"I Don't Think RAI Applies to My Model" – Engaging Non-champions with Sticky Stories for Responsible AI Work

NADIA NAHAR, Carnegie Mellon University, United States
CHENYANG YANG, Carnegie Mellon University, United States
YANXIN CHEN, Carnegie Mellon University, United States
WESLEY HANWEN DENG, Carnegie Mellon University, United States
KEN HOLSTEIN, Carnegie Mellon University, United States
MOTAHHARE ESLAMI, Carnegie Mellon University, United States
CHRISTIAN KÄSTNER, Carnegie Mellon University, United States

Responsible AI (RAI) tools—checklists, templates, and governance processes—often engage RAI champions, individuals intrinsically motivated to advocate ethical practices, but fail to reach non-champions, who frequently dismiss them as bureaucratic tasks. To explore this gap, we shadowed meetings and interviewed data scientists at an organization, finding that practitioners perceived RAI as irrelevant to their work. Building on these insights and theoretical foundations, we derived design principles for engaging non-champions, and introduced sticky stories—narratives of unexpected ML harms designed to be concrete, severe, surprising, diverse, and relevant, unlike widely circulated media to which practitioners are desensitized. Using a compound AI system, we generated and evaluated sticky stories through human and LLM assessments at scale, confirming they embodied the intended qualities. In a study with 29 practitioners, we found that, compared to regular stories, sticky stories significantly increased time spent on harm identification, broadened the range of harms recognized, and fostered deeper reflection.

ACM Reference Format:
Nadia Nahar, Chenyang Yang, Yanxin Chen, Wesley Hanwen Deng, Ken Holstein, Motahhare Eslami, and Christian Kästner. 2026. "I Don't Think RAI Applies to My Model" – Engaging Non-champions with Sticky Stories for Responsible AI Work. In *Proceedings of CHI conference on Human Factors in Computing Systems (CHI'26).* ACM, New York, NY, USA, 34 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

"David was the only employee who used the building's accessibility elevator. When the company deployed predictive maintenance AI, the algorithm learned his elevator had extremely low usage and began delaying its repairs to prioritize busy main elevators.

David's elevator grew unreliable, but his complaints were ignored—the system marked it "low priority." During the annual shareholder meeting, it broke down completely, trapping him for hours while VIPs toured the building.

Authors' Contact Information: Nadia Nahar, nadian@andrew.cmu.edu, Carnegie Mellon University, United States; Chenyang Yang, Carnegie Mellon University, United States; Yanxin Chen, Carnegie Mellon University, United States; Wesley Hanwen Deng, Carnegie Mellon University, United States; Ken Holstein, Carnegie Mellon University, United States; Motahhare Eslami, Carnegie Mellon University, United States; Christian Kästner, Carnegie Mellon University, United States.

Manuscript submitted to ACM

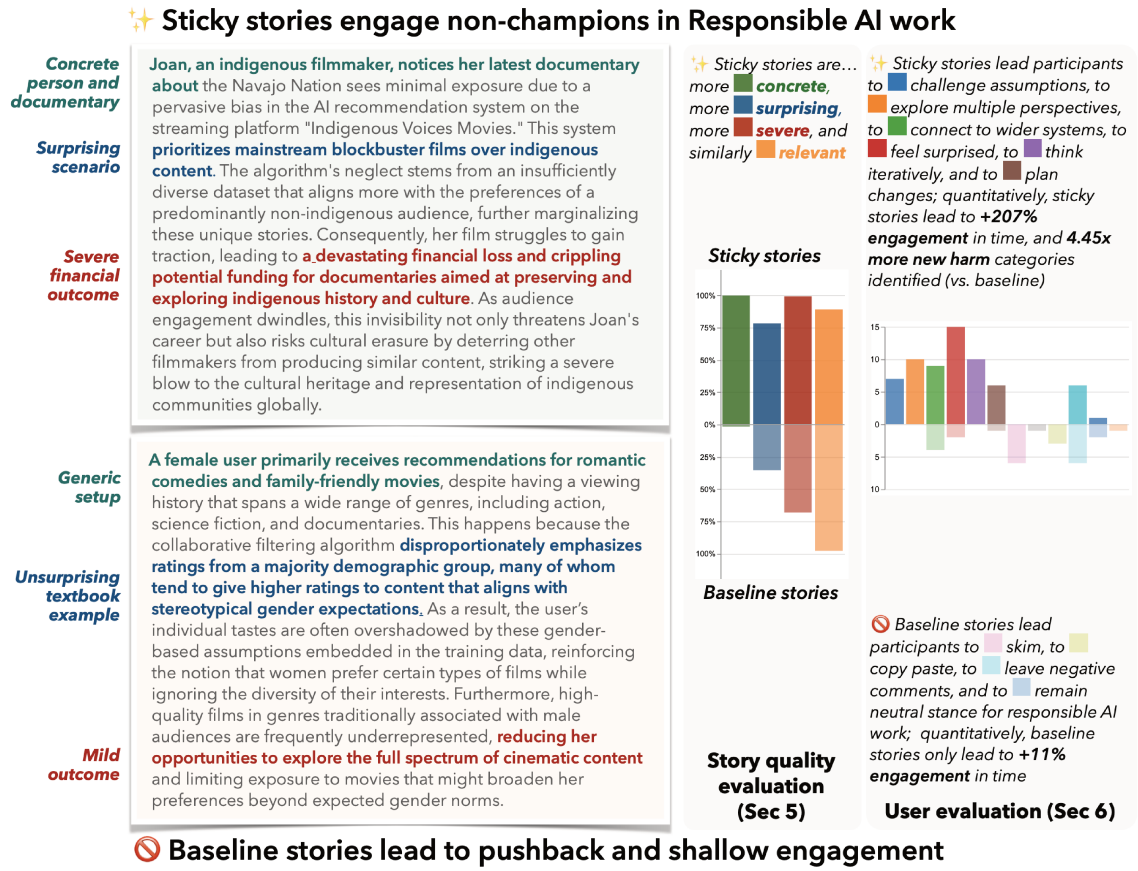

Fig. 1. Examples of different types of stories: (Bottom) Baseline story, which tends to be generic and straightforward but less memorable, and (Top) Sticky story, which incorporates elements such as surprise and emotional resonance to encourage deeper engagement and recall. We design (Section 4) and validate (Section 5) the sticky stories to be concrete, surprising, and severe. Our user evaluation (Section 6) with RAI practitioners demonstrates that sticky stories can better engage non-champions and lead them to more critical reflections.

David shared his experience on social media, and the story quickly went viral. Local news stations picked up the story, framing it as "AI bias leaves disabled worker trapped." The coverage triggered an ADA investigation revealing the AI had systematically neglected accessibility infrastructure based on usage metrics. Worse, the same system was deployed across dozens of buildings, all with severely under-maintained accessibility elevators now requiring emergency repairs. The company faced millions in fines, settlements, and unexpected maintenance costs."

We open with this LLM-generated vignette to make a point central to this paper: Surprising stories relevant to our own work cut through attention fatigue. As we will show, practitioners in "non-critical" domains are often desensitized by a steady stream of algorithmic bias headlines (e.g., biased hiring algorithms [22], bias in predictive policing and recidivism assessment [82, 134], autonomous vehicle crashes [19]) that seem bad but not relevant to them; as a result, fairness discussions remain abstract and easy to deprioritize. Despite a surge of checklists [64], templates [37, 76], games [4, 56, 68, 115, 117], and governance processes [93] for Responsible AI (RAI), prior research [46, 64, 87, 93, 94] has

primarily examined their effects on *RAI champions*—individuals intrinsically motivated to advocate for ethical practices within their organizations (sometimes in formalized roles [116])). Much less is known about the majority of practitioners—*non-champions*, those without prior motivation or formal RAI roles—and whether they would meaningfully engage with these resources or integrate them into their work. In this paper, we focus on *engaging non-champions to deliberate about RAI.*

In a formative study with a partner technology organization—where we shadowed meetings and interviewed governance members and data scientists—we observed a sharp contrast: Governance champions actively advanced RAI and designed governance structures, whereas many data scientists remained dismissive and disengaged regarding RAI concerns. Even when required to participate in activities, non-champions checked the boxes with minimal engagement. The key bottleneck was not the availability of guidance, processes, or tools, but the challenge of motivating these practitioners to meaningfully engage. This gap is critical, as RAI initiatives and tools cannot achieve impact if practitioners do not engage with them in a meaningful way. This observation motivated our research question: **How can we design interventions that foster deep engagement among non-champion practitioners in RAI processes?** By *deep engagement*, we mean more than just ticking boxes or going through the motions—it involves practitioners *critically reflecting* on potential harms, considering trade-offs thoughtfully, and retrospecting on their past experiences.

In this paper, we introduce an intervention designed to specifically engage *non-champion* practitioners by presenting them narrative-based scenarios, like the elevator-maintenance one above, that illustrate possible *unexpected* and *severe* real-world harms arising from *their own* ML systems. Drawing on theories from psychology and business communication [44, 70], we developed an LLM-based system that generates scenarios to embody five qualities that are known to drive engagement and memorability: Scenarios should be *concrete*, *severe*, *surprising*, *diverse*, and *relevant.* Inspired by the framework in *Made to Stick [44]*, we refer to these narrative scenarios as **sticky stories**. Unlike conventional documentation [75, 76] or checklists [64] or generic vignettes [17], which often feel abstract or disconnected from the realities of product development, our sticky stories are designed to evoke curiosity and spark critical reflection.

To evaluate the effectiveness of our *sticky stories*, we conducted two complementary evaluations. First, we assessed the "*stickiness*" of the generated stories themselves, that is, whether they embodied the five key qualities. The results showed that *sticky stories* significantly outperformed stories generated by zero-shot prompts used in past work across most qualities (e.g., severity: 99% vs 68%, and surprisingness 78% vs 35%) (cf. table 2). Second, we conducted a user study with 29 practitioners, mostly non champions, to measure the practical impact of these stories on engagement. Practitioners exposed to sticky stories spent significantly more time on harm identification (10 times more), identified a larger number of harm categories and subcategories (five times as many new categories and 3.5 times as many subcategories), and engaged in deeper critical reflection compared to those seeing baseline stories. Practitioners also exhibit distinct trajectories in shifting their attitudes from initial indifference or resistance toward a more engaged stance on RAI.

Contributions. This work makes the following contributions (see the study overview in Fig. 2).

- Our formative study characterizes the engagement gap among non-champion practitioners with Responsible AI efforts, and shows that existing governance tools and templates are insufficient to motivate meaningful participation.
- The design of *sticky stories* based on key qualities that capture practitioners' attention, provoke reflection, and foster engagement.

- A scalable compound AI system to generate stories that meet these qualities and integrate them into an interactive tool for practitioner use.
- An empirical evaluation demonstrating that sticky stories increase engagement time, the number and diversity of harms identified, and promote critical reflection.
- Practitioner-specific engagement trajectories, highlighting how different champions and non-champions in different profiles respond differently.

2 Related Work

Early work seeking to understand industry RAI practices suggested that RAI efforts were often driven by "individual advocates" who are self-motivated to pursue RAI work [64, 94]. In recent days, many practitioners are now formally tasked by their organizations with considering RAI issues [66, 101, 122]. In this paper, we use "*RAI champion*," a term used by organizations such as Microsoft as a role title [98], to refer to both self-motivated advocates and formally designated and trained RAI roles. While there is an abundance of prior work focusing on challenges RAI champions face and how to support them, it remains unclear whether such approaches transfer to the broader group of practitioners that we refer to as *non-champions*, namely those who are not already motivated to lead RAI work but nonetheless encounter RAI concerns in their everyday roles.

2.1 State of Responsible AI in Industry

Prior research has extensively examined industry RAI practices and challenges. Within the CHI and broader HCI communities, there has been a strong push to better understand industry RAI practices, challenges, and needs [46, 87, 94, 114]. For instance, through interviews and surveys, Holstein et al. [46] identified challenges in fairness-aware data collection and introduced proactive auditing processes. Focusing on UX professionals, Liao et al. [62] and Wang et al. [123] emphasized the need for improved tools and prototyping methods to facilitate communication and collaboration with technical teams when addressing RAI concerns.

Prior research highlights organizational challenges and risks that can limit the meaningful implementation of responsible AI in industry. A large body of HCI and RAI research has shown that individuals frequently encounter pushback from leadership when advocating for more responsible technologies [2, 9, 53, 114, 127, 128]. In addition, despite the many RAI principles, guidelines, and frameworks published by technology companies, organizational studies of industry RAI practices have consistently highlighted how the profit-driven and fast-paced nature of industry work often demotivates practitioners from engaging in meaningful RAI efforts [66, 87, 105, 126]. As a result, multiple studies warn that RAI processes risk becoming bureaucratic "check-the-box" exercises rather than reflective, substantive practices [15, 58, 64, 119, 120]. For instance, RAI documentation is often reduced to a compliance task [18, 27, 64, 129], while fairness and explainability evaluations can become performative practices, sometimes criticized as ethics washing [3, 24, 65, 100].

2.2 Supports for RAI Harm Identification

An abundance of structured templates and frameworks exist to support practitioners in Responsible AI impact assessment. To support practitioners in identifying potential harms and ethical risks, a wide range of RAI assessment approaches have been introduced, often in the form of structured checklists or impact assessments. For example, Microsoft's *RAI Impact Assessment Template* provides guidelines for conducting impact reviews prior to deploying AI products [75];

Bogucka et al. [13] co-designed and evaluated an AI impact assessment template with practitioners and compliance experts grounded it in regulatory requirements; Deng et al. [23] developed a *Societal Impact Assessment* template focused on design considerations for effective adoption and adaptation; and Rismani et al. [102] argue for the use of established hazard engineering techniques to structure the analysis

More broadly, several tools aim to promote broader reflection on the consequences of technology. Nathan et al. [79] developed a tool to help practitioners envision long-term effects of interactive systems. Elsayed-Ali et al. [30] introduced *Responsible & Inclusive Cards*, an online card-based tool designed to encourage critical reflection on project impacts. Ehsan et al. [29] introduced *Seamful XAI* to allow stakeholders identify mistakes and enhance AI explainability. Documentation frameworks such as *Datasheets for Datasets* [31], originally intended to improve transparency in data collection, have also been shown to help surface ethical concerns [11].

This line of work emphasizes structured processes and tooling as a means to support developers in anticipating harms and fostering reflection.

Recent research has begun leveraging large language models (LLMs) to help AI practitioners reflect on potential risks and harms in their systems. Building on prior HCI work that demonstrated the potential of large language models (LLMs) to support brainstorming and reflexivity, researchers have begun exploring how to incorporate LLMs into RAI tools to support reflection around RAI concerns [17, 84, 124]. Approaches either (a) use LLMs to generate examples of possible harms for a system, following structured reasoning internally to create diverse harms, such as the vignettes generated by AHA! *[17]* and our own work in this paper, or (b) identify and present real-world reports about related systems from news stories as in Farsight [124] and BLIP [84]. Both strategies aim to guide analysis with realistic examples and broaden the range of consequences considered.

We find these tools promising and build on similar ideas. While their design may not explicitly target RAI champions, we suspect that they are more effective for developers already motivated for RAI work. Since motivated RAI practitioners are more likely to volunteer for evaluations of RAI tools (self-selection bias) and the participants' prior motivation was not controlled for in prior studies, we are curious about how effective such tools are for a broad range of practitioners.

2.3 Engaging Non-Champions

Research suggests that existing fairness and interpretability tools often fail to foster genuine engagement, instead encouraging superficial compliance and limiting meaningful understanding. As some organizations make RAI steps mandatory, either through explicit RAI audit gates [96, 97, 99] or by attaching RAI considerations to existing required privacy assessment steps [26], it remains to be seen whether non-champions engage in depth or just do the minimum amount of work to complete the necessary steps ("check-the-box compliance")

Research has found evidence of such a check-the-box culture: For instance, a participant in Balayn et al.'s study noted, "*Fairness for many companies is just a small checkbox, and sometimes people put their mark without any question…*" [3]. Similarly, Kaur et al. [54] found that interpretability tools, while designed to improve understanding of machine learning models, can sometimes impair it, with strategies aimed at promoting deliberation and engagement frequently failing to overcome this, and Omar et al. [81] found that structured policy guidance was not effective at engaging with user needs for explanation designs.

Recent research has begun exploring strategies to involve non-champions in responsible AI work, though significant gaps remain. Common strategies to attempt to engage practitioners for RAI work involve nudging [10], gamification [4, 56, 115, 117], and reframing fairness work in familiar quantitative terms [26]. For example, Bhat et al. introduced

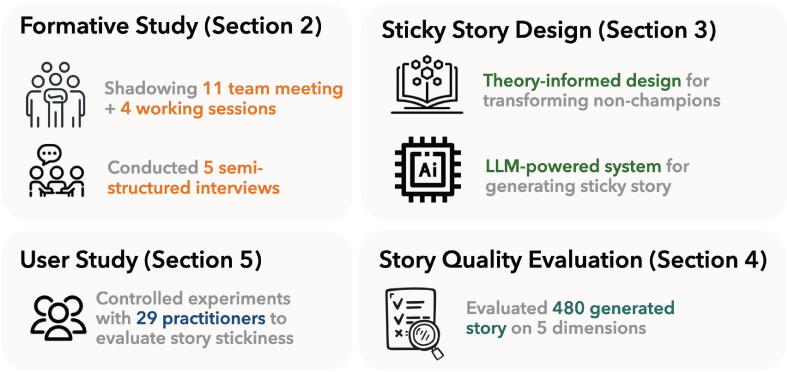

Fig. 2. Research overview. The work spans four parts: (1) a formative study, (2) a theory-informed, LLM-powered design for generating "sticky" stories, (3) a quality evaluation of sticky stories, and (4) a controlled user study.

a JupyterLab extension to nudge data scientists to complete and update model card documentation, particularly the ethics-related sections [10]. Ballard et al. proposed *Judgment Call*, which helps product teams surface ethical concerns using value-sensitive design and design fiction [4], and Kim et al. developed *The Desk: Dilemmas in AI Ethics*, a digital game-based approach to enhance learning about AI ethics [56].

While these strategies can capture short-term attention, it is not clear that they are sustainable to keep non-champions engaged. Nudges may increase initial actions, but cause longer-term behavior changes—and may even reduce follow-through in some cases [49, 90, 135]. Similarly, gamification can spike early engagement, yet motivation drops as the novelty wears off, and in some contexts may impair deeper learning or distract from authentic engagement [42, 103]. For example, Widder et al. [126] argued when evaluating one of these games, that hypothetical contexts created in the game are unlikely to be a viable mechanism for real world change.

In summary, prior research has explored barriers to RAI work and provided many processes and tools to support and engage practitioners in RAI work. However, as we have experienced and will describe next, this support is not equally effective for all practitioners and may fall short for non-champions who are not motivated to go beyond minimally required steps, if any. To the best of our knowledge, prior research focuses mostly (deliberately or not) on RAI champions and has not explored the differences between champions and non-champions. In this paper, we aim to fill this gap, by focusing specifically on how to engage non-champions for RAI activities.

3 Sticky Story Intervention Design

In a formative study with a partner organization, we found that, despite robust governance and support, most practitioners regarded RAI activities as irrelevant or bureaucratic, often engaging only for compliance. Templates and checklists were frequently ignored or treated as paperwork, highlighting a critical lack of motivational buy-in. We share our motivating findings from the formative study in the supplementary material/appendix for transparency, but otherwise focus on an intervention: Our goal is to engage "non-champions", that is stakeholders who may be indifferent or skeptical, to meaningfully engage in RAI efforts by making the consequences of neglecting its principles feel real and relevant. In this section, we discuss established theories about how practitioners' attitudes and engagement with RAI can be shifted or transformed, and how these theories inform the design decisions behind our intervention.

3.1 Theoretical Background and Design Principles

Transforming Non-Champions. Our initial aim was to motivate non-champions to care about RAI, and *transform* into champions. We drew on *Transformative Learning Theory* [72], which explains how people change underlying beliefs through *critical reflection*. Central to this, is the concept of *disorienting dilemma*—an event, challenge, or scenario that disrupts assumptions and compels self-examination. In practice, transformation is difficult and slow, requiring interventions that actively provoke reflection rather than passive exposure [73, 125]. *Cognitive Dissonance Theory [33]* complements *Transformative Learning Theory* by describing how psychological discomfort (*dissonance*) from conflicting beliefs can drive a new perspective.

Together, these theories suggest that effective interventions must present challenges to existing assumptions and support structured reflection. Prior research in domains such as ethics education, clinical training, and diversity initiatives demonstrates that reflective prompts, scenario-based exercises, and facilitated discussions can trigger these processes [67]. For example, in health professional education, curricula designed around transformative learning principles have been shown to improve practitioners' ethical reasoning, and critical reflection [104]. Similarly, in higher education, faculty engaged in action research grounded in transformative learning reported meaningful changes in their teaching practices [39]. However, lasting mindset change requires more than a single exposure—it depends on repeated reinforcement, supportive contexts, and engaging formats [89, 125]. The central question is how to design interventions that provoke these mechanisms to trigger such a transformation among RAI non-champions.

Our aim is to trigger such a transformation, not just "tricking" professionals to engage in short term (e.g., with nudging or gamification). Therefore, we sought to create moments of *disorienting dilemmas* to cause *cognitive dissonance*—points of discomfort strong enough to disrupt assumptions. However, not all disruptions are equally effective. Our formative findings revealed that practitioners had become desensitized to widely circulated media narratives of bias framed around gender and race, often dismissing as irrelevant to their projects. To overcome this resistance, we surfaced dilemmas that were likely unfamiliar to the practitioners and directly consequential within their work contexts.

Sticky Stories as Intervention. One way to capture practitioners' attention is by showing them stories of harm caused by their own ML systems. Prior work has demonstrated that vignettes and fictional scenarios can be effective for RAI champions [17], but we expect non-champions may require more than generic stories: They need something to disrupt and capture their attention, something that differs from common well-known media stories they already dismissed as irrelevant to their work. We aim to create stories that not only illustrate harms in the moment but also resonate deeply and are remembered later to seed an actual transformation.

To achieve such impact, we turned to marketing theory, grounding our approach in the principles of *Made to Stick* [44], which summarizes characteristics of memorable and persuasive ideas. While *Transformative Learning Theory* and *Cognitive Dissonance Theory* describe the stages of transformation and the need to create reflective moments, they do not specify how to craft materials that consistently capture attention and sustain engagement. *Made to Stick* offers a practical framework, suggesting strategies that can trigger these mechanisms. We operationalized these principles into five key story qualities to guide the construction and evaluation of our **sticky stories**:

- 😮 **Surprisingness.** Stories should present harms in ways that disrupt default assumptions, that are counterintuitive or non-obvious. This quality captures attention, which is a prerequisite for reflection and potential attitude change. Evidence from cognitive science shows that unexpected stimuli attract attention and trigger deeper processing [52]. By making harms surprising, practitioners are more likely to notice risks that would otherwise be overlooked and experience disorienting dilemmas.

- **Concreteness.** Stories should include tangible details, such as named roles, real-world analogues, or observable system behaviors. Concreteness helps audiences visualize harms and form emotional connections, making abstract principles more understandable and memorable. Dual Coding Theory [28] supports this approach, showing that concrete information is encoded both verbally and visually, improving recall compared to abstract descriptions.
- **Severity.** Severity emphasizes the magnitude and scope of potential harm, highlighting why certain outcomes demand attention and action. Perceived severity motivates engagement, as individuals are more likely to respond to risks they judge serious. Research on the affect heuristic [113] demonstrates that people's risk judgments are strongly influenced by the emotional weight of outcomes, with severe consequences eliciting stronger reactions and prompting protective or corrective behaviors.
- **Relevance.** Stories should align with domain-specific experiences and stakeholder concerns, ensuring that messages feel applicable to practitioners' work. Relevance increases the likelihood that examples are processed deeply and influence attitudes or behaviors [69]. By situating lessons in authentic professional contexts, practitioners can connect the scenarios to real decisions, enhancing engagement and reflection.
- **Diversity.** A broad range of stakeholders, harm types, and system behaviors can avoid narrow or stereotypical portrayals. Evidence from narrative transportation research indicates that encountering multiple perspectives enhances engagement and supports attitude change [40]. Diverse stories help practitioners anticipate harms across contexts rather than focusing on isolated examples.

Capturing Early Signs of Change: Critical Reflection. Our intervention used *sticky stories* to create *disorienting dilemmas*, aiming to prompt reflective thinking. While our ultimate goal is *transformation*, genuine perspective shifts are difficult to observe without extended observation windows over multiple years; immediate responses may not indicate lasting change. To keep the scope of our research manageable, we focus on *early indicators of transformation*, observing moments when *disorienting dilemmas* triggered *critical reflection*.

To observe signs of critical reflection, beyond just short-term measures of engagement, we focused on concrete behavioral signs that participants were moving beyond surface-level reactions. Prior work defines *critical reflection* as moving beyond descriptive or casual reflection to actively examine one's assumptions, beliefs, and actions, evaluating their validity and potential consequences [72]. In the context of responsible AI, this would involve questioning default practices, recognizing ethical risks, and considering how one's work may contribute to harm. To systematically detect these moments of *critical reflection*, we identified *concrete behavioral indicators,* summarized in Table 1. We will use these indicators in our evaluation in Sec. 6, tracing how these moments manifested during participants' engagement with the stories.

3.2 Generating Sticky Stories

Generating high-quality *sticky stories* is non-trivial. We found that single zero-shot or few-shot prompts with LLMs were not very effective in generating stories that meet all five of our criteria, as they often produce outputs that are overly generic, repeat common tropes (e.g., race and gender only), and fail to capture surprising or contextually relevant harms. Without structured guidance, LLMs struggle to systematically combine diverse harm types, non-obvious stakeholders, and generate stories that effectively provoke reflection. This motivated our design of a systematic and scalable compound AI system [118] that combined prompt engineering techniques with programmatic control to ensure that each story consistently embodied the five desired qualities. We broke down the task into a coherent sequence

Table 1. Signs of Critical Reflection in Non-Champions

Sign	Description	Indicators	Example
Challenges assumptions [72]	Practitioners critically examine or dispute underlying beliefs, premises, or taken-for-granted assumptions that might otherwise go unquestioned.	Identifies own beliefs that may not hold; contrasts story assumptions with their own understanding, experience, or evidence; expresses skepticism toward "default" ways of thinking or doing things.	***P23***: *"I think human evaluators are really important – it's something that I am kind of understanding now. We haven't done this. But looking at the stories, I think, it's a really important for us"*
Explores multiple perspectives [16]	Practitioners consider scenarios from multiple angles or stakeholders, compare alternatives, or expand the scope beyond their immediate perspective.	Weighing different interpretations; comparing different parts of a story; acknowledging multiple voices.	***P15***: *"As a data scientist, I can retrain the model, and I can test it. But what can a government or the person in FDA do? They don't know this."*
Connects to wider systems or past incidents [16]	Practitioners connect the story to broader organizational, social, or technical systems, often moving beyond the immediate prompt.	Linking story implications to other domains, contexts, or systemic risks; recalling real-world events.	***P4***: *"[...] like scholarship distribution and funding distribution specifically, whatever was coming to any charity. How are they distribute it amongst like schools or old age homes and other organizations where the funding has to go to."*
Expresses surprise [72]	Practitioners show surprise, novelty, or realization about aspects of the scenario, signaling a shift in understanding.	Expressions of being surprised, not having thought of it before, finding something new, unexpected, or revealing.	***P19***: *"oh, that's mind blowing [...] I didn't know this can happen."*
Engages in iterative thinking [107]	Practitioners demonstrate a back-and-forth reasoning process, revising or expanding their views as they talk.	Evidence of reconsideration, self-correction, extended reflection, or stepwise elaboration.	***P26***: *"Now that I'm looking at the word like demeaning. I feel like I can probably think of a couple more examples."*
Plans intentional change [71]	Practitioners translate reflection into concrete plans for action or acknowledge the need to modify future behavior, processes, or designs.	Commitment to follow-up action; expressing intent to discuss with others; specific takeaways for their own projects.	***P30***: *"I think, filtration of the training data is pretty simple. It could be done with a simple Regex based filtration, for at least getting rid of such harmful potential comments. And post filtration for the same should be simple, too. Yeah. So I think that's pretty painless to deploy. I would be motivated to do that."*

of logical steps and iterated through multiple design cycles to refine each component. This resulted in the following eight-step pipeline (cf. Fig.3):

- **Input (🎯).** To make the stories relevant to non-champions, we grounded them in the specific projects participants were working on. As inputs, we collected the ML system's description, its intended purpose, and a representative stakeholder use case. These inputs seeded the rest of the pipeline, ensuring that the generated stories were *relevant*—directly tied to the participant's own ML system rather than abstract or generic examples.
- **Step 1: Pre-define Harm Types (🎨).** To ensure diversity and coverage of edge cases, we began by specifying a set of harm categories. Pre-defining these categories grounds story generation in well-theorized frameworks of harm rather than ad hoc examples. For this, we drew on the taxonomy of fairness-related harms from prior studies of harm categories [17, 110] and fairness goals from Microsoft's RAI assessment guide [75]. The set included cultural misrepresentation, reinforcement of biases, unequal access to opportunities, and erasure of minorities.

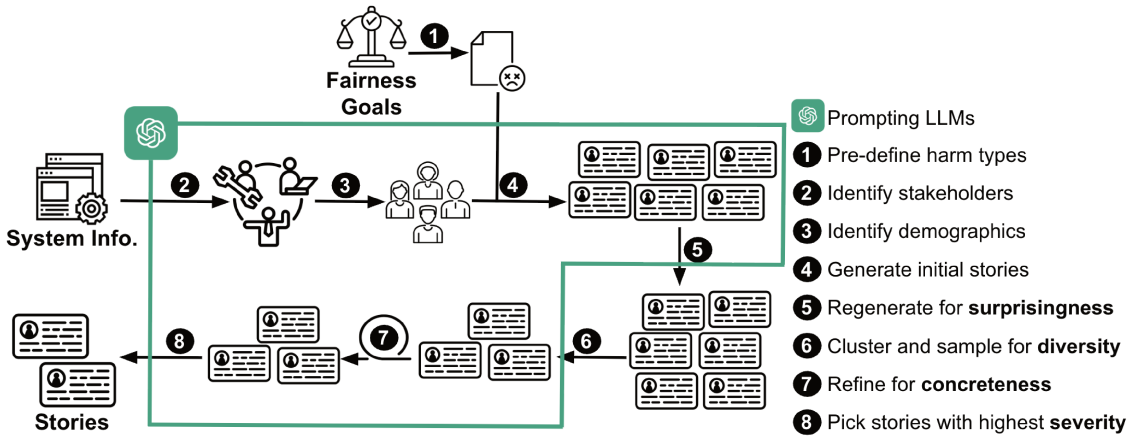

Fig. 3. Pipeline for *Sticky Story* Generation

- **Step 2: Identify Stakeholders (,).** To generate stories that capture diverse harms, it is also essential to identify stakeholders whom practitioners might otherwise overlook. For example, in a movie recommendation system, practitioners often focus on obvious stakeholders such as movie watchers, but may miss stakeholders such as movie producers, whose livelihoods depend on whether their films are surfaced. To account for these cases, we go beyond conventional direct and indirect stakeholders by introducing *direct-surprising* and *indirect-surprising* categories. These highlight marginalized or non-obvious groups, following the principles of *Design Justice* [21], which emphasizes attention to those often overlooked in design processes. We prompt an LLM to generate these different sets of stakeholders based on the product description. By surfacing unexpected stakeholders, the stories are more likely to create *disorienting dilemmas* that challenge practitioners' default assumptions.
- **Step 3: Identify Demographics (,).** To ground stories in concrete contexts, and relevant to the user, we then generate possible demographic attributes for each stakeholder (e.g., age, gender, ethnicity). This step facilitates subsequent steps in tailoring harms to marginalized or contextually relevant groups.
- **Step 4: Generate Initial Stories ().** To increase diversity in our stories, for each harm–stakeholder combination, we generate an initial set of harm stories, forming a matrix of harms and affected users. This matrix-based approach, inspired by prior work [17], ensures broad coverage and combinatorial richness. By systematically exploring combinations, we reduce the risk of narrow or stereotypical examples, and instead highlight harms that may not surface through ad-hoc brainstorming.
- **Step 5: Regenerate for Surprisingness ().** Given that the LLM might default to producing stories that align with patterns it frequently encounters, to avoid bland or generic outputs, we prompt the model to regenerate stories using earlier outputs as counterexamples—encouraging less typical, more striking, and surprising narratives.
- **Step 6: Cluster and Sample for Diversity ().** To further increase diversity and avoid redundancy, we transform the stories into sentence embeddings, apply K-means clustering (k=10), and sample from the five least-populated clusters—those most likely to contain unique narratives.

- **Step 7: Refine for Concreteness and Severity (🪪, ☠️).** Concreteness makes harms tangible and easier to visualize, increasing practitioners' emotional engagement. To make sure the stories are concrete and severe, we employ a two-stage refinement loop: one model refines stories for concreteness and severity, and a second evaluates the output. Stories lacking specificity or clarity are iteratively revised (up to three times) to meet our concreteness standard.
- **Step 8: Pick Stories with the Highest Severity (☠️).** Rather than enumerate every harm comprehensively (as other tools [17, 124] and hazard analysis [61, 102] pursue), we aim to provoke and persuade with a small number of high-impact stories. Therefore, in a final step, we prompt an LLM to select the two stories with the greatest magnitude and scope of harm, while ensuring they satisfy all five qualities.

3.3 Sticky Story Integration in a Tool

To demonstrate how sticky stories could be presented to practitioners and to run our evaluation study, we integrated the pipeline into an interactive tool. The tool is designed to replicate Microsoft's RAI assessment guide [75] (Fig. 4), which is typically completed as a static, text-based template.

Users begin by entering a brief description of the ML system, its intended purpose, a user story, and system stakeholders (Fig. 5-A (1, 2)). Subsequently, during the fairness assessment step (Fig. 5-B and C), users can request brainstorming assistance, which presents the previously generated sticky stories (Fig. 5-D) for the current assessment step.

To reduce potential delays and maintain a smooth user experience, these stories are often generated in earlier screens of the tool—based on the user's description of the ML system—and stored for retrieval in the subsequent fairness brainstorming step. This approach can help ensure that stories appear quickly when requested (story generation with the GPT-4o model typically takes 3–4 minutes) helping users focus on the task without waiting—though in practice the timing may vary depending on system load and context. The tool shows two stories by default, but users can request up to three more stories, provide feedback, and regenerate stories.

4 Evaluation I: Evaluating Stickiness of Harm Stories

We first conducted an offline evaluation to understand the quality and cost of generating sticky stories with our designed pipeline. In the evaluation, we curate diverse AI application scenarios and run our pipeline and an ablated version to generate the harm stories. We then measure the quality of the generated stories with the five desired qualities of sticky stories: **concrete (🪪), severe (☠️), surprising (😮), diverse (🧠), and relevant (🎯),** as well as the cost of generating these stories in terms of token usage and time elapsed.

4.1 Experiment Setups

4.1.1 Data. We collected diverse AI application scenarios from the Internet (e.g., [77, 91, 92]) and randomly sampled 15 scenarios (e.g., *voice assistants*, *image search*, *email monitoring*, and *demand forecasting*) for our evaluation.

We used an LLM (`gpt-4o`) to process the searched content into more detailed descriptions, similar to what a user would have input to our system, and we manually verified that these generated descriptions are valid.

4.1.2 Methods. For evaluating the generation capabilities, we implement most of our pipeline with `gpt-4o`, as it demonstrates strong writing capabilities [83]. In step 7, however, we use a smaller model `gpt-4o-mini` as the evaluator

Fairness considerations

2.4 *For each Fairness Goal that applies to the system, 1) identify the relevant stakeholder(s) (e.g., system user, person impacted by the system); 2) identify any demographic groups, including marginalized groups, that may require fairness considerations; and 3) prioritize these groups for fairness consideration and explain how the fairness consideration applies. If the Fairness Goal does not apply to the system, enter "N/A" in the first column.*

Goal F1: Quality of service

This Goal applies to AI systems when system users or people impacted by the system with different demographic characteristics might experience differences in quality of service that can be remedied by building the system differently. If this Goal applies to the system, complete the table below describing the appropriate stakeholders for this intended use.

Which stakeholder(s) will be affected?	For affected stakeholder(s) which demographic groups are you prioritizing for this Goal?	Explain how each demographic group might be affected.

Goal F2: Allocation of resources and opportunities

This Goal applies to AI systems that generate outputs that directly affect the allocation of resources or opportunities relating to finance, education, employment, healthcare, housing, insurance, or social welfare. If this Goal applies to the system, complete the table below describing the appropriate stakeholders for this intended use.

Which stakeholder(s) will be affected?	For affected stakeholder(s) which demographic groups are you prioritizing for this Goal?	Explain how each demographic group might be affected.

Fig. 4. *Snapshot of Microsoft's Responsible AI assessment template [75]*

to reduce cost, as validation usually requires less capabilities than generation. We use `mxbai-embed-large-v1` to produce sentence embeddings for K-means clustering.

Baseline. We compare our pipeline approach to a zero-shot prompting baseline, in line with prior work [17]. The prompt directly instructs an LLM (`gpt-4o`) to generate scenarios that illustrate harm to relevant stakeholders, and we instruct that the baseline prompts generate stories around 175 words, which is the average length we observe from the stories generated by our pipeline. We share all prompts used in our supplementary material.

Story generation. For each AI application scenario, we generate two stories for each of the two fairness goals (*Quality of Service*, and *Allocation of Resources and Opportunities*). In total, we curated 120 sticky stories and 120 baseline stories. This sample size allows us to draw conclusions with 90% confidence level with 8% margin of error.

4.1.3 Metrics. We evaluate the quality of the sticky stories and baseline stories and the cost of the generation method. For cost, we measure the time it takes to run the pipeline and the number of tokens it costs. For quality, we evaluate

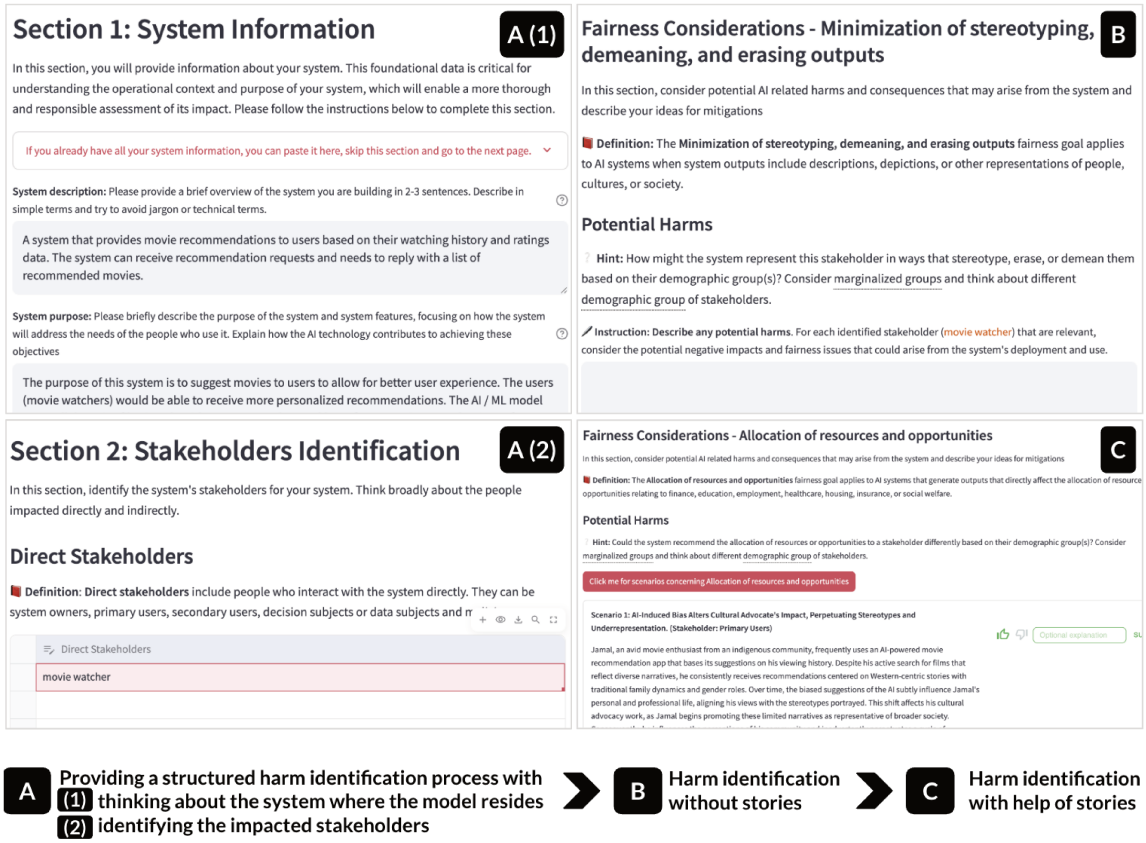

Fig. 5. Snapshots of the Tool Integrating the Sticky Story Generation Pipeline

the stories on the five desired qualities of *sticky stories*. Four quality metrics (*concrete*, *severe*, *surprising*, *relevant*) are evaluated with a two-phase evaluation involving both human raters and LLMs, with diversity evaluated by a separate established distance metric. The final evaluation covered **240 stories** produced by both methods.

Human Annotation and Inter-Rater Reliability. Three researchers independently evaluated a set of 20 harm stories using binary (yes/no) judgments across the four qualities of stickiness. We conducted multiple calibration rounds, during which we refined the definitions of the qualities and clarified edge cases based on observed disagreements. After we reached consensus, we developed the finalized rubric (see supplementary material) for a larger-scale evaluation.

Scalable Evaluation Using LLM-as-a-Judge. To scale the evaluation to the full set of generated stories, we adapted our rubric for an automated evaluation using `gpt-4o`. The prompt included the story text, plain-language definitions of each of the qualities, and a binary decision task for each dimension (0 = no, 1 = yes). Four of the five dimensions were rated using binary LLM judgments.

To validate the reliability of the LLM judgments, one researcher independently annotated 30 additional stories. The Cohen's Kappa values for agreement between human and GPT-4 annotations were: Concrete: 1.000, Severity: 0.4783, Surprising: 0.9356, and Relevant: 0.8387. All disagreements on Severity stemmed from the LLM being overly generous

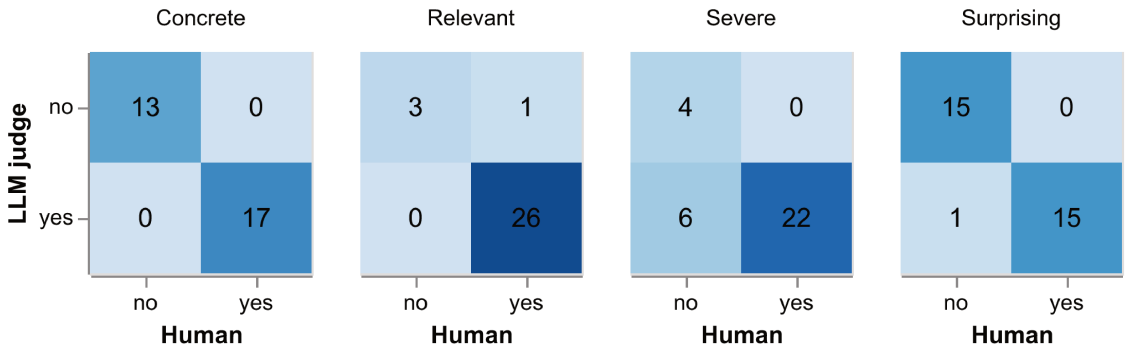

Fig. 6. Humans agree well with LLM judges on whether a story is concrete, relevant, or surprising. While there is more disagreement on severity, we found they all stemmed from LLM being overly generous in rating baseline stories as severe.

Table 2. Offline evaluation results of the generated stories.

	Quality					Cost	
	Severity	Surprising	Concrete	Relevant	Diversity	Token	Time/s
Pipeline	**0.992**	**0.783**	**1.000**	0.892	**0.156**	56665	50
Baseline	0.683	0.354	0.017	**0.979**	0.098	1232	9

in rating baseline stories as severe, thus overinflating the severity results of the baseline (see Figure 6) – hence, results regarding severity likely underestimate the real effect.

Measuring Diversity with Embedding Distances. For diversity, we computed cosine distances between semantic embeddings of the story titles (higher distance indicates higher diversity), as it is a well-established distance measurement in the literature [95].

4.1.4 Limitations (Threats to Validity). Despite extensive validation, internal validity may be affected by biases in LLM-based judgements, especially on severity. To mitigate the potential verbosity biases of LLM judges [132], we controlled for story length by generating texts of comparable length across conditions. Our binary ratings for each quality captures only big differences and may not represent more nuanced quality differences. External validity is constrained by the small, curated set of 15 scenarios, which may not represent all AI applications.

4.2 Results

Overall, we found that our pipeline is able to generate sticky stories that are more concrete (+98.3%), more severe (+30.9%), and more surprising (+42.9%) than baseline stories, demonstrating the effectiveness of our design (see Table 2 for more details). The sticky stories are also generally more diverse (+5.8%), due to our clustering approach. However, we do observe a small trade-off in relevance (-8.7%), as sometimes the generated sticky stories are overly dramatic and can be hard to relate. In addition, generating sticky stories requires more resources (5.5x time and 46x token usage) than the zero-shot generation of baseline stories. As we will show next, this cost is likely acceptable, as the sticky stories

are indeed more effective at engaging practitioners and inspiring them to think of RAI harms beyond their existing mindsets.

5 Evaluation II: User Study

After showing that sticky stories indeed embody the five desired qualities (Sec. 5), we now assess their practical value, that is, whether sticky stories actually lead to greater engagement from (non-champion) practitioners. We conducted a user study that explored how non-champion practitioners engage with harm identification tasks with and without stories. We analyzed differences in terms of (1) the time participants spent identifying harms, (2) the number of new harm categories they surfaced, and (3) the depth and nature of their *critical reflections* on the harms or stories.

5.1 Study Design

To evaluate the impact of *sticky stories* on practitioner engagement, we conducted a **mixed-design user study** that combined both **within-subject** and **between-subject** elements. Unlike prior work that often focuses only on champions, we deliberately sought to include non-champions, and ended up with a range of practitioners with varied levels of RAI motivation. Each participant completed two harm identification tasks under two of three conditions: *no stories*, *baseline stories*, or *sticky stories*. This design allowed us to disentangle the effect of story presence from the distinct qualities of sticky stories, while also partially controlling for potential learning and ordering effects. We assessed engagement through multiple indicators, including time spent, number and diversity of harms surfaced, and qualitative signs of critical reflection in response to the stories.

5.1.1 Participants and Recruitment. Unlike prior studies that did not account for self-selection bias, which likely led to primarily recruiting participants already motivated by RAI concerns, we sought to recruit *non-champion* practitioners—those less inclined to prioritize RAI in their work. Recruiting this group was inherently difficult, as they are unlikely to volunteer for a study framed around Responsible AI. To overcome this, we strategically oversampled through broad advertisements that emphasized ML evaluation broadly, and probed participants' preferences for different evaluation techniques in a screening survey. This design choice helped attract a broader and more neutral practitioner audience, including individuals who are less inclined toward fairness assessments and thus more representative of non-champions whom our intervention seeks to engage. Recruitment and study protocols were approved by our Institutional Review Board (IRB).

We recruited participants through professional platforms such as LinkedIn, Twitter, and a large Slack community for data scientists. Interested individuals completed the screening survey, which included questions regarding their familiarity with concepts such as model training, model evaluation, model fairness, AI ethics, and MLOps, and how useful they think various evaluation activities are, including in-distribution data evaluation, out-of-distribution data evaluation, model red-teaming, and responsible AI auditing (e.g., fairness). This enabled us to identify practitioners who do not prioritize RAI in their work. We received a total of 291 responses. We excluded submissions that indicated low engagement or fraudulent behavior—such as vague project descriptions, suspicious email addresses, or missing LinkedIn profiles—and filtered participants based on their stated level of high RAI interest. We conducted 5 pilot experiments to test and refine the study protocol. In the pilot experiments, we observed fairly strong effects, suggesting that we could reliably detect true effects with moderate numbers of participants. Afterward, we recruited 31 participants, but due to a tool malfunction in which the baseline stories failed to generate, data from two participants could not be analyzed. That

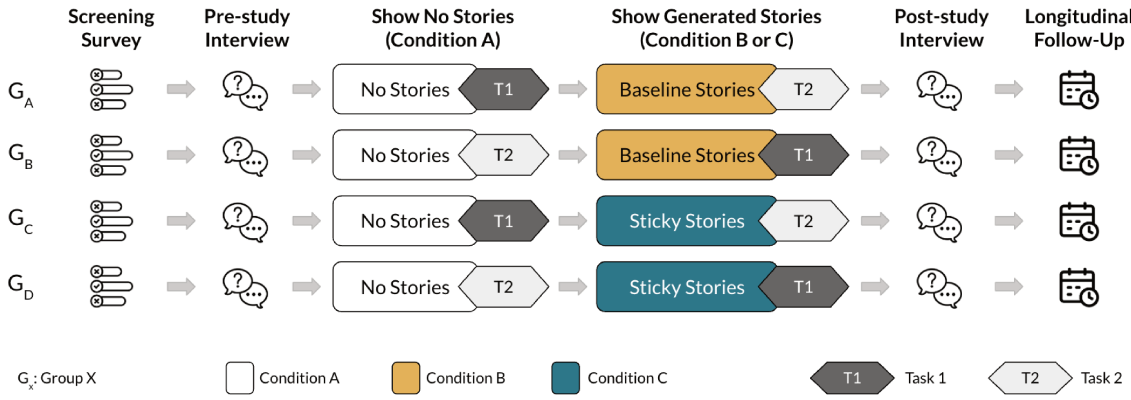

Fig. 7. The mixed study design for the user study that combines both within-subject and between-subject elements, to evaluate the effectiveness sticky stories compared to no stories and baseline stories

is, we successfully conducted the study with **29 participants**. Each participant received as compensation a $35 gift card.

5.1.2 Experimental Conditions. To evaluate the impact of the sticky story intervention on practitioner engagement, we designed three experimental conditions. Engagement could potentially be influenced either by the presence of any illustrative story or specifically by the *sticky stories*. To disentangle these effects, we implemented two control conditions and one treatment condition.

- **Condition A (No Stories - Control 1):** Participants complete a section of the RAI assessment without being shown any stories.
- **Condition B (Baseline Stories - Control 2):** Participants complete a section of the RAI assessment while being shown two baseline harm stories (see Sec. 5.1.2; zero-shot prompting, matched for length of sticky stories).
- **Condition C (Sticky Stories - Intervention):** Participants complete a section of the RAI assessment while being shown two *sticky stories*.

5.1.3 Tasks. Each participant completed two tasks focused on harm identification and mitigation planning. To ensure ecological validity [55] and help participants engage more deeply with the task that is realistic and personally relevant, participants analyze *their own projects*. Each task asks the participant to identify harms and possible mitigations for one of two fairness goals from Microsoft's RAI assessment guide [75]—a well-established framework developed independently of our study:

- **Task 1:** Analyze fairness regarding allocation of resources and opportunities.
- **Task 2:** Analyze fairness regarding stereotyping, demeaning, and erasing outputs.

Each participant worked on both tasks in a random order. Analyzing their own projects ensures that the evaluation reflects realistic, personally relevant contexts.

5.1.4 Groups. We randomly assigned participants to one of four groups. Each group completed their first task in the no story condition (Condition A), followed by the other task in either the baseline (Condition B) or sticky story condition (Condition C), see Fig 7. This design enables both **within-subject** comparison (e.g., engagement with vs. without stories) and **between-subject** comparison (baseline vs. sticky stories). In addition, when participants worked

on the first task, we could already generate stories for the second task in the background. By randomizing assignment and counterbalancing the order of fairness goals, we reduce confounds related to task complexity and learning effects.

5.1.5 Study Protocol. Each participant completed a pre-study survey, answered a few questions to establish their background, worked on the two tasks with and without stories, and finally debriefed with the facilitator. This usually took about 60 minutes. Two months after the experiment, we sent a follow up survey.

Pre-Study: Participant Background and RAI Orientation. We collected background information to understand each participant's experience and relationship with RAI practices (e.g., their exposure to RAI), both through a short survey and a brief verbal discussion (see *Interview Guide* in supplementary material). Drawing on stated choice research [63], to increase reliability, we ask questions about behaviors (revealed preferences) rather than preferences (stated preferences). Due to the sensitive nature of responsible AI and non-championship, we intentionally did not collect details about user study participants' demographics, following prior work on responsible AI practices in industry [25, 46, 64, 65].

During the Study: Think-Aloud Harm Identification. While the participants worked on the two tasks, we employed a **think-aloud protocol** to capture participants' real-time thought processes and reasoning while engaging with harm identification tasks, allowing us to understand not just what they identify but how they interpret and respond to different stories. We issued only minimal prompts to preserve the validity of participants' cognitive processes during harm identification tasks [14, 31]. The facilitator maintained a neutral stance and refrained from influencing participants' reasoning—intervening only when they misunderstood the task or deviated significantly from it. In conditions where participants were exposed to stories (either *baseline* or *sticky*), we deliberately avoided leading questions to minimize bias. Instead, we used minimal, open-ended prompts (e.g., "*What are your thoughts as you read this*?" or "*Feel free to keep thinking aloud*") to encourage continued verbalization and self-reflection.

Post-Study: Reflections and Intentions. Immediately after completing the tasks, we invited participants to reflect on their experience by answering open-ended questions about how the task influenced their understanding of fairness-related harms and whether they intended to take any concrete actions or revisit decisions in their own projects. These reflections allowed us to capture participants' intentions to act—serving as a proxy for how compelling, relevant, or actionable they found the experience.

Post-Study Follow-Up (2 Months Later). To assess whether the intervention led to sustained engagement or reflection, we conducted a follow-up survey two months after the main study. In the survey, participants were asked two open-ended questions:

- *Have you reacted to or done anything based on the findings from our session (e.g., changed anything in your past or new projects directly or indirectly based on concerns raised during our discussion)?*
- *Have you had any discussions—positive or negative—about responsible AI with your peers since the session?*

While insufficient to measure long-term transformations (which may require years), this still captures some concrete **behavioral and social impact over time**, even if modest, indicating whether participants engaged with RAI concepts in their professional communities beyond our study.

5.2 Data Analysis

To evaluate how practitioners engage differently across study conditions, we defined specific goals and aligned our data sources accordingly. Our primary goal was to understand engagement and reasoning in response to baseline

versus sticky stories. We operationalized this goal using targeted questions paired with corresponding metrics (Table 3), following the established Goal-Question-Metric (GQM) approach [6]. We combined quantitative metrics (e.g., harm diversity, time spent, user characteristics) with qualitative coding to characterize participants' engagement and reasoning processes (see the finalized variables in Table 4).

Table 3. Goal–Question–Metric (GQM) alignment for evaluating practitioner engagement and reasoning.

Goal	Question	Metric	Data Source
Understand engagement with stories	Do participants identify more potential harm categories and invest more time when exposed to sticky vs baseline stories?	Number of harms flagged, diversity of harms (coded according to harm taxonomy), time spent per task	Tool interaction logs
Understand reasoning processes	How do participants justify harm identification under baseline vs sticky story conditions?	Comparative coding for reasoning patterns	Think-aloud transcripts
Assess reflection and follow-up	Do participants act on insights or discuss them with peers?	Presence/absence of reported actions, types of discussions with peers	Follow-up survey responses

Quantitative Analysis. To evaluate whether practitioners invest more time when exposed to sticky versus baseline stories—a common proxy for engagement, e.g., [35]—we extracted the total time spent in each section along with the free-text harms from tool logs. To measure harm diversity, each harm was manually assigned to a category and subcategory from an established RAI taxonomy [17]. We assigned each participant a score regarding their prior RAI awareness and championship, based on self-reported experience in the pre-study survey and our assessment of their answers to our initial questions (revealed preferences), which we used as controls in our analysis. Table 4 summarizes the variables used in our study. We use ANOVA to analyze the influence of the experimental condition on the dependent variables. Model diagnostics—including checks for normality, homogeneity of variance, and influential points—were performed to ensure the validity of the analyses.

Qualitative Analysis. To understand *how* participants engaged with harm identification tasks—beyond what quantitative metrics could capture—we conducted *qualitative content analysis* [47, 60], a method that affords quantitative analysis of qualitatively coded data. We used a carefully designed codebook combining theory-driven indicators and patterns emerging from the transcripts (Table 1). In particular, deep engagement was coded using the indicators of critical reflection described in Section 4.1, supplemented with additional codes for shallow engagement. Two independent coders applied the codebook to the think-aloud transcripts, and inter-rater reliability was calculated to refine the codes, yielding a Cohen's κ of 0.72, which indicates substantial agreement. Codes were then applied systematically with each participant's transcript as a chunk of analysis, that is, multiple mentions by the same participant were not counted repeatedly. To analyze evolution of participant behaviors and identify trajectories, we used inductive qualitative coding: the first author systematically reviewed session notes and interaction logs, identified recurring patterns, and synthesized these into the practitioner profiles described in the paper. This data-driven approach enabled us to confidently assign all participants to profiles based on their observed behaviors, even though the process was exploratory rather than codebook-based.

Table 4. Study variables with type, operationalization, and data source. Icons indicate data source: ✎ = self-reported, 🖩 = derived, 🏷 = coded.

Type	Variable	Operationalization	Source
Independent	Story condition	No story (A) vs. baseline story (B) vs. sticky story (C)	–
Dependent (Quant.)	Time spent on harm identification	Duration of engagement, extracted from tool logs	🖩
Dependent (Quant.)	Number of harms	Number of identified harms, extracted and counted from tool logs	🖩
Dependent (Quant.)	Distinct harm categories	Number of unique categories/subcategories identified, coded with RAI harm taxonomy [17]	🏷
Dependent (Qual. & Quant.)	Engagement behaviors	Frequency of coded signs of critical reflection (Table 1), from think-aloud transcripts	🏷
Dependent (Qual.)	Self-reported planned actions	Reported and intended actions from follow-up survey responses	✎
Dependent (Qual.)	Transformation trajectories	Shifts in perspective, identified from transcripts and video recordings	🏷
Control	RAI awareness score	0–2 scale, based on in-session remarks: 0 = unaware, 1 = partially aware, 2 = aware	🏷
Control	Championship score	0–5 scale of advocacy for RAI principles in work 0 = unaware, 1 = actively opposed, 2 = acknowledges importance but takes no action, 3 = follows processes reluctantly, 4 = willing to contribute but not advocating, 5 = actively promotes RAI practices among peers	🏷
Control	Prior AI/ML experience	Self-reported in pre-screening survey: 1-2 years, 3-5 years, 6-10 years, more than 10 years	✎
Control	Task/fairness goal order	Randomly assigned	-

Follow-up survey responses were also examined to capture participants' self-reported takeaways and any indications—planned or already undertaken—of applying these insights in their own projects.

5.3 Limitations (Threats to Validity/Credibility)

As every study, ours needs to make tradeoffs and has limitations. Our study captures short-term engagement rather than lasting behavior change; even with a two-month follow-up, we cannot establish long-term effects. While our sample size is sufficiently powered to detect large effects, it is not large enough to explore differences across domains and organizations. Biases are possible at several levels: social desirability (despite neutral prompts, avoidance of "like/dislike" questions, and focusing on revealed preferences), bias in participant selection (even with broad, neutral framing, oversampling, screening for non-champions, and limited snowballing), and researcher experiences (including our formative study) shaping questions we ask and how we code data . Internally, the mixed design places the no-story control first for all participants, so any story is confounded with appearing second; learning, priming, fatigue, and carryover from the first fairness goal may influence observed gains. The think-aloud protocol can also alter problem-solving strategies and make engagement appear higher than in ordinary work. Our engagement proxies—time on task and counts of distinct harm categories—are imperfect; time can reflect confusion, and breadth can trade off with depth. Finally, while anchoring tasks in participants' own projects improves realism, it introduces heterogeneity that

Condition	Task	Time (min)	Harms	Harm categories	Harm subcategories
Baseline condition	Task 1 (no stories)	7.5 ± 4.3	1.46 ± 1.05	1.23 ± 0.73	1.38 ± 0.77
	Task 2 (baseline stories)	8.3 ± 3.6	1.31 ± 0.95	0.31 ± 0.48 (↑)	0.54 ± 0.66 (↑)
Sticky condition	Task 1 (no stories)	5.4 ± 2.7	1.31 ± 1.30	0.81 ± 0.66	1.06 ± 0.93
	Task 2 (sticky stories)	16.6 ± 7.4	2.31 ± 1.01	1.38 ± 0.62 (↑)	1.88 ± 0.62 (↑)

Table 5. Descriptive statistics for time on tasks, number of harms identified, and diversity of harms (categories and subcategories) across conditions. Values are reported as mean ± standard deviation (SD). Task 2 "Harm categories" and "Harm subcategories" indicate additional counts relative to Task 1 (↑)

	Relative time (tast 2/task 1)		Task 1 time		Task 2 time (ANCOVA)	
Source	F	p	F	p	F	p
Story condition (baseline vs sticky)	31.37***	<0.001	4.39*	0.047	28.47***	<0.001
Task/fairness goal order	0.60	0.445	1.01	0.324	2.48	0.130
RAI awareness score	1.07	0.313	2.13	0.158	5.11*	0.034
Championship score	0.03	0.855	0.20	0.656	2.81	0.108
Prior AI/ML experience	0.24	0.629	0.34	0.568	0.82	0.376
Task 1 time (cov.)	—	—	—	—	18.25***	0.0003

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6. Two-way ANOVA/ANCOVA results for time spent on Task 1 and Task 2. F-values and p-values are reported for all sources. Partial eta squared (η_p^2) for the main Story condition effect was 0.60 for Relative time, 0.08 for Task 1 time, and 0.33 for Task 2 time. Significant effects are indicated with stars.

can mask or mimic effects, and the study examines a single researcher-facilitated exposure rather than repeated use in everyday workflows. We deliberately designed the study, trading off various qualities and accepting some limitations; readers should interpret our results accordingly.

5.4 Findings

5.4.1 Finding 1: Practitioners spent significantly more time on harm identification tasks when sticky stories were shown. Participants with *sticky stories* for their second task spent substantially more time on harm identification tasks compared to those with *baseline stories*, with very large effects observed. Participants with baseline stories increased the time spent modestly over their task in the *no-story condition* (+11%, from 7.5 to 8.3 minutes), whereas participants with *sticky stories* spent nearly triple the time compared to their first task in the *no-stories condition* (+207%, from 5.4 to 16.6 minutes); see Table 5, Fig. 8.

Statistical analyses controlling for task order, experience, awareness, and championship confirm a very large effect of story type on the relative time increase between the first (no story) and second task (see Table 6). An additional analysis suggests that the time participants spend on the first task influences the time they spend on the second task (i.e., some participants are generally slower/more thorough than others), but story type accounted for substantially more variance, confirming that sticky stories associated clearly with more time spent on the second task.

While we anticipated some increase in time for sticky story participants, the magnitude of the increase even over baseline stories was surprising. Qualitative examination of tool logs, think-aloud transcripts, and video recordings revealed distinct patterns of engagement. Participants with baseline stories generally skimmed the stories, often moving quickly through the task without much deliberation. In contrast, participants with sticky stories typically processed each

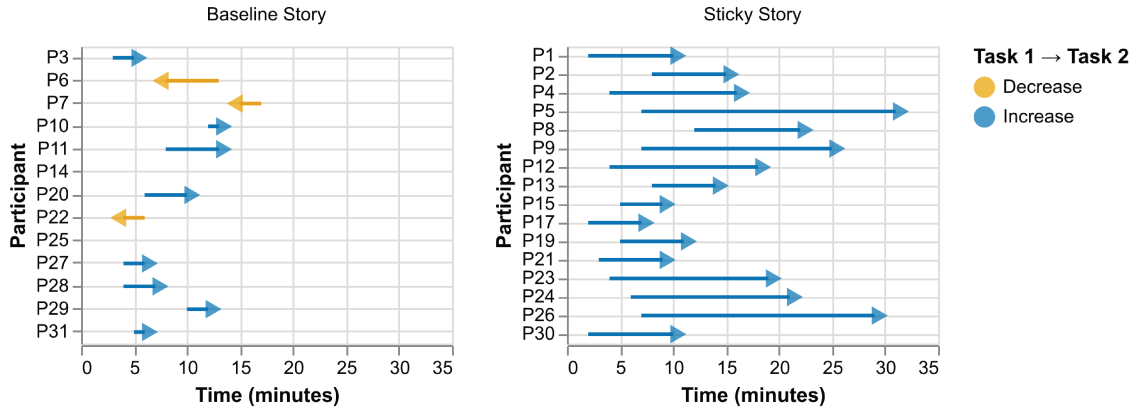

Fig. 8. We measured the time participants spent on the harm identification task 1 (no stories) and task 2 (baseline or sticky stories). We found that participants who read sticky stories consistently spent more time on the harm identification task, with much higher increases.

Source	Task 1 tax.		Task 2 tax. (ANCOVA)		Task 1 subtax.		Task 2 subtax. (ANCOVA)	
	F	p	F	p	F	p	F	p
Story condition (baseline vs sticky)	1.74	0.200	19.43***	0.0002	0.73	0.403	22.39***	0.0001
Task/fairness goal order	0.041	0.841	0.001	0.979	0.174	0.681	0.026	0.873
RAI awareness score	0.057	0.813	0.663	0.424	0.000003	0.999	0.004	0.950
Championship score	0.257	0.617	3.57	0.072	0.005	0.947	0.44	0.514
Prior AI/ML experience	0.119	0.734	0.000003	0.999	0.074	0.787	0.004	0.952
Task 1 tax.(cov.)	–	–	18.25***	0.0003	–	–	–	–
Task 1 subtax. (cov.)	–	–	–	–	–	–	0.018	0.896

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 7. ANOVA and ANCOVA results for taxonomy (tax.) and sub-taxonomy (subtax.) measures. F-values and p-values are reported for all sources. Partial η_p^2 for the main Condition effect are 0.09 (Task 1 tax.), 0.49 (Task 2 tax.), 0.036 (Task 1 subtax.), and 0.54 (Task 2 subtax.).

story sequentially, critically evaluating its applicability. Many paused to reflect on past incidents, consider stakeholder perspectives, or connect the story to broader systemic issues, indicating sustained cognitive engagement rather than superficial completion. We discuss this further in *Finding 4 and 5.*

5.4.2 Finding 2: Practitioners identified significantly more diverse harms when sticky stories were shown. Participants generally listed small numbers of harms as result from each task (typically 0 to 5 harms, cf. Table 5, Fig. 9). While we see an increase in the number of harms identified when sticky stories were provided (but not for baseline stories), we do not find counting the number of harms itself very meaningful, as participants might repeat very similar harms. Instead, we focused on the diversity of harms—measured as the number of distinct harm categories.

Participants exposed to *sticky stories* identified substantially more harm categories and subcategories than those in the *baseline stories condition*, corresponding to roughly 4.5× more new categories and 3.5× more new subcategories (see Table 5, Fig. 10). These differences remained statistically significant after controlling for task order, awareness, prior

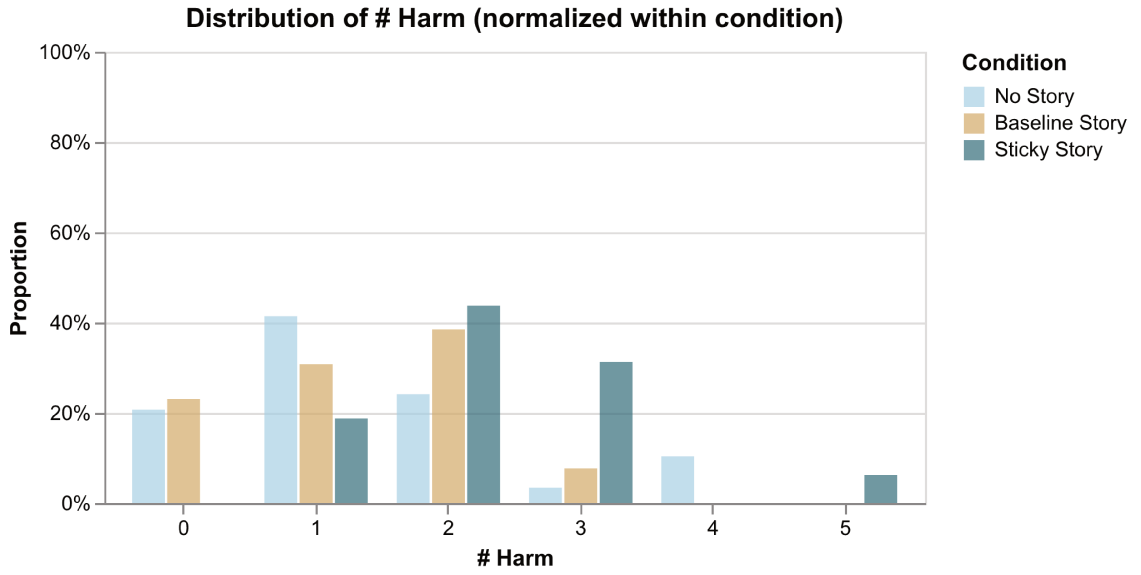

Fig. 9. The distribution of the number of harms identified before and after the intervention.

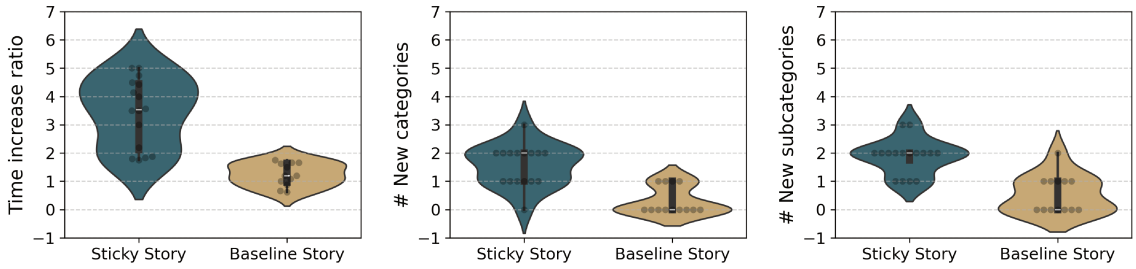

Fig. 10. We found that participants spent more time and identified more new harms in distinct categories or subcategories, after they read sticky stories.

championship, and experience in an ANOVA, suggesting that the observed increases were robust across participant backgrounds and prior familiarity with fairness concepts.

To understand the reasons behind the differences, we analyzed participants' responses before and after exposure to the stories, alongside their recorded interactions with them. Participants in the *sticky story condition* were more likely to surface new harms and expand across categories: of the six participants (***P4***, ***P15***, ***P17***, ***P21***, ***P25***, ***P30***) who left the harms section blank in the task 1 (*no-story condition)*, all but the *baseline* participant, added harms in task 2. Several also reconsidered fairness goals they had initially dismissed. For instance, ***P1*** began by rejecting the goal of minimizing stereotyping—"To be honest, I don't think this applies to this system"—but after reading a story, exclaimed, *"Oh, man, this is topical! My wife's [country] [...] this could be true*," and added stereotyping-related harms. All participants who expanded into multiple new harm categories (***P1***, ***P2***, ***P4***, ***P13***, ***P17***, ***P21***, ***P23***, ***P30***) were in the *sticky story condition*. In some cases, *sticky stories* also helped participants who were otherwise "stuck" in a single harm category—for instance,

P5 initially listed four harms exclusively under *Allocational Harms*, but in task 2, added one under Quality of Service Harms. By contrast, participants in the *baseline condition* often disengaged in task 2, leaving the section blank (**P25**, **P27**) or typing perfunctory responses (e.g., **P22**: "It is as was mentioned in the generated scenarios"). Even many who recorded harms (**P3**, **P10**, **P14**, **P19**, **P20**, **P22**, **P25**, **P27**, **P28**, **P31**) largely repeated categories they had already noted, showing little expansion of perspective.

5.4.3 Finding 3: Practitioners critically reflected on the sticky stories more, but only skimmed the baseline stories. Based on the distribution of critical reflection codes, participants in the sticky story group engaged more deeply and reflected more often than those in the baseline group (Fig. 11). Among the indicators of critical reflection (Table 1), the most common code was *expressing surprise or enthusiasm*, observed in 17 participants overall—15 of whom were from the *sticky story condition*. Participants expressed surprise with remarks such as *"oh, wow"* (**P15**, **P19**), *"Oh, my God, I mean, that could have caused an issue"* (**P9**), or *"I didn't really think about the fairness or responsible AI aspect of my system… I was really surprised"* (**P31**).

The second most frequent code was *connecting to the wider system or past incidents*, which also occurred more often in the *sticky story group* (9 participants) than in the baseline group (4 participants). Participants made these connections either by recalling past incidents—e.g., *"If I deploy [the biased model] then [...] it's gonna cost millions of dollars for internal systems. It happened some time back [...] [company] got shut down [...] millions of dollars of impact"* (**P23**) or anticipating wider implications.

Notably, behaviors such as *challenging assumptions (*e.g., **P26**: *"Now that I'm looking at the word demeaning [...] there's definitely some of this happening*"), *exploring multiple perspectives (*e.g., **P15**: "*how can a government or the FDA respond?*"*)*, and *iterative thinking* appeared exclusively in the *sticky story group*.

In contrast, signs of shallow engagement—such as skimming or dismissing the stories, or attempting to copy them directly as harms without further reasoning—were only observed in the baseline condition (e.g., **P7**, **P10**, **P25**, **P27**, **P28**, **P29**). These participants often spent minimal time with the scenarios and focused on completing the task quickly (consistent with *Finding 1*).

Six participants across both groups voiced some negative reactions to at least one story. However, the tone and follow-up differed markedly. Baseline participants tended to be dismissive—e.g., *"I find them very general"* (**P27**)—whereas sticky story participants more often contextualized or justified their critiques. For instance, **P26** initially found one story *"a little far-fetched"* but immediately connected it to their own experience of harassment, and **P5** expressed doubt but then qualified it with technical reasoning about their system's design.

5.4.4 Finding 4: Follow-up survey revealed more post-study actions than the sticky story group. We sent follow-up emails to all 29 participants. Nine participants responded—six from the sticky story group (**P1**, **P2**, **P9**, **P17**, **P19**, **P24**) and three from the baseline group (**P3**, **P14**, **P22**). While these responses are not sufficient to support statistical conclusions, we report them here for completeness and transparency.

Responses from the sticky story group consistently described concrete changes to practice, particularly in data collection and quality processes. These included recruiting more diverse participants for data collection (*P1*), conducting thorough safety checks before adding data for fine-tuning or training models using LLM-as-a-judge [132] (**P17**), enhancing data quality pipelines with expert review (**P19**), and adopting stricter labeling standards and rigorous annotation quality checks (**P24**). Many also reported ongoing discussions with colleagues about fairness, harm mitigation, and ethical risks.

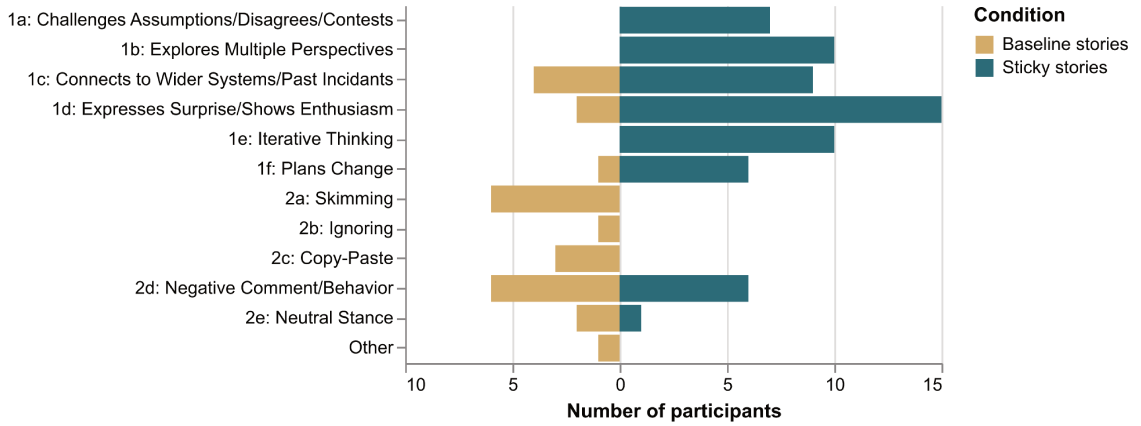

Fig. 11. Comparison of code frequencies per participant across baseline and sticky story conditions, highlighting increased reflection in the sticky story condition.

In contrast, baseline group responses were fewer and generally less action-oriented. One participant described an intention to "*use AI more responsibly*" without a concrete plan (*P3*), another reported having made no changes (*P22*), and a third (*P14*) mentioned future plans but had not yet acted, though they had shared key insights from the study with their organization's responsible AI team.

5.4.5 Finding 5: Practitioners exhibit distinct trajectories in shifting from initial indifference or resistance toward a more engaged stance on RAI. By examining participants' behaviors and narratives using an inductive coding approach, we posit five distinct engagement profiles. These profiles emerged from consistent patterns in how individuals responded to sticky stories, reflected on RAI issues, and engaged with harm identification tasks.

- *Resistors*: Explicitly push back against RAI as irrelevant or hype (***P17***, ***P30***).
- *Indifferents*: Acknowledge RAI is important but show little follow-through (***P2***, *P3*, ***P5***, *P6*, ***P21***, ***P26***).
- *Followers*: Follow RAI practices as their company enforces it (***P1***, *P14*, ***P19***, ***P23***, ***P24***, *P25*, *P27*).
- *Learners*: Have limited prior knowledge of RAI, but eager to learn (*P7*, ***P8***, ***P9***, *P10*, ***P15***, *P20*, *P22*, *P28*, *P31*).
- *Champions*: Already motivated and advocate of RAI (***P4***, *P11*, ***P12***, ***P13***, *P29*)

Resistors. We identified only two participants (***P17***, ***P30***) who were explicitly dismissive of RAI, describing it as hype and irrelevant to their work. This small number was not surprising, given the likelihood of *social desirability bias*—many who held similar sentiments may have instead presented themselves as *Indifferents*. Both resistors were in the sticky story group, meaning we cannot draw conclusions about how baseline participants in this category might have behaved. Notably, however, both individuals demonstrated a notable trajectory:

> Outright dismissal ⟶ Skepticism ⟶ Recognition of overlooked risks ⟶ Reframing as RAI-relevant issues (expanded awareness)

This trajectory illustrates how *sticky stories* can move even resistant practitioners toward expanded awareness, with *skepticism* emerging as the stories create *disorienting dilemmas* that challenge participants' existing views.

To illustrate this trajectory, we highlight the case of ***P17***, who initially rejected RAI as unrelated to their domain, saying, "*I don't really believe in it, to be honest. In my area of research, the societal harms are very low. [. . .] I'll give the cliché example—the prison system predicting recidivism across races. But that has nothing to do with code generation.*" This perspective carried into the *no-story condition*, where they spent only two minutes and recorded no harm. The introduction of *sticky stories* shifted this pattern. Although initially skeptical—*"It's hypothetically possible, but very implausible"*—they subsequently acknowledged overlooked risks, such as bias toward non-English codebases: *"Sure, you can overfit to English. This is actually true."* By the end, ***P17*** conceded that what they had considered "just tool issues" could fall under RAI: *"I didn't know that certain things fall under RAI."* This shows how sticky stories can provoke expanded awareness even among practitioners who begin with strong resistance.

Indifferents. Indifferents are participants whose stated preference in responsible AI (RAI) contrasted sharply with their revealed preference (i.e., they rate it as important in a survey, but their described past behaviors do not suggest active engagement). We speculate that they may acknowledge RAI's importance mostly in a general sense or due to social desirability. Within this profile, we observed distinct patterns between the *sticky story* and *baseline groups*. Participants in the *baseline group* (*P3*, *P6*) demonstrated minimal, superficial engagement, whereas those exposed to *sticky stories* (***P2***, ***P5***, ***P21***, ***P26***) engaged more deeply, reflected on the scenarios, and generated concrete, actionable plans, following the trajectory:

Indifference ⟶ Curiosity sparked ⟶ Critical reflection ⟶ Planned concrete actions

This curiosity may again stem from *disorienting dilemmas*, triggered by *diverse* (🎨) and *surprising* (😮) cases they had not anticipated.

To illustrate this trajectory, we highlight the case of ***P5***. Initially, without stories, ***P5*** spent 7 minutes on the first task, engaging mechanically and offering broad, vague assessments of potential harms, such as unreliable model performance or general stakeholder impacts. Their proposed mitigations were equally generic, such as improving system accuracy. After exposure to *sticky stories*, ***P5***'s engagement deepened substantially, spending 31 minutes carefully working through five stories (two initial and three additional ones). They revisited each story multiple times, expressed surprise and curiosity, and began considering a wider range of potential risks and systemic consequences. Through this process, ***P5*** transitioned from indifference to critical reflection, ultimately generating actionable plans they could realistically implement, showing how *sticky stories* can spark genuine engagement and meaningful planning even among initially indifferent participants.

Followers. These practitioners operate in organizations with established RAI infrastructures—mandatory harm assessment, governance teams, and formal review processes. Their engagement with RAI was often shaped more by institutional requirements than personal motivation, which often made the processes feel bureaucratic and burdensome. In the baseline group (*P14*, *P25*, *P27*), this compliance-driven stance generally persisted, with participants identifying only surface-level harms and routine mitigations (e.g., *P25*: "Okay, is this the end of it?"). By contrast, we saw a noticeable shift in all participants in the sticky story group (***P1***, ***P19***, ***P23***, ***P24***): The stories encouraged them to move beyond "checking the boxes" of oversight toward deeper reflection on risks, systemic consequences, and their own role in addressing them (e.g., ***P23***: *"I wasn't expecting it to go this deep… I really liked it."*). For some, this shift translated into a new sense of ownership and proactive responsibility, as they began connecting stories to gaps in their work and articulating concrete actions they intended to take. This reflects a trajectory similar to that of the Indifferents:

Compliance mindset ⟶ Assumptions challenged / Recognition of stake ⟶ Critical reflection ⟶ Proactive responsibility

We conjecture that *severity* (☠), *relevance* (🎯), and *concreteness* (🪪) of harm stories are particularly important to move participants beyond their routine view of RAI.

🎓 *Learners.* Participants in this profile entered with only a vague awareness of Responsible AI. Many had heard of fairness concerns in the abstract, but lacked concrete exposure to what such issues actually look like in practice. Within this profile, we saw two orientations: (a) some participants did not initially recognize the gaps in their knowledge or the need to deepen it (*P7*, ***P8***, ***P15***, *P31*), while (b) others openly admitted limited familiarity but expressed curiosity and willingness to explore (***P9***, *P10*, *P20*, *P22*, *P28*). Although the sample size is too small for statistical claims, anecdotally, participants in the sticky story group showed a larger learning shift compared to those in the baseline group. The engagement of learners suggested the following trajectory:

Unawareness ⟶ Recognition of gaps (group a) / Broadening of perspective (group b) ⟶ Proactive involvement

This mirrors the trajectory from unconscious incompetence to conscious incompetence in pegagogy literature [34]. We suspect *surprisingness* (😮) and *diversity* (🎨) of stories is particularly important here.

👤+ Champions. While our study primarily aimed to recruit non-champions, we classified some of our participants as champions when we talked to them (***P4***, *P11*, ***P12***, ***P13***, *P29*). These individuals did not inherit RAI responsibilities through a formal role but still were intrinsically motivated to pursue RAI, often driving initiatives without organizational incentives. Among this group, we did *not* observe substantial differences in harm identification across conditions. In both the baseline and sticky story conditions, most of them engaged with the stories and critically reflected on them. As we expected, champions are already motivated to engage deeply with RAI assessments, whereas non-champions (👤−, 👤, 👷, 🎓) seem to need a little push to spark meaningful reflection and action.

6 Discussion

Prior research in RAI has overwhelmingly focused on supporting "champions"—practitioners already motivated to advocate for and advance RAI principles within their organizations [46, 64, 87, 93, 94]. In line with this work, we show once more that AI-generated stories can support champions in fairness assessments. However, we also show, as our formative study and emerging HCI literature suggest [3, 81], that our baseline stories that mirror existing interventions are most effective for individuals who are already receptive or formally responsible for RAI work, but often fall short in engaging others.

Our sticky story intervention is designed to address this gap by operating at the psychological level: Aiming to capture attention, spark curiosity, and encourage critical self-examination. By shifting the focus from procedure to motivation and mindset, we envision sticky stories as a means to reframe RAI work from a bureaucratic obligation to a more vivid and memorable experience. In a way, our approach complements existing RAI tools and techniques, helping them reach and resonate with a wider range of practitioners, engaging them to a level where subsequently more traditional RAI tools or frameworks become effective.

6.1 Designing for Attitude Shifts: RAI Meets Psychology

Affecting long-term change is difficult. Past efforts to engage reluctant stakeholders have turned to nudges, gamification, or simplified documentation workflows [4, 10, 56, 115, 117]. While these strategies can improve initial uptake, some research suggests they do not reliably foster deeper shifts in mindset or sustained behavioral change [49, 90, 135]. In our work, we build on *Transformative Learning Theory* [72] and *Cognitive Dissonance Theory* [33] that suggest fundamental underlying strategies to achieve long-term change through *disorienting dilemmas* and *cognitive dissonance*—conditions necessary for critical reflection and potential transformation. While these theories explain *why* disorienting dilemmas enable transformation, *Made to Stick* complements them by explaining *how* to design such dilemmas so they reliably capture attention and remain memorable. *Sticky stories* intervene precisely at this level: they present practitioners with novel, surprising, and severe consequences grounded in their own systems, eliciting the discomfort and curiosity that precede deeper reflection.

While we cannot actually measure long-term effects with our experiments, we can observe signs of exactly the kind of mechanisms and trajectories when participants engage with sticky stories that the theories predict: Transformation, according to Transformative Learning Theory [72], is a multi-stage long journey beginning with a (a) *disorienting dilemma* that (b) *challenges prior assumptions*, followed by (c) *self-examination* and (d) *critical assessment* of those assumptions, (e) *recognition of shared experiences* with others, (f) *exploration of new roles*, (g) *planning a course of action*, (h) *trying out new roles*, (i) *building competence* in those roles, and (j) finally *reintegration* of these changes into one's perspective and behavior. We observed participant actions aligning with (a) to (f) repeatedly and saw signs of (g), which is an encouraging sign that the underlying mechanisms for long-term change might also work here. Research also suggests that sustaining mindset change typically requires repeated reinforcement, ongoing organizational support, and integration into routine practices [89, 125], which goes beyond the scope of our research.

6.2 Who Benefits Most from the Sticky Stories, and How?

The few champions (👤+) in our study were already highly engaged with RAI tasks – they benefited from sticky stories, but were also just as motivated to work through the baseline stories and engaged deeply in the task even without stories. In contrast, non-champions (👤−, 👤, 👤, 🤔) showed noticeable changes in attitudes and behaviors when interacting with sticky stories, with increases in time spent, harms identified, and critical reflection. This suggests that while existing interventions may be effective among already motivated individuals, non-champions benefit from additional prompts to trigger *disorienting dilemmas*—and sticky stories appear to provide that push.

Moreover, the observed trajectories suggest that different story qualities, such as severity and surprisingness, may resonate differently with different kinds of non-champions. *Learners* (🤔) may benefit most from diverse (🎨) stories that expose them to new possibilities; *followers* (👤) may be more influenced by severe (☠) stories that highlight potential consequences; *indifferents* (👤) may respond primarily to surprising (😮) stories that challenge expectations; and *resistors* (👤−) may require a combination of surprising (😮) and relevant (🎯) stories that create both dissonance and relevance to shift their engagement. We do not imply that any single quality is sufficient—transformation may depend on multiple qualities acting together—but certain qualities may have stronger effects depending on practitioner type. This points to an important direction for future work: Systematically evaluating how individual story qualities affect different practitioner profiles.

6.3 Practical Risks and Limitations in Story Design

While sticky stories show clear benefits for attention and engagement in our experiment, they may introduce new ethical, practical, and methodological risks, and so their use must be critically examined on several fronts.

6.3.1 Do Ethics Need to be Extreme to Elicit Reflection? A recurring tension in our findings concerns the role of drama or emotional salience in prompting ethical reflection. Within the RAI and applied ethics literature, several scholars emphasize that meaningful ethical practice often unfolds through routine, quiet, and procedural deliberation rather than dramatic events or extreme scenarios [11, 64, 88, 108]. From this perspective, extreme interventions may not be necessary, and an overemphasis on them could risk obscuring the subtle, cumulative, and structural harms that characterize many real-world AI failures.

At the same time, research in psychology and cognitive science consistently shows that attention, memory, and sensemaking are disproportionately shaped by vividness, novelty, and emotional resonance [7, 113, 133]. Work in communication, journalism, and marketing similarly suggests that people are more likely to notice, recall, and act upon messages that contain surprising or consequential elements [8, 41, 44]. These insights raise important questions for RAI practice: In fast-paced professional environments where attention is scarce, do vivid or "dramatic" scenarios trigger effective deeper engagement or are they ultimately distracting? Reliance on extreme stories is also likely not an effective long-term strategy for regular interventions, as practitioners likely grow numb and fatigued to such stories. Ideally, extreme introductions would capture attention initially and change minds for the long term, thus enabling a nuanced and detail-oriented, and often drama-free, subsequent analysis beyond the initial stories.

6.3.2 Representational and Emotional Risks of Narrative. Narrative interventions raise critical questions about representation, power, and emotional impact. HCI and AI ethics scholars have long warned that narratives can unintentionally sensationalize harms, reproduce stereotypes, or reinforce dominant perspectives while obscuring more structural or less visible forms of injustice [21, 110]. Similarly, design and computing research has shown that storytelling practices can marginalize certain voices or frame communities through deficit-oriented lenses [21]. When narrative generation is delegated to LLMs, these risks are amplified: Generative systems are known to encode and reproduce structural biases in their training data, privileging some identities, harms, and cultural frames while rendering others less visible [36, 59, 111]. As a consequence, harm stories created with LLMs may skew toward familiar tropes, overemphasize dramatic but already highly visible cases, or underrepresent the slow, infrastructural, and context-specific damages documented in sociotechnical and critical data studies [38, 74]. This pattern raises important questions about whose experience is centered or erased in AI harm assessment processes, and how automated storytelling might perpetuate these imbalances.

In our sticky story pipeline, we sought to reduce these risks by using structured prompting and deliberately steering generation away from default narrative paths—such as by using initial LLM outputs as counter-examples for "surprisingness," and systematically varying stakeholders, demographic attributes, and harm categories. This strategy broadened the narrative space and reduced surface-level repetition. Nonetheless, such approaches can only partially mitigate systemic tendencies baked into LLMs: Even with targeted prompting, generative models may still gravitate toward dominant cultural logics or overlook subtle and infrastructural harms [45, 74, 112, 131]. Therefore, we argue that embedding narrative interventions in RAI practice requires ongoing, deliberate efforts—including participatory curation, engagement with affected communities, and the development of robust ethical criteria for story inclusion [20, 48, 80, 86]—to ensure these tools do not simply reproduce, but instead challenge, dominant patterns of harm

and representation. As narrative-based tools become more widespread, addressing these risks is essential to avoid reinforcing the very injustices RAI seeks to address.

6.3.3 The Double-Edged Sword of Severity and Concreteness. While leveraging severity and concreteness in narrative interventions can heighten attention and promote reflection, stories that stray too far from practitioners' real-world experience or drift into science fiction territory risk undermining engagement. In our user study, some participants found the consequences depicted in some sticky stories too unrealistic to take seriously. For example, P26 remarked that one story "seems a little far-fetched," and P17 described another as "very implausible...wouldn't happen in practice." Unfortunately, the technical nature of LLMs makes it very challenging to ensure realism and avoid "hallucinations.". That is, independent of whether a story is severe or dramatic, an unrealistic story poses a risk not only to credibility, but to the perceived relevance of RAI generally.

The underlying problem is technical and hard to overcome, but an active area of LLM research. Many strategies have been explored to reduce "hallucinations" and keep generated text more realistic, such as retrieval-augmented generation (RAG) to induce real-world knowledge and context into the generation process (as done in Farsight [124]). In addition, better story generation pipelines could incorporate plausibility-checking agents, human-in-the-loop evaluation, or automated fact-checking workflows to help filter out implausible scenarios before presentation. Combining generated sticky stories with links to actual news articles, if found, or case studies might further help practitioners anchor reflection in real-world stakes, reinforcing connections to prior work on scenario-based harm assessment.

Importantly, not every story needs to perfectly map to real failures. While improving realism can help ensure narratives are accepted as relevant prompts for reflection, rather than dismissed as speculative fiction, at the same time, even less realistic stories may retain value as creative catalysts, highlighting overlooked risk categories and opening space for broader ethical consideration. In fact, we observed frequently how participants dismissed a specific story, but used it as a jumping-off point to explore a related but more realistic concern in their specific project.

Looking ahead, we see value not only in technical work, such as enhancing the realism, contextual relevance, and fact-grounding of LLM-generated narratives, but also in conceptual research to better understand how the presentation of stories to practitioners influences how they react, as well as exploring how factors like severity, concreteness, or surprisingness influence practitioner engagement and decision-making. Identifying the sweet spot between realism, attention-grabbing, and not-quite-real but useful inspiration for creative analysis will be a useful direction, requiring both HCI and ML research.

6.3.4 Workflow and Productivity Concerns. Our study found that integrating sticky stories into harm reflection substantially increased the time practitioners spent on these tasks. This finding echoes prior HCI and CSCW research suggesting that interventions fostering deeper engagement or reflection often come with a trade-off in efficiency or throughput [5, 106, 109]. While deeper reflection is widely viewed as desirable in ethics and RAI [12, 32, 50, 85], in industry, incentives often align more with high-velocity production and quick deployment, which slows adoption of RAI practices to the deep frustration of RAI champions [1, 51, 94, 121]. It is not obvious which role tools like ours can play in an environment with misaligned incentives. We optimistically hope that sticky stories change minds and get developers to personally buy into RAI practices; these developers might further use the sticky stories to convince others in their team or management chain to invest into RAI practices. At that point, existing complementary approaches to supporting responsible AI practices in organizations [43, 57, 78, 130] could take over.

Overall, in recognizing these various risks, we advocate for the careful and ethically-informed integration of sticky stories into RAI practice. This calls for ongoing curation of story pipelines, ethical oversight, and empirical study

of practitioner well-being and organizational impact. Stories should be paired with concrete "what now?" options, ensuring that practitioners can move from recognition to action. Importantly, sticky stories are most effective when they complement existing tools and processes—helping practitioners appreciate the significance of ethical issues across the full spectrum of ML systems.

References

[1] Ali, S.J., Christin, A., Smart, A. and Katila, R. 2023. Walking the walk of AI ethics: Organizational challenges and the individualization of risk among ethics entrepreneurs. *arXiv [cs.CY]*.

[2] Baker-Brunnbauer, J. 2021. Management perspective of ethics in artificial intelligence. *AI and ethics*. 1, 2 (2021), 173–181.

[3] Balayn, A., Yurrita, M., Yang, J. and Gadiraju, U. 2023. "Fairness toolkits, A checkbox culture?" on the factors that fragment developer practices in handling algorithmic harms. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2023), 482–495.

[4] Ballard, S., Chappell, K.M. and Kennedy, K. 2019. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. *Proceedings of the 2019 on Designing Interactive Systems Conference* (2019), 421–433.

[5] Bardzell, J. and Bardzell, S. 2013. What is "critical" about critical design? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2013).

[6] Basili, V., Heidrich, J., Lindvall, M., Münch, J., Regardie, M., Rombach, D., Seaman, C. and Trendowicz, A. 2014. GQM+strategies: A comprehensive methodology for aligning business strategies with software measurement. *arXiv [cs.SE]*.

[7] Baumeister, R.F., Bratslavsky, E., Finkenauer, C. and Vohs, K.D. 2001. Bad is stronger than good. *Review of general psychology: journal of Division 1, of the American Psychological Association*. 5, 4 (Dec. 2001), 323–370.

[8] Berger, J. and Milkman, K.L. 2012. What makes online content viral? *JMR, Journal of marketing research*. 49, 2 (Apr. 2012), 192–205.

[9] Bessen, J., Impink, S.M. and Seamans, R. 2022. The cost of ethical AI development for AI startups. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), 92–106.

[10] Bhat, A., Coursey, A., Hu, G., Li, S., Nahar, N., Zhou, S., Kästner, C. and Guo, J.L.C. 2023. Aspirations and Practice of ML Model Documentation: Moving the Needle with Nudging and Traceability. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), 1–17.

[11] Bietti, E. 2021. From ethics washing to ethics bashing: A moral philosophy view on tech ethics. *Journal of social computing*. 2, 3 (Sep. 2021), 266–283.

[12] Binns, R. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. *Conference on Fairness, Accountability and Transparency* (Jan. 2018), 149–159.

[13] Bogucka, E., Constantinides, M., Šćepanović, S. and Quercia, D. 2024. Co-designing an AI impact assessment report template with AI practitioners and AI compliance experts. *arXiv [cs.HC]*.

[14] Boren, T. and Ramey, J. 2000. Thinking aloud: reconciling theory and practice. *IEEE transactions on professional communication*. 43, 3 (2000), 261–278.

[15] Boyd, K. 2022. Designing up with value-sensitive design: Building a field guide for ethical ML development. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 2069–2082.

[16] Brookfield, S.D. 2017. *Becoming a critically reflective teacher*. Jossey-Bass.

[17] Buçinca, Z., Pham, C.M., Jakesch, M., Ribeiro, M.T., Olteanu, A. and Amershi, S. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. *arXiv [cs.HC]*.

[18] Chang, J. and Custis, C. 2022. Understanding implementation challenges in machine learning documentation. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (2022), 1–8.

[19] Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian: *https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf*.

[20] Commons-based Data Set Governance for AI: *https://openfuture.eu/publication/commons-based-data-set-governance-for-ai*. Accessed: 2025-12-05.

[21] Costanza-Chock, S. 2020. *Design Justice: Community-Led Practices to Build the Worlds We Need*. MIT Press.

[22] Dastin, J. 2022. Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women. *Ethics of Data and Analytics*. Auerbach Publications. 296–299.

[23] Deng, W.H., Barocas, S. and Wortman Vaughan, J. 2025. Supporting industry computing researchers in assessing, articulating, and addressing the potential negative societal impact of their work. *Proceedings of the ACM on human-computer interaction*. 9, 2 (2025), 1–37.

[24] Deng, W.H., Guo, B.B., Devos, A., Shen, H., Eslami, M. and Holstein, K. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), 1–18.

[25] Deng, W.H., Nagireddy, M., Lee, M.S.A., Singh, J., Wu, Z.S., Holstein, K. and Zhu, H. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022), 473–484.

[26] Deng, W.H., Yildirim, N., Chang, M., Eslami, M., Holstein, K. and Madaio, M. 2023. Investigating practices and opportunities for cross-functional collaboration around AI fairness in industry practice. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. (2023), 705–716.

[27] Dominique, B., Maghraoui, K.E., Piorkowski, D. and Herger, L. 2023. FactSheets for hardware-aware AI models: A case study of analog in memory computing AI models. *Proceedings of the 2023 IEEE International Conference on Software Services Engineering (SSE)* (2023), 148–158.

[28] Dual-coding theory: 2025. *https://en.wikipedia.org/wiki/Dual-coding_theory*.

[29] Ehsan, U., Liao, Q.V., Passi, S., Riedl, M.O. and Daumé, H., III 2024. Seamful XAI: Operationalizing seamful design in Explainable AI. *Proceedings of the ACM on human-computer interaction*. 8, CSCW1 (2024), 1–29.

[30] Elsayed-Ali, S., Berger, S. E., Santana, V. F. D., & Becerra Sandoval, J. C. 2023. Responsible & inclusive cards: An online card tool to promote critical reflection in technology industry work practices. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), 1–14.

[31] Ericsson, K.A. and Simon, H.A. 1993. *Protocol Analysis*. MIT Press.

[32] F. Clancy, R., Zhu, Q. and Majumdar, S. 2025. Exploring AI ethics in global contexts: a culturally responsive, psychologically realist approach. *AI and ethics*. 5, 6 (Dec. 2025), 6329–6338.

[33] Festinger, L. 1957. *A Theory of Cognitive Dissonance*. Stanford University Press.

[34] Four stages of competence: 2025. *https://en.wikipedia.org/wiki/Four_stages_of_competence*.

[35] Fredricks, J.A., Blumenfeld, P.C. and Paris, A.H. 2004. School engagement: Potential of the concept, state of the evidence. *Review of educational research*. 74, 1 (2004), 59–109.

[36] Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R. and Ahmed, N.K. 2024. Bias and fairness in large language models: A survey. *Computational linguistics (Association for Computational Linguistics)*. (Aug. 2024), 1–83.

[37] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D. and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*. 64, 12 (2021), 86–92.

[38] Ghosh, S., Venkit, P.N., Gautam, S., Wilson, S. and Caliskan, A. 2024. Do generative AI models output harm while representing non-Western cultures: Evidence from A community-centered approach. *arXiv [cs.CY]*.

[39] Gravett, S. 2002. Transformative Learning through Action Research: A Case Study from South Africa. *Adult Education Research Conference* (2002).

[40] Green, M.C. and Appel, M. 2024. Narrative transportation: How stories shape how we see ourselves and the world. *Advances in Experimental Social Psychology*. Elsevier. 1–82.

[41] Green, M.C. and Brock, T.C. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology*. 79, 5 (Nov. 2000), 701–721.

[42] Hadi Mogavi, R., Guo, B., Zhang, Y., Haq, E.-U., Hui, P. and Ma, X. 2022. When gamification spoils your learning: A qualitative case study of gamification misuse in a language-learning app. *Proceedings of the Ninth ACM Conference on Learning @ Scale* (2022).

[43] Hadley, E., Blatecky, A. and Comfort, M. 2025. Investigating algorithm review boards for organizational responsible artificial intelligence governance. *AI and ethics*. 5, 3 (Jun. 2025), 2485–2495.

[44] Heath, D. and Heath, C. 2009. *Made to Stick*. Random House Trade.

[45] Hida, R., Kaneko, M. and Okazaki, N. 2025. Social bias evaluation for large language models requires prompt variations. *Findings of the Association for Computational Linguistics: EMNLP 2025* (Stroudsburg, PA, USA, 2025), 14507–14530.

[46] Holstein, K., Vaughan, J.W., Daumé, H., III, Dudík, M. and Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019), 1–16.

[47] Hsieh, H.-F. and Shannon, S.E. 2005. Three approaches to qualitative content analysis. *Qualitative health research*. 15, 9 (2005), 1277–1288.

[48] Hsu, Y.-C., Huang, T.-H. 'kenneth', Verma, H., Mauri, A., Nourbakhsh, I. and Bozzon, A. 2022. Empowering local communities using artificial intelligence. *Patterns (New York, N.Y.)*. 3, 3 (Mar. 2022), 100449.

[49] Hummel, D. and Maedche, A. 2019. How effective is nudging? A quantitative review on the effect sizes and limits of empirical nudging studies. *Journal of behavioral and experimental economics*. 80, (Jun. 2019), 47–58.

[50] Ibitoye, A.O., Nkwo, M.S. and Orji, R. 2025. Rethinking responsible AI from ethical pillars to sociotechnical practice. *AI and ethics*. 5, 6 (Dec. 2025), 6207–6223.

[51] Işık, Ö. and Goswami, A. 2025. The Three Obstacles Slowing Responsible AI. *MIT Sloan Management Review*. (Oct. 2025).

[52] Itti, L. and Baldi, P. 2009. Bayesian surprise attracts human attention. *Vision research*. 49, 10 (2009), 1295–1306.

[53] Kallina, E., Bohné, T. and Singh, J. 2025. Stakeholder participation for responsible AI development: Disconnects between guidance and current practice. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (2025), 1060–1079.

[54] Kaur, H., Conrad, M.R., Rule, D., Lampe, C. and Gilbert, E. 2024. Interpretability gone bad: The role of bounded rationality in how practitioners understand machine learning. *Proceedings of the ACM on human-computer interaction*. 8, CSCW1 (2024), 1–34.

[55] Kihlstrom, J.F. 2021. Ecological validity and "ecological validity." *Perspectives on psychological science: a journal of the Association for Psychological Science*. 16, 2 (2021), 466–471.

[56] Kim, S.-E., Kim, K., Lee, J., Ko, Y., Wang, Y. and So, H.-J. 2025. Dilemmas in AI ethics: A digital game for moral reasoning and collective decision-making. *Proceedings of the International Conference on Artificial Intelligence in Education* (Cham, 2025), 434–447.

[57] Laine, J., Minkkinen, M. and Mäntymäki, M. 2024. Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. *Information & management*. 61, 5 (Jul. 2024), 103969.

[58] Lanne, M., Nieminen, M. and Leikas, J. 2025. Organisational tensions in introducing socially sustainable AI. *AI & society*. (2025). DOI:https://doi.org/10.1007/s00146-025-02293-y.

[59] Large Language Models generate biased content, warn researchers: 2024. *https://www.ucl.ac.uk/news/2024/apr/large-language-models-generate-biased-content-warn-researchers*. Accessed: 2025-12-05.

[60] Lazar, J., Feng, J.H. and Hochheiser, H. 2017. *Research Methods in Human-Computer Interaction*. Morgan Kaufmann.

[61] Leveson, N.G. 2016. *Engineering a safer world: Systems thinking applied to safety*. The MIT Press.

[62] Liao, Q.V., Subramonyam, H., Wang, J. and Wortman Vaughan, J. 2023. Designerly understanding: Information needs for model transparency to support design ideation for AI-powered user experience. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), 1–21.

[63] Louviere, J.J., Hensher, D.A. and Swait, J.D. 2014. *Stated choice methods*. Cambridge University Press.

[64] Madaio, M.A., Stark, L., Wortman Vaughan, J. and Wallach, H. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (2020), 1–14.

[65] Madaio, M., Egede, L., Subramonyam, H., Wortman Vaughan, J. and Wallach, H. 2022. Assessing the fairness of AI systems: AI practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on human-computer interaction*. 6, CSCW1 (2022), 1–26.

[66] Madaio, M., Kapania, S., Qadri, R., Wang, D., Zaldivar, A., Denton, R. and Wilcox, L. 2024. Learning about responsible AI on-the-job: Learning pathways, orientations, and aspirations. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024), 1544–1558.

[67] Mann, K., Gordon, J. and MacLeod, A. 2009. Reflection and reflective practice in health professions education: a systematic review. *Advances in health sciences education: theory and practice*. 14, 4 (2009), 595–621.

[68] Martelaro, N. and Ju, W. 2020. What could go wrong? Exploring the downsides of autonomous vehicles. *Proceedings of the 12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (2020).

[69] Massaro, D.W., Petty, R.E. and Cacioppo, J.T. 1988. Communication and Persuasion: Central and Peripheral Routes to Attitude Change. *The American journal of psychology*. 101, 1 (1988), 155.

[70] McAdams, D.P. 2011. Narrative Identity. *Handbook of Identity Theory and Research*. Springer New York. 99–115.

[71] Mezirow, J. 2000. *Learning as transformation: Critical perspectives on a theory in progress*. Jossey-Bass.

[72] Mezirow, J. 1991. *Transformative dimensions of adult learning*. Jossey-Bass.

[73] Mezirow, J. 2018. Transformative learning theory. *Contemporary Theories of Learning*. Routledge. 114–128.

[74] Mickel, J., De-Arteaga, M., Liu, L. and Tian, K. 2025. More of the same: Persistent representational harms under increased representation. *arXiv [cs.CL]*.

[75] Microsoft RAI Impact Assessment Template: *https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-RAI-Impact-Assessment-Template.pdf*.

[76] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D. and Gebru, T. 2019. Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), 220–229.

[77] ml-practical-usecases: A database of 450 Machine Learning (ML) system design case studies from 100+ companies: *https://github.com/mallahyari/ml-practical-usecases*. Accessed: 2025-09-09.

[78] Mokander, J. and Floridi, L. 2024. Operationalising AI governance through ethics-based auditing: An industry case study. *arXiv [cs.CY]*.

[79] Nathan, L.P., Friedman, B., Klasnja, P., Kane, S.K. and Miller, J.K. 2008. Envisioning systemic effects on persons and society throughout interactive system design. *Proceedings of the 7th ACM conference on Designing interactive systems* (2008), 1–10.

[80] Odhiambo, J.M. and Ondimu, K. 2025. A framework for ethical AI-generated content governance. *Preprints*.

[81] Omar, Z.A., Nahar, N., Tjaden, J., Gilles, I.M., Mekonnen, F., Hsieh, J., Kästner, C. and Menon, A. 2025. Beyond Accuracy, SHAP, and Anchors – On the difficulty of designing effective end-user explanations. *arXiv [cs.HC]*.

[82] O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

[83] OpenAI et al. 2023. GPT-4 Technical Report. *arXiv [cs.CL]*.

[84] Pang, R.Y., Santy, S., Just, R. and Reinecke, K. 2024. BLIP: Facilitating the exploration of undesirable consequences of digital technologies. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), 1–18.

[85] Pant, A., Hoda, R., Spiegler, S.V., Tantithamthavorn, C. and Turhan, B. 2023. Ethics in the Age of AI: An Analysis of AI Practitioners' Awareness and Challenges. *ACM Transactions on Software Engineering and Methodology*. (Dec. 2023). DOI:https://doi.org/10.1145/3635715.

[86] Parthasarathy, A., Phalnikar, A., Jauhar, A., Somayajula, D., Krishnan, G.S. and Ravindran, B. 2024. Participatory approaches in AI development and governance: A principled approach. *arXiv [cs.CY]*.

[87] Passi, S. and Barocas, S. 2019. Problem Formulation and Fairness. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, Jan. 2019), 39–48.

[88] Passi, S. and Sengers, P. 2020. Making data science systems work. *Big Data & Society*. 7, 2 (Jul. 2020), 2053951720939605.

[89] Paunesku, D., Walton, G.M., Romero, C., Smith, E.N., Yeager, D.S. and Dweck, C.S. 2015. Mind-set interventions are a scalable treatment for academic underachievement. *Psychological science*. 26, 6 (2015), 784–793.

[90] Polman, E. and Maglio, S.J. 2024. The Problem With Behavioral Nudges. *The Wall Street Journal*. The Wall Street Journal.

[91] Popular Machine Learning Applications and Use Cases in our Daily Life: 2019. *https://www.analyticsvidhya.com/blog/2019/07/ultimate-list-popular-machine-learning-use-cases/*. Accessed: 2025-09-09.

[92] ProjectPro, B.Y. 2021. 15 Machine Learning Use Cases and Applications in 2025. *ProjectPro*.

[93] Raji, I.D., Smart, A., White, R.N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020), 33–44.

[94] Rakova, B., Yang, J., Cramer, H. and Chowdhury, R. 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), 1–13.

[95] Reimers, N. and Gurevych, I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019).

[96] Responsible AI: *https://www.ibm.com/trust/responsible-ai*. Accessed: 2025-09-11.

[97] Responsible AI: Ethical policies and practices: *https://www.microsoft.com/en-us/ai/responsible-ai*. Accessed: 2025-09-11.

[98] Responsible AI Maturity Model: 2023. *https://www.microsoft.com/en-us/research/publication/responsible-ai-maturity-model/*. Accessed: 2025-09-11.

[99] Responsible AI Transparency Report: *https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/msc/documents/presentations/CSR/Responsible-AI-Transparency-Report-2024.pdf*. Accessed: 2025-09-11.

[100] Richardson, B., Garcia-Gathright, J., Way, S.F., Thom, J. and Cramer, H. 2021. Towards fairness in practice: A practitioner-oriented rubric for evaluating fair ML toolkits. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–13.

[101] Rismani, S. and Moon, A. 2023. What does it mean to be a responsible AI practitioner: An ontology of roles and skills. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2023), 584–595.

[102] Rismani, S., Shelby, R., Smart, A., Jatho, E., Kroll, J., Moon, A. and Rostamzadeh, N. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), 1–18.

[103] Rodrigues, L., Pereira, F.D., Toda, A.M., Palomino, P.T., Pessoa, M., Carvalho, L.S.G., Fernandes, D., Oliveira, E.H.T., Cristea, A.I. and Isotani, S. 2022. Gamification suffers from the novelty effect but benefits from the familiarization effect: Findings from a longitudinal study. *International journal of educational technology in higher education*. 19, 1 (2022), 13.

[104] Ryan, C.L., Cant, R., McAllister, M.M., Vanderburg, R. and Batty, C. 2022. Transformative learning theory applications in health professional and nursing education: An umbrella review. *Nurse education today*. 119, 105604 (Dec. 2022), 105604.

[105] Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P. and Aroyo, L.M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1–15.

[106] Schiavo, G., Mich, O., Ferron, M. and Mana, N. 2022. Trade-offs in the design of multimodal interaction for older adults. *Behaviour & information technology*. 41, 5 (Apr. 2022), 1035–1051.

[107] Schoen, D.A. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

[108] Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, Jan. 2019), 59–68.

[109] Sengers, P., Boehner, K., David, S. and Kaye, J. 'jofish' 2005. Reflective design. *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility* (New York, NY, USA, Aug. 2005), 49–58.

[110] Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., Yilla-Akbari, N. 'mah, Gallegos, J., Smart, A., Garcia, E. and Virk, G. 2023. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (2023), 723–741.

[111] Shieh, E., Vassel, F.M., Sugimoto, C.R. and Monroe-White, T. 2025. Laissez-faire harms: Algorithmic biases in generative language models (extended abstract). *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 8, 3 (Oct. 2025), 2373–2374.

[112] Sivakumar, N., Mackraz, N., Khorshidi, S., Patel, K., Theobald, B.-J., Zappella, L. and Apostoloff, N. 2025. Bias after prompting: Persistent discrimination in large language models. *Findings of the Association for Computational Linguistics: EMNLP 2025* (Stroudsburg, PA, USA, 2025), 18568–18593.

[113] Slovic, P., Finucane, M.L., Peters, E. and MacGregor, D.G. 2007. The affect heuristic. *European journal of operational research*. 177, 3 (2007), 1333–1352.

[114] Smith, J.J., Madaio, M., Burke, R. and Fiesler, C. 2025. Pragmatic fairness: Evaluating ML fairness within the constraints of industry. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (2025), 628–638.

[115] The board game: 2025. *https://tethics.eu/the-board-game/*. Accessed: 2025-09-09.

[116] The building blocks of Microsoft's responsible AI program: 2021. *https://blogs.microsoft.com/on-the-issues/2021/01/19/microsoft-responsible-ai-program/*. Accessed: 2025-08-04.

[117] The Ethical Dilemmas Board Game: *https://cfrr.worldbank.org/publications/ethical-dilemmas-board-game*. Accessed: 2025-09-09.

[118] The Shift from Models to Compound AI Systems: *http://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/*. Accessed: 2025-08-27.

[119] Vaast, E. 2025. Experiencing and addressing the moral ambivalence of developing digital technology: Insights from artificial intelligence developers. *Proceedings of the Annual Hawaii International Conference on System Sciences* (2025).

[120] Vakkuri, V., Kemell, K.-K., Tolvanen, J., Jantunen, M., Halme, E. and Abrahamsson, P. 2022. How do software companies deal with artificial intelligence ethics? A gap analysis. *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering* (2022), 100–109.

[121] Varanasi, R.A. and Goyal, N. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 2023), 1–17.

[122] Wang, A., Datta, T. and Dickerson, J.P. 2024. Strategies for increasing corporate responsible AI prioritization. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2024), 1514–1526.

[123] Wang, Q., Madaio, M., Kane, S., Kapania, S., Terry, M. and Wilcox, L. 2023. Designing responsible AI: Adaptations of UX practice to meet responsible AI challenges. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (2023), 1–16.

[124] Wang, Z.J., Kulkarni, C., Wilcox, L., Terry, M. and Madaio, M. 2024. Farsight: Fostering responsible AI awareness during AI application prototyping. *Proceedings of the CHI Conference on Human Factors in Computing Systems* (2024), 1–40.

[125] Watkins-Hayes, C. 2019. *Remaking a life*. University of California Press.

[126] Widder, D.G., Dabbish, L., Herbsleb, J.D. and Martelaro, N. 2024. Power and play: Investigating "license to critique" in teams' AI ethics discussions. *Proceedings of the ACM on human-computer interaction*. 8, CSCW2 (2024), 1–23.

[127] Widder, D.G. and Nafus, D. 2023. Dislocated accountabilities in the "AI supply chain": Modularity and developers' notions of responsibility. *Big data & society*. 10, 1 (2023). DOI:https://doi.org/10.1177/20539517231177620.

[128] Winecoff, A.A. and Watkins, E.A. 2022. Artificial concepts of artificial intelligence. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), 788–799.

[129] Winecoff, A. and Bogen, M. 2025. Improving governance outcomes through AI documentation: Bridging theory and practice. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (2025), 1–18.

[130] Xia, B., Lu, Q., Perera, H., Zhu, L., Xing, Z., Liu, Y. and Whittle, J. 2023. Towards concrete and connected AI risk assessment (C2AIRA): A systematic mapping study. *arXiv [cs.SE]*.

[131] Yang, X., Zhan, R., Wong, D.F., Yang, S., Wu, J. and Chao, L.S. 2025. Rethinking prompt-based debiasing in large language models. *arXiv [cs.CL]*.

[132] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E.P., Zhang, H., Gonzalez, J.E. and Stoica, I. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Advances in neural information processing systems*. 36, (2023), 46595–46623.

[133] Zillmann, D. 2006. Exemplification effects in the promotion of safety and health. *The Journal of communication*. 56, suppl_1 (Aug. 2006), S221–S237.

[134] Ziosi, M. and Pruss, D. 2024. Evidence of what, for whom? The socially contested role of algorithmic bias in a predictive policing tool. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (2024), 1596–1608.

[135] 2024. The problem with the nudge effect: it can make you buy more carrots – but it can't make you eat them. *The Guardian*. The Guardian.