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

MUZHE WU*, Carnegie Mellon University, 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

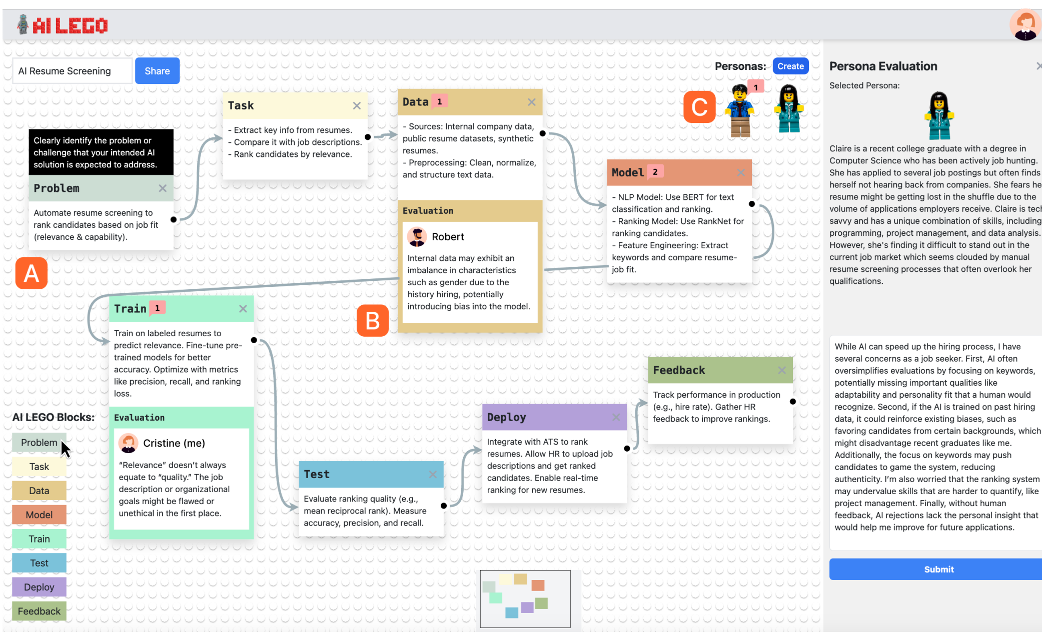

Fig. 1. AI LEGO is a web tool that scaffolds cross-functional collaboration (esp., knowledge handoff) among industrial AI practitioners to proactively plan AI system development and identify problematic design choices. With AI LEGO, technical AI developers can first sketch out a development plan using (A) *Lifecycle Blocks*, each corresponding to a specific stage in AI development, along with a tailored prompt. They then hand off the plan to non-technical/user-facing roles, who conduct (B) *Stage-centered Evaluation* to systematically identify and address design choices that could lead to potential harm at each stage of development. Finally, the (C) *Persona-centered Evaluation* feature helps them surface edge cases in harm identification by generating personas of impacted stakeholders and simulating their perspectives.

*Both authors contributed equally to this research.
†Now at Google DeepMind.

Authors' addresses: Muzhe Wu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, muzhew@andrew.cmu.edu; Yanzhi Zhao, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, zyz199874@gmail.com; Shuyi Han, Northwestern University, Evanston, Illinois, USA, shuyihan2026@u.northwestern.edu; Michael Xieyang Liu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, xieyangl@cs.cmu.edu; Hong Shen, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, hongs@andrew.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 knowledge handoff challenge: the difficulty of transferring high-level, early-stage technical design rationales from technical experts to non-technical or user-facing roles for ethical evaluation and harm identification. Through literature review and a co-design 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 supports 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.

ACM Reference Format:
Muzhe Wu, Yanzhi Zhao, Shuyi Han, Michael Xieyang Liu, and Hong Shen. 2025. AI LEGO: Scaffolding Cross-Functional Collaboration in Industrial Responsible AI Practices during Early Design Stages. 1, 1 (May 2025), 26 pages. https://doi.org/XXXXXXX.XXXXXXX

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 (e.g., [27, 31, 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 (e.g., [9, 16, 64, 67]). High-level policy frameworks have also been proposed to ensure the responsible development of AI, such as NIST's AI Risk Management Framework (e.g., [2, 11, 37]).

In light of these concerns and policy initiatives, researchers in Responsible AI (RAI) and Human-Computer Interaction (HCI) have developed various technical tools and interactive systems to help AI teams better evaluate and de-bias their models (e.g., [10, 13]). 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 [22, 36]; and (2) being conducted only after the system is built or deployed, when significant harms had already occurred [23, 69]. Recently, there has been increasing recognition on the socio-technical nature of AI [18, 34, 36] and how involving non-technical/user-facing roles in industry cross-functional AI teams (e.g., UI/UX designers and project/product managers) [24], often without advanced technical expertise, can at early design stages contribute in uncovering problematic design choices that may lead to downstream harms in a wide range of AI-powered systems [24, 41, 70].

XXXX-XXXX/2025/5-ART
https://doi.org/XXXXXXX.XXXXXXX

Previous research has identified several high-level challenges faced by cross-functional AI practitioners around communication gaps between technical developers and non-technical or user-facing roles [24, 57, 60, 70, 73]. One persistent issue is the "knowledge handoff" challenge, where technical roles struggle to transfer high-level design rationales and technical constraints to non-technical roles for ethical evaluation and harm identification [24]. To better understand the barriers to effective communication in early-stage cross-functional collaboration and to identify specific design opportunities, we conducted a co-design study with eight cross-functional AI practitioners. Our findings unpack how this barrier manifests — technical design choices is 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 systematically harm evaluation. Based on these insights, we derived three key design goals to scaffold early-stage RAI collaboration: (1) Facilitating effective knowledge handoff from technical roles on AI development planning; (2) Streamlining collaborative workflow with customizable visual tools; and (3) Scaffolding structured harm identification processes by non-technical/user-facing roles.

To address these challenges identified by prior work as well as our formative co-design study, we developed AI LEGO (Fig. 1), an interactive web tool 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 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 15% higher likelihood of occurring, compared to using Google Docs and Harm Modeling framework [50] baseline. Participants also reported that AI LEGO was effective in bridging the gaps in knowledge handoff providing flexibility, and inspiring multi-perspective, structured harm evaluation. The contributions described in this work include:

- **Design goals** from a formative co-design study and literature review to tackle the challenges of scaffolding cross-functional collaborations (esp., knowledge handoff from technical roles to non-technical roles) in RAI work during the early design stage.
- **AI LEGO, an interactive web tool** to support cross-functional AI practitioners in effectively facilitating knowledge handoff and identify 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 effectively cross-functional collaboration in conducting RAI work, especially in the early-stage AI design.

2 RELATED WORK

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

Over the years, the RAI community in HCI and 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 (e.g., see [4, 19, 36, 47, 71]). Several interactive tools and systems have been developed to help practitioners understand, evaluate, and mitigate RAI-related concerns, with a strong emphasis on bias and fairness (e.g., see [10, 13, 72]). Despite their contributions, this early line of research has faced criticism for (1) failing to uncover certain "blind spots," often due to the lack of diversity in *technical* AI teams [22, 36], and (2) being conducted after the system is built, when challenges like biased dataset or problematic proxies become harder to resolve, or significant harms have already occurred [23, 69].

Recently, with increasing emphasis on the importance of the early design process [75–77], researchers have begun developing tools and systems to support AI developers in RAI work at earlier stages, before models, systems, and applications are fully developed. For example, Farsight [71] helps anticipate potential harms in LLM-powered applications during the early prototyping phase. Lam et al. [40] developed a technical framework to support ML practitioners iteratively authoring functional approximations of a model's decision-making logic. However, these tools are typically designed to support either ML practitioners or AI prototypers and rarely serve as a communication channel between technical AI developers, who design ML models and applications, and non-technical/user-facing roles, who have a deeper understanding of the users. This gap in communication and feedback mechanisms limits opportunities for model refinement, despite prior research highlighting the critical importance of cross-functional collaboration in RAI [24].

We contribute to this line of research by designing interactive tools that push RAI work to occur *much earlier* in the design process and to be *more interdisciplinary.* AI LEGO develops a few scaffolding techniques (e.g., eight-stage blocks and worksheets) to support the identification and communication of high-level, critical design decisions both across different roles in industrial AI teams and across the entire AI development lifecycle, from task definition to feedback collection [52, 61, 78]. This approach allows cross-functional AI teams to collectively identify harmful problems and iterate on high-level designs much earlier, before investing time and effort into implementing the system.

2.2 Cross-functional collaboration in Responsible AI

With the growing consensus that AI systems are fundamentally *socio-tehnical* (e.g., [18, 20, 34, 36, 41, 63, 65]), more RAI work in industry settings requires both technical roles and non-technical/user-facing roles throughout the AI design and development process — a practice often referred to as "cross-functional collaboration" [24]. Previous work in HCI and CSCW has provided valuable insights into the challenges of cross-functional collaboration, both in RAI and beyond (e.g., [3, 26, 38, 42, 45, 48, 53, 56, 68, 78]). For example, Deng et al. [24] conducted interviews and design workshops, reporting that practitioners from different roles often engage in "bridging" work to overcome frictions, such as interpreting fairness vocabularies and reconciling qualitative data from user research with quantitative metrics, when addressing fairness-related concerns. These specific frictions are part of a broader pattern of "communication gaps," a recurring challenge highlighted across prior work often stemming from knowledge disparities [24, 57] and misaligned workflows, such as involving non-developers only after the system or model has already been built [60, 70, 73]. Madaio et al. [46] further identified several nuanced needs of cross-functional AI teams when contextualizing fairness checklists, including better integration of checklists into existing collaborative workflows and managing tensions between individual ownership of fairness decisions and the presence of a designated approver. All these challenges reflect longstanding themes in the CSCW literature on planning, negotiation, and coordination in organizational work practices [66].

Despite this growing emphasis and recognition of challenges, there is still a lack of tools to support cross-functional AI practitioners in RAI work [3, 23, 24, 73]. Previous work (e.g., [17, 26, 75]) developed design resources to support cross-functional collaboration. For example, the *AI Brainstorming kit* [75] aims to scaffold AI concept ideation within multidisciplinary teams, with a goal to identify low-risk, high value AI concepts. Constantinides et al. [20] developed an interactive webpage that provides role-specific guidelines grounded in regulations for individual uses. However, currently, no tools have been developed to address the inherently collaborative nature of industrial practices, specifically focusing on the early design stages [46].

In this paper, we introduce AI LEGO with the goal of helping bridge the gap between technical AI developers and non-technical/user-facing roles in industrial RAI work and push this work into the early design stage. Inspired by prior literature on sharing and reusing summarized knowledge [42], our tool enables AI developers to hand off high-level AI designs to non-developers, using non-technical terms, which has been highlighted as useful in helping UX designers and product managers gain a practical understanding of complex AI systems [76]. In addition, it also supports non-developers in more effectively identifying problematic AI designs at the early design stage and communicating their feedback back in a format that is easy for technical roles to comprehend.

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, 29, 44]. 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 (e.g., [8, 25, 30, 47, 63]). For example, Madaio et al. [47] co-designed a set of AI fairness checklists with 48 industry practitioners to operationalize high-level ethical principles. They found that these checklists could serve as a formalized organizational framework, transforming ad-hoc processes into structured approaches and empowering individual advocates to promote ethical AI practices more effectively. Raji et al. [59] 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.[58] 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.

With the rise of generative AI models, increasing attention has been given to how they can be leveraged to contextualize, accelerate, and inspire human evaluation in RAI work. Buçinca et al. [15] introduced a general framework that assists AI practitioners and decision-makers in anticipating potential harms of AI systems, using large language models (LLMs) to generate descriptions of potential harms for various stakeholders. In a similar vein, Farsight [71] 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, drawing inspiration from [39], we developed a stage-based checklist to support harm anticipation, which covers key design choices across the entire AI development lifecycle. Second, building on [15, 63, 71], we incorporated LLMs to assist non-AI developers in brainstorming potentially impacted stakeholders and gaining a deeper understanding of their lived experiences.

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 co-design study[1] as detailed below.

[1]The formative study along with the evaluation study were approved by our institution's IRB (STUDY2023_00000435).

3.1 Co-design Study

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 product managers, and 2 UI/UX designers) [20] and work at companies of varying sizes[2]. 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 [46], 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 co-design study session lasted approximately 60 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, followed by a semi-structured formative interview to explore participants' current workflows and challenges they have faced in cross-functional collaboration between technical and non-technical/user-facing roles, helping us understand their practices and pain points. We then presented five preliminary design artifacts (Fig. 4 in Appendix A), 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 [21], and three UIs for brainstorming potential harms: a survey adapted from Value Cards [63], 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. The design rationales of each artifact are provided in Appendix A.

3.2 Findings

We analyzed participants' (P1-8) notes with affinity diagrams [43] and derived the following insights:

Communication Challenges. Resonating with the literature [24, 57], participants consistently reported inefficiencies in communication between technical and non-technical/user-facing roles ($n = 5$). 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), all ($n = 8$) agreed that initiating an AI development plan is best led by technical roles or practitioners with strong technical expertise to ensure feasibility. P3 noted, "My PM often overestimates what the model can achieve." However, some participants ($n = 3$) also acknowledged the challenge of translating technical specifications into high-level design choices that non-technical or user-facing roles can understand and effectively evaluate. Most of the communication management tools used by participants, such as JIRA [6], Asana [5] and Google Docs [33], were primarily designed for fast turnarounds in product development or generic usages but either lack space for extending communication to RAI values or "take a lot of time to set up" (P8). To establish a mutual understanding, both technical and non-technical/user-facing roles often need to spend extra efforts on learning from guidelines or technical documentation, possibly through synchronous sessions, which many participants found infeasible and unnecessary ($n = 4$).

Scaffolding Early-Stage AI Development Planning. Comparing the two artifacts, all participants ($n = 8$) preferred stage-based scaffolding, acknowledging its effectiveness in enabling flexible planning and cross-functional comprehension. 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,

[2]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.

"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.

Scaffolding Potential Harms Identification. Participants (n = 4) appreciated the straightforwardness and in-depth investigation of the survey adapted from Value Card [63], 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 literature review and co-design study, we establish the following design goals:

- **G1: Facilitate effective knowledge handoff from technical roles.** To support effective RAI work, the system should bridge communication gaps by first enabling technical roles to "hand off" their understandings of AI development tasks at high levels so that non-technical/user-facing roles can meaningfully engage in evaluating the AI system's development.
- **G2: Streamline collaborative workflow with customizable visual tools.** The system should provide a flexible, visual workflow, breaking down AI development tasks into modular, manageable components that both technical and non-technical team members can easily understand and interact with.
- **G3: Scaffold structured evaluation by non-technical roles.** The system should integrate structured evaluation methods that guide non-technical members through assessing potential AI harms at every stage of the AI development process while fostering the awareness of stakeholders and open-ended brainstorming to overcome "blind-spot."

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 raises concerns about its impact on accuracy. To explore further, Emma uses the *Persona-centered*

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?

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

The core of AI LEGO in scaffolding cross-functional collaboration in RAI work is *Eight-Stage Worksheet* and *Checklist* (Table 1). *Eight-Stage Worksheet* and *Checklist* consist 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 the co-design study on the perceived effectiveness of "stages," we developed *Eight-Stage Checklist* and *Checklist* to address the eight key stages — originally proposed by Cramer et al. [21] as representing the typical AI development lifecycle further validated as comprehensively reflecting participants' workflows in our formative study. The curation of prompt questions took an iterative process to ensure that they could inspire social-technical discussions around crucial aspects in each stage.

4.3 User Interface

AI LEGO further develops a visual, interactive interface with the following components (see Fig. 1):

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 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 address the "expert blind spot" issue encountered in existing harm identification RAI tools [22, 36], AI LEGO introduces *Persona-centered Evaluation*, a feature leveraging LLMs to scaffold brainstorming. 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 (e.g., [15]). These simulated reactions are saved and displayed when hovering over the persona icon.

The changes of UI states due to user interactions are detailed in Fig. 5 in Appendix B.

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[3] package. While the frontend is hosted on Vercel, the backend relies on Google Firebase for user authentication and data persistence. The *Persona-centered Evaluation* functionality leverages the OpenAI GPT-4 model with zero-shot prompting techniques (see Appendix C).

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 replicate real-world cross-functional collaborative settings in a controlled environment, we organized participants into teams and divided the study sessions into two phases. To understand how individual features of AI LEGO can address the existing challenges differently, we adopted a within-subject design and created two variants of our tool: AI LEGO LITE and AI LEGO FULL. Both

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

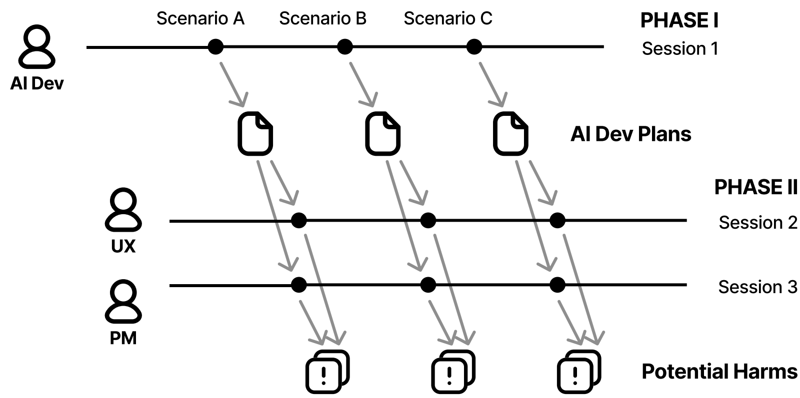

Fig. 2. Two-phase team-based evaluation study task flow.

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 project/product managers (PMs), and 6 UI/UX designers[4]. 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 two-phase task flow with three one-on-one study sessions for each team is shown in Fig. 2. In **Phase I**, the technical AI developer within a team 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 scenario (see Section 5.2.3) and leveraging a different tool corresponding to the conditions (see Section 5.2.2). In **Phase II**, the product manager and UX

[4]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).

designer 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. The study follows a within-subject design, with each participant experiencing 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 [50], 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. Compared to AI LEGO FULL, AI LEGO LITE did not support *Persona-centered Evaluation*. This allowed us to assess the effectiveness of this LLM-based feature in Phase II. Hence there is no difference between AI LEGO LITE and AI LEGO FULL in Phase I of drafting AI development plans[5]. The order of conditions was fixed for participants within each team and was counterbalanced across teams through Latin squares.

5.2.3 AI System Development Scenarios. To mimic typical industrial early-stage AI design processes, we created five scenarios (see Appendix D) where AI systems are to be developed to address real-world problems. These scenarios were sourced and adapted from the AI Incident Database [49]. 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 E.

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 semi-structured interview, 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 the baseline Google Doc 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 [71], we analyzed the count, severity, and likelihood of identified harms to evaluate the effectiveness of AI LEGO (RQ1). After collecting the identified

[5]There are effectively two conditions in Phase I, i.e., Google Doc and AI LEGO LITE (twice).

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. After the initial ratings, the authors resolved discrepancies through discussion and developed a grading rubric based on Sociotechnical Harms Taxonomy [62]. The initial ratings showed a fair to moderate agreement with Cohen's Kappa scores of 0.30 for severity and 0.53 for likelihood. The final grading rubric with examples are provided in Table 2 in Appendix F.

The interval data (count of identified harms) was analyzed using a repeated ANOVA, followed by Tukey's HSD for pairwise comparisons. This approach was appropriate as the Shapiro-Wilk test indicated no significant deviation from normality ($p > .05$). The ordinal data (severity, likelihood of identified harms, SUS scores, average preference rankings) was primarily analyzed with non-parametric Kruskal-Wallis tests, followed by Dunn's post-hoc tests. For SUS scores and preference rankings in Phase I, the Mann-Whitney U test was applied instead as there were only two conditions.

5.3.2 Qualitative Analysis. We transcribed the semi-structured interview recordings and applied thematic analysis [32] to identify key themes, patterns, and insights from user feedback on how our tools impact RAI practices (RQ2). Recurring expressions related to user experiences, tool functionality, usability, and effectiveness were coded. These codes were grouped into potential themes, which were reviewed, refined, and supported by re-analyzing the relevant extracts.

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 FULL enables the cross-functional team to uncover more problematic design choices. The repeated ANOVA test revealed a significant main effect of conditions ($F(2, 33) = 16.26, p < .001$) on the number of problematic design choices identified. A post-hoc Tukey HSD test further showed that participants using AI LEGO FULL ($m = 5.6$) identified significantly more (195%) harms compared to those in the Google Doc condition ($m = 1.9, p < .001$). This suggests that AI LEGO FULL equips cross-functional teams with the ability to explore and surface more potential harms during the AI development process, offering a broader perspective on possible design flaws compared to traditional methods.

6.1.2 AI LEGO FULL allows cross-functional teams to identify more pressing and relevant AI system design flaws. The Kruskal-Wallis test revealed a significant main effect of conditions on the likelihood of identified harms ($H = 7.33, p = .03$). Post-hoc Dunn's tests showed that the harms identified using AI LEGO FULL ($mdn = 3, m_R = 81.3$) were rated as 15% more likely to occur compared to those identified in the Baseline Google Doc condition ($mdn = 3, m_R = 57.6, p = .02$). Although no significant differences were found in the other pairwise comparisons, this result underscores that AI LEGO FULL enables teams to focus on more pressing and likely design flaws, improving their ability to prioritize relevant risks in the development process. However, 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.

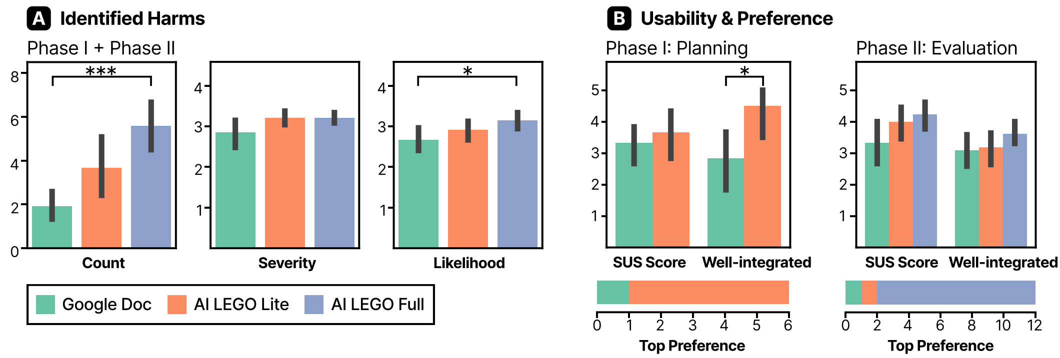

Fig. 3. Effect of conditions on (A) Count, Severity, and Likelihood of identified harms and (B) Usability and Preference. AI LEGO Full allowed AI teams to identify more potential harms with a higher likelihood; it was more preferred and considered well-integrated, especially in the planning phase. Significance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

6.1.3 *AI LEGO tools are more usable and preferred.*

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. Notably, participants provided significantly higher ratings for AI LEGO Lite in the planning phase on the question "I found the various functions in this system were well integrated" compared to Google Doc ($U = 4.00, p = .02$). For the planning phase, participants gave AI LEGO a significantly higher average ranking than Google Doc ($U = 30.0, p = .03$). Kruskal-Wallis H test also showed a main effect on average ranking for the evaluation phase ($H = 15.6, p < .001$). Post-hoc Dunn tests revealed that participants preferred AI LEGO Full over Google Doc baseline ($p < .001$).

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, as illustrated in the following quotes:

> *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. This consequently makes communication more explicit and actionable, rather than being ambiguous and unapproachable ($n = 4$), as noted by T4AI:

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

6.2.2 *AI LEGO presents the workflow through an intuitive and engaging layout (G2).*

Most participants ($n = 11$) found the tree taxonomy from by interconnected *Lifecycle Blocks* in AI LEGO suitable for the AI product design process. The overall layout was described as visually appealing ($n = 2$), which led to a sense of playfulness and increased motivation, as T4UX described:

> *AI LEGO gives a playful vibe in the theme of LEGO with the colorful blocks and minifigures.* (T4PM)

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. T3AI noted:

> *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)

Besides, AI LEGO was reported to contribute to direct manipulation, a key concept around the sense of control in HCI, through the seamless 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 neat layout and intuitive UI features consequently resulted in a reduced learning curve, as noted by 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)

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 ($p = .02$). Most of the non-technical participants ($n = 8$) attributed the enhanced evaluation efficacy primarily to the simulated persona feature. 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 more efficiently and from multiple viewpoints ($n = 6$). This was illustrated by the following quotes:

> *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)
> *Personas could stimulate your thinking process when you work alone by providing mocked data.* (T6UX)

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).

7 DISCUSSION AND DESIGN IMPLICATIONS

7.1 Design opportunities: Towards Early-stage, Cross-functional collaboration in RAI Practices

Supporting Early-stage RAI Envisioning. Compared to the few existing exploratory tools [40, 71], AI LEGO introduces a novel solution to structure communication between technical and non-technical roles within cross-functional teams, pushing the process of identifying potential harmful design decisions both *much earlier* and *more interdisciplinary*. Notably, AI LEGO simplifies high-level and early-stage RAI efforts by breaking down complex AI system development into modular, manageable components. The *Eight-Stage Worksheet* and *Checklist* — prompts tailored to the key stages of AI development lifecycle — proved effective in scaffolding communication across different roles during AI product planning. In particular, they highlighted the importance of "abstraction" [74] in collaboratively sketching complicated AI systems before actual implementation. AI LEGO further incorporates them into interactive UI components that can be edited as well as intuitively dragged and connected. Its overall tree-typology structure was highly appreciated and perceived as an appropriate information architecture by our participants, echoing the findings of prior work [71]. This design enables AI practitioners to engage with system planning in a more flexible yet formalized and standardized manner [46]. We envision future work could continue building on the "abstraction" and "standardization" provided by the AI LEGO concept to support RAI practices at the early stage of AI design.

Towards Human-AI complementarity in RAI Practices. On top of *Eight-Stage Worksheet* and *Checklist*, AI LEGO introduces the LLM-based personas to assist users in brainstorming potential stakeholders who might be negatively affected and simulate their perspectives. This design fundamentally raises the challenge between preserving human autonomy in ethical deliberation and using LLM to support brainstorming. Previously, harm-envisioning or auditing in RAI has relied largely on human intelligence [12, 22, 47]. With the advent of LLMs, recent work has explored how to use LLMs to better support humans in this process [15, 71]. Compared to tools like Farsight [71] that delegate subtasks of evaluation to LLMs, AI LEGO explores an alternative paradigm where users retain full control over the harm envisioning process while being able to actively inquire LLMs as an additional resource to surface edge cases. The design of *Persona-centered Evaluation* was based on the assumption that LLMs should only support participants in brainstorming the perspectives of stakeholders rather than replace human judgment or real-world user experiences to uphold engagement, accountability, and mitigate potential model biases [1]. Our results show that participants with this feature were able to identify more problematic AI design choices with a higher likelihood of risk and reported greater usability. However, participants also noted that the persona might not fully capture the complexity of real-world user experience. To some extent, this reflects a tradeoff between improving work efficiency — a core value in many industry settings — and making time for human reviewers to engage in contextualizing AI fairness [46], contemplating, and discerning potential risks in nuanced cases. We encourage future research to further explore how to better support such human-AI complementarity [35] in RAI practices, as well as better understand its limitations.

Integrating into Industrial Cross-functional Product Workflows. Prior research [24, 57] has examined the importance and complexities of industrial cross-functional RAI practices. Compared to other solutions (e.g., [8, 51]), AI LEGO aims to support such cross-functional collaboration by integrating into early product design processes, blending product planning with harm anticipation without imposing excessive workloads. The results of our evaluation demonstrated the feasibility of this concept. However, AI LEGO does not yet fully align with the diverse dynamics of industrial cross-functional collaboration. Real-world AI product development can sometimes follow a more agile and flexible approach, and is often influenced by other factors such as company size, team composition, and product evolution. Towards this end, the AI LEGO concept offered in this study can be further developed to add features found in other early-stage product management tools, such as support for work allocation and milestone tracking, increasing its practicality for supporting RAI practices in industrial cross-functional teams.

7.2 Limitations and Future Work

As an initial effort to support early-stage, cross-functional RAI work, our approach has a number of limitations. First off, 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 [40], 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 7.1, 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 [1, 28]. 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. This setup cannot fully capture the complexity of real-world industrial collaborative processes. 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 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.

REFERENCES

[1] 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*. 1–12.

[2] NIST AI. 2023. Artificial Intelligence Risk Management Framework (AI RMF 1.0).

[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*. 871–875.

[5] Asana, Inc. 2025. Asana Project Management Tool. https://asana.com Accessed: 2025-01-18.

[6] Atlassian. 2025. JIRA Software. https://www.atlassian.com/software/jira Accessed: 2025-01-18.

[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*. 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] Joseph R Biden. 2023. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence.

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

[13] 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. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[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 preprint arXiv:2306.03280* (2023).

[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] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2021. Onboarding Materials as Cross-functional Boundary Objects for Developing AI Assistants. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[18] Stevie Chancellor. 2023. Toward practices for human-centered machine learning. *Commun. ACM* 66, 3 (2023), 78–85.

[19] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why is my classifier discriminatory?. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 3539–3550.

[20] 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. https://doi.org/10.1145/3686927

[21] 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* Tutorial* 2 (2019).

[22] 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*. 1–18.

[23] 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*. 473–484.

[24] 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*. 705–716.

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

[26] Graham Dove and Anne-Laure Fayard. 2020. Monsters, metaphors, and machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–17.

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

[28] Xianzhe Fan, Qing Xiao, Xuhui Zhou, Jiaxin Pei, Maarten Sap, Zhicong Lu, and Hong Shen. 2024. User-Driven Value Alignment: Understanding Users' Perceptions and Strategies for Addressing Biased and Discriminatory Statements in AI Companions. *arXiv preprint arXiv:2409.00862* (2024).

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

[30] 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*. 1145–1148.

[31] Tarleton Gillespie. 2014. The relevance of algorithms. (2014).

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

[33] Google. 2025. Google Docs. https://docs.google.com Accessed: 2025-01-18.

[34] 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*. 19–31.

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

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

[37] White House. 2022. *Blueprint for an ai bill of rights: Making automated systems work for the american people.* Nimble Books.

[38] Sean Kross and Philip Guo. 2021. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–28.

[39] 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*. 1–17.

[40] 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*. 1–24.

[41] 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.

[42] 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.

https://doi.org/10.1145/3449240

[43] 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, 231–248.

[44] 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. https://doi.org/10.1145/223904.223922

[45] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach. 2022. Assessing the fairness of ai systems: Ai practitioners' processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[46] 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. https://doi.org/10.1145/3653705

[47] 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*. 1–14.

[48] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.

[49] 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.

[50] Microsoft. 2022. Harms Modeling - Azure Application Architecture Guide. https://learn.microsoft.com/en-us/azure/architecture/guide/responsibleinnovation/harms-modeling/.

[51] Microsoft Azure Architecture Center. 2022. Responsible Innovation Community Jury Guide. https://learn.microsoft.com/en-us/azure/architecture/guide/responsible-innovation/community-jury/.

[52] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[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*. 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.

[55] Safiya Umoja Noble. 2018. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

[56] 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.

[57] 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.

[58] 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.

[59] 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. https://doi.org/10.1145/3351095.3372873

[60] 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.

[61] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.

[62] 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.

https://doi.org/10.1145/3600211.3604673
[63] 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*. 850–861.
[64] 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.
[65] Hong Shen, Leijie Wang, Wesley H Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The model card authoring toolkit: Toward community-centered, deliberation-driven AI design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 440–451.
[66] Anselm Strauss. 1988. The articulation of project work: An organizational process. *Sociological Quarterly* 29, 2 (1988), 163–178.
[67] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.
[68] Rama Adithya Varanasi and Nitesh Goyal. 2023. "It is currently hodgepodge": Examining AI/ML Practitioners' Challenges during Co-production of Responsible AI Values. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–17.
[69] 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*. 1–14.
[70] 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*. 1–16.
[71] 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*. 1–40.
[72] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.
[73] 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.
[74] 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*. 1–12.
[75] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, et al. 2023. Creating design resources to scaffold the ideation of AI concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2326–2346.
[76] Nur Yildirim, Mahima Pushkarna, Nitesh Goyal, Martin Wattenberg, and Fernanda Viégas. 2023. Investigating how practitioners use human-ai guidelines: A case study on the people+ ai guidebook. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.
[77] 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*. 1–18.
[78] 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 CO-DESIGN STUDY

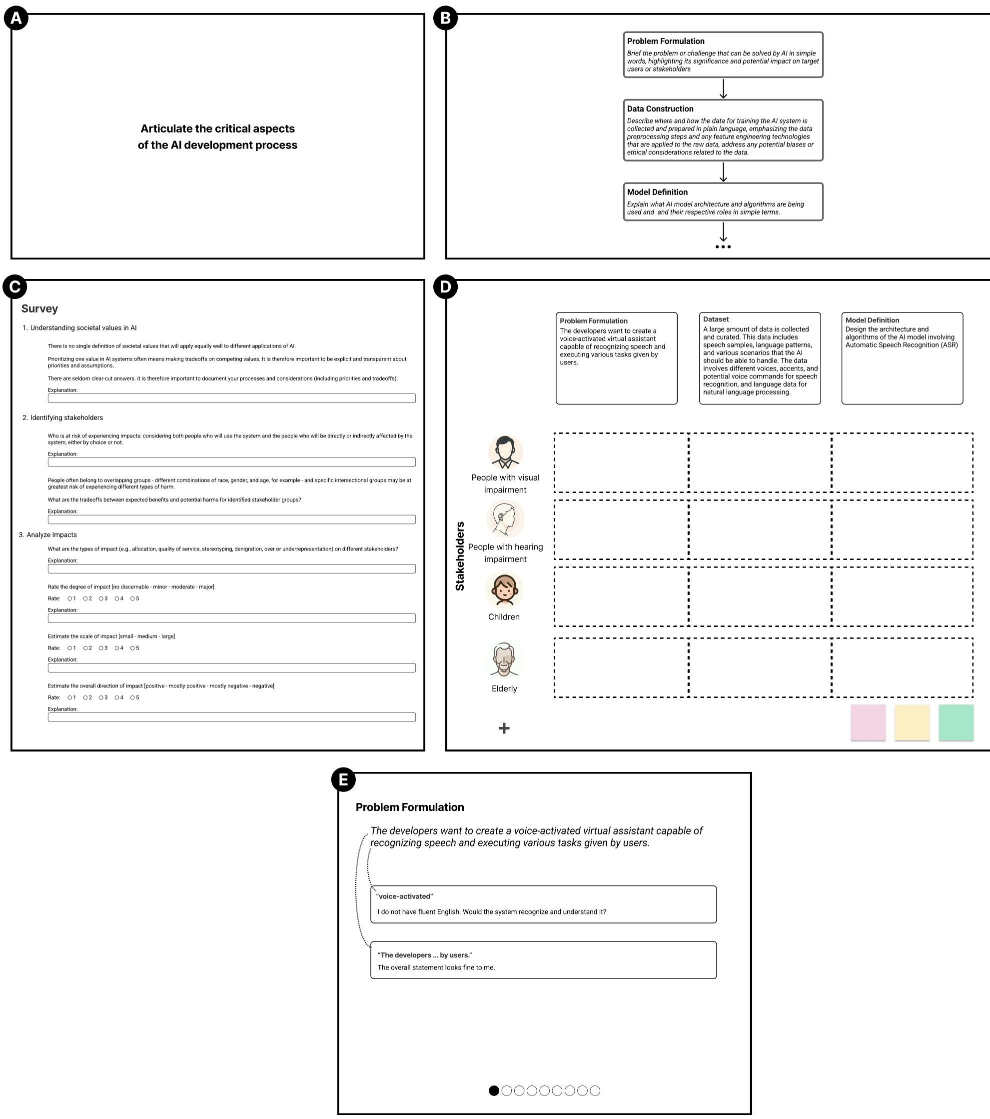

Fig. 4. Concepts & Artifacts explored in the co-design study. Planning: (A) Plain description presents currently unstructured approaches to articulating AI development plans in cross-functional teams; (B) Stage-based scaffolding breaks down and connects AI development stages while providing prompts, as inspired by AI storyboard [21]. Harm identification: (C) The survey was adapted from Value Cards [63] 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.

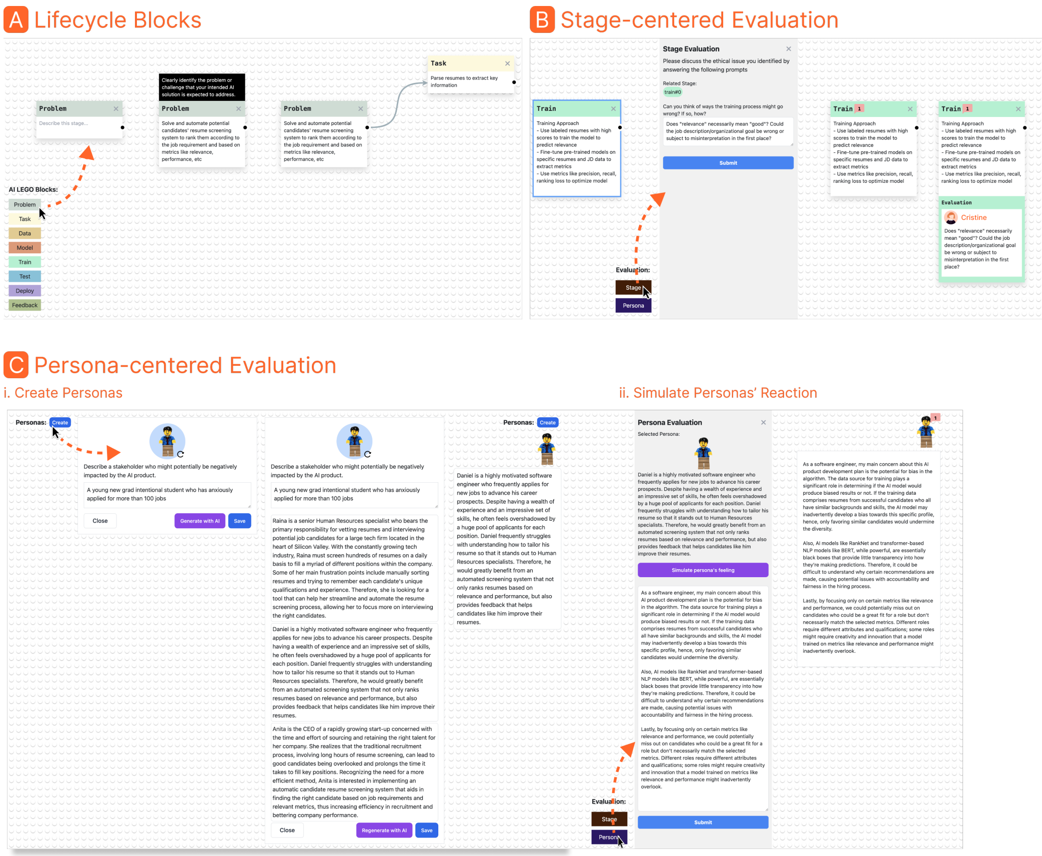

Fig. 5. Key components and transitions in AI LEGO user interface.

B AI LEGO UI COMPONENTS

C 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.

D AI SYSTEM DEVELOPMENT SCENARIOS

D.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.

D.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.

D.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

D.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.

D.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.

E EVALUATION STUDY SCENARIO & CONDITION ORDERS

	Block 1	Block 2	Block 3
T1	Image Generation Baseline	Gunshot Detection AI LEGO Lite	Resume Screening AI LEGO Full
T2	Image Generation AI LEGO Full	Vaccine Allocation Baseline	Credit Assessment AI LEGO Lite
T3	Resume Screening AI LEGO Lite	Vaccine Allocation AI LEGO Full	Gunshot Detection Baseline
T4	Credit Assessment Baseline	Gunshot Detection AI LEGO Full	Image Generation AI LEGO Lite
T5	Vaccine Allocation AI LEGO Lite	Resume Screening Baseline	Credit Assessment AI LEGO Full
T6	Gunshot Detection AI LEGO Full	Credit Assessment AI LEGO Lite	Image Generation Baseline

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

F HARM RATING RUBRICS

Table 2. Grading rubrics for harm ratings, developed using the Socio-technical Harm Taxonomy [62]. 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)