AI LEGO: Scaffolding Cross-Functional Collaboration in Industrial Responsible AI Practices during Early Design Stages

MUZHE WU*, Carnegie Mellon University, USA and University of Michigan, USA
YANZHI ZHAO*, Carnegie Mellon University, USA
SHUYI HAN, Northwestern University, USA
MICHAEL XIEYANG LIU, Carnegie Mellon University, USA
HONG SHEN, Carnegie Mellon University, USA

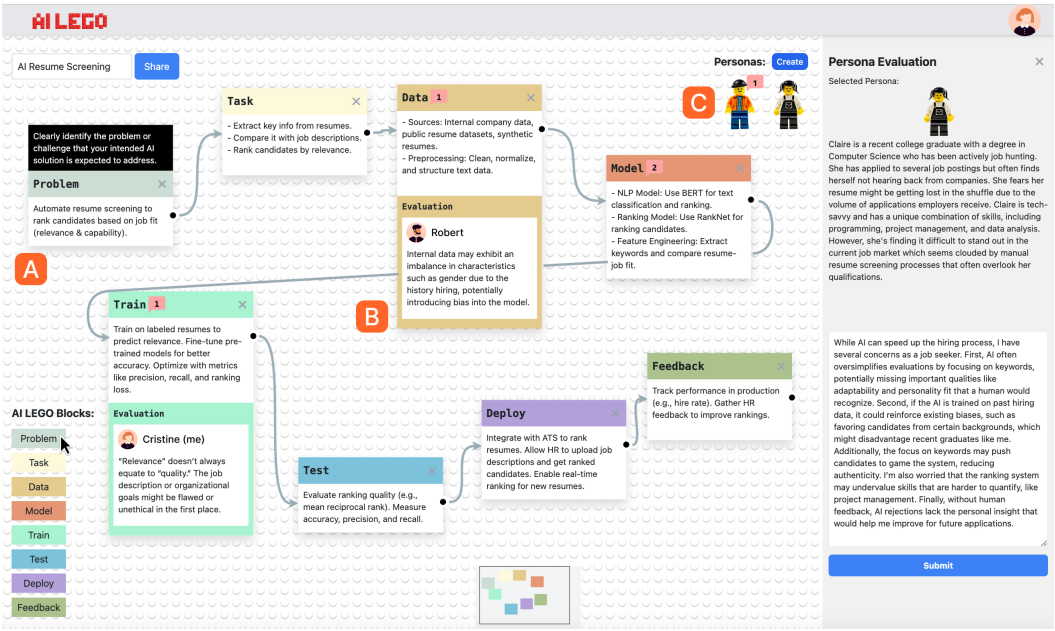

Fig. 1. AI LEGO is a web tool that scaffolds cross-functional collaboration in industrial AI development planning and proactive harm identification. With AI LEGO, technical AI developers can sketch out a development plan using (A) *Lifecycle Blocks*, each corresponding to a specific AI development stage with a tailored prompt. They hand off the plan to non-technical/user-facing roles, who conduct (B) *Stage-centered Evaluation* to systematically identify and address potentially harmful design choices at each stage. Lastly, (C) *Persona-centered Evaluation* feature helps generate stakeholder personas to surface edge cases by simulating diverse perspectives. Minifigure images © seewhatmitchsee / iStock by Getty Images. Used under license.

*Both authors contributed equally to this research.

Authors' Contact Information: Muzhe Wu, Carnegie Mellon University, Pittsburgh, PA, USA and University of Michigan, Ann Arbor, MI, USA, muzhewu@gmail.com; Yanzhi Zhao, Carnegie Mellon University, Pittsburgh, PA, USA, yanzhiz@alumni.cmu.edu; Shuyi Han, Northwestern University, Evanston, IL, USA, shuyihan2026@u.northwestern.edu; Michael Xieyang Liu, Carnegie Mellon University, Pittsburgh, PA, USA, lxieyang.cmu@gmail.com; Hong Shen, Carnegie Mellon University, Pittsburgh, PA, USA, hongs@cs.cmu.edu.

Responsible AI (RAI) efforts increasingly emphasize the importance of addressing potential harms early in the AI development lifecycle through social-technical lenses. However, in cross-functional industry teams, this work is often stalled by a persistent coordination challenge: how technical roles hand off technical intent, how teams establish shared structures for collaboration, and how non-technical roles are supported in systematically evaluating harms. Through literature review and a semi-structured interview study with 8 practitioners, we unpack how this challenge manifests—technical design choices are rarely handed off in ways that support meaningful engagement by non-technical roles; collaborative workflows lack shared, visual structures to support mutual understanding; and non-technical practitioners are left without scaffolds for systematic harm evaluation. Existing tools like JIRA or Google Docs, while useful for product tracking, are ill-suited for supporting joint harm identification across roles, often requiring significant extra effort to align understanding. To address this, we developed AI LEGO, a web-based prototype that operationalizes the boundary object theory to support cross-functional AI practitioners in effectively facilitating knowledge handoff and identifying harmful design choices in the early design stages. Technical roles use interactive blocks to draft development plans, while non-technical roles engage with those blocks through stage-specific checklists and LLM-driven persona simulations to surface potential harms. In a study with 18 cross-functional practitioners, AI LEGO increased the volume and likelihood of harms identified compared to baseline worksheets. Participants found that its modular structure and persona prompts made harm identification more accessible, fostering clearer and more collaborative RAI practices in early design.

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**; *Empirical studies in collaborative and social computing*; *Interactive systems and tools*.

Additional Key Words and Phrases: Responsible AI, Human-AI Collaboration, Cross-functional Collaboration, AI Evaluation

ACM Reference Format:
Muzhe Wu, Yanzhi Zhao, Shuyi Han, Michael Xieyang Liu, and Hong Shen. 2026. AI LEGO: Scaffolding Cross-Functional Collaboration in Industrial Responsible AI Practices during Early Design Stages. *Proc. ACM Hum.-Comput. Interact.* 10, 6, Article CSCW049 (October 2026), 33 pages. https://doi.org/10.1145/3816897

1 Introduction

Artificial Intelligence (AI)-driven systems, powered by Machine Learning (ML) techniques, have exercised power over many aspects of our everyday lives, from deciding which social media posts we see, the quality of healthcare we receive, to the ways we are represented to potential employers [29, 34, 54, 55]. The rise of Generative AI has further amplified this influence. Yet they have also raised pressing concerns, as public outcries and scholarly work have increasingly documented the ways these systems can, whether inadvertently or intentionally, perpetuate or even amplify existing societal biases or create new harmful impacts [9, 16, 68, 72].

Alongside the establishment of high-level governmental policies and frameworks (e.g., NIST's AI Risk Management Framework in the United States [73]), researchers in Responsible AI (RAI) and Human-Computer Interaction (HCI) have developed various technical tools and interactive systems to support model evaluation, auditing, and debiasing [10, 12]. Despite their significant contribution, this line of work has often been criticized for (1) failing to surface certain "blind spots", often due to the lack of diversity in *technical* AI teams [25, 40]; and (2) being conducted only after the system is built or deployed, when significant harms had already occurred [26, 74].

Recent CSCW and HCI research has increasingly emphasized the socio-technical nature of AI systems and the importance of involving non-technical/user-facing roles, such as designers, product/project managers (PMs), and policy stakeholders, in RAI work [18, 37, 44]. In industry, such work is often carried out by *cross-functional AI teams*, where participants bring diverse forms

ACM 2573-0142/2026/10-ARTCSCW049
https://doi.org/10.1145/3816897

of expertise [27]. Prior studies show that involving non-technical roles in these teams can help surface problematic design assumptions in the early development process [75, 78].

However, existing work also indicates that early-stage collaboration in cross-functional AI teams is not simply a matter of participation, but of coordination [1, 57, 64, 81]. When AI designs are still evolving and only partially specified, teams must collaboratively reason about technical intent, constraints, and potential downstream impacts across roles with asymmetric expertise. While prior research has identified communication challenges in such settings [3, 27, 53, 58], how collaborative work can be structured to support this form of early-stage RAI coordination, before implementation decisions become fixed, remains insufficiently specified [48, 61, 78].

Through a formative interview study with eight cross-functional AI practitioners, we examine how early-stage RAI collaboration breaks down in practice. When AI designs are still evolving and only partially specified, teams must collaboratively reason about technical intent, design trade-offs, and potential downstream harms across roles with asymmetric expertise. Our study surfaced three recurring challenges that shape this coordination work. First, technical design rationales and constraints are often difficult to externalize in ways that support meaningful engagement by non-technical collaborators, leading to breakdowns in knowledge handoff [27, 45]. Second, collaborative practices frequently rely on ad hoc documentation and communication, leaving teams without shared structure to support discussion, reflection, and iteration [58, 61]. Third, even when design information is available, non-technical practitioners often lack scaffolds for systematically interpreting design choices and evaluating potential downstream harms [75, 78]. Together, these challenges point to a broader coordination gap in early-stage cross-functional RAI collaboration.

Building on CSCW theories of articulation work [64] and boundary objects [69], we view these challenges as intertwined aspects of a coordination problem rather than isolated breakdowns. From this perspective, harm identification is not a separate or downstream activity, but depends on whether technical intent can be articulated, shared, and interrogated across roles. Supporting early-stage RAI collaboration therefore calls for tools that integrate articulation, shared collaborative structure, and role-appropriate scaffolds for interpretation and evaluation within a single workflow.

This interpretation leads us to ask the following question: **How can boundary objects be designed to support early-stage cross-functional RAI collaboration across roles with asymmetric technical expertise?** We derived three design goals to address the challenges: (1) supporting articulation and handoff of technical design intent, (2) providing shared and flexible structures to organize cross-functional collaboration, and (3) scaffolding systematic harm evaluation by non-technical/user-facing roles. We instantiated these goals in AI LEGO (Fig. 1), an interactive web tool that operationalizes the boundary object theory [69] to help cross-functional AI practitioners more effectively communicate high-level design ideas and identify potential problematic design choices that may lead to downstream harms in the early design stage.

We further conducted an evaluation user study with 18 cross-functional AI practitioners grouped into 6 teams to investigate (1) the effectiveness of our tool in helping cross-functional practitioners with RAI work in the early design stage; (2) the value of our tool over existing RAI practices for cross-functional AI practitioners. The results show that AI LEGO helped AI practitioners identify an average of 195% more problematic design choices that would lead to downstream harms with a higher likelihood of occurring (1.25-point increase in median on a 4-point scale), compared to using Google Docs and Harm Modeling framework [51] baseline. Participants also reported that AI LEGO was effective in bridging gaps in knowledge handoff, providing a flexible collaborative structure, and supporting multi-perspective harm evaluation.

The contributions of this work include:

- **Design goals** from literature review and a formative semi-structured interview study to tackle the challenges of scaffolding early-stage cross-functional RAI collaborations, including

knowledge handoff from technical roles to non-technical/user-facing roles, lack of shared collaborative structure, and limited scaffolds for harm evaluation.
- **AI LEGO, an interactive web tool** to support cross-functional AI practitioners in effectively facilitating knowledge handoff and identifying harmful design choices in the early design stage, when issues are easier to fix without causing actual harm.
- **Empirical findings** from an evaluation study on using AI LEGO during early-stage cross-functional RAI collaboration practices. AI LEGO helped participants identify more high-quality potential harms in AI designs and is more useful compared to existing resources.
- **Design implications** on supporting effective cross-functional collaboration in conducting RAI work, especially in the early-stage AI design.

2 Related Work

We situate our work at the intersection of (1) RAI tools and methods that support anticipating and addressing harms, and (2) CSCW research on coordination across heterogeneous roles through shared representations, articulation work, and boundary objects. We first review RAI toolkits and early-stage design methods, then examine cross-functional RAI collaboration as a coordination problem, and finally discuss harm identification and stakeholder reasoning approaches that inform our system design.

2.1 Developing techniques, toolkits and systems for Responsible AI in industry

Over the years, the RAI community in HCI/CSCW has made significant contributions in developing techniques, toolkits, and systems to support industry practitioners in conducting RAI-related work across a wide range of AI systems [4, 20, 40, 49, 76]. Many early tools focused on supporting technically trained practitioners in evaluating trained models, with a particular emphasis on fairness, bias, and performance trade-offs [10, 12, 77]. While these tools have been influential in operationalizing RAI principles, prior work has noted that they often surface issues late in the development lifecycle, when core design choices, such as data sources, proxies, or task definitions, are already difficult to revise [26, 74].

Beyond model-centric tools, a parallel line of work has developed ethics toolkits and design methods aimed at supporting broader, cross-functional reflection on ethical and social impacts. These include card-based and workshop-oriented approaches such as Envisioning Cards [33], Value Cards [67], and practitioner-facing toolkits such as IDEO's AI Ethics Cards [41]. Such methods explicitly seek to engage designers, developers, and business stakeholders in collaborative discussions of ethical considerations early in the design process. However, prior work also cautions that ethics toolkits can implicitly encode assumptions about who performs ethics work and how ethical concerns translate into accountable, coordinated action, potentially leaving gaps around workflow integration, ownership, and power dynamics [78].

Recent work has begun exploring earlier-stage RAI interventions that move upstream in the design process. For example, Farsight [76] supports harm anticipation during early prototyping of LLM-powered applications, and Lam et al. [43] introduce frameworks for iteratively authoring approximations of model decision logic. While these approaches push RAI considerations earlier, they are typically designed for technical practitioners and do not explicitly structure handoff or evaluation across cross-functional roles. Our work builds on these efforts by focusing on how early-stage design artifacts can be structured to support collaborative reasoning and evaluation across roles with asymmetric expertise.

2.2 Cross-functional collaboration in Responsible AI

CSCW has long examined how collaborative work is coordinated across heterogeneous roles, emphasizing articulation work, shared information spaces, and the persistent negotiation required to align specialized forms of expertise [1, 64, 70, 71]. In organizational settings, cross-functional collaboration often depends on shared representations that allow participants to coordinate despite differences in goals, vocabularies, and evidentiary standards.

Empirical studies of industry AI development show that such coordination is particularly challenging in AI work, where design decisions span technical, organizational, and social considerations. Passi and Jackson describe how corporate data science projects rely on continual "translation" work to make technical decisions intelligible and accountable to non-technical stakeholders [57]. Zhang et al. [81] and Piorkowski et al. [58] further document how AI teams depend on a patchwork of tools, documents, and meetings to coordinate across roles, often adapting general-purpose artifacts (e.g., tickets, shared documents) to communicate design intent and constraints. However, these adaptations are fragile: they frequently leave key assumptions implicit, privilege technical framings, or require substantial interpretive labor from non-technical collaborators.

Prior literature highlights that cross-functional collaboration around fairness and ethics introduces additional layers of complexity. Studies show that fairness concepts are interpreted differently across roles, that evidence takes multiple forms (e.g., metrics, anecdotes, lived experience), and that responsibility for ethical decisions is often diffuse or contested [27, 44, 61]. Wong et al. [78] further argue that many RAI tools implicitly assume how ethical work should proceed, without adequately accounting for how teams actually coordinate this work over time. Together, this literature suggests that participation alone is insufficient: effective early-stage RAI work depends on how collaboration is structured to support articulation, interpretation, and negotiation across roles with asymmetric expertise.

A core CSCW concept for understanding such coordination is *boundary objects*—shared yet flexible artifacts that enable collaboration across epistemic and professional boundaries [17, 69]. CSCW work has examined how boundary objects support coordination in settings such as healthcare and collaborative planning, where artifacts must remain interpretable and actionable for participants with different expertise [62, 63]. Importantly, boundary objects are not "solutions" on their own: they often require ongoing articulation work to keep representations meaningful and aligned with situated practice [1, 64]. In RAI contexts, boundary objects have been used to support shared reflection and accountability. For example, Madaio et al. [48] have proposed to contextualize fairness checklists into cross-functional AI teams' collaborative workflows and balance individual ownership of fairness decisions with the oversight of a designated approver. However, existing work rarely specifies how these artifacts are actually worked through over time—how they are handed off, interpreted, and evaluated as AI design decisions evolve across development stages and roles.

We developed AI LEGO to further explore and operationalize boundary objects as part of a structured, staged workflow that explicitly supports knowledge handoff [21, 45], collaborative evaluation, and iteration in early-stage RAI work. Rather than treating coordination as an implicit outcome of discussion, we examine how coordination can be designed for through shared artifacts that evolve as design decisions are examined across roles.

2.3 Identifying harmful problems in AI systems, products and services

Identifying harmful or unethical issues in technology has been a long-standing concern in HCI [7, 32, 47]. As AI systems have become more integrated into society, addressing these concerns has taken on even greater importance. Over the years, a wide range of design methods and toolkits have been developed to support designers, developers, and data scientists in recognizing and anticipating

potential harms in AI systems [8, 28, 33, 49, 67]. For example, Raji et al. [60] define a meta-level, end-to-end framework for AI system audit, with the goal to ensure the accountability of the system prior to deployment. On the other hand, Qiang et al. [59] conducted a series of comparative studies and found that different AI ethics frameworks serve complementary purposes and are most effective when applied in combination, depending on the context.

One prominent approach to supporting RAI evaluation is the use of personas and stakeholder representations [19, 23, 67]. Personas provide concrete stand-ins for affected groups and serve as boundary objects that help teams reason about social impacts and lived experiences [23]. Prior work, such as Value Cards toolkit [67], incorporates Persona Cards alongside Model Cards and Checklist Cards, demonstrating effectiveness in facilitating discussions on the social impacts of ML models. With the rise of generative AI models like LLMs, this approach can be further augmented to accelerate and inspire human evaluation in RAI work. For example, Buçinca et al. [15] introduced a general framework that assists AI practitioners and decision-makers in anticipating potential harms of AI systems, using LLMs to generate descriptions of potential harms for various stakeholders. In a similar vein, Farsight [76] leverages LLMs to help practitioners build LLM-based applications to ideate different use cases, identify stakeholders, and consider potential harms.

We build on and extend this body of work by integrating it into the context of cross-functional collaboration at the early design stage. AI LEGO introduces two scaffolding techniques specifically designed to help non-AI developers more effectively identify problematic AI design choices early in the process. First, we developed a stage-based checklist to support harm anticipation, which covers key design choices across the entire AI development lifecycle for its effectiveness [42]. Second, building on prior work [15, 76], we incorporated LLMs to assist non-AI developers in brainstorming potentially impacted stakeholders and gaining a deeper understanding of their lived experiences. By treating these scaffolds as boundary objects, integrated into a shared workflow that structures how AI design decisions are examined across roles, we explore how they can support interpretive alignment while also surfacing tensions between machine-generated suggestions and human ethical reasoning.

3 Formative Study and Design Goals

To better understand (1) the challenges that interdisciplinary teams face when identifying potential harms in the early stage of AI development, and (2) how features in cross-functional collaboration tools or frameworks can be developed to support this process, we conducted a formative semi-structure interview study[1] as detailed below.

3.1 Semi-Structured Interview

We recruited 8 industrial AI practitioners (4 female, 4 male) via social media and mailing lists. These participants held diverse roles across three major categories (4 AI developers[2], 2 PMs, and 2 UI/UX designers) [22] and work at companies of varying sizes[3]. Participants were screened and selected based on their experience in cross-functional AI development. They were not expected to have prior experience in RAI, as informed by prior work [48], reflecting the typical demographics of AI practitioners aiming for broader understanding. All participants were based in the United States due to logistical constraints and to ensure alignment with regional industry practices. Each participant was compensated with a $35 USD Amazon gift card. Each study session lasted approximately 60

[1]The formative study along with the evaluation study were approved by our institution's IRB (STUDY2023_00000435).
[2]In this work, we broadly use the term "AI developers" to refer to practitioners who implement, train, or maintain AI/ML systems, including roles such as AI/ML Engineer, Data Scientist, and AI/ML Research Scientist.
[3]Four from companies with 5,000–24,999 employees, and one each from companies with 1,000–4,999, 250–999, 50–249, and fewer than 50 employees.

minutes and focused on identifying specific hurdles a participant has faced in interdisciplinary teamwork and exploring how different tool designs can enable efficient early-stage collaboration to mitigate problematic AI design. We began by introducing the goal of our study, then asked participants to describe their current workflows and challenges encountered in cross-functional collaboration between technical and non-technical/user-facing roles. This helped us understand their existing practices and pain points. We then presented five preliminary design artifacts (see Appendix A for the designs and their rationales), created by the researchers prior to the session, to the participants. These artifacts included two approaches for surfacing AI development plans: plain description and stage-based scaffolding inspired by AI storyboarding [24], and three UIs for brainstorming potential harms: a survey adapted from Value Cards [67], stakeholder table, and commenting. Participants were asked for their opinions on whether and how these designs could support their current workflows, their perceived usability, and suggestions for improvement (see Appendix B for the full semi-structured interview protocol).

3.2 Findings

We analyzed participants' (P1–8) notes with affinity diagrams [46]. Two authors independently reviewed all collected materials and generated initial observations by clustering related comments, sketches, and annotations. Through an iterative process, the authors collaboratively grouped these observations into three higher-level clusters:

3.2.1 Articulating and Handing Off Technical Intent. Participants consistently described early-stage cross-functional collaboration as challenging not simply due to a lack of communication, but due to difficulties in articulating technical intent in forms that non-technical collaborators could meaningfully engage with. While participants noted that communication flows and patterns within AI teams can vary depending on factors such as team size (P2), organizational or business type (P4, P8), and individual differences (P3), most of them ($n = 8$) agreed that initiating an AI development plan is best led by technical roles or practitioners with strong technical expertise to ensure feasibility (e.g., P3 noted, "*My PM often overestimates what the model can achieve.*") This arrangement, however, would often result in breakdowns during handoff to non-technical or user-facing roles ($n = 5$).

Several participants ($n = 3$) emphasized that existing documentation and planning artifacts were either overly technical or fragmented, making it difficult for non-technical collaborators to understand how design decisions were made or where potential risks might arise. For example, P6 noted that "*explicating data construction in plain language would be a prerequisite for success*," reflecting a broader need for accessible representations of technical decisions. In the absence of such representations, some non-technical participants reported a strong dependence on technical roles for interpretation and decision-making ($n = 2$), limiting their autonomy in contributing to early design discussions.

Participants also highlighted how commonly used coordination tools—such as JIRA [6], Asana [5], and Google Docs [36]—were poorly suited for this form of early-stage RAI work. These tools were described as either too rigid to support fast iteration or too generic to accommodate discussions of values and harms. As P8 remarked, "*JIRA takes a lot of time to set up and isn't suitable for fast turnarounds.*" As a result, teams often relied on ad hoc explanations or synchronous meetings to establish shared understanding, which several participants ($n = 4$) found inefficient or unnecessary.

3.2.2 Establishing Shared Structure for Early-stage AI Planning. Participants expressed a strong preference for structured yet flexible representations to support early-stage AI development planning across roles. Comparing the two artifacts, all participants ($n = 8$) preferred stage-based scaffolding, consistently viewing it as an effective structure for coordinating work and establishing a shared frame of reference. Participants believed that stage-based scaffolding allowed technical developers

to clearly outline tasks and dependencies within AI development, while non-technical members benefited from its clear and modularized visuals. According to P2, "stage-based representation standardizes the language of AI development planning", thus being more easily graspable for both technical and non-technical/user-facing roles. The stage-level breakdown also "facilitates frequent iteration" (P1), which is especially helpful in early-stage AI practices.

3.2.3 Scaffolding Interpretive Work in Harm Identification. Participants ($n = 4$) appreciated the straightforwardness and in-depth investigation of the survey adapted from Value Card [67], albeit the amount of reading appeared discouraging (P1, P2). In contrast, the stakeholder table had a clear information visualization ($n = 4$), allowing for "both horizontal and vertical comparisons" (P1) across stages and stakeholders for the comprehensiveness of evaluation. Commenting was highlighted for its ability to foster open-ended discussions and effectively link identified problems to specific design choices ($n = 3$), facilitating faster iteration. Ideal harm identification scaffolding should integrate these characteristics.

3.3 Design goals

Combining the insights from the literature review and the semi-structured interview study, we establish the following design goals:

- **G1 Support articulation and handoff of technical intent.** The system should enable technical roles to externalize early AI development decisions at an appropriate level of abstraction, allowing non-technical and user-facing collaborators to meaningfully engage with and evaluate.
- **G2 Provide shared, flexible structure for collaborative planning.** The system should offer visual, modular representations that organize early AI development work while remaining adaptable to different project contexts and collaboration styles.
- **G3 Scaffold interpretive evaluation by non-technical and user-facing roles.** The system should integrate structured yet open-ended evaluation supports that help non-technical practitioners reason about potential harms across development stages and stakeholder perspectives, while preserving space for exploratory and situated judgment.

4 AI LEGO

Guided by the design goals, we developed AI LEGO, an interactive web tool that helps cross-functional AI practitioners collaboratively identify harmful design choices in the early stages of AI design. We begin by showcasing a sample scenario of a cross-functional team collaborating on an RAI project to demonstrate the AI LEGO workflow.

4.1 Example Usage Scenario

A startup movie platform is developing a personalized recommendation model. John, an AI developer, is tasked with creating the development plan but struggles to present it to the rest of the team. Emma, the product manager, is worried about potential biases, like recommendations influenced by gender or age, associated with problematic design decisions.

They decide to use AI LEGO for support. John starts outlining the key AI development stages in the project using AI LEGO's *Lifecycle Blocks* (Fig. 1A). The embedded *Eight-Stage Worksheet* prompts him to input details for each stage, creating a clear visual map of the development plan. When John shares the plan with Emma, the clear visual framework and tailored prompts immediately provide Emma with an overarching understanding of the development plan. Emma then uses AI LEGO's *Stage-centered Evaluation* (Fig. 1B) to review key stages like "Dataset Construction" and "Model Definition." The tool's *Eight-Stage Checklist* guides her in identifying risks, such as sensitive data like gender potentially causing biased recommendations. She suggests anonymizing this data but

Table 1. The Eight-Stage Worksheet and Checklists offer prompts tailored to different stages of AI development, guiding AI developers in drafting the development plan and assisting non-technical/user-facing roles in identifying problematic design choices.

Stage Name	Worksheet Prompt	Checklist Prompt
Problem Formulation	Clearly identify the problem or challenge that your intended AI solution is expected to address.	Is the problem or challenge itself ethical? Can the intended AI solution provide a viable solution to the identified problem?
Task Definition	Detail the specific task(s) that your intended AI solution is designed to perform to solve the previously identified problem.	What are the boundaries and limitations of what the AI solution is expected to achieve?
Dataset Construction	Describe where and how the data for training your intended AI solution is collected and prepared, and explain the data preprocessing steps and any feature engineering technologies that are applied to the raw data. If the intended AI solution is a generative AI, describe the fine-tuning dataset you're using.	Are there any potential biases, privacy concerns, and other ethical considerations in data handling?
Model Definition	Describe the AI model architecture, the proxies selected, and the algorithms used, explaining their roles and functionalities. If the intended AI solution is a generative AI, describe the base model you're using, the prompting techniques, and the fine-tuning methods.	Is the choice of models, proxies, and algorithms appropriate for the task identified? If the intended AI solution is a generative AI, is the choice of the base model, the prompting techniques, and the fine-tuning methods appropriate for the task identified?
Training	Describe how the AI solution is trained using the curated data, and clarify the process of how it improves its performance.	Can you think of ways the training process might go wrong? If so, how?
Testing	Explain how the AI solution is evaluated and highlight the key performance metrics used to assess the AI.	What would it mean for the AI to be successful? Are the performance metrics sufficient for evaluating the success of AI?
Feedback	Outline how feedback is collected, specifying who the feedback is gathered from and how often the feedback is collected.	Does the deployment of the AI fit with the real-world practical use environments?

raises concerns about its impact on accuracy. To explore further, Emma uses the *Persona-centered Evaluation* (Fig. 1C) to assess stakeholder impact. She creates a few personas, including a teenage movie enthusiast, which helps her realize the need for age-appropriate recommendations. This prompts a reassessment of how sensitive data is handled to balance fairness and accuracy.

With these insights, John revises the development plan, adjusting the dataset and model design to address ethical concerns. After a few iterations, John and Emma finalized a well-rounded plan ready for actual development.

4.2 Eight-stage Worksheet & Checklist

Eight-Stage Worksheet and *Checklist* (Table 1) are both lists of prompt questions designed to, respectively, elicit high-level descriptions of key design choices in AI development from technical roles and support non-technical/user-facing roles in potential harm investigation. Based on findings from prior research [64, 65] and our formative study findings on the effectiveness of "staging", we developed prompts that address eight stages of the AI development lifecycle: problem formulation, task definition, dataset construction, model definition, training, testing, and feedback. These stages are adapted from Cramer et al.'s seven-stage workflow [24], with the addition of an explicit problem formulation stage. Prior work [42] has similarly incorporated this stage and highlights its critical role in shaping downstream technical decisions in RAI practice [56]. The curation of prompt questions took an iterative process, ensuring that they not only capture socio-technical aspects,

but also scaffold effective articulation of technical intent or non-technical/user-facing concerns and meaningful reflection around crucial aspects of each stage.

4.3 User Interface

AI LEGO further integrates *Eight-Stage Worksheet* and *Checklist* into a visual, interactive interface consisting of the following components (see Fig. 1; UI transitions detailed in Fig. 5 in Appendix C):

Lifecycle Blocks. The *Lifecycle Blocks* are stage-based visual building blocks used to map out the development plan of an intended AI system (G2). Technical AI developers can create blocks in the central canvas by clicking on the stage button in the bottom left corner (Fig. 1A). When they click on the block's textbox, a corresponding prompt from the *Eight-Stage Worksheet* will appear above, guiding them to fill in the relevant descriptions. Once the descriptions are completed, they can link the blocks and reposition them using drag-and-drop functionality to establish a flow or structure. The canvas expands infinitely in all directions, allowing developers to add as many blocks and branches as needed for maximum flexibility.

Stage-centered Evaluation. *Stage-centered Evaluation* enables non-technical/user-facing roles to examine the individual stages in a drafted plan to identify potential problematic design choices (G3). Users start by selecting a block and pressing the "Stage" button in the bottom-right corner. This generates a sidebar displaying the corresponding prompt from the *Eight-Stage Checklist* and allowing users to describe the problematic designs. The completed evaluation is automatically saved, attached to the corresponding stage block, and shared within the team (Fig. 1B).

Persona-centered Evaluation. To support perspective-taking and reflective brainstorming during early-stage harm evaluation, AI LEGO introduces *Persona-centered Evaluation*, a feature that uses personas as scaffolds for exploring potential stakeholder viewpoints. During the evaluation process, non-technical roles can envision any stakeholders who might be affected by the intended AI system and create personas with specified characteristics such as personal and professional backgrounds using text descriptions. To balance user control and diverse representation, *Persona-centered Evaluation* supports a combination of manual input and LLM-generated examples. These personas are saved and visualized in the top-right corner as LEGO Minifigure icons (Fig. 1C). After identifying the impacted stakeholders, *Persona-centered Evaluation* can further inspire evaluators by simulating personas' reactions to the proposed AI development plan. Non-technical roles can choose a saved persona and inquire about their subjective feelings. To enable more open-ended discussions, we prompt LLMs to simulate the concerns stakeholders might articulate based on their backgrounds, rather than directly identifying harms [15]. These simulated reactions are saved and displayed when hovering over the persona icon.

4.4 Implementation Details

The AI LEGO web application is built in HTML, JavaScript, and CSS with the React library. Its node-based interface is enabled with vanilla React plus react-draggable[4] package. While the frontend is hosted on Vercel, the backend uses Google Firebase through API calls for user authentication and data persistence (per state change). The *Persona-centered Evaluation* functionality leverages the OpenAI GPT-4 model with zero-shot prompting techniques (see Appendix D).

5 Evaluation User Study

We conducted a user study to evaluate the effectiveness of AI LEGO in helping cross-functional AI practitioners communicate and identify problematic design choices of AI systems at an early stage. To approximate real-world cross-functional collaborative settings while maintaining a controlled

[4]https://www.npmjs.com/package/react-draggable

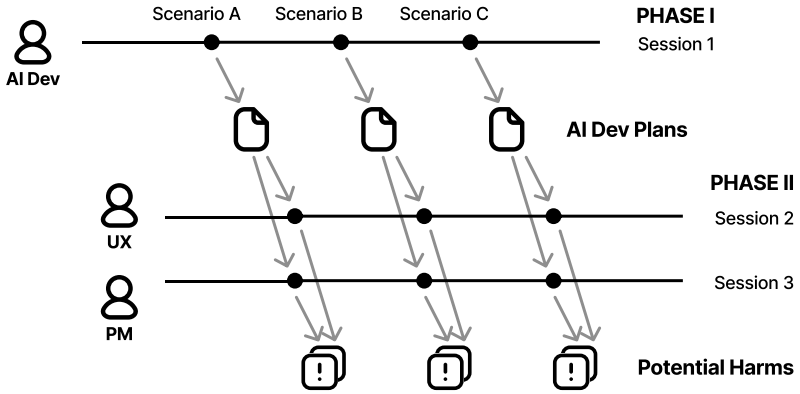

Fig. 2. Two-phase team-based evaluation study task flow.

environment for an initial empirical understanding of our tool, we organized participants into teams and divided the study sessions into two phases. We adopted a within-subject design and created two variants of our tool, AI LEGO Lite and AI LEGO Full. Both variants allow practitioners to draft AI development plans with *Lifecycle Blocks*. Yet AI LEGO Lite only supports *Stage-centered Evaluation* while AI LEGO Full supports both *Stage-centered* and *Persona-centered Evaluation*.

Our study was guided by the following Research Questions: ***RQ1***: How effective are AI LEGO Full and AI LEGO Lite in assisting cross-functional AI practitioners?; ***RQ2***: Do AI LEGO Full and AI LEGO Lite offer value over existing RAI practices for cross-functional AI practitioners?

5.1 Participants & Team Formation

We recruited 18 industry AI practitioners (6 female, 12 male) through social media posting and snowball sampling. None of these participants were involved in our previous study. Similarly, all participants were physically located in the United States and were required to have prior industry experience working on AI-infused products in cross-functional teams, but were not expected to have worked on RAI. In total, we received 101 voluntary responses. Based on pre-screening survey responses about their roles in past AI projects, we selected 6 AI developers, 6 PMs, and 6 UI/UX designers[5]. Participants were randomly assigned to one of six teams before the study. Each team included one technical role (AI developer) and two non-technical/user-facing roles (one PM and one UI/UX designer), who participated in different tasks during the study (see Section 5.2). Participants within each team were unidentified/unacquainted with each other due to the recruitment procedure. Future work could investigate the dynamics with acquaintances. Each participant received a $35 USD Amazon gift card after the study session.

5.2 Study Design

We designed a within-subject two-phase study to evaluate the effectiveness of AI LEGO in cross-functional RAI practices:

5.2.1 Two-phase Design. The overall task flow within a team (3 sessions) is shown in Fig. 2. In **Phase I**, the technical AI developer initiated the process by drafting AI development plans. In three blocks, they were tasked with creating three AI development plans, each focusing on a specific

[5]A total of 5 participants worked at companies with 25,000 or more employees. Others came from companies with 5,000–24,999 employees (3), 1,000–4,999 (2), 250–999 (1), 50–249 (1), and fewer than 50 (4).

scenario (see Section 5.2.3) and leveraging a different tool corresponding to the conditions (see Section 5.2.2). In **Phase II**, the PM and UX designer separately reviewed the plans created by the AI developer within their team in the prior session and were tasked with reviewing the plan to identify potential downstream harms that arose from problematic design choices. They were presented with the plans created by the AI developer and the scenario description, applying tools corresponding to the conditions.

5.2.2 Conditions. Each participant experienced all three different conditions/tools: Baseline (Google Docs), AI LEGO Lite, and AI LEGO Full. Google Docs is one of the most popular online collaboration and documentation tools commonly used in industry settings. It facilitates a high degree of freedom in text input and styling. We further complement Harm Modeling [51], a popular risk identification framework (as an editable table) in the doc. Together, they provide a reasonably strong baseline for scaffolding RAI practices including planning and identifying potential risks. AI LEGO Lite and AI LEGO Full are identical in Phase I for scaffolding AI development plan drafting. In Phase II, AI LEGO Lite supports only *Stage-centered Evaluation*, whereas AI LEGO Full supports both *Stage-centered Evaluation* and *Persona-centered Evaluation*. This contrast was intentionally designed to examine the incremental contribution of *Persona-centered Evaluation* to the harm identification process when layered on top of the structured, stage-centered evaluation workflow. The order of conditions was fixed for participants within each team and was counterbalanced across teams through Latin squares [13].

5.2.3 AI System Development Scenarios. To mimic typical industrial early-stage AI design processes, we created five scenarios (see Appendix E) where AI systems are to be developed to address real-world problems. These scenarios were sourced and adapted from the AI Incident Database [50]. All sampled incidents had significant media coverage across diverse AI application domains. Each scenario includes information about the background, goals, requirements, evaluation metrics, key features, deployment details, and target users—representing a typical starting point for early-stage AI design. We removed all identifiable entity names from the materials to avoid biases. We ensured that all team members were unfamiliar with the original incident through a pre-screening survey. Each team was assigned three out of the five scenarios in a fixed order. Across all teams, each scenario was assigned to each condition/tool at least twice. The detailed scenarios and conditions assignment is shown in Fig. 6 in Appendix G.

5.2.4 Session Procedure. Each session lasted approximately 60 minutes and was conducted one-on-one via remote video conferencing. At the start, participants were introduced to the goal of the study and provided their consent for recording their screen activity and audio. Participants then progressed through the three blocks, where they performed tasks according to their roles in designated conditions. Each block took approximately 15 minutes and ended when participants reported reaching thought exhaustion. At the end of a block, they evaluated the tool's usability by filling the System Usability Scale [14]. For Phase I, SUS scores for AI LEGO Lite were recorded only for the first intervention. Participants were offered a voluntary 2-minute break between conditions. The session concluded with a 10-minute semi-structured interview (see Appendix F for the protocol), where they shared their preferences over the conditions, reasons behind their choices, and additional feedback.

5.3 Data Analysis

Data analysis aims to assess the effectiveness and usability of AI LEGO Full and AI LEGO Lite compared to Google Docs in early-stage RAI practices. We analyzed both quantitative (ratings of

identified harm, SUS scores, Preference) and qualitative data (semi-structured interview notes) with methodologies described as follows.

5.3.1 Quantitative Analysis. Following [76], we analyzed the count, severity, and likelihood of identified harms to evaluate the effectiveness of AI LEGO (RQ1). After collecting the identified potential harms through system logging and recording, two authors independently rated the severity ("How severe is the harm's impact in this scenario?") and likelihood ("How likely is this harm to occur in this scenario?") of all the harms using a 4-point scale. Both authors were graduate students with academic & industry experiences in RAI (unlike many of the participants). They were blinded to the conditions under which harms were identified but had access to the scenarios for contextual understanding. The initial ratings showed a fair to moderate agreement with weighted (quadratic) Cohen's Kappa scores of 0.51 for severity and 0.64 for likelihood. After the initial ratings, the authors resolved discrepancies through discussion and developed a grading rubric based on Sociotechnical Harms Taxonomy [66]. The final grading rubric with examples are provided in Table 2 in Appendix H.

The interval data (count of identified harms) were aggregated at participant level using the mean and analyzed using a repeated measures ANOVA, followed by paired t-tests with Holm correction for pairwise comparisons. This approach was appropriate as the Shapiro–Wilk test indicated no significant deviation from normality ($p > .05$). Mauchly's test indicated a violation of the sphericity assumption ($p = .02$) and thus Greenhouse–Geisser correction was applied. The ordinal data (severity, likelihood of identified harms, SUS scores, average preference rankings) was aggregated at participant level using the median and analyzed with non-parametric Friedman tests, followed by Wilcoxon signed-rank tests with Holm correction for post-hoc pairwise comparisons. For SUS scores and preference rankings in Phase I, the Wilcoxon signed-rank test was applied instead as there were only two conditions.

5.3.2 Qualitative Analysis. We transcribed the semi-structured interview recordings and applied inductive thematic analysis [35]. Two authors independently conducted an initial open coding pass over the transcripts, focusing on participants' descriptions of their experiences with the tools, perceived utility, limitations, and impacts on cross-functional RAI practices (RQ2). The authors then met to compare and reconcile their codes, iteratively grouping related codes into higher-level themes. These themes were refined through repeated examination of the transcripts to ensure they captured recurring patterns across participants while preserving different perspectives.

5.3.3 Power and Sensitivity Analysis. To assess whether our study design was capable of detecting meaningful within-subject effects, we conducted a sensitivity analysis using G*Power for a one-group repeated-measures ANOVA with three conditions ($\alpha = .05$). We focus this analysis on Phase II, which constitutes the primary confirmatory evaluation of system effects, while Phase I is intended to support exploratory design insights. Assuming a moderate correlation among repeated measures ($r = .5$) and a conservative Greenhouse–Geisser correction ($\epsilon = .65$, consistent with our observed data), the Phase II design ($n = 12$) was sensitive to effects of approximately Cohen's $f \geq 0.46$ with 80% power. This indicates that the study was sufficiently sensitive to detect moderate-to-large condition effects, though smaller effects may not have been reliably detected. Limitations, including demographic composition and implications for generalizability, are discussed in Section 7.4.

6 Findings

6.1 Effectiveness of AI LEGO in Assisting Cross-Functional AI Practitioners (RQ1)

Quantitative results demonstrated varied effectiveness of AI LEGO Full and AI LEGO Lite in helping cross-functional teams proactively identify problematic design choices to prevent potential

downstream harms (Fig. 3A). Participants also rated AI LEGO Lite and AI LEGO Full to be more usable and preferable compared to Google Doc (Fig. 3B).

6.1.1 AI LEGO enables the cross-functional team to uncover more problematic design choices. The repeated measures ANOVA test revealed a significant main effect of conditions ($F(2, 22) = 16.26$, $p_{GG} < .001$) on the number of problematic design choices identified. Post-hoc Holm-corrected paired t-tests revealed significant differences between all conditions. Compared to Google Docs ($m = 1.9$), participants identified significantly more problematic design choices when using both AI LEGO Full ($m = 5.6, p < .001$) and AI LEGO Lite ($m = 3.7, p = .03$). This indicates that AI LEGO equips cross-functional teams with the ability to explore and surface more potential harms during the AI development process. Moreover, participants identified significantly more problematic design choices when using AI LEGO Full than AI LEGO Lite ($p = .04$), suggesting that the additional *Persona-centered Evaluation* feature further supports brainstorming by encouraging a more comprehensive exploration of possible design flaws.

6.1.2 AI LEGO allows cross-functional teams to identify more likely and relevant AI system design flaws. The Friedman test revealed a significant main effect of condition on the likelihood of identified harms ($\chi^2 = 9.72, p < .01$). Post-hoc Holm-corrected Wilcoxon signed-rank tests showed that, compared to Google Docs ($mdn = 2$), the harms identified using both AI LEGO Full ($mdn = 3.25$, $p = .03$) and AI LEGO Lite ($mdn = 3$, $p = .03$) were rated as significantly more relevant. No significant difference was observed between AI LEGO Lite and AI LEGO Full ($p = .2$). This result underscores that AI LEGO enables teams to focus on more likely design flaws, improving their ability to prioritize relevant risks in the development process. No significant main effect was found regarding the severity of identified harms, indicating that AI LEGO Full primarily enhances teams' capacity to detect risks that are more probable rather than more severe.

6.1.3 AI LEGO tools are more preferred and show comparable usability. Although participants rated the highest average overall SUS score for AI LEGO Lite in the planning phase and AI LEGO Full in the evaluation phase, statistical tests revealed no significant main effect, indicating the margins were negligible. For individual SUS items in the planning phase, Holm-corrected Wilcoxon signed-rank tests showed that participants rated AI LEGO Lite significantly lower on the item "I found the tool very cumbersome to use" ($W = 0.00, p = .04$) and significantly higher on the item "I found the various functions in this system were well integrated" ($W = 0.00, p = .03$) compared to Google Doc. For the planning phase, 5 out of 6 participants preferred AI LEGO to Google Docs. For the evaluation phase, the Friedman test showed a significant main effect on average ranking ($\chi^2 = 10.67, p < .01$). Post-hoc Holm-corrected Wilcoxon signed-rank tests showed that participants ranked AI LEGO Full ($mdn = 1$) significantly higher than Google Doc ($mdn = 3, p < .001$).

6.2 Values Over Existing RAI Practices (RQ2)

Interview notes from participants across teams (T1–6) with different roles (AI, UX, PM) revealed varied effectiveness in AI LEGO's features in achieving the design goals, while also identifying opportunities for future improvements.

6.2.1 Eight-stage Worksheet & Checklist serve as an effective channel for knowledge handoff in cross-functional teams (G1). Participants reported that they could more easily elaborate their ideas under the scaffolding provided by *Eight-Stage Worksheet & Checklist*. Both technical ($n = 3$) and non-technical/user-facing roles ($n = 7$) considered them being helpful in structuring their thoughts within the framework of the AI lifecycle. Compared to the "free-form" communication style of Google Docs (T2AI), this structured method helped participants focus on individual AI development stages, facilitating design choice elaboration and idea comprehension:

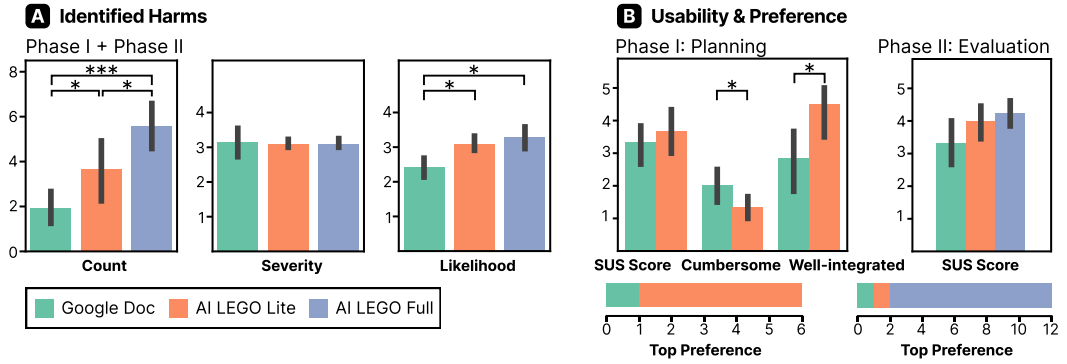

Fig. 3. Effect of conditions on (A) Count, Severity, and Likelihood of identified harms and (B) Usability and Preference. AI LEGO Full allowed AI teams to identify more potential harms with a higher likelihood; it was more preferred and considered well-integrated, especially in the planning phase. Significance levels: *** $p < .001$, ** $p < .01$, * $p < .05$. Error bars indicate 95% CI.

> *I can easily follow the prompts (from Eight-Stage Worksheet) to plan around the typical workflow of (AI system) development, whereas in Google Docs I have no clue of what to describe and where to start from.* (T5AI)
> *Following the clear structure by AI LEGO allows me to easily comprehend my teammate's planning.* (T6UX)

Additionally, the eight-stage mechanism breaks down tasks for individuals, which makes communication more explicit and actionable, rather than being ambiguous and unapproachable (T3PM, T4AI, T4UX, T6UX):

> *AI LEGO provides better support in writing out the plan and specifies the tasks I should focus on more clearly.* (T4AI)

Participants T1AI and T2AI explicitly reported feeling encouraged to think more broadly about critical components that they may otherwise have overlooked via Google Docs. They also described a shift towards more proactive risk identification from a reactive, "fix-it-later" approach (T2AI):

> *The "Feedback" stage in AI LEGO prompted me to consider input from a broader range of sources, including users, law enforcement, and other public entities.* (T1AI)

The easy-to-grasp stage-based inputs provided by AI developers subsequently enabled non-technical/user-facing roles to engage more actively and take greater responsibility in the AI development process (T3PM, T4UX, T5UX), in contrast to traditional models where technical teams hold primary power in defining the project. As T3PM noted, instead of being a "passive recipient of information", they became "an active partner in the responsible AI process".

6.2.2 AI LEGO presents the workflow through an intuitive and engaging layout (G2). Most participants ($n = 11$) found the tree taxonomy of interconnected *Lifecycle Blocks* in AI LEGO suitable for the AI product design process. Compared to plain text in Google Docs, the tree taxonomy allowed participants to view the entire flow of AI development at a glance. Instead of scrolling up and down to locate relevant notes, they could easily navigate the workflow and focus on specific stages. This design enabled participants to track their progress more effectively while maintaining a continuous thought process:

> *The tree taxonomy visualization allows me to keep track of my progress when drafting the plan. However, with Google Docs, I struggled to get an overview and frequently had to navigate back and forth to reference earlier designs.* (T3AI)

This flow-based information architecture serves as a shared visual language, fostering a more organized and targeted approach to collaboration. Participants T1UX, T5UX and T6UX emphasized that it made them feel like integrated contributors to the development lifecycle rather than participants in a one-off feedback session:

> *The interaction mechanism of AI LEGO where you can hover on a block to see the prompt is very intuitive.* (T1UX)
> *I particularly like this kind of flow-based format…and it's a good communication model…for example, you can leave comments on a specific block, and then you can have an async conversation with the developer about a specific piece of content.* (T5UX)

Building on this structure, AI LEGO further supports seamless user interaction. Participants praised AI LEGO for facilitating direct manipulation, a key concept around the sense of control in HCI, through integration between *Eight-Stage Worksheet/Checklist* and specific UI features. For instance, participants appreciated that they could drag and drop stage blocks (T2UX) and hover over them to view prompts (T1UX, T4AI):

> *The interaction mechanism of AI LEGO where you can hover on a block to see the prompt is very intuitive.* (T1UX)
> *With the drag and drop feature (over the stage blocks), the experience is interactive and playful.* (T2UX)

The intuitive and visually appealing overall layout led to a reduced learning curve and a sense of playfulness (T2UX, T4UX):

> *Even without an explanation of the persona-based evaluation feature, I can still learn how to use it through the provided prompts and self-exploration.* (T4UX)
> *AI LEGO gives a playful vibe in the theme of LEGO with the colorful blocks and minifigures.* (T4PM)

6.2.3 AI LEGO helps support the evaluation process by non-technical roles (G3). Compared to Google Docs, AI LEGO enabled non-technical roles to identify more potential harms of AI products ($p < .001$) and with a higher likelihood ($p = .02$). When using Google Docs, non-technical participants frequently encountered inconsistent structure (T1PM, T3UX, T5UX) and unclear technical jargon (T1PM, T2PM, T3UX, T4UX, T5UX) in the drafted plans; in contrast, no such obfuscation was reported when using AI LEGO. The integrated and scaffolded design of AI LEGO also reduces the cognitive load of switching between different information sources scattered across the document (T3UX, T5UX). This design helps "shift" (T3UX) the focus of ethical analysis from abstract concepts to concrete, stage-specific problems, leading to a more deliberate and contextualized approach to harm analysis, which is more precise and actionable than the generalized, scattershot analysis often seen in Google Docs + Harm Envisioning (T1PM, T2PM).

Most of the non-technical participants ($n = 8$) attributed the enhanced evaluation efficacy primarily to the *Persona-centered Evaluation*. By engaging with the mocked personas, participants were able to "rethink" each stage of the AI development plan through the lens of identified stakeholders, allowing them to generate a broader range of ideas from multiple viewpoints ($n = 6$) and directly link abstract technical plans to "real-world human impact" (T5PM, T4PM):

> *The generated personas allow me to think about a case more deeply based on their characteristics, which is really helpful in finding the edge cases.* (T1PM)

> *...the persona function helps me broaden my point of view by providing stakeholders' perspectives.* (T2PM)

According to participants, this approach was particularly effective in unfamiliar scenarios (T3UX, T4UX) or when working independently without access to collaborative input (T3PM, T5UX, T6UX). For examples,

> *The persona feature is particularly helpful when I am unfamiliar with specific scenarios, as it allows me to approach the problem from different standpoints and levels of granularity.* (T3UX)
> *The persona feature helps me think about things like labor markets, financial impact...things that are not very obvious.* (T4UX)
> *Personas could stimulate your thinking process when you work alone by providing mocked data.* (T6UX)

Interestingly, some participants in non-technical/user-facing roles mentioned a shift in power imbalance present in current cross-functional teams (identified in the formative study). Participants described *Persona-centered Evaluation* feature as a resourceful "intelligent assistant" (T5UX) or a "creativity stimulant" (T6PM) that simulates different stakeholder perspectives, which could give them a stronger set of arguments to use when questioning or challenging a developer's plan (T5UX, T6UX). Meanwhile, *Eight-Stage Worksheet* facilitated clear articulation of AI design choices from AI developers, reducing non-technical roles' needs for frequent clarification from developers, enabling them to work more autonomously without worrying they were constantly bothering technical counterparts:

> *(Without AI LEGO) I would probably need to ask a few questions...things that I didn't understand, or things that they (AI developers) didn't explain clearly.* (T4UX)

6.2.4 Future improvements to better achieve the design goals. While participants generally appreciated AI LEGO for its enhanced efficiency and user experience, they also offered suggestions for improvements regarding the identified design goals. To enhance the effectiveness of knowledge handoff (G1), participants proposed adding prompts tailored to specific AI products or scenarios in *Eight-Stage Worksheet* and *Checklist* (T3AI) and enabling customizable blocks for additional notes. T4UX suggested enhancing the intuitiveness of AI LEGO's visualization by incorporating formatted text similar to Google Docs (G2). For *Persona-centered Evaluation* (G3), participants recommended exploring features such as backend automation to trigger evaluations upon detecting issues (T3UX) and incorporating harm severity ratings to aid prioritization (T4UX). Participants T4UX and T6UX expressed concerns about the AI-simulated stakeholders' lack of real-world context or accuracy:

> *I'm a little concerned because it's all AI-generated...I don't know the truth or the priority (of the harms) without doing my own research.* (T4UX)

Mechanisms should thus be developed to prevent AI teams from over-relying on *Persona-centered Evaluation* replacing the need for user-facing professionals' expertise and research to validate and prioritize the identified harms. We further discuss this in Section 7.2.

7 Discussion and Design Implications

Our findings highlight the effectiveness of AI LEGO in scaffolding cross-functional AI teams in handing off knowledge of high-level AI system designs and conducting more comprehensive early-stage harm evaluation. Below, we explore the role of AI LEGO within three critical aspects of RAI, highlighting its contributions, limitations, and implications for fostering future RAI practices.

7.1 Supporting Early-stage Cross-functional RAI Envisioning

Our findings highlight early-stage RAI envisioning as a form of cross-functional coordination work rather than a purely technical or reflective activity. Compared to prior early-stage tools that primarily support individual practitioners or narrow design moments [43, 76], AI LEGO focuses on structuring how technical and non-technical roles collaboratively articulate, examine, and iterate on AI design decisions before implementation. Concretely, AI LEGO introduces stage-based lifecycle blocks, associated prompts, and linked evaluation artifacts as shared representations through which teams coordinate this work. In doing so, it foregrounds the collaborative labor involved in anticipating harms when system designs are still fluid and underspecified.

A key contribution of AI LEGO lies in its use of abstraction to make early AI design work collectively tractable. By breaking down AI development into modular, stage-based components, the *Eight-Stage Worksheet* and accompanying checklists enable participants to reason about complex systems at an appropriate level of granularity during planning. This aligns with prior work on sketching and abstraction in early design [79, 80], suggesting that early-stage representations need not be complete or precise to support meaningful collaboration. Instead, they must be interpretable and negotiable across roles

Participants' responses further illustrate a central CSCW tension between structure and openness in collaborative work. While many participants valued the stage-based structure and tree-like organization of AI LEGO as a shared reference point, providing common vocabularies and locus for discussion, some (T2AI and T6PM) preferred the baseline's free-form documentation for its flexibility and ability to support open-ended exploration. Rather than indicating a failure of structure, this tension echoes longstanding CSCW findings that coordination mechanisms must balance formalization with opportunities for situated interpretation. Our findings suggest that effective early-stage RAI tools should allow teams to fluidly move between structured articulation and open reflection, depending on the phase and purpose of collaboration.

These findings reinforce the role of boundary objects in early-stage RAI work. The lifecycle stages and modular blocks in AI LEGO function as boundary objects that are simultaneously stable enough to anchor discussion and flexible enough to accommodate diverse perspectives [48, 69]. By externalizing high-level design intent into shared, manipulable artifacts, AI LEGO supports articulation work across roles and enables harm identification to be treated as an integral part of collaborative planning rather than a downstream audit task. Future work should further explore how such boundary objects can be adapted to different organizational contexts and collaboration styles. In particular, there is an opportunity to investigate how personalization, partial structure, or role-specific views can coexist with shared grounding, enabling teams to tailor early-stage RAI envisioning without fragmenting collective understanding.

7.2 Towards Human-AI complementarity in RAI Practices

The use of LLM-generated personas in AI LEGO raises broader questions about how human and AI capabilities should be combined in ethical deliberation and harm identification. Traditionally, RAI evaluation has relied on human expertise through methods such as expert review, crowdsourcing, and structured audits [11, 25, 49]. Recent work has begun to explore how LLMs can assist in these processes by generating candidate harms or stakeholder perspectives [15, 76].

Rather than delegating evaluative authority to LLMs, AI LEGO adopts a complementary design in which participants retain control over harm identification while using LLMs as a generative resource. In this configuration, persona-centered evaluation supports ideation and perspective-taking without replacing human judgment. This design aligns with the perspectives on human-AI collaboration that emphasize the importance of preserving human agency, accountability, and contextual reasoning

when introducing AI assistance into collaborative work [30, 39]. Importantly, the LLM-generated personas in AI LEGO function not merely as informational outputs, but as boundary objects that provide a shared narrative reference that participants can collectively interpret, contest, and refine. For non-technical practitioners, these personas help surface edge cases and prompt reflection on unfamiliar technical decisions. For teams as a whole, they could offer a common frame for discussing potential harms across roles with asymmetric expertise.

At the same time, participants' concerns about the realism and completeness of LLM-generated personas highlight a critical tradeoff. While such personas can increase the elasticity of boundary objects, enabling teams to rapidly explore a broader range of scenarios, they may also introduce synthetic abstraction that risks obscuring real-world complexity. Moreover, because these personas are generated based on generic language models (e.g., GPT-4) and limited contextual prompts, they may reproduce prevailing assumptions or biases present in the training data, and are unlikely to surface "blind spots" rooted in lived experience, cultural specificity, or structural marginalization. This tension reflects a broader challenge in machine-mediated collaboration: balancing efficiency and generativity with the need for grounded, accountable representations [48].

We view this tension not as a limitation of AI LEGO alone, but as an open design space for CSCW research. Future work should further investigate how human-AI complementarity can be operationalized in RAI practices, including when and how AI-generated artifacts should be surfaced, constrained, or contextualized to support ethical deliberation without displacing human responsibility [38].

7.3 Integrating into Industrial Cross-functional Product Workflows

The scope of this work is to understand the primary challenges in conducting early-stage cross-functional RAI work, as well as the effectiveness of AI LEGO in addressing them, specifically, knowledge handoff by technical roles and harm envisioning by non-technical/user-facing roles. Compared to other solutions that remain at a high level of representation (e.g., guidelines [8, 52]), AI LEGO was purposefully built as a collaborative web tool. The embodiment of our tools aims to integrate a less-formalized socio-technical activity into early product design processes, blending product planning with harm anticipation without imposing excessive workloads. The evaluation results demonstrated the feasibility of this concept in enhancing involvement across roles and improving overall evaluation performance. Participants T2PM, T3PM and T4UX specifically praised how AI LEGO modularizes RAI work into stepwise tasks with clear instructions, noting its alignment with industrial practices. For example,

> *Given that industrial practices invest more in pushing the product, I think AI LEGO could work as it scales up the evaluation process through collaboration and AI generation without adding significant extra work.* (T4UX)

While our system and study designs were largely grounded in realistic settings, we must acknowledge that industrial cross-functional practices involve additional factors and complexities that were not fully captured in this work [27, 58]. For example, different teams and organizations use varied communication channels, project planning tools, and collaboration policies that shape their working practices. While AI LEGO can be further optimized for minimal onboarding effort or adapted into lightweight plugins or a fully featured tool for practical values (e.g., with work allocation and milestone tracking features), this would require a deeper understanding of organizational preferences, constraints, and potential frictions that may limit adoption. For instance, T6UX noted:

> *The tool…is to come up with potential issues or potentially, you know, places of bias. But that's also kind of coming at it from a, like, "deficit mindset". I hadn't even thought of,*

> *like, who in my organization would want something like this because of the way the tool was framed.* (T6UX)

Additionally, AI LEGO focuses on facilitating knowledge articulation from technical AI developers to non-technical/user-facing roles, based on insights from our formative study that tasks related to AI system planning and harm evaluation should be delegated to developers and non-technical/user-facing roles, respectively. However, the underlying assumption that these roles are always cooperative (e.g., developers readily integrating feedback into system designs) does not always hold. Our qualitative results suggest that AI LEGO shows promise in addressing the prevailing power imbalance, enabling non-technical and user-facing roles to participate more autonomously and actively in responsible AI planning rather than relying on developers' voices. Nevertheless, more systematic investigations in future studies would be necessary to validate this, particularly through field deployments of AI LEGO in uncontrolled, real-world industrial team environments, to examine how AI LEGO may support, mitigate, or reshape broader organizational dynamics (e.g., power, accountability) as reflected in actual collective behaviors.

7.4 Limitations and Future Work

As an initial effort to support early-stage, cross-functional RAI work, our approach has a number of limitations. First, AI LEGO was developed primarily to address the challenge of knowledge handoff from technical to non-technical/user-facing roles in cross-functional RAI work—a challenge well-documented in prior literature and surfaced in our formative study as a common issue in industry (See Section 3). Toward more human-centered RAI practices, we acknowledge that future work should also investigate alternative workflows—particularly knowledge flows from non-technical roles to technical ones. These workflows are critical for incorporating contextual, ethical, and user-centered insights into technical decision-making. Additional areas for future exploration include concept mapping [43], as well as collective sensemaking and cross-role decision-making.

Second, as a proof-of-concept prototype, AI LEGO would benefit from further refinement to enhance user experience, performance, and workflow integration (see Section 6.2). In particular, as discussed in detail in Section 7, the LLM-simulated *Persona-centered Evaluation* holds both promises and risks and would require more systematic future investigation (e.g., higher versus lower levels of LLM autonomy in harm assessment or user inquiries). Please note that our LLM-based persona is intended only as a starting point—not a substitute for real-world stakeholder engagement [2, 31]. Future work should explore how to more effectively integrate real-world users and stakeholders into the pipeline we introduce here. More nuanced data that reflect real end-user interactions and feedback is needed to provide stronger stakeholder-based evaluations in RAI practices.

Third, while our evaluation study aimed to approximate real-world cross-functional collaboration settings, many simplifications were made in the study design for practical reasons. For example, we only recruited participants based in the US and formulated teams, each consisting of three typical roles, which limited our understanding of how the effectiveness of our tool might vary across different cultures, regions, and team setups. AI LEGO was evaluated end-to-end in a single pass, reflecting a minimal level of team coordination and offering limited opportunities for dynamic feedback and idea exchange. Future work should explore less controlled experimental settings, such as deployment studies, to more accurately mirror everyday practices.

Finally, the six AI product scenarios used in the evaluation were adapted from popular AI incidents. Although we counterbalanced their selection and order, we did not fully control for each scenario's complexity. Several participants noted that more complex scenarios appeared to heighten the tool's perceived value. Future investigations could thus delve deeper into how scenario complexity influences the tool's effectiveness.

8 Conclusion

We introduce AI LEGO, an interactive tool designed to help cross-functional AI practitioners better communicate high-level AI designs throughout the development lifecycle and identify potential harmful design choices early on. We evaluated AI LEGO with 18 industry practitioners from various roles, including AI developers, UI/UX designers, and PMs. Our results show that AI LEGO helped participants identify more problematic design choices and those with higher likelihood. This implies that the specific design of the tool would impact the effectiveness of facilitating cross-functional knowledge handoff and harm envisioning, such as the stage-level prompts, block-based layout, and LLM-simulated personas. We discuss the challenges and opportunities for supporting cross-functional collaboration in early-stage RAI work.

Availability

AI LEGO is live at https://ailego.vercel.app/ (source code: https://github.com/henrw/ai-lego).

References

[1] Mark S. Ackerman. 2000. The intellectual challenge of CSCW: the gap between social requirements and technical feasibility. *Hum.-Comput. Interact.* 15, 2 (Sept. 2000), 179–203. doi:10.1207/S15327051HCI1523_5

[2] William Agnew, A. Stevie Bergman, Jennifer Chien, Mark Díaz, Seliem El-Sayed, Jaylen Pittman, Shakir Mohamed, and Kevin R. McKee. 2024. The Illusion of Artificial Inclusion. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[3] Jumana Almahmoud, Robert DeLine, and Steven M Drucker. 2021. How teams communicate about the quality of ML models: a case study at an international technology company. *Proceedings of the ACM on Human-Computer Interaction* 5, GROUP (2021), 1–24.

[4] Rico Angell, Brittany Johnson, Yuriy Brun, and Alexandra Meliou. 2018. Themis: Automatically testing software for discrimination. In *Proceedings of the 2018 26th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. Association for Computing Machinery, New York, NY, USA, 871–875.

[5] Asana, Inc. [n. d.]. Asana Project Management Tool. https://asana.com

[6] Atlassian. [n. d.]. Jira Software. https://www.atlassian.com/software/jira

[7] James Auger. 2013. Speculative design: crafting the speculation. *Digital Creativity* 24, 1 (2013), 11–35.

[8] Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. 2019. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. Association for Computing Machinery, New York, NY, USA, 421–433.

[9] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.

[10] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63, 4/5 (2019), 4–1.

[11] Jeffrey P Bigham, Michael S Bernstein, and Eytan Adar. 2015. Human-computer interaction and collective intelligence. *Handbook of collective intelligence* 57, 4 (2015).

[12] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft.

[13] James V Bradley. 1958. Complete counterbalancing of immediate sequential effects in a Latin square design. *J. Amer. Statist. Assoc.* 53, 282 (1958), 525–528.

[14] J Brooke. 1996. SUS: A quick and dirty usability scale. *Usability Evaluation in Industry* (1996).

[15] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. Aha!: Facilitating ai impact assessment by generating examples of harms. arXiv:2306.03280

[16] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[17] Paul R Carlile. 2002. A pragmatic view of knowledge and boundaries: Boundary objects in new product development. *Organization science* 13, 4 (2002), 442–455.

[18] Stevie Chancellor. 2023. Toward Practices for Human-Centered Machine Learning. *Commun. ACM* 66, 3 (Feb. 2023), 78–85. doi:10.1145/3530987

[19] Chaoran Chen, Weijun Li, Wenxin Song, Yanfang Ye, Yaxing Yao, and Toby Jia-Jun Li. 2024. An Empathy-Based Sandbox Approach to Bridge the Privacy Gap among Attitudes, Goals, Knowledge, and Behaviors. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 234, 28 pages. doi:10.1145/3613904.3642363

[20] Irene Y. Chen, Fredrik D. Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., Red Hook, NY, USA, 3543–3554.

[21] Christoph Clases and Theo Wehner. 2002. Steps across the border–Cooperation, knowledge production and systems design. *Computer Supported Cooperative Work (CSCW)* 11, 1 (2002), 39–54.

[22] Marios Constantinides, Edyta Bogucka, Daniele Quercia, Susanna Kallio, and Mohammad Tahaei. 2024. RAI Guidelines: Method for Generating Responsible AI Guidelines Grounded in Regulations and Usable by (Non-)Technical Roles. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 388 (Nov. 2024), 28 pages. doi:10.1145/3686927

[23] Alan Cooper and Robert Reimann. 2003. About Face 2.0. *The Essentials of Interaction Design* (2003).

[24] Henriette Cramer, Jenn Wortman Vaughan, Ken Holstein, Hanna Wallach, Jean Garcia-Gathright, Hal Daumé III, Miroslav Dudík, and Sravana Reddy. 2019. Challenges of incorporating algorithmic fairness into industry practice. FAT* 2019 Tutorial. Tutorial.

[25] Wesley Hanwen Deng, Boyuan Guo, Alicia Devrio, Hong Shen, Motahhare Eslami, and Kenneth Holstein. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[26] Wesley Hanwen Deng, Manish Nagireddy, Michelle Seng Ah Lee, Jatinder Singh, Zhiwei Steven Wu, Kenneth Holstein, and Haiyi Zhu. 2022. Exploring how machine learning practitioners (try to) use fairness toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 473–484.

[27] Wesley Hanwen Deng, Nur Yildirim, Monica Chang, Motahhare Eslami, Kenneth Holstein, and Michael Madaio. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, New York, NY, USA, 705–716.

[28] Consequence Scanning Doteveryone. 2020. An Agile Practice for Responsible Innovators.

[29] Virginia Eubanks. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor.* St. Martin's Press, New York, NY, USA.

[30] Mingming Fan, Xianyou Yang, TszTung Yu, Q. Vera Liao, and Jian Zhao. 2022. Human-AI Collaboration for UX Evaluation: Effects of Explanation and Synchronization. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 96 (April 2022), 32 pages. doi:10.1145/3512943

[31] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2025. User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in AI companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[32] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[33] Batya Friedman and David Hendry. 2012. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1145–1148.

[34] Tarleton Gillespie. 2014. The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, Tarleton Gillespie, Pablo J. Boczkowski, and Kirsten A. Foot (Eds.). MIT Press, Cambridge, MA, USA, 167–194. doi:10.7551/mitpress/9780262525374.003.0009

[35] Cliff Goddard. 2011. *Semantic analysis: A practical introduction.* Oxford University Press, USA.

[36] Google. [n. d.]. Google Docs. https://docs.google.com

[37] Ben Green and Salomé Viljoen. 2020. Algorithmic realism: expanding the boundaries of algorithmic thought. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. Association for Computing Machinery, New York, NY, USA, 19–31.

[38] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. 2021. Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review. *PACIS* (2021), 78.

[39] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2023. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 152 (April 2023), 20 pages. doi:10.1145/3579628

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

[41] IDEO. 2019. AI Ethics Cards: Collaborative Activities for Designers. https://page.ideo.com/download-ai-ethics-cards. A set of reflective design activities to guide ethically responsible, human-centered design with data-driven systems.
[42] Tzu-Sheng Kuo, Hong Shen, Jisoo Geum, Nev Jones, Jason I Hong, Haiyi Zhu, and Kenneth Holstein. 2023. Understanding Frontline Workers' and Unhoused Individuals' Perspectives on AI Used in Homeless Services. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–17.
[43] Michelle S Lam, Zixian Ma, Anne Li, Izequiel Freitas, Dakuo Wang, James A Landay, and Michael S Bernstein. 2023. Model sketching: centering concepts in early-stage machine learning model design. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–24.
[44] Q Vera Liao, Mihaela Vorvoreanu, Hari Subramonyam, and Lauren Wilcox. 2024. UX Matters: The Critical Role of UX in Responsible AI. *Interactions* 31, 4 (2024), 22–27.
[45] Michael Xieyang Liu, Aniket Kittur, and Brad A. Myers. 2021. To Reuse or Not To Reuse? A Framework and System for Evaluating Summarized Knowledge. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 166 (April 2021), 35 pages. doi:10.1145/3449240
[46] Andrés Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II 15*. Springer, Springer, Cham, Switzerland, 231–248.
[47] Wendy E. Mackay. 1995. Ethics, Lies and Videotape.... In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 138–145. doi:10.1145/223904.223922
[48] Michael A. Madaio, Jingya Chen, Hanna Wallach, and Jennifer Wortman Vaughan. 2024. Tinker, Tailor, Configure, Customize: The Articulation Work of Contextualizing an AI Fairness Checklist. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 214 (April 2024), 20 pages. doi:10.1145/3653705
[49] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–14.
[50] Sean McGregor. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 15458–15463.
[51] Microsoft. 2022. Harms Modeling - Azure Application Architecture Guide. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/harms-modeling/.
[52] Microsoft Azure Architecture Center. 2022. Responsible Innovation Community Jury Guide. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/.
[53] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In *Proceedings of the 44th international conference on software engineering*. Association for Computing Machinery, New York, NY, USA, 413–425.
[54] Arvind Narayanan and Sayash Kapoor. 2024. *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference*. Princeton University Press, Princeton, NJ, USA.
[55] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. http://www.jstor.org/stable/j.ctt1pwt9w5
[56] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. doi:10.1145/3287560.3287567
[57] Samir Passi and Steven J Jackson. 2018. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–28.
[58] David Piorkowski, Soya Park, April Yi Wang, Dakuo Wang, Michael Muller, and Felix Portnoy. 2021. How ai developers overcome communication challenges in a multidisciplinary team: A case study. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–25.
[59] Vivian Qiang, Jimin Rhim, and AJung Moon. 2024. No such thing as one-size-fits-all in AI ethics frameworks: a comparative case study. *AI & SOCIETY* 39, 4 (2024), 1975–1994.
[60] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. doi:10.1145/3351095.3372873
[61] Bogdana Rakova, Jingying Yang, Henriette Cramer, and Rumman Chowdhury. 2021. Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–23.

[62] Madhu C Reddy, Paul Dourish, and Wanda Pratt. 2001. Coordinating heterogeneous work: Information and representation in medical care. In *ECSCW 2001: Proceedings of the Seventh European Conference on Computer Supported Cooperative Work 16–20 September 2001, Bonn, Germany*. Springer, 239–258.

[63] Kjeld Schmidt. 1997. Of maps and scripts—the status of formal constructs in cooperative work. In *Proceedings of the 1997 ACM International Conference on Supporting Group Work* (Phoenix, Arizona, USA) *(GROUP '97)*. Association for Computing Machinery, New York, NY, USA, 138–147. doi:10.1145/266838.266887

[64] Kjeld Schmidt and Liam Bannon. 1992. Taking CSCW seriously: Supporting articulation work. *Computer supported cooperative work (CSCW)* 1, 1 (1992), 7–40.

[65] Kjeld Schmidt and Carla Simone. 1996. Coordination mechanisms: Towards a conceptual foundation of CSCW systems design. *Computer Supported Cooperative Work (CSCW)* 5, 2 (1996), 155–200.

[66] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (Montréal, QC, Canada) *(AIES '23)*. Association for Computing Machinery, New York, NY, USA, 723–741. doi:10.1145/3600211.3604673

[67] Hong Shen, Wesley H Deng, Aditi Chattopadhyay, Zhiwei Steven Wu, Xu Wang, and Haiyi Zhu. 2021. Value cards: An educational toolkit for teaching social impacts of machine learning through deliberation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. Association for Computing Machinery, New York, NY, USA, 850–861.

[68] Hong Shen, Alicia DeVos, Motahhare Eslami, and Kenneth Holstein. 2021. Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–29.

[69] Susan Leigh Star and James R. Griesemer. 1989. Institutional Ecology, 'Translations' and Boundary Objects: Amateurs and Professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science* 19, 3 (1989), 387–420. http://www.jstor.org/stable/285080

[70] Anselm Strauss. 1988. The articulation of project work: An organizational process. *Sociological Quarterly* 29, 2 (1988), 163–178.

[71] Lucille Alice Suchman. 1987. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press.

[72] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[73] Elham Tabassi. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0). doi:10.6028/NIST.AI.100-1

[74] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[75] Qiaosi Wang, Michael Madaio, Shaun Kane, Shivani Kapania, Michael Terry, and Lauren Wilcox. 2023. Designing responsible ai: Adaptations of ux practice to meet responsible ai challenges. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–16.

[76] Zijie J Wang, Chinmay Kulkarni, Lauren Wilcox, Michael Terry, and Michael Madaio. 2024. Farsight: Fostering Responsible AI Awareness During AI Application Prototyping. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–40.

[77] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65. doi:10.1109/TVCG.2019.2934619

[78] Richmond Y Wong, Michael A Madaio, and Nick Merrill. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27.

[79] Qian Yang, Justin Cranshaw, Saleema Amershi, Shamsi T Iqbal, and Jaime Teevan. 2019. Sketching nlp: A case study of exploring the right things to design with language intelligence. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[80] Nur Yildirim, Susanna Zlotnikov, Deniz Sayar, Jeremy M Kahn, Leigh A Bukowski, Sher Shah Amin, Kathryn A Riman, Billie S Davis, John S Minturn, Andrew J King, et al. 2024. Sketching AI Concepts with Capabilities and Examples: AI Innovation in the Intensive Care Unit. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[81] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.

A Initial UI Designs in Formative Study

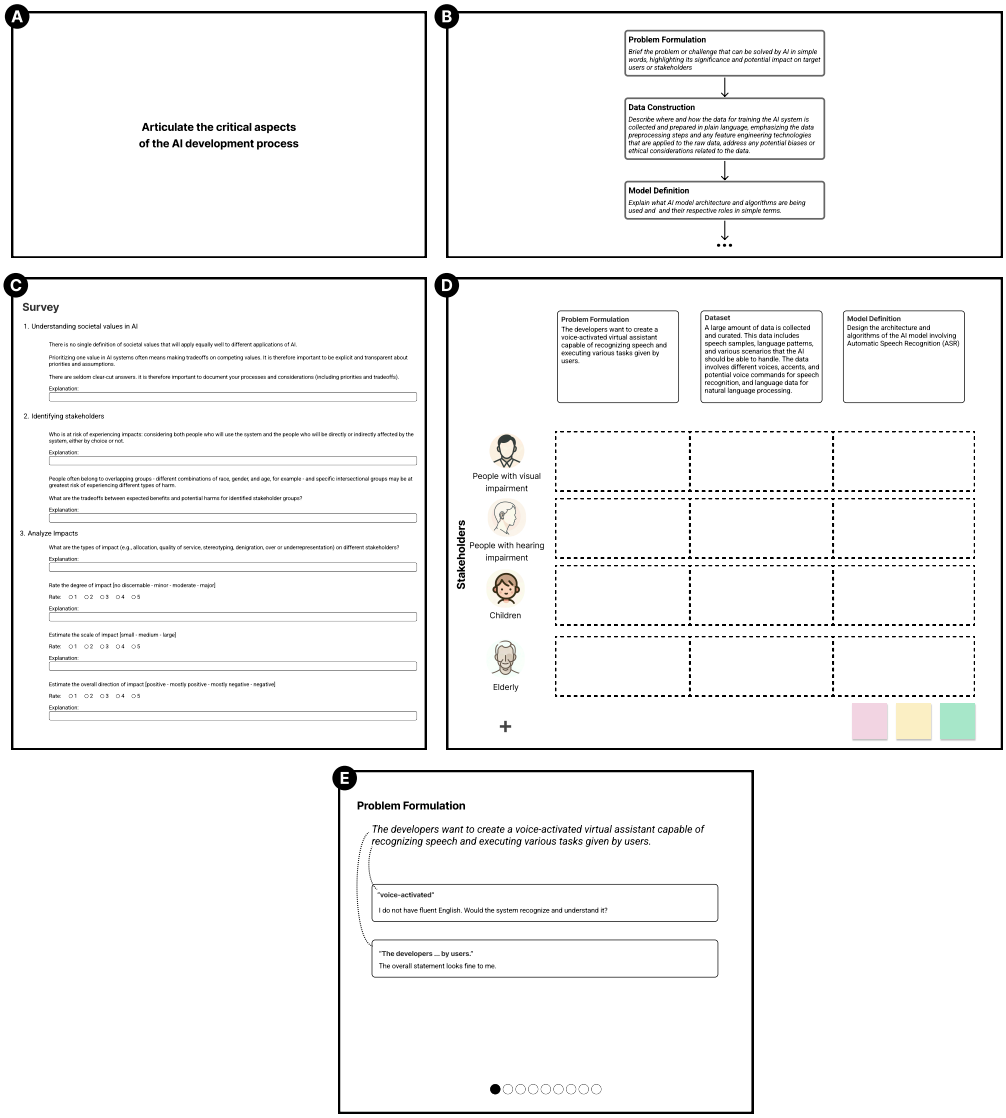

Fig. 4. Concepts & Artifacts explored in the formative semi-structured interview study. Planning: (A) Plain description presents currently unstructured approaches to articulating AI development plans in cross-functional teams; (B) Stage-based scaffolding breaks down and connects AI development stages while providing prompts, as inspired by AI storyboard [24]. Harm identification: (C) The survey was adapted from Value Cards [67] investigating different social-technical dimensions, (D) Stakeholder table visualizes AI development stages versus stakeholders for better sensemaking processes, (E) Commenting represents common features in asynchronous communication tools.

B Formative Study Semi-Structured Interview Protocol

Note: Follow-up questions and probes were used throughout the sessions to clarify responses.

Background and Current Practice

Goal: Understand participants' roles, existing workflows, and baseline coordination practices.

(1) **Role and context**
- Can you briefly describe your current role and responsibilities?
- Do you primarily work in a technical, non-technical, or mixed role when developing AI systems?

(2) **Current tools and workflows**
- What tools do you currently use for project tracking, planning, or interdisciplinary collaboration (e.g., JIRA, Asana, Google Docs)?
- How do these tools support collaboration between technical and non-technical or user-facing roles?

(3) **Demonstration and reflection**
- If possible, could you walk us through how you used one of these tools in a recent project?
- What aspects of the tool worked well? Where did you encounter friction or limitations?
- Please take a moment to write down any thoughts, preferences, or frustrations you have with these tools.

B. Experiences with Ethical Issues and Cross-functional Communication

Goal: Surface concrete breakdowns in articulation, handoff, and harm-related reasoning.

(1) **Concrete experience**
- Can you recall a specific instance where ethical or harm-related concerns arose during an AI project?
- How were these concerns identified, discussed, and addressed within your team?

(2) **Communication challenges**
- Were there moments when communication between technical and non-technical roles was difficult? Why?
- What kinds of information were hardest to convey or understand across roles?

(3) **Scenario-based probe (if no prior example)**
- Imagine your team is building an AI system for a given application scenario.
- How would your team plan the system and identify potential risks or harms?
- What challenges do you anticipate in this process?

AI Development Lifecycle and Planning Representations

Goal: Probe reactions to different representations of early-stage AI planning.

(1) **Lifecycle reflection**
- (Show AI development lifecycle figure [24])
- Does this lifecycle resemble how AI development is planned in your work?
- Are there stages you would add, remove, or merge?

(2) **Artifact comparison: planning**
- (Show Artifacts A and B)
- How effective is each approach for explaining key design choices?
- Which representation would better support collaboration across roles, and why?
- How might these representations affect early identification of potential issues?

D. Harm Identification and Evaluation Scaffolds

Goal: Understand how participants reason about harms and evaluate different scaffolding approaches.

(1) **Artifact comparison: harm identification**
- (Show Artifacts C, D and E)
- How effective is each approach for harm identification?
- How would you use each of these in your current workflow?
- What do you find helpful or challenging about each approach?

(2) **Comparative reflection**
- Which artifact(s) best support systematic evaluation of potential harms?
- Which better support open-ended discussion or iteration?
- How do these approaches differ in supporting technical versus non-technical roles?

C AI LEGO UI Components

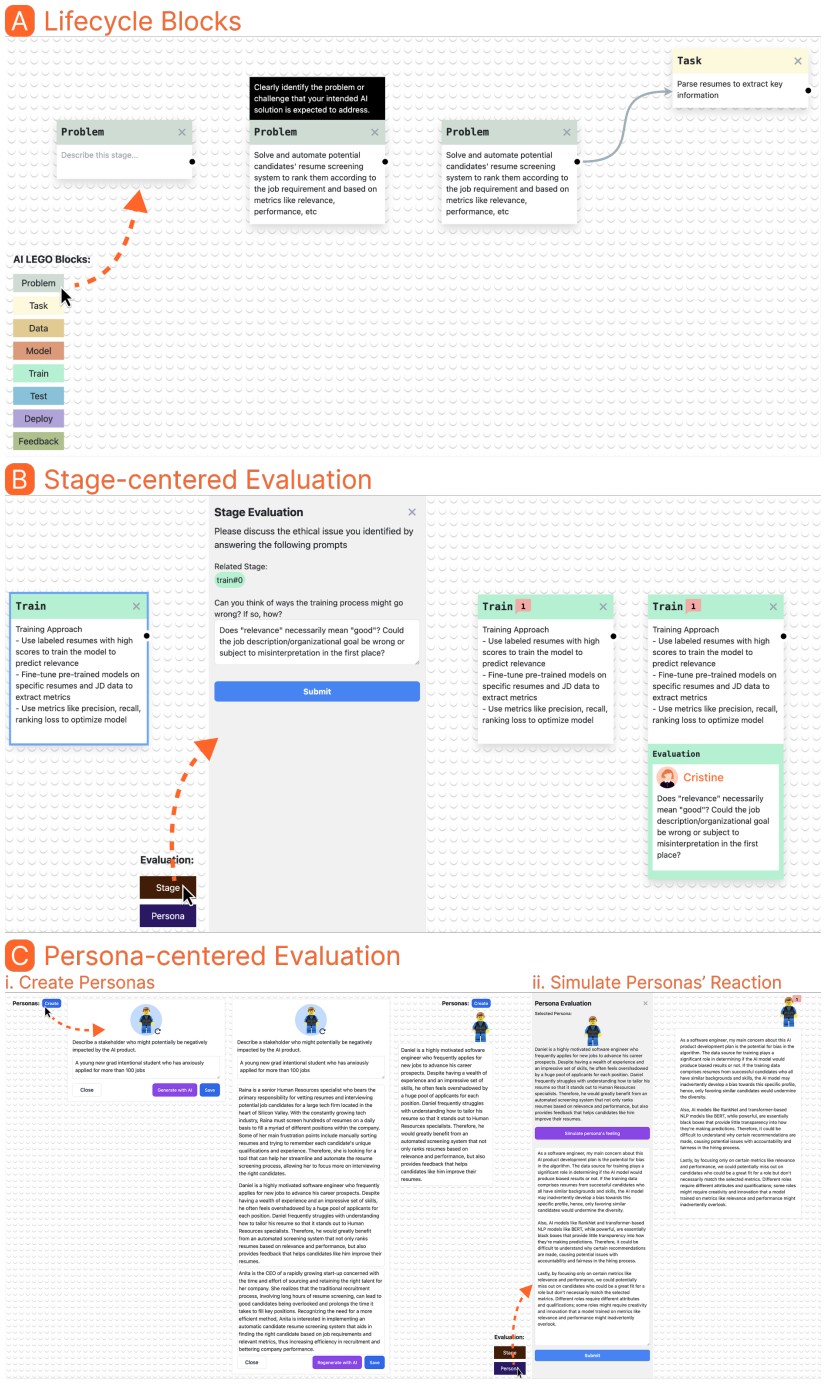

Fig. 5. Key components and transitions in AI LEGO user interface.

D Prompts & Examples for Persona-based Evaluation

Persona Generation:

```
  Generate three distinct personas, each described in a concise one-paragraph
format. These personas should reflect individuals who are either directly
or indirectly impacted by the identified problem(s): {problem descriptions}.
Additionally, they should complement the Existing Personas by representing a
broader spectrum of stakeholders. Each paragraph should begin with a brief summary
of the persona, followed by a detailed account of their background, including
relevant personal and professional aspects, motivations, and how they engage with
the problem. Existing Personas: {personas}
```

Example Output: Samantha is a Human Resource Manager at a fast-growing tech startup. In charge of recruitment and talent management, she has an average of 200 resumes to screen for each open position. Samantha has over a decade of experience in HR which has equipped her with a good eye for filtering resumes. But due to the surge in applications, her workload has significantly increased which is slowing down the hiring process considerably. The stress of potentially missing out on quality candidates due to a manual and time-consuming process affects her greatly. She is in search of an automated resume screening tool that can drastically reduce the time taken for this step and improve efficiency, allowing her to focus more on the subsequent interview and negotiation stages.

Stakeholders' Feeling Simulation:

```
  Assume you are {persona description}. You will be presented with a development
plan for an AI product and asked to express your concerns about its potential
negative impact on you. Your response should be clear, concise, and well-reasoned,
aligning with your persona's background, lived experience, and perspective. AI
Development Plan:
{{stage name}: {stage description} for all stages}
```

Example Output: While the proposed AI implementation certainly presents an innovative and efficient approach towards streamlining the hiring process, it raises certain concerns and implications from an HR perspective.

One of the primary concerns is accuracy. The AI performance would heavily depend on how well the algorithms are trained and fine-tuned. It might fail to detect subtlety, creativity, or potential in candidates' resumes that human judgment can ascertain. Therefore, we could lose out on good candidates because of the reductionist approach the AI might take to screen resumes.

Secondly, the customization of models based on the resume-job fit is concerning. Given the dynamic nature of each job role and its requirements, the AI might not accommodate the nuances and specifics of each individual role.

Thirdly, bias is a significant concern while deploying AI models for recruitment. If the current employees' data is used for training these models, there's a risk of replicating biases that might exist in the present hiring process. The AI model could be biased against non-traditional career paths, gaps in employment, or non-conventional educational backgrounds.

Finally, the ethical dimension of data privacy must be considered. Using public resume datasets or synthetic data shown during model training emphasizes the need for clear, transparent communication with applicants about how their personal information is being used.

E AI System Development Scenarios

E.1 Scenario 1: Image Generation

Background: Creating custom visual content efficiently is essential for rapid prototyping, creative exploration, and personalized media production today. This service can also significantly boost a company's revenue.
Project goal: Develop an AI-powered image generation tool that creates high-quality, contextually relevant images from text prompts.
Evaluation metrics: Image quality; Generation speed; Model robustness.
Key features: Text-to-image conversion using advanced AI models; Customizable image style options.
Requirements: Utilize transformer-based models for image generation; Get the model trained from the selected dataset.
Deployment: Use cloud-based services for scalability and reliability. Regularly update the model with new data to improve quality.
Target Users: Graphic designers; Content creators; General consumers.

E.2 Scenario 2: Gunshot Detection

Background: Quicker and more accurate responses to gun-related incidents are crucial for enhancing urban public safety.
Project goal: Develop an AI-powered gunshot detection system that identifies and alerts gunshot sounds in real time.
Evaluation metrics: Detection accuracy; Response time; Reliability under different environments.
Key features: Real-time detection of gunshots; Automatic alerts to local police station; Integration with existing infrastructure.
Requirements: Utilize a proper AI model for gunshot detection; Dataset needs to be created and marked by experts.
Deployment: Implement edge computing for real-time processing and scalability; Cloud-based monitoring.
Target Users: Law enforcement agencies; Security Companies; Police.

E.3 Scenario 3: Resume Screening

Background: Manually processing a high volume of applications can be both time-consuming and resource-intensive in large organizations.
Project goal: Develop an AI-powered resume screening tool that automatically evaluates and ranks job candidates based on their resumes.
Evaluation metrics: Accuracy in extracting relevant information and ranking candidates.
Key features: Automatic extraction and analysis of key resume information.
Requirements: Use NLP models for parsing and analyzing resume content. Datasets are given from internal sources from the company. Rank the candidates from multiple metrics.
Deployment: Integrate the AI tool with existing ATS or HR software platforms. Use cloud-based deployment for scalability and ease of updates
Target Users: HR Departments; Recruitment Agencies

E.4 Scenario 4: Vaccine Allocation

Background: Equitable and needs-based distribution of COVID-19 vaccines would maximize the safety of people.

Project goal: Develop an AI-powered vaccine allocation tool that prioritizes individuals based on urgency and need, ensuring equitable distribution during pandemics.
Evaluation metrics: Accuracy of risk stratification; Fairness in prioritization; Adaptability to new variants; Effectiveness of update.
Key features: Risk assessment based on health data, demographics, and exposure risk. Rank individuals by vaccination urgency; Real-time updates.
Requirements: Use ML models for risk stratification and prioritization; Datasets are given from personal medical records and epidemiological data; Rank the candidates from multiple metrics.
Deployment: Integration with existing health information systems (EHRs, public health databases); Use cloud-based deployment for real-time updates.
Target Users: Public health organizations; Government agencies; Vaccine providers.

E.5 Scenario 5: Credit Assessment

Background: Traditionally, the process of determining creditworthiness has been complex and time-consuming, often resulting in lengthy waits for users.
Project goal: develop an AI-powered tool that evaluates and recommends credit card lines based on individual financial behavior and risk assessment.
Evaluation metrics: Accuracy compares to historical data (backtesting).
Key features: Assessment of creditworthiness using historical financial data; Risk scoring tool.
Requirements: Datasets are given from personal credit records and bank agent-approved data; Estimate the credit line based on multiple metrics from applicants' historical finance and credit data.
Deployment: Use secure cloud-based deployment for real-time updates; Integration with existing banking systems and customer profiles.
Target Users: Credit card issuers.

F Evaluation Study Semi-structured Interview Protocols

Overall comparison. Rank these tools from most to least effective for this task, and why? Were there moments when one tool made the task easier or harder?
Understanding and handoff. How easy was it to create (AI) or understand (UX, PM) the development plan in each condition? Did any tool reduce the need to ask for clarification?
Workflow structure. How did the organization and layout of each tool affect how you worked through the task?
Harm identification (non-AI only). Which tool best supported you in identifying potential harms? What features contributed most to this?
Persona-centered evaluation (non-AI only). If applicable, how did the persona feature influence your thinking? Were there cases where it felt unhelpful or unrealistic?
Limitations and improvements. What aspects of AI LEGO would you improve to better support RAI work in practice?

G Evaluation Study Scenario & Condition Orders

Fig. 6. Scenario & Condition orders in the user study.

H Harm Rating Rubrics

Table 2. Grading rubrics for harm ratings, developed using the Socio-technical Harm Taxonomy [66]. Specific ratings were determined based on the details and nuances of individual cases.

Harm Type	Specific Harm	Severity Range	Likelihood Range	Example
Representational Harms	Unverified/Outdated/Inappropriate dataset	3–4	3–4	There could be biased data in the resume pool dataset for attributes like gender, races. The model trained on this will make biased decisions, which will reinforce people's biases. (G3PM)
	Uniformity of model metric	2–3	2–3	The choices of metrics would determine the social impacts (not hiring strong applicants for some other reasons; or hiring people who can only make money,) which might be unethical,l but fit well to the companies goal or social role. (G3UX)
Allocative Harms	Inappropriate objectives in training leading to imbalanced inference	3–4	3–4	AUC score is not an appropriate metrics for "success" but instead should evaluate the model with human feedback. (G6UX)
Interpersonal Harms	Privacy of service	2–3	3	(Gunshot detector) may cause privacy and social inequality for a certain neighborhood. (G4UX)
	Deploy data steal	1	4	Using a cloud-based platform may lead to user data leaks. (G4UX)
	Inaccurate decision	3–4	2–4	Innocent people (may be) identified as the shooter due to inaccurate gunshot detection (G3UX)
	Harm to specific users	4	4	Image generator might not be able to generate images tailored to specific needs of customers (like logo designing for companies) if it uses a very general dataset of images. (G2UX)
	Data leak in training	1	3	When training the model (for credit line assessment), there might be a data leak of personal information. (G2PM)
Social System Harms	Explainability/Accountability	3–4	3	Model lack of explanation, too simple model(decision tree), may cause trustworthy issue for whoever using it. (G2PM)
	Copyright	3	3	Data sources might have copyright issues for commercial uses that are stolen in the first place. (G2UX)
	Changed behavior	1–4	1–4	Some users might change their financial behaviour based on the result of credit information gotten from AI tool. (G4PM)
	Insufficient feedback	1–3	1–3	(The system should) have user input feedback rather than just thumb up and down. (G2PM)

Received 13 May 2025; revised 13 January 2026; accepted 17 March 2026