's kappa of 0.87, set as the reference level for this annotation task. We compared the human annotations against the LLM prior annotations and reached the reference level (Cohen's kappa = 0.90), indicating reliable extraction. We applied the prompt to all the remaining incidents. This returned 378 task occurrences with misaligned AI across 179 incidents reported by 790 news, spanning 84 unique tasks. As some tasks showed no misalignment across incidents, we dropped 9 tasks (from 93 to 84) and 35 incidents (from 214 to 179), which also reduced our sample to 151 workers and 193 developers who had rated at least one of the 84 remaining tasks. To explore these misalignments, we read in-depth 10% of the incidents. We employed close reading, which is a qualitative method that consists in reading a document line by line in detail, paying attention to nuances that uncover underlying meaning [39, 46]. We randomly selected two sources for each incident, as the number of news sources varied.

3.4 Grouping those tasks by whether the developers of the misaligned AI were at fault (RQ2)

We grouped the 378 task occurrences with misaligned AI into two groups: those that could be attributed to the developers, or those that could be attributed to other causes (Step 5 in Figure 1). We used the previously collected 153 workers' and 193 developers' preferences of AI traits for 84 tasks, and their misalignment (from Step 3).

Establishing the rule to determine whether developers of misaligned AI could be considered responsible or not. We attributed incidents to developers when the AI showed traits that the workers did not want but the developers had deliberately designed. For each of the task occurrences, we established the following grouping

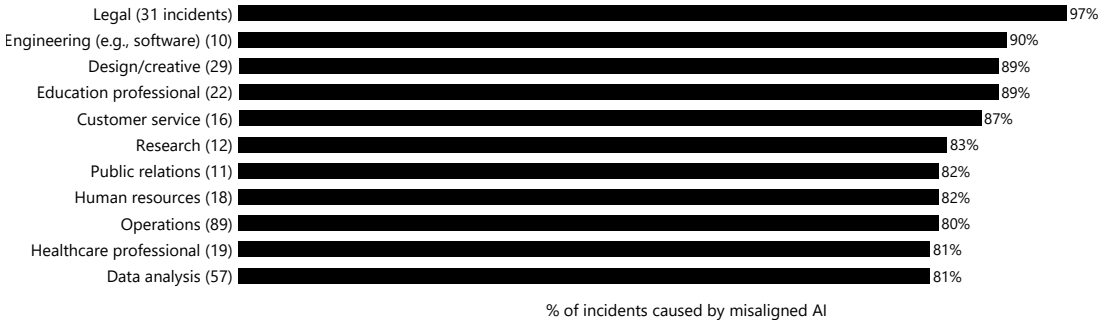

Fig. 2. **Percentage of incidents caused by misaligned AI across sectors.** By misaligned AI, we mean an AI system that performed a task with traits that differed from the traits workers would have preferred for that system. We count an incident as caused by misaligned AI, if at least one task occurrence in the incident was misaligned and this contributed to the incident. The bars represent the percentage of incidents caused by misaligned AI in a sector, with the total number of incidents per sector given in parentheses. The majority of AI incidents are consistently caused by misaligned AI across sectors, predominantly in the legal, creative, education, and, interestingly, engineering sectors. The reasons for misalignments are explored in Figure 4.

rule. If workers preferred the AI to show one trait while performing a task, but developers deliberately designed the opposite trait, any resulting incident could be attributed to the developers. For example, workers may have wanted the AI to be practical rather than imaginative when drafting a report, whereas developers may have designed it to be imaginative. In this case, the fault lies in the design stage. Otherwise, if both the worker and the developer preferred the AI to show the same trait (e.g., be practical rather than imaginative) and this resulted in an incident, then the incident could not be attributed to the developer. In this case, the cause is harder to trace.

Applying the grouping rule to group the tasks. Applying the previously defined rule resulted in two groups: 293 task occurrences where the developers may be responsible, and 85 task occurrences where the developers may not be responsible. In line with the previous section (§3.3), we performed close reading of the documented incidents and their news sources to better understand the context of the incidents.

4 Results

4.1 Prevalence and patterns of misaligned AI causing incidents in the workplace (RQ1)

To answer RQ1, we examined 214 workplace incidents and their corresponding 482 task occurrences, identifying those caused by misaligned AI (Step 3 in Figure 1).

Misaligned AI plays a major role in causing incidents in the workplace. RQ1 asked 'To what extent are workplace AI incidents caused by misaligned AI?' By misaligned AI, we mean an AI system that performed a task with traits that differed from the traits workers would have preferred for that system. Worryingly, we found an overwhelming majority of the incidents in the workplace were caused by misaligned AI: 83.6% (179 of the 214 workplace AI incidents). These range from 81% for data analysis to 97% of the incidents for the legal sector (Figure 2). The legal sector has the highest share (97%), followed by engineering (e.g., software) (90%), and the creative and educational sectors (both 89%).

In most cases, workers want systems that are precise, insightful, or personal, but instead receive systems that are basic, simple, or general. Globally, we found that the pair of misaligned AI traits that most frequently caused incidents involved the AI being *basic* but the worker needing it to be *precise* (across 58.8% of

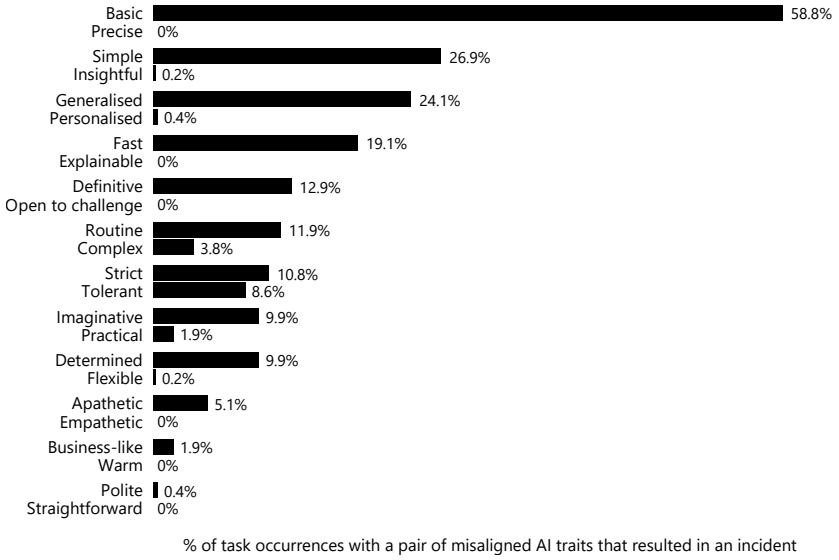

Fig. 3. **Percentage of task occurrences exhibiting a given misaligned AI trait, out of all task occurrences with a pair of misaligned AI traits that resulted in an incident.** A task occurrence is counted each time a task from our set of 93 tasks appears in one of the 214 incidents; since an incident may involve multiple tasks, the same task can be counted more than once. For each pair of AI traits, the two bars show the percentage of tasks where the AI displayed one trait (for example, basic or precise), while workers preferred the opposite trait, and that resulted in an incident. For 58.8% of task occurrences with AI misalignment, the AI provided *basic* responses rather than the *precise* answers workers expected. A task occurrence may involve AI misalignment on multiple pairs of traits. The traits linked to the most incidents were caused by AI systems that were too basic, too simple, too impersonal, too oversimplified, and too fast. Results are mostly asymmetric, except for the strict *vs.* tolerant pair.

task occurrences) (first row in Figure 3). This was followed by the AI being too simple, while the worker needed it to be insightful (26.9%), and the AI being general, and treating everyone similarly, but the worker needed it to be personalised and adjust based on the individual (24.1%).

The pairs of traits in misaligned AI that cause incidents tend to show an asymmetry (Figure 3). This is, given a pair of AI traits, incidents mostly occur when the AI shows one trait (such as basic) while workers consistently preferred the opposite (such as precise), but rarely the reverse. This does not mean the reverse misalignment cannot exist, but, when it does, it seems less likely to result in an incident. There are some exceptions for some pairs of traits, strict *vs.* tolerant (10.8% *vs.* 8.6%), routine *vs.* complex (11.9% *vs.* 3.8%), and, partially, imaginative *vs.* practical (9.9% *vs.* 1.9%).

Noticeably, we found some traits did not cause any incidents (Figure 3). No incidents involved an empathetic AI (0%), focused on addressing human needs and emotions, when the worker just wanted an AI doing data handling; neither a warm AI (0%), showing care, when the worker preferred it to remain business-like. Further, no incidents happened because the AI was explainable (0%), making decisions that are easy for people to understand, but the worker needed a fast AI, with automatic decisions without explanations. Finally, incidents did not happen because an AI that was too open to challenge, but the worker preferred a definitive AI. It is also remarkable that the pair of traits of AI being too polite, even if the system is not fully honest, or too straightforward, almost none caused any incident in any direction (0.4%/0%). Yet, incidents involving AI being too straightforward have been

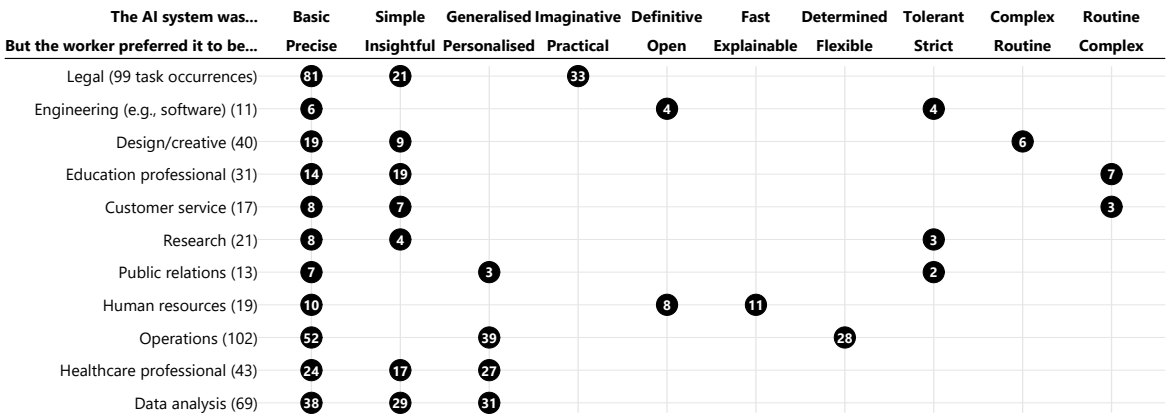

Fig. 4. **Top three most frequent reasons for misalignment in each sector.** By misaligned AI, we mean an AI system that performed a task with traits that differed from the traits workers would have preferred for that system (first two rows in the graph). The numbers in the black circles represent the number of the sector's task occurrences where the AI exhibited a given trait and the worker preferred the opposite trait, out of all the sector's task occurrences with AI misalignment (in parentheses). A task occurrence may involve AI misalignment on multiple pairs of traits. In the legal sector, imaginative AI was responsible for incidents totalling 33 task occurrences, while in human resources, fast AI accounted for 11.

documented outside the workplace, such as in mental health contexts, where overly direct AI responses have caused harm; we reflect on this in the discussion (§5).

We then honed in the most frequent reasons for misalignment in each sector, and we found key differences. We explored the top three pairs of misaligned traits across task occurrences that resulted in an incident (Figure 5). This analysis revealed that the sectors of healthcare professional and data analysis coincide with the global top misalignments of basic, simple, and generalised systems, but others do not. Sectors in which the top misalignments do not coincide with the global ones are the legal sector, where imaginative AI instead of being practical creatively making up references (33 task occurrences); the human resources sector with AI systems used to recruit being too fast and definitive, but not explainable (11); the educational sector, with a routine AI used with students, when it should be more complex (6) (Figure 4).

Finally, we examined if misalignment changed over time (Figure 6). Misalignment consistently dominates task occurrences across the observation period (2014-2025), suggesting misalignment with workers persists despite rapid AI advances. Yet, trait-level trends diverge beneath. Fast AI decreasingly causes incidents over time, possibly reflecting advances in explainable AI that have made ML systems more predictable (Appendix B). Conversely, imaginative AI increasingly causes incidents post-2022, likely driven by generative AI systems producing unreliable outputs. This suggests that as one form of misalignment declines, another emerges, which has kept the total rate consistently high.

Now we turn to the close reading of incidents with the most misaligned tasks (Figure 5), starting with the task 'Gather and analyse research data, such as decisions, articles and codes' in the legal sector, (third row)." The main trait misalignment happened when the lawyers needed a practical system (as indicated in our survey), but the AI system was imaginative (in 93% of the incidents with the task). This was followed by the case when the worker needed a precise system that in turn was basic (80%). Across incidents, the AI tool developed for the legal sector was making up fictitious references, and even when summarising the results it was not being precise enough. For example, a South African legal team used Legal Genius AI, an AI tool, and discovered this misalignment when

		AI MISALIGNMENT			
SECTOR	OCCUPATIONAL TASK	WORKER PREFERRED	THE AI SHOWED	% INCIDENTS WITH TRAIT MISALIGNMENT	INCIDENT IDs
Healthcare professional	Use data and social work to plan patient care and ensure efficacy (n=5 incidents, 28% of incidents in sector)	Personalised	Generalised	100%	110, 124, 603, 608, 699
		Insightful	Simple	100%	110, 124, 603, 608, 699
Research	Conduct internet-based and library research (n=6, 8%)	Precise	Basic	100%	470, 506, 614, 852, 1084, 1193
		Insightful	Simple	50%	470, 1084, 1193
Legal	Gather and analyze research data, such as decisions, articles and codes (n=15 incidents, 48%)	Practical	Imaginative	93%	541, 615, 623, 960, 1027, 1073, 1074, 1099, 1137...
		Precise	Basic	80%	541, 615, 623, 960, 1027, 1073, 1074, 1099, 1137...
Education professional	Observe and evaluate students' performance, behavior, social development, and physical health (n=9, 43%)	Tolerant	Strict	78%	96, 111, 116, 119, 192, 239, 355
		Personalised	Generalised	67%	96, 111, 116, 119, 192, 355, 386
Human resources	Analyze employment-related data and prepare required reports (n=10, 63%)	Explainable	Fast	70%	131, 301, 808, 1167, 1177, 1213, 1215
		Open	Definitive	70%	131, 301, 1167, 1177, 1213, 1215

Fig. 5. **The tasks with a higher fraction of incidents caused by misaligned AI.** For each task, we show the two most prevalent traits where an AI system performed a task with a trait that differed from the trait the workers would have preferred for that system. We report the percentage of incidents associated with a given task in which the AI showed a given trait and was misaligned, out of all incidents linked to the task (we provide incident IDs). An incident can involve the AI showing many misaligned traits for a given task. We limited the analysis to tasks across at least five incidents. In general, when AI systems are too basic, incidents may well occur.

the system generated non-existent case law in an urgent court filing, despite allegedly being trained on South African legal precedents (ID 1139) [21].

Another example comes from the human resources (HR) sector, involving the task 'Analyse employment-related data and prepare required reports', which occurs in 63% of the sector's incidents (n=10) (fourth row in Figure 5). In our study, HR workers indicated that, when AI augments this task, they need the system to be explainable and open. Yet in the incidents we analysed, AI systems were designed to be the opposite: fast (causing 70% of incidents with the task) and definitive instead (70%), misaligned with what workers needed to perform their task. Amazon's algorithmic management system for Flex drivers illustrates this misalignment can cause harm (ID 111). The system automated firing decisions based primarily on punctuality metrics, with algorithms that "scan the gusher of incoming data for performance patterns," while "human feedback is rare" [86]. Rather than supporting HR workers with an explainable system, this fast design model prioritized measurable metrics and simple rules for efficiency. As a result, the harm fell on the drivers that HR workers were supposed to oversee. The misalignment resulted in wrongful terminations, with drivers penalized for traffic, weather, or route complexity factors the system's definitive logic could not accommodate.

4.2 Developers' design decisions resulting in AI incidents in the workplace (RQ2)

To answer RQ2, we examined 179 incidents and 378 tasks occurrences, and identified whether the developers could be considered responsible for the misalignment, or whether it was due to other causes (Step 4 in Figure 1).

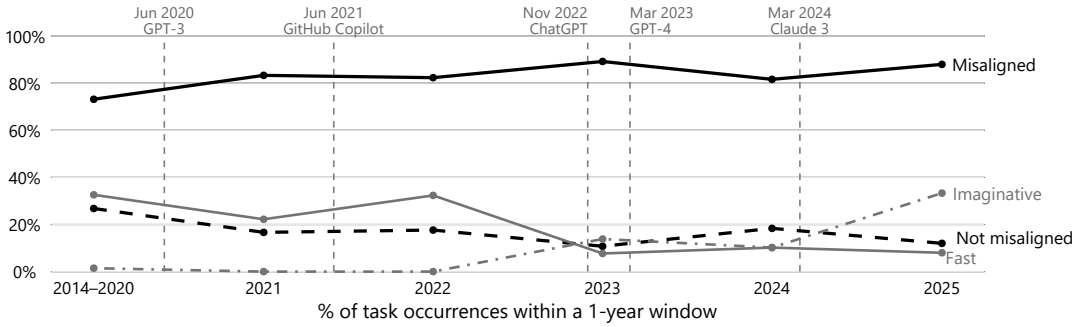

Fig. 6. **Percentage of task occurrences involving incidents of misaligned AI (solid line), or incidents of not misaligned AI (dashed line) over time.** Each data point represents the percentage of task occurrences classified as specified in the label within a year. We merged 2014-2020 as there were significantly fewer data points. Vertical lines mark milestones in AI research and deployment for context [63]. Misalignment consistently remained the dominant cause of incidents, although the frequency of misalignments involving specific traits increased or decreased over time: fast AI used to cause incidents, yet, since 2022, imaginative AI started to do so (Appendix C reports the results for all traits).

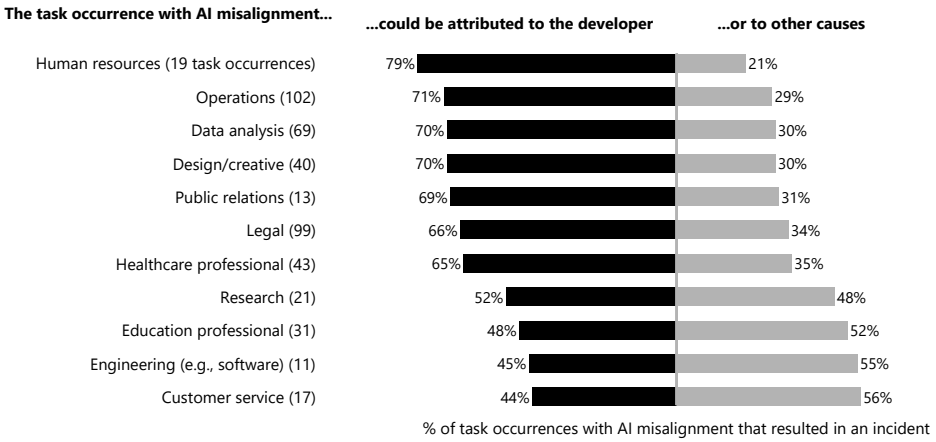

Fig. 7. **Percentage of task occurrences in incidents with AI misalignment attributable to developers (black bar) or to other causes (gray bar), out of all task occurrences with AI misalignment within each sector (total in parentheses).** We attribute the responsibility of a task occurrence with AI misalignment to the developer, if the developer designed the AI to perform a task with at least one trait that differed from the traits workers would have preferred for that system. Developers' design intentions may account for more than half of the task occurrences with AI misalignment that resulted in incidents (74%), mainly in working-facing sectors such as human resources. The reasons for misalignments are explored in Figure 8.

Developers' may be responsible for more than half of the tasks occurrences with misaligned AI, mainly in the human resources and public relations sectors. RQ2 asked 'When trait misalignment results in incidents, how often could it be attributed to developers *vs.* other causes?' We found 73.6% of the task occurrences with AI misalignment could be attributed to the developer. This is, when the developers design an AI system to perform a task with traits that differ from the traits that workers would have preferred for the system, and that

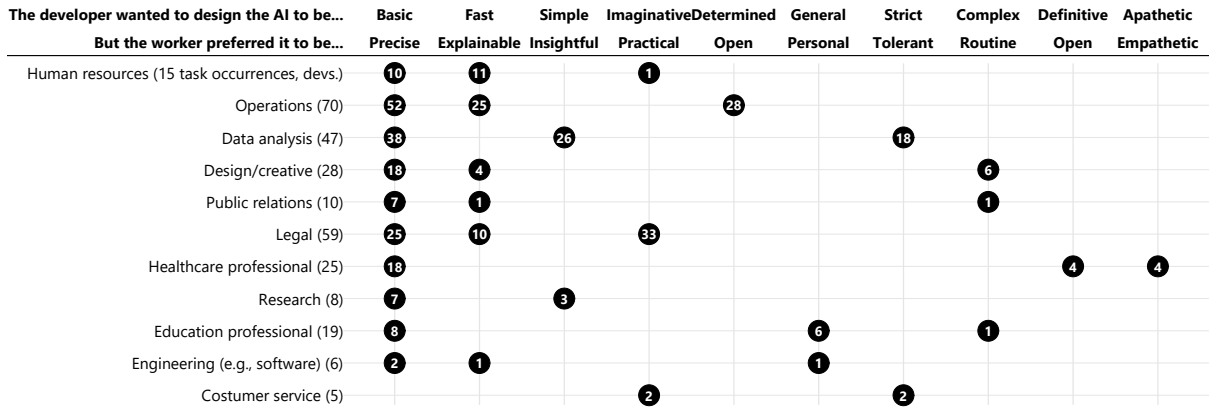

Fig. 8. **Top three most frequent misalignment attributable to developers in each sector.** We attribute the responsibility of a task occurrence with AI misalignment to the developer, when the developer designed the AI system to perform a task with traits that differed from the traits workers would have preferred for that system (first two rows in the figure). The numbers in the black circles represent the number of the sector's task occurrences where the developer designed for a given trait and the worker preferred the opposite trait, out of all the sector's task occurrences with AI misalignment that could be attributed to the developer (in parentheses). A task occurrence may involve AI misalignment on multiple pairs of traits. Developers are most misaligned with workers on traits that prioritize efficiency and speed (basic and fast).

resulted in incidents. The remaining 26.4% tasks involved traits where both developers and workers agreed on all the traits that the AI should have for a given task. In this case, the misaligned AI occurred due to other reasons, which also resulted in an incident.

Figure 7 shows the percentage of task occurrences with misaligned AI that could be attributed to developers or to other causes, out of all task occurrences with AI misalignment within in each sector. We find some sectors where misalignment at the design stage largely translates into workplace incidents, with human resources and operations accounting for 79% and 71% of the tasks occurrences with AI misalignment. In the sectors of costumer service and engineering, more incidents caused by misalignment were attributed to other causes than to developers, who accounted for 44% and 45% of the tasks occurrences. Notably, in some sectors the attribution of misalignment is more balanced between developers and other causes, like in research and education.

Developers differ most from workers on traits that prioritize efficiency and speed. We found the most misaligned traits across sectors are basic, fast, imaginative, determined, generalised, strict, tolerant, complex, and apathetic, but workers preferred precise, explainable, practical, open, personalised, tolerant, strict, routine, and empathetic AI (Figure 8). All those pairs of traits also appeared in the most frequent trait misalignment (Figure 4), but apathetic. These traits range from one sector, e.g., apathetic AI healthcare in, to ten sectors, e.g., basic AI.

A pattern appears: systems were designed for efficiency-driven traits (basic or fast) while workers needed traits for the contextual complexity and stakes of their tasks. One case previously discussed (§4.1) involves an HR task for the analysis of employment records, where developers rated fast as more important than workers, who preferred explainability. This misalignment resulted in incidents involving Amazon's Flex driver systems (ID 111, ID 116). The AI automatically processed workers' performance data and output termination decisions with minimal to no human oversight. Statements from developers explain misaligned design choices. For example, in ID 111, an engineer said "Inside Amazon, the Flex program is considered a great success, whose benefits far outweigh the collateral damage" [86] suggesting that speed and automation were prioritized over interpretability.

A similar pattern emerges in incidents involving AI agents supervising technical workers. In our study, engineers preferred open over definitive, precise over basic, and complex over simple. In ID 194, an AI system made field deployment decisions that could not be overridden, despite producing false positives. Engineers noted that "the automation team was unable to turn off the bot" in cases that "only a human operator could understand" [89]. These failures reflect design choices that prioritised definitive and simplified decision-making where workers needed agency and flexibility.

5 Discussion

In this section, we contextualise our findings in relation to prior literature (§ 5.1), their outline the implications of our research (§ 5.2), and list the main limitations, while offering directions for future work (§ 5.3).

5.1 Overview of our findings in relation to prior literature

Consistency with prior literature. We discuss two findings that are consistent with the literature. First, our findings empirically demonstrate that trait (mis)alignment is contextual and task-specific, rather than universal. While some of the AI traits affect all sectors similarly (like basic AI, simple AI, and general AI) other traits matter more for some sectors (Figure 4). This resonates with Suchman's concept of *situated action* [88], where an interaction is contingent on the context. For instance, we found that imaginative AI is problematic in the legal sector, and apathetic AI is so in the healthcare sector (Figure 8). A recent study found that people feared the same misaligned traits we identified such as excessive imagination when augmenting a lawyer's tasks, or not addressing human needs in medical tasks [25].

Second, by exploring where trait misalignment originates in the AI lifecycle, we demonstrate that many misalignments already exist in developer design. We found developers are most misaligned with the worker-facing sector of HR (Figure 8). Prior work shows developers often lack familiarity with the sectors they build for, resulting in a 'failure of imagination' [16, 41, 50]. Developer demographics partly explain this: in our study, they were mostly U.S.-based, computer science-trained, and male, often misaligned with responsible AI practices [45, 71]. Yet, our close reading revealed that developers' misalignment also reflected organizational priorities, such as speed when workers needed explainability (§4.2). Developers' decisions are not solely their own choices, but are shaped by the wider environment in which they work, prioritising employers over worker needs [37, 75, 95].

Novelty compared to prior literature. We discuss two findings that contribute to the literature. First, our analysis demonstrated that AI trait misalignment with worker needs plays a key role in workplace incidents. Worryingly, it accounts for 83.4% of the incidents in our dataset (Figure 2). Prior work had found negative effects of misaligned AI. For example, clinicians struggled to uptake LLMs to summarise their clinical notes, as the AI was too strict for their open-ended, 'messy' notes [53]. However, no study had shown misalignment leads to incidents in the workplace. We provide the first macro-level assessment of trait misalignment across 12 sectors, grounded in real-world incidents.

Second, we developed a novel framework to analyse workplace AI incidents through the lenses of HCI, enabling researchers to systematically link AI failures to design choices and prevent them at design stage [50]. This approach is well-established in safety-critical fields such as aviation and healthcare. Methods in aviation include gathering detailed incident reports, and testing how user interfaces behave under edge-case conditions [90].

5.2 Implications

Trait misalignment is a major risk factor, and should be incorporated into risk assessments. Responsible AI risk assessments should incorporate HCI indicators, like AI trait misalignment, to anticipate deployment contexts and worker needs, following HCI design guidelines accounting for user mental models [3, 68, 94]. For

example, while the EU AI Act mandates risk assessments only for high-risk systems [36], lower-risk workplace AI may still cause harm through trait misalignment [14], a concern the Act does not currently address.

Preventing misalignment requires structural interventions at the design stage. Developer teams should incorporate participatory approaches from the earliest stages of development. These approaches can draw from HCI and CSCW methodologies, including co-design workshops, and iterative feedback with workers [62, 79]. Yet, to be effective, interventions must be structural by altering the cultural context [13]. As such, interventions should address the drivers of developers' misalignment at three levels: micro, meso, and macro. At the micro level, they should tackle developers' assumptions stemming from their technical training, and intersectional experience [12, 24, 35, 71]. At the meso level, they should address the companies where developers work, and how they envision innovation and for *whom* (e.g., do edtech startups genuinely centre school teachers?) [41, 49, 92]. At the macro level, they should challenge how geopolitical or economic pressures push agendas when discussing worker needs (e.g., framing automation as a national imperative for global competition) [27, 28]. Without meaningful structural interventions, we risk *participation washing* [85].

There are additional sources of failure that the literature should still study. We observed cases where developers were aligned, yet AI failed to exhibit the traits required to perform the task. AI capabilities may have limitations for alignment [42, 51]. For example, research has shown how trying to personalise AI for fair assessments may not work in the real-world, e.g., *fairML* [38, 48]. Despite the AGI hype on AI capabilities, future work should explore its limitations.

5.3 Limitations

Our study has four main limitations. First, our incident data from the AIID mostly capture harms that are newsworthy, and US-focused [23, 78], potentially missing everyday worker experiences. For example, polite AI rarely caused incidents (0.4%) (Figure 3). Though this may seem positive, caution is due. Sycophantic AI, too agreeable at the expense of accuracy [84], causes severe harms in mental health [20]. Our finding could be explained for two reasons. First, our data is not exhaustive, and polite AI incidents may not be newsworthy. Second, these failed interactions take time to surface. Future work should explore the implications of polite AI at work, with longitudinal studies or ethnographies.

Second, our task-level approach does not fully operationalise a job or an AI product [6, 91]. Future work could consider additional dimensions of jobs, such as informal activities, contextual organisational fit, workers' skills, and AI literacy; and additional dimensions of AI products, such as AI type, and company [6].

Third, our counterfactual approach (traits are causal if their absence would not result in incidents) cannot definitively establish causation as we would in controlled experiments. While we rely on analyst judgment informed by incident descriptions, validated with high inter-rater reliability (Cohen's Kappa=0.89), further work could explore other contextual causes of the incident.

Our last limitation concerns the user study compatibility. Our incident analysis yields classifications on whether the trait caused incident (yes/no), while the user study we draw upon measured misalignment on a scale. We bridged this through an established threshold of 0.5 Likert point difference between developers and workers [76].

6 Conclusion

We systematically analyzed 214 workplace AI incidents through 12 widely-used pairs of psychological traits to assess AI misalignment. Worryingly, we found an overwhelming majority of incidents (83%) involve trait misalignment. We also found that 74% of the misaligned tasks already existed in the developers design choices. Our findings show the importance of accounting for worker needs from the earliest stages of design. As Suchman noted, "too often, assumptions are made as to how tasks are performed rather than unearthing the underlying work practices" [87]. By learning from incidents, developers can create AI systems aligned with worker needs.

7 Endmatter statements

7.1 Generative AI Usage Statement

In this paper we made use of generative AI for three tasks, data analysis, code assistant, and editing, which we explain in detail next.

(1) **Data analysis.** We used Open AI GPT-5.1 to assist us with data classification. This was used to identify workplace incidents, extract their job tasks, and analyse the misaligned AI traits that contributed to the incident. We explain all the steps in the Research Design section (§3).
(2) **Code assistant.** We used the free version of ChatGPT by OpenAI and Claude by Anthropic to assist with coding for creating the plots and the Overleaf tables. We followed similar practices than those from Stack Overflow, by looking for help when necessary, rather than generating all the content.
(3) **Editing.** Finally, we used again the free versions of ChatGPT and Claude to assist with text editing. This was minimally used in a way in which the full text was not rewritten, but rather we asked for recommendations on *which* words to cut down, and which sentences might come across as repetitive.

7.2 Ethical Considerations Statement

This research studied AI systems in the workplace and the misalignment between AI system traits and worker needs. All the data we used was drawn from publicly available sources, including the AI Incident Database [64] and prior user studies [76], and no personally identifiable information was collected or analyzed. Our analysis emphasizes ethical AI design by highlighting how developers' design choices can, even if not intentionally, contribute to workplace harms, with the aim of supporting safer and more worker-centered AI systems.

References

[1] Daron Acemoglu, David Autor, and Simon Johnson. 2023. Can we have pro-worker AI. *Choosing a path* (2023).
[2] AI, Algorithmic, and Automation Incidents and Controversies (AIAAIC). 2026. AIAAIC Repository of AI, Algorithmic, and Automation Incidents and Controversies. https://www.aiaaic.org/. Accessed: 2026-01-06.
[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, Glasgow Scotland Uk, 1–13. doi:10.1145/3290605.3300233
[4] Elske Ammenwerth, Carola Iller, and Cornelia Mahler. 2006. IT-adoption and the interaction of task, technology and individuals: a fit framework and a case study. *BMC Medical Informatics and Decision Making* 6, 1 (Jan. 2006), 3. doi:10.1186/1472-6947-6-3
[5] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861* (2021).
[6] David H. Autor. 2013. The "task approach" to labor markets: an overview. *Journal for Labour Market Research* 46, 3 (Sept. 2013), 185–199. doi:10.1007/s12651-013-0128-z
[7] Ezra Awumey, Sauvik Das, and Jodi Forlizzi. 2024. A systematic review of biometric monitoring in the workplace: analyzing socio-technical harms in development, deployment and use. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*. 920–932.
[8] Ezra Awumey, Sauvik Das, and Jodi Forlizzi. 2024. A Systematic Review of Biometric Monitoring in the Workplace: Analyzing Socio-technical Harms in Development, Deployment and Use. In *The 2024 ACM Conference on Fairness Accountability and Transparency*. ACM, Rio de Janeiro Brazil, 920–932. doi:10.1145/3630106.3658945
[9] Jascha Bareis and Christian Katzenbach. 2022. Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values* 47, 5 (2022), 855–881.
[10] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 610–623. doi:10.1145/3442188.3445922
[11] Federico Bianchi, Amanda Cercas Curry, and Dirk Hovy. 2023. Artificial intelligence accidents waiting to happen? *Journal of Artificial Intelligence Research* 76 (2023), 193–199.
[12] Abeba Birhane, Elayne Ruane, Thomas Laurent, Matthew S. Brown, Johnathan Flowers, Anthony Ventresque, and Christopher L. Dancy. 2022. The forgotten margins of AI ethics. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*. 948–958.

[13] Kim M Blankenship, Samuel R Friedman, Shari Dworkin, and Joanne E Mantell. 2006. Structural interventions: concepts, challenges and opportunities for research. *Journal of Urban Health* 83, 1 (2006), 59–72.

[14] Edyta Bogucka, Sanja Šćepanović, and Daniele Quercia. 2024. Atlas of AI Risks: Enhancing Public Understanding of AI Risks. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 12 (Oct. 2024), 33–43. doi:10.1609/hcomp.v12i1.31598

[15] Edyta Paulina Bogucka, Marios Constantinides, Julia De Miguel Velazquez, Sanja Scepanovic, Daniele Quercia, and Andrés Gvirtz. 2024. The Atlas of AI Incidents in Mobile Computing: Visualizing the Risks and Benefits of AI Gone Mobile. In *Adjunct Proceedings of the 26th International Conference on Mobile Human-Computer Interaction* (Melbourne, VIC, Australia) *(MobileHCI '24 Adjunct)*. Association for Computing Machinery, New York, NY, USA, Article 26, 6 pages. doi:10.1145/3640471.3680447

[16] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming failures of imagination in AI infused system development and deployment. *arXiv preprint arXiv:2011.13416* (2020).

[17] Michelle Brachman, Amina El-Ashry, Casey Dugan, and Werner Geyer. 2025. Current and future use of large language models for knowledge work. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–24.

[18] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[19] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2025. Generative AI at work. *The Quarterly Journal of Economics* 140, 2 (2025), 889–942.

[20] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. Social sycophancy: A broader understanding of llm sycophancy. *arXiv preprint arXiv:2505.13995* (2025).

[21] Cliffe Dekker Hofmeyr. 2025. Another episode of fabricated citations, real repercussions: South African courts show no tolerance for AI-hallucinated cases. https://www.cliffedekkerhofmeyr.com/en/news/publications/2025/Practice/Employment-Law/combined-employment-and-knowledge-management-alert-4-july-Another-episode-of-fabricated-citations-real-repercussions-South-African-courts-show-no-tolerance-for-AI-hallucinated-cases. Accessed: 2026-03-19.

[22] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[23] Julia De Miguel Velázquez, Sanja Šćepanović, Andrés Gvirtz, and Daniele Quercia. 2024. Decoding Real-World Artificial Intelligence Incidents. *Computer* 57, 11 (2024), 71–81. doi:10.1109/MC.2024.3432492

[24] Catherine D'ignazio and Lauren F Klein. 2023. *Data feminism.* MIT press.

[25] Mengchen Dong, Jane Rebecca Conway, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2024. Fears about artificial intelligence across 20 countries and six domains of application. *American Psychologist* (2024).

[26] Mengchen Dong, Jane Rebecca Conway, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2024. Fears about artificial intelligence across 20 countries and six domains of application. *American Psychologist* (2024). doi:10.1037/amp0001454 Place: US Publisher: American Psychological Association.

[27] Madeleine Clare Elish. 2019. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society* 5 (March 2019), 40–60. doi:10.17351/ests2019.260

[28] Madeleine Clare Elish and Danah Boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication monographs* 85, 1 (2018), 57–80.

[29] Michael Feffer, Nikolas Martelaro, and Hoda Heidari. 2023. The AI Incident Database as an Educational Tool to Raise Awareness of AI Harms: A Classroom Exploration of Efficacy, Limitations, & Future Improvements. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3617694.3623223

[30] Diana E Forsythe. 1993. Engineering knowledge: The construction of knowledge in artificial intelligence. *Social studies of science* 23, 3 (1993), 445–477.

[31] Sarah E. Fox, Vera Khovanskaya, Clara Crivellaro, Niloufar Salehi, Lynn Dombrowski, Chinmay Kulkarni, Lilly Irani, and Jodi Forlizzi. 2020. Worker-Centered Design: Expanding HCI Methods for Supporting Labor. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3334480.3375157

[32] Sarah E. Fox, Samantha Shorey, Esther Y. Kang, Dominique Montiel Valle, and Estefania Rodriguez. 2023. Patchwork: The Hidden, Human Labor of AI Integration within Essential Work. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1 (April 2023), 81:1–81:20. doi:10.1145/3579514

[33] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, Seliem El-Sayed, Sasha Brown, Canfer Akbulut, Andrew Trask, Edward Hughes, A. Stevie Bergman, Renee Shelby, Nahema Marchal, Conor Griffin, Juan Mateos-Garcia, Laura Weidinger, Winnie Street, Benjamin Lange, Alex Ingerman, Alison Lentz, Reed Enger, Andrew Barakat, Victoria Krakovna, John Oliver Siy, Zeb Kurth-Nelson, Amanda McCroskery, Vijay Bolina, Harry Law, Murray Shanahan, Lize Alberts, Borja Balle, Sarah de Haas, Yetunde Ibitoye, Allan Dafoe, Beth Goldberg, Sébastien Krier, Alexander Reese, Sims Witherspoon, Will Hawkins, Maribeth Rauh, Don Wallace, Matija Franklin, Josh A. Goldstein, Joel Lehman, Michael Klenk, Shannon Vallor, Courtney Biles, Meredith Ringel Morris, Helen King, Blaise Agüera y Arcas, William Isaac, and James Manyika. 2024.

The Ethics of Advanced AI Assistants. doi:10.48550/arXiv.2404.16244 arXiv:2404.16244 [cs].

[34] Anna Gausen, Bhaskar Mitra, and Siân Lindley. 2024. A Framework for Exploring the Consequences of AI-Mediated Enterprise Knowledge Access and Identifying Risks to Workers. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Rio de Janeiro Brazil, 207–220. doi:10.1145/3630106.3658900

[35] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (Dec. 2021), 86–92. doi:10.1145/3458723

[36] Delaram Golpayegani, Harshvardhan J Pandit, and Dave Lewis. 2023. To be high-risk, or not to be—semantic specifications and implications of the ai act's high-risk ai applications and harmonised standards. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 905–915.

[37] Sinem Görücü, Yuheng Ren, Gabrielle Samuel, and Georgia Panagiotidou. 2025. " As an individual, I suppose you can't really do much": Environmental Sustainability Perceptions of Machine Learning Practitioners. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 1312–1324.

[38] Eileen Guo, Jeroen van Raalte, Justin-Casimir Braun, Gabriel Geiger, Amanda Silverman, Eva Constantaras, Melissa Heikkilä, Tahmeed Shafiq, Alice Milliken, Crofton Black, and Daniel Howden. 2025. The Limits of Ethical AI. https://www.lighthousereports.com/investigation/the-limits-of-ethical-ai/. Accessed: 2026-01-14.

[39] Sacha Gutierrez, Dennis Nguyen, and Karin van Es. 2025. Tool, companion or a catalyst force? Exploring sociotechnical imaginaries Within AI livestreams' communities of practice. *Big Data & Society* 12, 4 (Dec. 2025), 20539517251381663. doi:10.1177/20539517251381663 Publisher: SAGE Publications Ltd.

[40] Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, et al. 2025. Which Economic Tasks are Performed with AI. *Evidence from Millions of Claude Conversations* (2025).

[41] Emma Harvey, Allison Koenecke, and Rene F. Kizilcec. 2025. "Don't Forget the Teachers": Towards an Educator-Centered Understanding of Harms from Large Language Models in Education. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–19. doi:10.1145/3706598.3713210

[42] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. 2019. Improving fairness in machine learning systems: What do industry practitioners need?. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–16.

[43] Alexis Shore Ingber and Nazanin Andalibi. 2025. Emotion AI in Job Interviews: Injustice, Emotional Labor, Identity, and Privacy. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens Greece, 1–17. doi:10.1145/3715275.3732002

[44] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational Psychiatry* 11, 1 (2021), 108. doi:10.1038/s41398-021-01224-x

[45] Maurice Jakesch, Zana Buçinca, Saleema Amershi, and Alexandra Olteanu. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 310–323. doi:10.1145/3531146.3533097

[46] Stefan Jänicke, Greta Franzini, Muhammad Faisal Cheema, Gerik Scheuermann, et al. 2015. On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. *EuroVis (STARs)* 2015 (2015), 83–103.

[47] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Borong Zhang, Donghai Hong, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Lukas Vierling, Zhaowei Zhang, Fanzhi Zeng, Juntao Dai, Xuehai Pan, Hua Xu, Aidan O'Gara, Kwan Ng, Brian Tse, Jie Fu, Stephen Mcaleer, Yanfeng Wang, Mingchuan Yang, Yunhuai Liu, Yizhou Wang, Song-Chun Zhu, Yike Guo, Yaodong Yang, and Wen Gao. 2025. AI Alignment: A Contemporary Survey. *ACM Comput. Surv.* 58, 5, Article 132 (Nov. 2025), 38 pages. doi:10.1145/3770749

[48] Mackenzie Jorgensen, Hannah Richert, Elizabeth Black, Natalia Criado, and Jose Such. 2023. Not so fair: The impact of presumably fair machine learning models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 297–311.

[49] Nadia Karizat, Alexandra H Vinson, Shobita Parthasarathy, and Nazanin Andalibi. 2024. Patent applications as glimpses into the sociotechnical imaginary: ethical speculation on the imagined futures of emotion AI for mental health monitoring and detection. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–43.

[50] Anna Kawakami, Jordan Taylor, Sarah Fox, Haiyi Zhu, and Kenneth Holstein. 2026. AI failure loops in devalued work: The confluence of overconfidence in AI and underconfidence in worker expertise. *Big Data & Society* 13, 1 (2026), 20539517261424164.

[51] Os Keyes, Jevan Hutson, and Meredith Durbin. 2019. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3290607.3310433

[52] Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. 2025. Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications* 12, 1 (2025), 1–9.

[53] Kristina L. Kupferschmidt, Kieran O'Doherty, and Joshua A. Skorburg. 2025. Write on Paper, Wrong in Practice: Why LLMs Still Struggle with Writing Clinical Notes. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 2 (Oct. 2025), 1524–1534.

doi:10.1609/aies.v8i2.36651
[54] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.
[55] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. 2025. The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*. 1–22.
[56] Hao-Ping Lee, Yu-Ju Yang, Thomas Serban Von Davier, Jodi Forlizzi, and Sauvik Das. 2024. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.
[57] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.
[58] Nancy G. Leveson. 2011. *Engineering a safer world: systems thinking applied to safety.* MIT press, Cambridge (Mass.).
[59] Megan Li, Wendy Bickersteth, Ningjing Tang, Lorrie Cranor, Jason Hong, Hong Shen, and Hoda Heidari. 2025. A Closer Look at the Existing Risks of Generative AI: Mapping the Who, What, and How of Real-World Incidents. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 8, 2 (Oct. 2025), 1561–1573. doi:10.1609/aies.v8i2.36655
[60] Isabella Loaiza and Roberto Rigobon. 2024. The EPOCH of AI: Human-Machine Complementarities at Work. doi:10.2139/ssrn.5028371
[61] Jonathan Lynn, Rachel Y. Kim, Sicun Gao, Daniel Schneider, Sachin S. Pandya, and Min Kyung Lee. 2025. Regulating Algorithmic Management: A Multi-Stakeholder Study of Challenges in Aligning Software and the Law for Workplace Scheduling. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Athens Greece, 547–572. doi:10.1145/3715275.3732037
[62] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376445
[63] Nestor Maslej, Loredana Fattorini, Raymond Perrault, Yolanda Gil, Vanessa Parli, Njenga Kariuki, Emily Capstick, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, Toby Walsh, Armin Hamrah, Lapo Santarlasci, Julia Betts Lotufo, Alexandra Rome, Andrew Shi, and Sukrut Oak. 2025. The AI Index 2025 Annual Report. doi:10.48550/arXiv.2504.07139
[64] Sean McGregor. 2021. Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 17 (May 2021), 15458–15463. doi:10.1609/aaai.v35i17.17817 Number: 17.
[65] Kevin R McKee, Xuechunzi Bai, and Susan T Fiske. 2024. Warmth and competence in human-agent cooperation. *Autonomous Agents and Multi-Agent Systems* 38, 1 (2024), 23.
[66] National Center for O*NET Development. 2026. O*NET Database. https://www.onetcenter.org/database.html. Accessed: 2026-03-24.
[67] Nataliya Nedzhvetskaya and JS Tan. 2024. No Simple Fix: How AI Harms Reflect Power and Jurisdiction in the Workplace. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 422–432. doi:10.1145/3630106.3658915
[68] Don Norman. 2013. *The design of everyday things: Revised and expanded edition.* Basic books.
[69] OECD. 2025. *Towards a common reporting framework for AI incidents.* OECD Artificial Intelligence Papers. doi:10.1787/f326d4ac-en Edition: 34 Series: OECD Artificial Intelligence Papers.
[70] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLOS ONE* 15, 2 (2020), e0229132.
[71] Lauren Olson, Ricarda Anna-Lena Fischer, Florian Kunneman, and Emitzá Guzmán. 2025. Who Speaks for Ethics? How Demographics Shape Ethical Advocacy in Software Development. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 2847–2862.
[72] OpenAI. 2023. GPT-5: Large Language Model. https://openai.com/research/gpt-5. Accessed: 2026-03-24.
[73] Charles Perrow. 1984. *Normal Accidents: Living with High-Risk Technologies.* Basic Books.
[74] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.
[75] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.
[76] Jaspreet Ranjit, Ke Zhou, Swabha Swayamdipta, and Daniele Quercia. 2026. Are We Automating the Joy Out of Work? Designing AI to Augment Work, Not Meaning. (2026).
[77] Jens Rasmussen. 1997. Risk management in a dynamic society: A modelling problem. *Safety Science* 27, 2-3 (1997), 183–213.
[78] Isabel Richards, Claire Benn, and Miri Zilka. 2025. From Incidents to Insights: Patterns of Responsibility following AI Harms. In *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 151–169.
[79] Shadan Sadeghian, Migle Bareikyte, Marcus Burkhardt, and Marc Hassenzahl. 2025. WorkAI: A Toolkit for the Design of AI-driven Future of Work. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–27.

[80] Nadine B Sarter, David D Woods, and Charles E Billings. 1997. Automation surprises. *Handbook of Human Factors and Ergonomics* 2 (1997), 1926–1943.

[81] Kristen M. Scott, Sonja Mei Wang, Milagros Miceli, Pieter Delobelle, Karolina Sztandar-Sztanderska, and Bettina Berendt. 2022. Algorithmic Tools in Public Employment Services: Towards a Jobseeker-Centric Perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2138–2148. doi:10.1145/3531146.3534631

[82] Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 59–68. doi:10.1145/3287560.3287598

[83] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. 2025. Future of Work with AI Agents: Auditing Automation and Augmentation Potential across the U.S. Workforce. doi:10.48550/arXiv.2506.06576 arXiv:2506.06576 [cs].

[84] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548* (2023).

[85] Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–6.

[86] Spencer Soper. 2021. *Fired by Bot at Amazon: It's You Against the Machine.* https://www.bloomberg.com/news/features/2021-06-28/fired-by-bot-amazon-turns-to-machine-managers-and-workers-are-losing-out Bloomberg.

[87] Lucy Suchman. 1995. Making work visible. *Commun. ACM* 38, 9 (Sept. 1995), 56–64. doi:10.1145/223248.223263

[88] Lucy Suchman. 2007. *Human-machine reconfigurations: Plans and situated actions*. Cambridge university press.

[89] Anand Tamboli. 2020. *A Lesson Worth $11 Million.* https://medium.com/tomorrow-plus-plus/a-lesson-worth-11-million-7851be19921f Medium, tomorrow++.

[90] Violet Turri and Rachel Dzombak. 2023. Why We Need to Know More: Exploring the State of AI Incident Documentation Practices. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society (AIES '23)*. Association for Computing Machinery, New York, NY, USA, 576–583. doi:10.1145/3600211.3604700

[91] Judy Wajcman and Emily Rose. 2011. Constant connectivity: Rethinking interruptions at work. *Organization studies* 32, 7 (2011), 941–961.

[92] J Wajcman, E Young, D Kampmann, and J De Miguel Velazquez. 2024. Rebalancing Innovation: Women, AI and Venture Capital in the UK. (2024).

[93] Sonja Mei Wang, Kristen M Scott, Margarita Artemenko, Milagros Miceli, and Bettina Berendt. 2023. "We try to empower them" - Exploring Future Technologies to Support Migrant Jobseekers. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 972–983. doi:10.1145/3593013.3594056

[94] Justin D. Weisz, Jessica He, Michael Muller, Gabriela Hoefer, Rachel Miles, and Werner Geyer. 2024. Design Principles for Generative AI Applications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–22. doi:10.1145/3613904.3642466

[95] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 467–479.

[96] David D. Woods, Sidney Dekker, Richard Cook, Leila Johannesen, and Nadine Sarter. 2010. *Behind Human Error* (2nd ed.). Ashgate, Farnham, UK. First edition 2006.

[97] David D Woods and Nadine B Sarter. 2000. Learning from automation surprises and "going sour" accidents. In *Cognitive Engineering in the Aviation Domain*. Lawrence Erlbaum, 327–353.

A Appendix

A.1 LLM prompt to classify workplace-related incidents

Prompt

System: You are an experienced researcher studying how AI systems augment job tasks in the workplace.

User: You will see an incident ID, TITLE, and DESCRIPTION.

Return True **only if** the incident describes a worker (contracted, self-employed, gig-worker), or an organisation (profit, non-profit, government, gig-platform) using AI **to augment, automate or manage a job task**. A task is a unit of work activity that produces a meaningful output. The physical location does not matter: remote, field, factory, and office-based work all count as workplace contexts.

Else, return False.

Exclude:
- Incidents where AI affects only end users, customers, or the public, without showing impact on workers performing job tasks.
- Social media, platform, or recommendation system errors unless they clearly affect workers performing their jobs.
- AI used for illegal, adversarial, or malicious purposes (e.g., fraud, blackmail, jailbreaking for illicit content).

Output a JSON object with a key of incident ID and its value, and a key of workplace and its value of True or False, e.g.:

{{
"incident_id": "18",
"workplace": "False"
}}
-

The ID of the incident: `{incident_id}`
The title of the incident: `{title}`
The description of the incident: `{description}`

Whether the incident occurred in the workplace:

A.2 LLM prompt to extract job tasks

Prompt

System: You are an experienced researcher studying how artificial intelligence tools augment human work tasks across different sectors.

User: You will receive a description of an AI incident and a list of job tasks (with IDs) in the industry sector of {{sector}}.

First, determine whether the incident involves an AI system augmenting/automating workplace activities within a position in {{sector}} or within the sector of {{sector}}. If yes, proceed to rate each task below. If no, assign a rating of 0 to all tasks.

For each task in the {{list_of_tasks}}, imagine an AI system augmenting or automating this specific job task. Answer to which extent this incident description mentions such an AI system by rating it as below.

Ratings:
3 The incident mentions the AI system augmenting the task and its application in the task was the main source of the incident.
2 The incident mentions the AI system augmenting the task and its application in the task was a partial source in the incident.
1 The incident mentions the AI system augmenting the task but it is unclear whether the its application in task might have caused the incident.
0 The incident does not mention the AI system augmenting the task.

Your output format should be:
A JSON object where each key is the task ID and the value is an object with the task description (as a string) and an object with "rating", e.g.

```
{{
“1” {“task_description” “Customer data analysis”, “rating” 0},
“2” {“task_description” “Interpret data and regulations”, “rating” 1},
“3” {“task_description” “Gather and analyse data", “rating” 3}
}}
-
```

Incident description: {{incident_description}}
Incident sources: {{incident_sources}}
List of job tasks with IDs, e.g., 1. Consumer data analysis: {{list_of_tasks}}

Question: On a scale from 0 (incident does not mention AI system) to 3 (incident mentions and involves AI system augmenting the task), rate the extent to which this incident involves AI augmenting/automating each specific task in {{sector}}.

A.3 LLM prompt to extract misaligned psychological traits

Part 1/2

System: You are an experienced researcher studying how AI systems augment job tasks in the workplace.

User: Your task:
You are given a workplace AI incident. The incident involves a worker using AI to augment a job task.
Your goal is to determine whether the incident involved TRAIT MISALIGNMENT: i.e., the AI system behaved differently from what the worker needed.
For EACH label below, output "Yes" or "No".
-
TRAIT MISALIGNMENT LABELS
1. Complex vs. routine work
- complex_sufficient: AI handled complex work, but worker needed routine work.
- complex_necessary: Would the incident still occur if AI were NOT handling complex work?
- routine_sufficient: AI handled routine work, but worker needed complex work.
- routine_necessary: Would the incident still occur if AI were NOT handling routine work?

2. Emotions vs. apathetic
- emotion_sufficient: AI focused on addressing human needs and emotions, but worker just needed data handling.
- emotion_necessary: Would the incident still occur if AI were NOT focusing on addressing emotions?
- apathetic_sufficient: AI just focused on data handling, but worker needed it to address human needs/emotions.
- apathetic_necessary: Would the incident still occur if AI were NOT just focusing on data handling?

3. Fast vs. explainable
- fast_sufficient: AI made fast, automatic decisions without explanation, but worker needed decisions that are easy to understand.
- fast_necessary: Would the incident still occur if AI were NOT making fast, automatic decisions without explanation?
- explainable_sufficient: AI was making decisions that are easy for people to understand, but worker needed fast, automatic decisions without explanations.
- explainable_necessary: Would the incident still occur if AI were NOT making decisions that are easy for people to understand?

4. Open to challenge vs. definitive
- open_sufficient: AI was open to challenge, but worker needed a final decision.
- open_necessary: Would the incident still occur if AI were NOT open to challenge?
- definitive_sufficient: AI treated decisions as final, but worker needed openness to challenge.
- definitive_necessary: Would the incident still occur if AI were NOT treating decisions as final?

5. Personalised vs. generalised
- personalised_sufficient: AI adjusted based on the individual it was helping, but the worker needed it to treat everyone similarly.
- personalised_necessary: Would the incident still occur if AI were NOT adjusting based on the individual it was helping?
- generalised_sufficient: AI treated everyone similarly, but the worker needed it to adjust based on the individual.
- generalised_necessary: Would the incident still occur if AI were NOT treating everyone too similarly?

6. Warm vs. business-like
- warm_sufficient: AI showed warmth and care, but the worker needed it to remain neutral and business-like.
- warm_necessary: Would the incident still occur if AI were NOT showing warmth and care?
- businesslike_sufficient: AI remained neutral and business-like, but the worker needed it to show warmth and care.
- businesslike_necessary: Would the incident still occur if AI were NOT remaining neutral and business-like?

7. Polite vs. straightforward
- polite_sufficient: AI was polite even if that meant not being fully honest, but the worker needed it to be sincere and straightforward.
- polite_necessary: Would the incident still occur if AI were NOT being polite even if that meant not being fully honest?
- straightforward_sufficient: AI was sincere and straightforward, but the worker needed it to be polite even if that meant not being fully honest.
- straightforward_necessary: Would the incident still occur if AI were NOT being sincere and straightforward?

8. Tolerant/open-minded vs. strict
- tolerant_sufficient: AI was tolerant and open-minded, but the worker needed it to be strict and follow rules exactly.
- tolerant_necessary: Would the incident still occur if AI were NOT being tolerant and open-minded?
- strict_sufficient: AI strictly followed rules, but the worker needed it to be tolerant and open-minded.
- strict_necessary: Would the incident still occur if AI were NOT strictly following rules?

Part 2/2

9. Precise vs. basic
- precise_sufficient: AI was highly skilled and precise, but the worker needed it to be fast and simple even if less perfect.
- precise_necessary: Would the incident still occur if AI were NOT being highly skilled and precise?
- basic_sufficient: AI was fast and simple even if less perfect, but the worker needed it to be highly skilled and precise.
- basic_necessary: Would the incident still occur if AI were NOT being fast and simple even if less perfect?

10. Flexible vs. determined
- flexible_sufficient: AI was flexible and willing to change course, but the worker needed it to be determined and persistent.
- flexible_necessary: Would the incident still occur if AI were NOT being flexible and willing to change course?
- determined_sufficient: AI was determined and persistent, but the worker needed it to be flexible and willing to change course.
- determined_necessary: Would the incident still occur if AI were NOT being determined and persistent?

11. Insightful/comprehensive vs. simple
- insightful_sufficient: AI showed comprehensiveness, deep understanding, and insight, but the worker needed it to keep things simple and straightforward.
- insightful_necessary: Would the incident still occur if AI were NOT showing comprehensiveness, deep understanding, and insight?
- simple_sufficient: AI kept things simple and straightforward, but the worker needed it to show comprehensiveness, deep understanding, and insight.
- simple_necessary: Would the incident still occur if AI were NOT keeping things simple and straightforward?

12. Practical vs. imaginative
- practical_sufficient: AI stayed practical and followed familiar approaches, but the worker needed it to be imaginative and bring new ideas.
- practical_necessary: Would the incident still occur if AI were NOT staying practical and following familiar approaches?
- imaginative_sufficient: AI was imaginative and brought new ideas, but the worker needed it to stay practical and follow familiar approaches.
- imaginative_necessary: Would the incident still occur if AI were NOT being imaginative and bringing new ideas?

—
OUTPUT FORMAT:
{{
"complex_sufficient": "Yes|No",
"complex_necessary": "Yes|No",
"routine_sufficient": "Yes|No",
...
"practical_necessary": "Yes|No",
"imaginative_sufficient": "Yes|No",
"imaginative_necessary": "Yes|No"
}}
—
INCIDENT DETAILS

INCIDENT_ID: `{incident_id}`
TASK: `{task_description}`
TITLE: `{title}`
DESCRIPTION: `{description}`
SOURCES: `{sources}`

Check your planned output before outputting it: if it contains any explanations besides the JSON string, omit the explanations. Make sure to output ONLY a correctly formatted JSON string and nothing else. Do not miss any of the traits.

B Survey details

Table 1. Characteristics of participants recruited from Prolific, recruited from [76]. The workers were highly familiar with the tasks in their corresponding sector. Developers work in software, data, IT, and ML/AI roles who actively engage with modern AI tools and contribute to AI enabled workflows across diverse organizational sectors.

Demographics	**Workers (N=202)**	**Developers (N=197)**
Mean age (SD)	42.63 (13.05)	36.68 (10.29)
Gender		
Male	34.03%	64.46%
Female	61.94%	28.42%
Non-binary / Other	0%	0%
Consent Revoked	4.03%	7.11%
Employment status		
Full-time	46.94%	69.54%
Part-time	16.29%	13.71%
Unemployed / Other	4.03%	1.52%
Consent Revoked / No data available	15.22%	23.80%

Table 2. Survey items on preference of AI traits

#	**Item**
1	Handle more complex work rather than routine work.
2	Focus more on addressing human needs and emotions rather than just data handling.
3	Make fast, automatic decisions without explanation rather than decisions that are easy for people to understand.
4	Be open to challenge or treat the decision as final.
5	Adjust based on the individual it's helping rather than treat everyone the same.
6	Show warmth and care rather than remain neutral and business-like.
7	Be polite even if that means not being fully honest, rather than being sincere and straightforward.
8	Be strict and follow the rules exactly rather than be tolerant and open-minded.
9	Be fast and simple even if less perfect, rather than highly skilled and precise.
10	Be determined and persistent rather than flexible and willing to change course.
11	Show comprehensiveness, deep understanding and insight rather than keep things simple and straightforward.
12	Be imaginative and bring new ideas rather than stay practical and follow familiar approaches.

Table 3. Definitions of sectors, adapted from [76].

Sector	Definition
Operations	Activities concerned with planning, coordinating, and executing an organization's core processes for producing goods or delivering services, including workflow management, logistics, pricing, and operational administration.
Human Resources	Functions related to managing an organization's workforce, including recruitment, employment recordkeeping, workforce analytics, and administration of employee lifecycle events in accordance with organizational and regulatory requirements.
Finance or Accounting	Activities focused on designing, implementing, and maintaining financial and accounting systems to track, manage, and report economic transactions, ensuring financial accuracy, regulatory compliance, and decision support.
Engineering (e.g., Software)	Technical activities involving the design, development, modification, supervision, and maintenance of engineered systems, particularly software systems, to ensure functionality, performance, and alignment with technical specifications.
Data Analysis	The systematic cleaning, processing, modeling, visualization, and interpretation of data to identify patterns, trends, and relationships that support evidence-based decision-making.
Research	Systematic investigation involving data collection, literature review, survey methods, and statistical analysis to generate, validate, and extend knowledge for scientific, policy, or organizational purposes.
Healthcare Professional	Work focused on the assessment, treatment, monitoring, and coordination of patient care, including clinical decision-making, documentation, patient advocacy, and supervision to support physical, mental, and social well-being.
Legal	Activities involving the interpretation, application, and analysis of laws and regulations, including legal research, document preparation, advisory services, case strategy development, and advocacy.
Education Professional	Activities related to designing, delivering, and evaluating instructional practices, including classroom management, student assessment, curriculum implementation, and use of educational technologies to support learning and development.
Design or Creative	Work centered on the conceptualization and production of visual, written, or artistic content, translating ideas, narratives, and aesthetic principles into tangible outputs using artistic and digital tools.
Public Relations / Communications	Activities focused on managing information flows between organizations and their audiences through strategic messaging, media relations, content creation, and reputation management.
Customer Experience / Support	Direct customer-facing activities aimed at providing information, resolving service or billing issues, and maintaining customer satisfaction through communication and problem resolution.
Account Management	Activities involving the administration and support of customer financial relationships, including processing applications, explaining financial products, and managing ongoing accounts to ensure accuracy, compliance, and service continuity.

C Chronological analysis

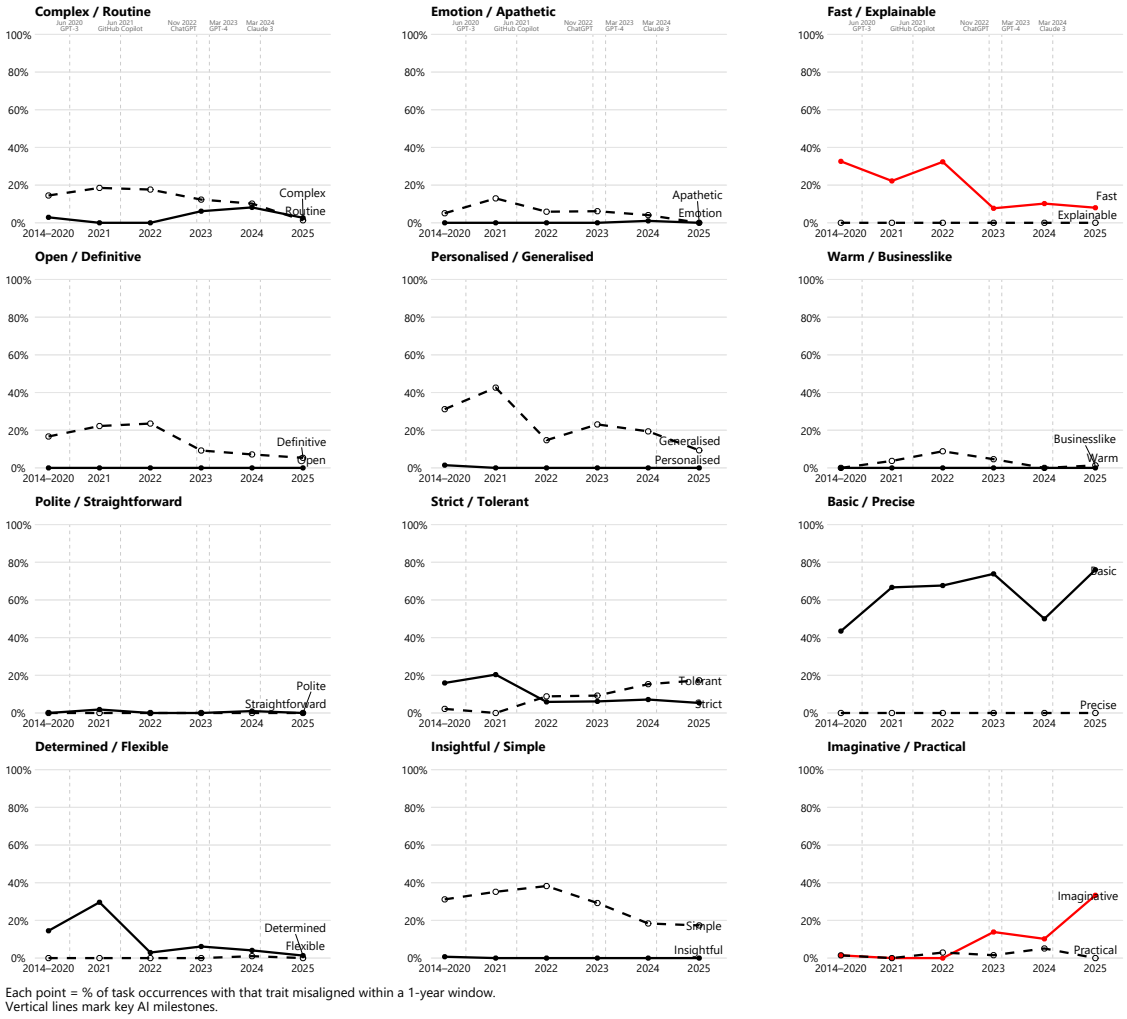

Fig. 9. **Percentage of task occurrences involving incidents of misaligned AI for each pair of traits over time.** Each data point represents the percentage of task occurrences classified as specified in the label within a year. We merged 2014-2020 as there were significantly fewer data points. For each pair of traits, the first trait listed (e.g., complex) is represented by the solid line, and the second trait listed (e.g., routine) is represented by the dashed line. Vertical lines mark milestones in AI research and deployment for context [63]. Interestingly, the frequency of misalignments involving specific traits increased or decreased over time: fast AI used to cause incidents, yet, since 2022, imaginative AI started to do so.

Table 4. Distribution of incidents that happened because of trait misalignment or other reason.

# Type	# incidents	% incidents
Misaligned	179	83.4
Aligned	35	16.6

Table 5. Distribution of the number of trait misalignments per incident.

# trait misalignments	# incidents	% incidents
0	37	17.3
1	32	15.0
2	33	15.4
3	26	12.1
4	24	11.2
5	22	10.3
6	19	8.88
7	12	5.61
8	4	1.87
9	4	1.87
10	1	0.47

Received 13 January 2026; revised 25 February 2026; accepted 15 April 2026