Making the Right Thing: Bridging HCI and Responsible AI in Early-Stage AI Concept Selection

Ji-Youn Jung
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jiyounj@andrew.cmu.edu

Devansh Saxena*
The Information School
University of Wisconsin-Madison
Madison, Wisconsin, USA
devansh.saxena@wisc.edu

Minjung Park
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
mpark2@andrew.cmu.edu

Jini Kim
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
jinik@andrew.cmu.edu

Jodi Forlizzi
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
forlizzi@cs.cmu.edu

Ken Holstein
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
kjholste@andrew.cmu.edu

John Zimmerman
Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
johnz@cs.cmu.edu

Abstract

AI projects often fail due to financial, technical, ethical, or user acceptance challenges—failures frequently rooted in early-stage decisions. While HCI and Responsible AI (RAI) research emphasize this, practical approaches for identifying promising concepts early remain limited. Drawing on Research through Design, this paper investigates how early-stage AI concept sorting in commercial settings can reflect RAI principles. Through three design experiments—including a probe study with industry practitioners—we explored methods for evaluating risks and benefits using multidisciplinary collaboration. Participants demonstrated strong receptivity to addressing RAI concerns early in the process and effectively identified low-risk, high-benefit AI concepts. Our findings highlight the potential of a design-led approach to embed ethical and service design thinking at the front end of AI innovation. By examining how practitioners reason about AI concepts, our study invites HCI and RAI communities to see early-stage innovation as a critical space for engaging ethical and commercial considerations together.

CCS Concepts

• **Human-centered computing** → **Interaction design process and methods**.

Keywords

AI innovation, Responsible AI, Early-stage innovation, Concept selection, Ideation

ACM Reference Format:
Ji-Youn Jung, Devansh Saxena, Minjung Park, Jini Kim, Jodi Forlizzi, Ken Holstein, and John Zimmerman. 2025. Making the Right Thing: Bridging HCI and Responsible AI in Early-Stage AI Concept Selection. In *Designing Interactive Systems Conference (DIS '25), July 05–09, 2025, Funchal, Portugal.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3715336.3735745

1 Introduction

Artificial intelligence (AI) offers immense potential to revolutionize industries and address global challenges. It creates value by automating work and increasing efficiency (e.g., warehouse and manufacturing robots, chatbots, document summarization); forecasting and providing critical insights (e.g., financial forecasting, smart inventory management, predictive maintenance); and accelerating breakthroughs in areas that have inhumanly large problem spaces like drug discovery [35, 61] and the creation of new materials for things like batteries and computer chips [59, 73]. The hype, hope, and promise surrounding AI have sparked widespread interest and huge investments. A recent survey showed that 40% of CEOs plan large AI investments, and they expect these to pay off with improved competitive advantage [64]. Hundreds of billions of dollars are being spent each year with the hope that AI produces rich rewards [51, 77].

*Author conducted this research as a Presidential Postdoctoral Fellow at Carnegie Mellon University.

DIS '25, Funchal, Portugal

ACM ISBN 979-8-4007-1485-6/25/07
https://doi.org/10.1145/3715336.3735745

The success and investment surrounding AI products and services imply an effective innovation process; however, recent human-computer interaction (HCI) research exposed many challenges. Currently, about 85% of AI initiatives fail [32, 47, 50, 92] because many systems: (i) cannot generate benefits that outweigh their development and operational costs, (ii) cannot build a model that achieves the minimal required level of performance to create real value, (iii) don't address real needs, so users won't adopt and use them, and/or (iv) face significant ethical issues such as biased data that cause unacceptable, unintended harm. In addition to the high failure rate, researchers note a lot of missed, low-hanging fruit [98, 100–102]. Organizations seem to ignore situations where simple AI implementations with moderate performance would be immediately useful for users and generate value for service providers. The hype surrounding AI seems to pressure organizations to rush into AI innovation, and to take big risks instead of a more measured and steady approach.

The Responsible AI (RAI) community has made significant progress in identifying sources of unintended harm and creating a variety of practical toolkits and frameworks. However, implementing these insights within organizations often faces resistance, driven by concerns about their perceived incompatibility with existing workflows (e.g., [28, 44, 56, 93]) or fear of potentially hindering innovation [84]. To address these challenges, prior research has assessed the state of adoption and explored ways to bridge gaps. For instance, Rakova et al. highlighted practitioners' aspirations to align RAI efforts with their organization's mission and values [76]. Despite these efforts, a critical knowledge gap remains: how to operationalize these aspirations and effectively integrate commercial objectives with RAI principles.

A growing body of RAI researchers found that trying to fix ethical issues after models have been trained and deployed is often infeasible or even impossible in practice [22, 31, 48]. This has led to a self-critique within the RAI community: while adept at identifying problems, the solutions often come too late in the development process. Researchers suggest that one way to avoid RAI challenges is to address concerns during the earliest stages of the innovation process [24, 49, 68, 75, 79, 81, 90, 102]. At the same time, HCI researchers suggest that improving ideation (the generation of many innovation concepts) might reduce the high AI failure rate by surfacing low-risk, high-value concepts. Recent work offers new ways to generate ideas that more effectively surface the harmonious intersection of what users need and what AI can reasonably produce [57, 102]. Despite the shared interest and a common vision from both RAI and HCI communities in improving early-stage AI innovation, little existing work explicitly bridges these domains. Specifically, there has been little exploration of RAI principles shaping ideation processes or AI innovation research adopting RAI insights.

Our research builds on insights from RAI advocating for the early consideration of ethical harms in innovation, and from HCI research focused on improving the ideation of AI concepts. We specifically focus on concept selection, and how innovation teams could assess concepts generated during brainstorming to surface low-risk, high-benefit opportunities. HCI currently does not offer a formal process for selecting the "best" concept. We inferred that scaffolding the process might help, given the complexity of simultaneously addressing technical, financial, user-acceptance, and ethical risks. We ask the following questions: *Can RAI concerns be integrated into a commercial AI innovation process at the point of project selection? Are AI innovators open to a more formal process for assessing and selecting low-risk, high-benefit opportunities?*

Following a Research through Design approach [106], we conducted three design experiments including a probe study with professional AI innovators. The probe explored a more structured approach to collaboratively assessing concepts and selecting what to innovate. We found that professionals were open to integrating RAI concerns into a commercial AI innovation process, and they could leverage the collective intelligence in their team to reach a consensus. Their reflections on current practices indicated they don't often brainstorm new product or service concepts, raising questions about who is doing this work.

Our paper makes four contributions. First, we present a probe study on early-stage AI innovation practices. Second, we advance our understanding of how to support more effective and responsible AI innovation in practice. Third, we offer insights with the goal of improving early-stage ideation and project selection. Finally, we pose new research questions for the HCI and RAI communities, advancing the discourse on AI innovation.

2 Related Work

Our work draws from and attempts to integrate two separate threads of research. We build on HCI research focused on lowering the risk of failure by improving the ideation of AI concepts. We also draw on RAI research and its intention of reducing AI's unintended harms. We are committed to RAI's goals of maximizing AI's benefits while minimizing its harms [69].

2.1 HCI Research on Improving AI Innovation

HCI research has a long history of discussing the interplay between *sketching* (ideation: the envisioning of many different things to make) and *prototyping* (iteration: the use of rapid prototyping to iteratively refine a system into being). Buxton, in his seminal work on user experience, discusses this as the difference between "making the **right thing**" (sketching) and "making the **thing right**" (prototyping) [18]. Recent HCI research on AI innovation suggests that the high failure rate likely stems from poor quality ideation [96, 98, 104]. Researchers note that data science teams often envision concepts customers don't want, while design teams envision concepts that cannot be built.

This body of research details several potential causes of poor-quality ideation. In many cases, HCI/UX is not invited to join projects until after deciding what to make has happened [29]. This is sort of like traveling back in time to the 1990s when designers were invited to join a project late in the process. They dismissively described the goal of slapping a "pretty" interface onto an application no one wants as "putting lipstick on a pig" [20, 21]. However, including HCI/UX from the start won't solve the failure problem. Many HCI/UX practitioners find it challenging to grasp AI's capabilities — what exactly can AI do [29]? The few "effective" HCI/UX practitioners employ internalized AI capability abstractions, allowing them to recognize situations when a capability might be valuable [97].

They also use a set of examples associated with these capabilities to communicate their ideas to other HCI/UX practitioners or AI collaborators. One challenge unique to AI is its high level of uncertainty. When ideating, AI innovators often struggle with understanding whether they can technically deliver a desired capability and the potential errors a system might make [98]. As a system is deployed and accumulates data, these factors can change, influencing the system's performance. Innovators also struggle to recognize AI's development and operational costs — is a concept expensive or cheap [101]? Collectively, these issues all present challenges to effective ideation and project selection.

Data science plays a key role in AI innovation; however, it lacks a strong connection to human-centered design processes like ideation. Today, most data science textbooks describe the innovation process as starting with problem formulation: defining a problem, identifying constraints and resources, and scoping a project that should create value [94]. Data science students are not typically taught to ideate, to think of a hundred different things they might create, and then assess their collection to find the best concept. HCI research exploring the emerging role of data science in the enterprise detailed several challenges that can impact the ideation and selection of viable, valuable projects. Data scientists shared that they struggle to connect business strategy, user needs, and AI opportunities [53, 67]. They also struggle to communicate the meaning of a model's performance, negatively impacting collaboration with non-data science stakeholders [105]. This challenge to define and communicate what is "good enough" creates tension around assessing how good an innovation concept might be.

In support of human-centered AI innovation, HCI researchers have developed resources including design guidelines and patterns [5, 7, 38]. Currently, these resources only support the prototyping phase [103]. Similarly, tools for machine learning practitioners often focus too narrowly on low-level technical details, neglecting the broader, conceptual exploration necessary during sketching [54]. Research on how HCI/UX teams employ these guidelines surfaced practitioners' requests for additional tools and resources that can help with ideation and selection [103]. Research also shows that HCI/UX practitioners have started creating their own resources by documenting AI capabilities in support of ideation [101, 102]. HCI researchers developed two methods to improve ideation. One approach casts HCI/UX in the role of a facilitator who gets data science teams and problem owners to brainstorm together [99, 102]. This work explicitly focused on helping innovators recognize situations where moderate model performance creates user value. It builds on the observation that making systems with moderate performance is much easier than making systems with excellent model performance. The other approach, AI Matchmaking [57], builds on a technology-centered innovation approach (matchmaking [13]). It centers user needs by including the problem owners in the matchmaking work. Both ideation approaches work to reduce the risks of technical viability and user acceptance. However, they do not address the financial or ethical risks.

The ideation phase typically involves two iterative steps: generating concepts and selecting the most promising ones. Bill Buxton emphasizes the importance of this selection process, aiming to "discard more than we keep. [18]" He highlights that effective selection isn't just about liking an idea, but making critical comparisons—asking "Do I want this rather than that, and why? [18]" This approach underscores the necessity of evaluating concepts in parallel to determine their relative merits [30, 87]. However, recent efforts in early-stage AI innovation have primarily focused on idea generation, leaving a gap in our understanding of how to effectively select the concepts with the most potential.

2.2 RAI Research on Mitigating Problems in Real-World Settings

The RAI community has seen significant growth over the past decade with the increasing use of AI systems across various domains and a subsequent rise in AI harm cases. To address the ethical concerns arising from harmful AI systems, the RAI community has developed tools and processes that refine existing systems, document their limitations [66], or facilitate audits and impact assessments [44, 95]. However, there is a growing recognition that some ethical concerns are inherent to a specific problem formulation, requiring a fundamental system redesign rather than post hoc refinements [15, 44, 75]. Some post hoc technical fixes such as "de-biasing" may inadvertently amplify the biases that they were supposed to address. In sum, much of the RAI work has focused on improving existing systems (i.e., making the thing right) rather than exploring which other AI concepts might have been more feasible in the first place (i.e., making the right thing). Recent studies investigating industry product teams' current practices and challenges around AI fairness, found that teams were most interested in finding ways to avoid ethical challenges in the first place.

The RAI community has also developed databases and taxonomies that capture ongoing and emergent forms of algorithmic harm and may further help AI practitioners anticipate these potential harms [1, 12, 16]. However, AI systems are deployed in heterogeneous social contexts and interplay with social, cultural, and/or organizational dynamics [82]. This makes it difficult to anticipate harm. A design approach can help develop practical tools that are theoretically grounded in these taxonomies, center critical, actionable dimensions, and guide AI developers in systematically identifying potential harms before committing to system development. Moreover, the RAI community has predominantly focused on ethical risks. AI practitioners must ensure that AI concepts also create value for stakeholders. This requires a more holistic evaluation that incorporates concerns like financial and user-acceptance risks.

Given these challenges, recent calls to action urge researchers to focus on the earliest phases of AI innovation, where teams can ideate more broadly (e.g., [15, 44, 68, 75, 91]). The RAI community has shown a growing interest in leveraging HCI and design methodologies to integrate ethical considerations from the start [27, 44, 55]. Although some recent studies have examined these early stages [49], to our knowledge none have compared multiple concepts in parallel, and few have evaluated AI ideas beyond ethical concerns. This study contributes to the emerging body of work on RAI. It showcases how early-stage design probes can assist interdisciplinary teams in integrating RAI considerations with commercial goals. Teams can evaluate multiple AI concepts using a holistic and structured framework that weaves ethical concerns into commercial criteria of desirability, feasibility, and viability.

3 Design Process Overview

We were interested in the handoff between ideation and iteration when AI innovation teams assess many AI concepts and select what to make. We wanted to explore a proactive approach that enhances the coordination of interdisciplinary teams to identify low-risk, high-benefit concepts efficiently. We aim to integrate RAI considerations into the commercial development process to improve receptivity and acceptance of RAI in the private sector. This exploration seeks to understand how AI practitioners perceive and adapt to the envisioned workflow.

For this exploratory research, we chose a Research through Design (RtD) approach [106]. RtD leverages the tools and processes of design practice to holistically explore problematic situations. It is useful for investigating new ways of designing. We conducted three interconnected design experiments. Design experiments are different from scientific experiments. Within the context of RtD research, "a design experiment includes [any] intentional actions design researchers take to further their understanding of the situation their work addresses" [107]. Design Experiment 1 (DE1) explored examples from design and other fields of assessing and making a selection from many options. This informed and inspired the design of our probe materials. Design Experiment 2 (DE2) developed the materials for the probe study. This constructive activity operationalized insights from DE1. Design Experiment 3 (DE3) executed our probe study. It brought professional AI innovators together and had them enact a more formal approach to assessing and selecting an AI concept.

HCI has a long history of using probes where researchers disrupt the current state in very intentional ways to observe participants' reactions. They are meant to offer insights that inspire much more than to answer terse research questions, and they have been taken up and used in wildly different ways since their introduction [14, 17, 33, 39, 63].

Across the three design experiments, we employed a highly dialogical approach, continuously sharing and refining our findings through feedback from internal and external collaborators. Internally, our multidisciplinary team–experts in HCI, design, RAI, computer science, and psychology–engaged in brainstorming and design activities. Externally, we shared progress with two research groups: 50 HCI researchers and 30 RAI researchers. We used their feedback to inform the design of the probe and the questions we explored during the study with professionals.

4 Design Experiment 1: Gathering Inspiration on Making Design Choices

Designers often seek inspiration by broadly searching for relevant examples where others have addressed a similar situation [65]. We built on this pattern by exploring examples of how design and other fields have approached the challenge of choosing one from many. We looked at examples from HCI, design, engineering, and business. We hoped to find ideas we could borrow in the design of our probe.

We began by brainstorming where to look. Next, we conducted a broad search across academic and non-academic sources. We started with the ACM Digital Library and Google Scholar, employing diverse keywords, including but not limited to "early-stage concept selection," "project selection," "concept convergence," "innovation evaluation instrument," "sorting techniques," and "ideation selection techniques." We next expanded our search to include widely adopted practices in the industry, identified through books, online media, and video demonstrations. We collectively reviewed the growing set of artifacts and methods using a process similar to critique where we talked about the usefulness of an example to our problematic situation. We repeatedly asked which examples supported consideration of technical, financial, user acceptance, and ethical risks and benefits for services, customers, users, and society.

Examples we drew inspiration from include:

- Dot Voting: supports the collective intelligence of the group by allowing members to vote with 'dots' on various options [37].
- Impact-Effort Matrix: 2x2 matrix for a group to evaluate and prioritize tasks or projects based on the level of effort required and the impact they will have [36, 40].
- Morphological Charts: systematically generates new ideas and selects optimal solutions by decomposing a problem into its functions and listing various options for each, from which the best can be chosen [83, 89].
- Pugh Analysis: supports numeric rating and comparison of many possibilities by evaluating each option against a baseline using specific criteria, facilitating the identification of the most viable solution [72].
- Harris Profile: visualizes and evaluates strengths and weaknesses of design concepts in relation to predefined design requirements [42, 89].
- Risk Assessment Matrix: A prioritization framework (low-med-high) to decide which risk factors to prioritize, based on the severity and the likelihood of the risk happening [6].
- Cooper's New Product Selection Model: Thirteen factors that indicate the product's potential for success, derived from 195 industrial product projects [23]

Our critique of the examples led to the following key considerations for the probe design:

- The process should strike a balance between having a formal structure and remaining designerly [86]. It should feel more like a practice method and less like a setup for a controlled study.
- Innovators should holistically consider the benefits and the four risks. They all interact with each other.
- The whole process must be fast. Teams should not spend more than 2 hours comparing 10 to 20 concepts to make a selection.
- The process should leverage the collective intelligence of the different disciplines and consideration of different organizational priorities.
- The process should compare multiple concepts in parallel, instead of focusing on one concept at a time [87].

Figure 1: Initial prototype combining dot-voting with the 2x2 matrix approach.

5 Design Experiment 2: Designing a Probe to Explore AI Concept Sorting and Responsible AI Integration

Building on the criteria from DE1, we focused on making a probe that was flexible and could be used for rapid assessment. We designed a probe to provide professionals with scaffolding for assessing and rating a set of AI concepts. The scaffolding was meant to guide them toward high-benefit, low-risk concepts and away from high-risk, low-benefit. With respect to the holistic experience of the probe, we took inspiration from experience prototyping [17], particularly how it delivers critical aspects of a possible future experience and then asks participants to reflect on what they really want.

5.1 Process Overview

The process comprised three main activities: (i) brainstorming probe designs, (ii) assessing impacts, and (iii) refining structure and flow. We conducted iterative pilot testing to enhance our design process, including 13 formal pilot tests simulating the envisioned probe process and numerous informal tests. Feedback from diverse academic researchers and practitioners further enriched our approach.

5.1.1 Brainstorming Probe Designs. The team held five brainstorming sessions, where researchers sketched probe designs based on Design Experiment 1 criteria and refined them through group feedback. Broader team input helped prioritize and shape the ideas.

5.1.2 Assessing Impacts. To select a direction, we evaluated each concept against key criteria: (i) speed and usability in early-stage design, (ii) effective scaffolding of discussions across four areas (financial, technical, user acceptance, and ethical), and (iii) the ability to surface agreements and disagreements within teams. After extensive discussions, we chose an approach that combined dot-voting and a 2x2 matrix. This was intended to allow team members to categorize concepts into high-benefit, low-risk or low-benefit, high-risk groups and to facilitate focused discussions on prioritizing the concepts they believe to be most viable through voting. This framework also provided a simple, intuitive three-level rating system (high-medium-low) to assess each concept's desirability, feasibility, and viability. Figure 1 shows the initial prototype. [1]

5.1.3 Refining the Form, Structure, and Flow. Feedback from pilot tests with both students and professional industry teams continuously led to refinements in structure. In developing our prompt questions, we utilized DE1 inspirations for foundational guidance, selectively incorporating and synthesizing financial insights from business literature [10, 43, 85, 88], technical insights from classical input-model-output representations [2], and both ethical and technical considerations from RAI literature [44, 49, 75, 78, 82, 90]. Researchers then brainstormed prompts together. After multiple discussions between us, we organized them under desirability, feasibility, and viability. Inspired by prior research (e.g., "Do-Reason-Know worksheet" [104]), we designed the Concept Card to outline the AI's input-model-output mechanism. We also included stakeholder analysis, informed by repeated findings and feedback during the pilot tests, to address broader ecosystem concerns. Facilitation processes were streamlined through additional tests, reducing inefficiencies.

5.2 Final Probe Design

The final probe includes four artifacts and a workflow for utilizing them. Below, we describe the artifacts, instructions for using the artifacts, and outline the flow of the workshop session.

5.2.1 Concept Card. The concept card (Figure 2) is a template completed prior to the workshop and includes four key components: (i) concept overview, (ii) stakeholder list, (iii) AI system description, and (iv) data.

The concept overview provides a brief title and a short description summarizing the concept. The stakeholder list identifies four key roles: payer, end-user, servicing party, and impacted individual. The payer is the entity funding the service, while the end-user regularly interacts with the system. The servicing party distributes and manages the service. Lastly, the impacted individual includes those

[1] In the DE3 workshop study, we modified the probe's design to improve the online facilitation process by replacing dot-voting with a drop-down button format. Although the form changed, the content and functionality of the probe stayed consistent throughout the paper.

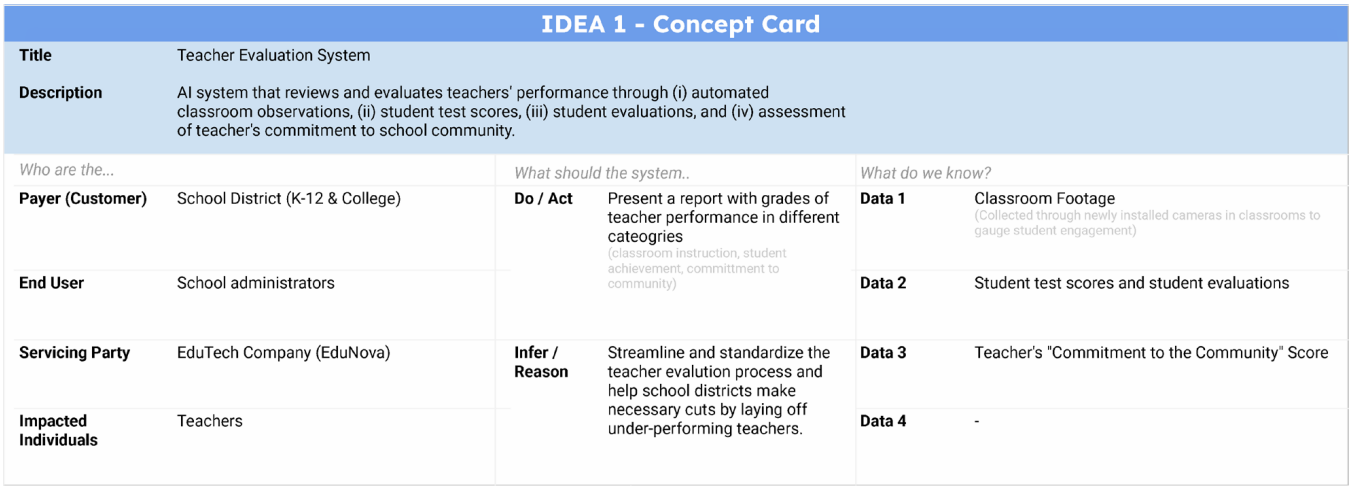


Figure 2: Concept Card with a top area for the title and description. The bottom area has three columns: stakeholders (left), system description (middle), and relevant datasets (right).

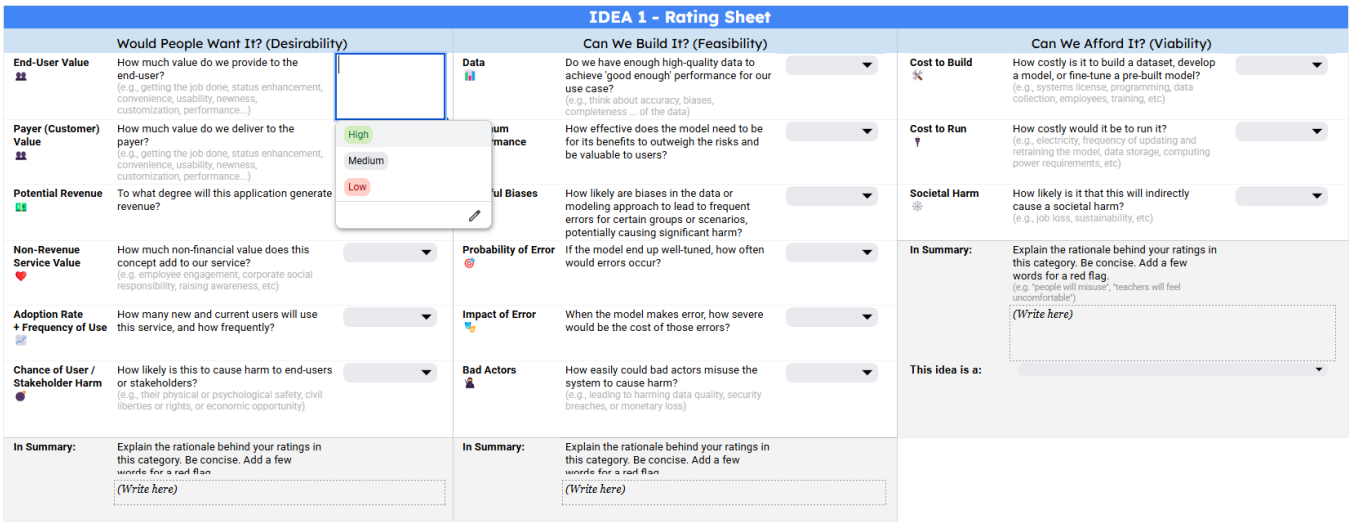


Figure 3: Individual Rating Sheet where participants rate each item as high, medium, or low using a dropdown menu. At the bottom of each column, they provide qualitative reasoning to support their quantitative ratings.

indirectly affected by the service, such as those experiencing unintended consequences. For example, in the case of Google Search, the end-user is the person searching for information, the payer is the advertisers, and impacted individuals might include website owners affected by search ranking algorithms. The system description adopts the "Do-Reason-Know" framework from Yildirim et al. (2024) [104]. It outlines (i) Do—the actions performed by the system (e.g., presenting recommended videos); (ii) Reason—the reasoning or inference process enabling the action (e.g., ranking videos by relevance while balancing diversity). Finally, the data column outlines (iii) Know—the computational data required to enable these actions (e.g., user data, metadata, or contextual information).

The Concept Card format facilitates a uniform presentation of AI concepts, aiding those assessing multiple ideas. Pilot testing revealed that a consistent layout helps users quickly locate essential information, making comparisons easier and more efficient.

5.2.2 Individual Rating Sheet. The individual rating sheet (Figure 3) helps people quickly assess AI innovation concepts based on four factors: financial, technical, user-acceptance, and ethical concerns. Designed to be completed in 5 minutes, it encourages informal, rapid evaluations using a High-Medium-Low scale paired with brief qualitative comments for context.

The sheet focuses on three main questions: (1) Desirability ("Would people want it?"), (2) Feasibility ("Can we build it?"), and (3) Viability ("Can we afford it?"), integrating ethical considerations into the framework. With 15 questions total (Table 1), participants conclude by categorizing the concept as Keeper, Maybe, or Showstopper.

Table 1: A list of prompting questions in an Individual Rating Sheet.

Category	Keyword	Prompt
Would People Want It? (Desirability)	End-User Value	How much value do we provide to the end-user? (e.g., getting the job done, status enhancement, convenience, usability, newness, customization, performance...)
	Payer (Customer) Value	How much value do we deliver to the payer? (e.g., getting the job done, status enhancement, convenience, usability, newness, customization, performance...)
	Potential Revenue	To what degree will this application generate revenue?
	Non-Revenue Service Value	How much non-financial value does this concept add to our service? (e.g. employee engagement, corporate social responsibility, raising awareness, etc)
	Adoption Rate + Frequency of Use	How many new and current users will use this service, and how frequently?
	Chance of User / Stakeholder Harm	How likely is this to cause harm to end-users or stakeholders? (e.g., their physical or psychological safety, civil liberties or rights, or economic opportunity)
Can We Build It? (Feasibility)	Data	Do we have enough high-quality data to achieve 'good enough' performance for our use case? (e.g., think about accuracy, biases, completeness ... of the data)
	Minimum Performance	How effective does the model need to be for its benefits to outweigh the risks and be valuable to users?
	Harmful Biases	How likely are biases in the data or modeling approach to lead to frequent errors for certain groups or scenarios, potentially causing significant harm?
	Probability of Error	If the model ends up well-tuned, how often would errors occur?
	Impact of Error	When the model makes errors, how severe would be the cost of those errors?
	Bad Actors	How easily could bad actors misuse the system to cause harm? (e.g., leading to harming data quality, security breaches, or monetary loss)
Can We Afford It? (Viability)	Cost to Build	How costly is it to build a dataset, develop a model, or fine-tune a pre-built model? (e.g., systems license, programming, data collection, employees, training, etc)
	Cost to Run	How costly would it be to run it? (e.g., electricity, frequency of updating and retraining the model, data storage, etc)
	Societal Harm	How likely is it that this will indirectly cause societal harm? (e.g., job loss, sustainability, etc)

5.2.3 *Team Response Overview.*

The Team Response Overview (Figure 4) visualizes team ratings for a concept using color codes—green (positive), red (negative), and gray (neutral). Each column represents a question, with individual responses displayed side-by-side. Facilitators and team members use this visualization to quickly scan across each other's perspectives as input for subsequent collective discussion. Here, the team can spot (dis)agreements, focusing discussions on critical issues for efficient and targeted decision-making.

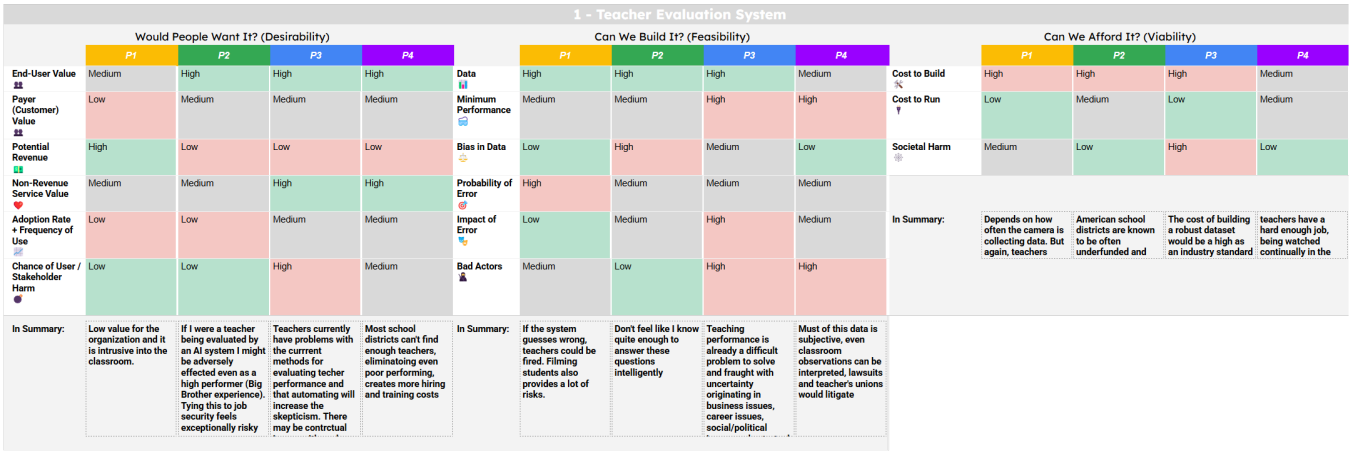
5.2.4 *Risk-Benefit Matrix.*

The Risk-Benefit Matrix (Figure 5) is a 2x2 matrix for evaluating concepts based on benefit (Y-axis) and risk (X-axis). Concepts in the top-left quadrant (high benefit, low risk) are prioritized, while those in the bottom-right (low benefit, high risk) are deprioritized or discarded. Quadrants in between prompt further discussion for refinement or strategic decision-making. Teams collaboratively place concepts on the matrix, with a facilitator guiding the process and referencing the Team Response Overview to address agreements and disagreements.

The Risk-Benefit Matrix was inspired by the 2×2 structure of the Impact-Effort Matrix. In the Impact-Effort Matrix, effort represents the work required, and impact reflects the potential positive outcomes. While widely used in industry [3, 36, 40] and HCI research [102, 103], it lacked the flexibility to address the broader range of risks and benefits needed for our project. For instance, ethical risks cannot be fully captured through the lens of effort. To address these limitations, we developed the Risk-Benefit Matrix as a more comprehensive framework.

5.2.5 *Workflow.*

Our probe's workflow is a one-hour workshop designed for an interdisciplinary team, ideally including representatives covering financial, technical, user acceptance, and ethical concerns. The workshop focuses on assessing, sorting, and selecting AI concepts. Figure 6 illustrates the overall workflow. Participants begin by reading the Concept Card (section 5.2.1) and individually evaluating a subset of concepts using the Individual Rating Sheet (section 5.2.2). These individual ratings are then aggregated into a Team Response Overview (section 5.2.3), which highlights areas of agreement and disagreement among participants.



1 - Teacher Evaluation System

	Would People Want It? (Desirability)			
	P1	P2	P3	P4
End-User Value	Medium	High	High	High
Payer (Customer) Value	Low	Medium	Medium	Medium
Potential Revenue	High	Low	Low	Low
Non-Revenue Service Value	Medium	Medium	High	High
Adoption Rate + Frequency of Use	Low	Low	Medium	Medium
Chance of User / Stakeholder Harm	Low	Low	High	Medium
In Summary:	Low value for the organization and it is intrusive into the classroom.	If I were a teacher being evaluated by an AI system I might be adversely effected even as a high performer (Big Brother experience). Tying this to job security feels exceptionally risky	Teachers currently have problems with the current methods for evaluating techer performance and that automating will increase the skepticism. There may be contrctual	Most school districts can't find enough teachers, eliminatoing even poor performing, creates more hiring and training costs

	Can We Build It? (Feasibility)			
	P1	P2	P3	P4
Data	High	High	High	Medium
Minimum Performance	Medium	Medium	High	High
Bias in Data	Low	High	Medium	Low
Probability of Error	High	Medium	Medium	Medium
Impact of Error	Low	Medium	High	Medium
Bad Actors	Medium	Low	High	High
In Summary:	If the system guesses wrong, teachers could be fired. Filming students also provides a lot of risks.	Don't feel like I know quite enough to answer these questions intelligently	Teaching performance is already a difficult problem to solve and fraught with uncertainty originating in business issues, career issues, social/political	Must of this data is subjective, even classroom observations can be interpreted, lawsuits and teacher's unions would litigate

	Can We Afford It? (Viability)			
	P1	P2	P3	P4
Cost to Build	High	High	High	Medium
Cost to Run	Low	Medium	Low	Medium
Societal Harm	Medium	Low	High	Low
In Summary:	Depends on how often the camera is collecting data. But again, teachers	American school districts are known to be often underfunded and	The cost of building a robust dataset would be a high as an industry standard	teachers have a hard enough job, being watched continually in the

Figure 4: Team Response Overview consolidates individual responses from the rating sheet into a single view. Example shown is for illustrative purposes only and does not reflect actual ratings.

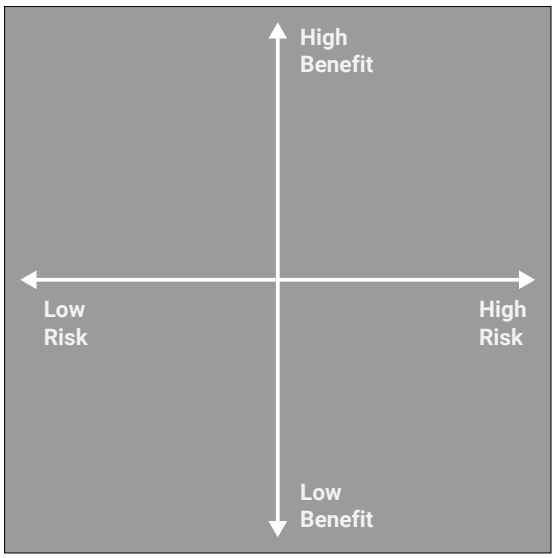


Figure 5: The Risk-Benefit Matrix, a 2x2 matrix with risk on the x-axis and benefit on the y-axis. The top-left quadrant marks the most desirable concepts, the bottom-right the least.

The facilitator plays a central role in managing the workshop, using the Team Response Overview to guide cross-domain discussions. They prioritize aspects with the most disagreements, encouraging dialogue to surface diverse perspectives and foster shared understanding. During team discussions, participants collaboratively map the concepts onto the Risk-Benefit Matrix (section 5.2.4), identifying high-benefit and low-risk concepts. This process is repeated for manageable subsets of concepts to maintain focus and ensure efficiency.

5.3 Reflection

Design Experiment 2 opened a window into the complexities of creating an early-stage AI design probe that could help teams sort and prioritize ideas with both confidence and care. Through this journey, we uncovered insights that shaped not only the probe itself but also the way we think about responsible innovation.

5.3.1 Clarity in Concept Definition. Early ideation thrives on creativity and abstraction, where ideas are fluid, bold, and unbound by constraints. Yet, when it came time to evaluate those ideas, participants struggled with vagueness. They needed something more concrete—details about technical dependencies, system actions, or clear definitions of stakeholders—to make informed decisions. This revealed a pivotal tension: how do you guide teams from the expansive freedom of brainstorming to the sharp clarity needed for meaningful evaluation? The Concept Card emerged as our bridge. We extensively explored what would be the bare minimum information to be able to rate the concept under four concerns.

5.3.2 Integrating Ethical Concerns into Traditional Innovation. As we developed our probe, we faced a key challenge: how to embed ethics into commercial innovation practice. Research revealed that practitioners often struggle to integrate ethics, citing reasons such as lack of guidance or institutional pressures, among others [71, 76, 80]. To address this, we wove ethics into the familiar framework of desirability, feasibility, and viability, making it a natural part of the evaluation process. This shift transformed ethics from a distant afterthought into a practical, integral component of innovation discussions, fostering more responsible and balanced decision-making.

5.3.3 The Power of Facilitation. Early pilot tests showed how critical the facilitator was in managing the interdisciplinary discussion. Without structure, discussions drifted, and participants spent far too long evaluating a single concept. By refining the facilitation process, we improved slow workshop progress (1 concept discussed for 1 hour) into a much more dynamic, collaborative experience.

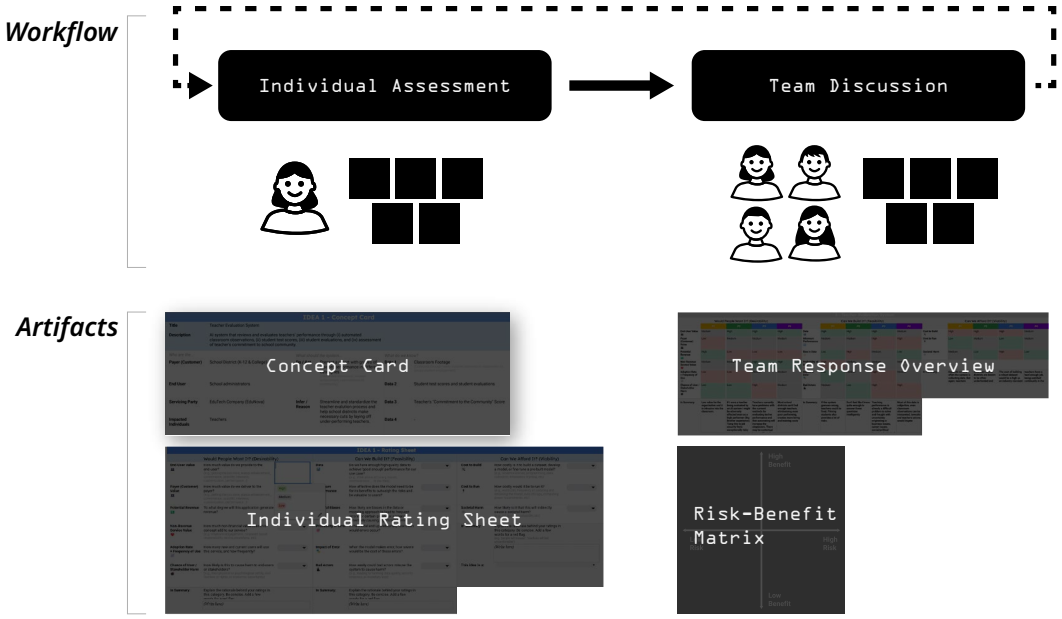


Figure 6: Visualized workflow of our probe. The top row illustrates the structure, beginning with individual assessments followed by team discussions on rated concepts. The bottom row highlights the artifacts used at each step.

Skilled facilitators managed time effectively, keeping discussions focused while allowing room for debate. Our experience mirrors prior work that promotes the role of designers working as facilitators during the ideation of AI concepts [91, 101].

6 Design Experiment 3: Probe Study

For DE3, we executed a probe study using the materials created in DE2. We wanted to explore resistance to and acceptance of a more formal assessment and selection process as well as the integration of RAI into what was meant to feel like a commercial development activity. We designed DE3 to help practitioners critically reflect on their future and present practices.

6.1 Study Design

We recruited 15 AI innovation professionals using email, LinkedIn, and Slack. We divided participants into four interdisciplinary teams, with 3 to 4 on each team. During the study, participants individually rated and collectively sorted 6 early-stage AI concepts while we observed and took notes. Afterward, we interviewed them to gather insights about their experiences and how the process differed from their existing practices. The study design was submitted to and approved by our university's Institutional Review Board (IRB).

Participants. We required all participants to be 18 years of age or older, and to live and work in the US. Recruitment focused on people with prior or current experience in AI innovation. We only accepted people with at least two years of experience innovating with AI. Recruitment materials described the study as an opportunity to try a tool for early-stage AI concept assessment, with a focus on selecting high-benefit, low-risk ideas. The recruitment materials did not frame the study in terms of ethics or Responsible AI. Participants came from diverse professional backgrounds, including design, data science, engineering, business strategy, and product management (see Table 2 for details). Each received a $30 gift card as compensation. All participants provided informed consent and signed consent forms prior to participation.

Table 2: List of participants who attended the probe study in DE 3.

ID	Team	Yrs of Professional Experience	Role
P1	1	6-10 yrs	UX Engineer
P2	1	2-5 yrs	UX Designer
P3	1	21-30 yrs	UX Designer
P4	1	16-20 yrs	Product Manager
P5	2	11-15 yrs	Instructional Designer
P6	2	11-15 yrs	Data Scientist
P7	2	31 yrs and above	Data Scientist
P8	2	2-5 yrs	Software Engineer
P9	3	2-5 yrs	UX Designer
P10	3	6-10 yrs	UX Designer
P11	3	2-5 yrs	UX Designer
P12	3	6-10 yrs	Business Strategist
P13	4	6-10 yrs	Learning Engineer
P14	4	6-10 yrs	UX Designer
P15	4	11-15 yrs	Business Strategist

Table 3: A list of concepts prepared prior to the workshop, with brief descriptions. Full details in the Concept Card format are provided in Appendix A.

#	Title	Description
1	Teacher Evaluation System	AI system that reviews and evaluates teachers' performance through (i) automated classroom observations, (ii) student test scores, (iii) student evaluations, and (iv) assessment of teacher's commitment to the school community.
2	AI-Driven Career Counseling	A career advisory service that uses AI to analyze a student's skills, interests, and job market trends to suggest suitable career paths and necessary courses or skills development. It takes a special account of future workforce forecasts.
3	LLM-Based Programming TA	LLM-powered programming assistant in "Introduction to Programming" courses that respond to students' conceptual questions about coding, without revealing direct code solutions. It is expected to provide ongoing and immediate support to students struggling with coding assignments, just like a TA during office hours.
4	Test Scoring for Essays & Open-Ended Questions for High School Students	An AI system that automates the grading of essays and provides the reason/rubric behind the grades and open-ended questions for high school assignments and exams.
5	AI Proctor in Classrooms	During exams in the physical classroom, assess each student's behavior by tracking their facial expressions, eye gaze, posture, and lip movement in real-time and abnormal sounds in the room to flag the possibility of cheating to the proctor. The video footage will be stored and accessible to teachers for further review to determine if cheating occurred.
6	AI-Powered Storyteller	An AI-driven application where middle-school kids can generate their own stories, read them, and share them with classmates and friends. The tool will read the story aloud and also generate illustrations for the stories. The stories will be stored in the system to allow users to continually update them.

Table 4: Workshop timetable that specifies the activities, time, and artifacts used in each step.

Activity	Artifact	Duration	Description
Introduction	-	20 min	Participants introduce themselves and build rapport with their team. The facilitator outlines the workshop goals and activities.
1st iteration - Individual	Individual Rating Sheet	15 min	Participants individually rate three concepts using the individual rating sheet.
1st iteration - Team	Risk-Benefit Matrix, Team Response Overview	10 min	The team collaboratively places each concept on the Risk-Benefit Matrix, guided by the facilitator.
2nd iteration - Individual	Individual Rating Sheet	15 min	Repeat the 1st iteration process
2nd iteration - Team	Risk-Benefit Matrix, Team Response Overview	10 min	Repeat the 1st iteration process
Post-workshop Interview	-	20 min	The facilitator asks participants reflective questions about their experience.

Concepts. We prepared 6 AI concepts in the domain of education using our Concept Card structure. We selected concepts that showcase a range of benefits and risks, drawing inspiration from existing academic and industry examples, such as the AIAAC incident database for high-risk concepts. A summarized overview of the concept list is presented in Table 3. The full details of the concepts are provided in Appendix A.

Workshop. We ran the 90-minute workshops over Zoom. One researcher facilitated the sessions, while at least one other observed and took notes. The workshop structure was consistent across all teams (see Table 4 for the schedule). As a warm-up, we used a hypothetical scenario and asked participants to roleplay being coworkers at a fictional educational technology company. We provided an overview of the fictional company's profile (see Figure 7) to help establish context.

Figure 7: Fictional company profile that was shown to participants.

Post-Workshop Interview. We reserved the final 20 minutes of each session for a semi-structured, group-based interview conducted with all participants immediately following the design activity. These focus group-style discussions explored participants' overall experience with assessing and selecting a concept, and how this process compared to their current practices. We also invited suggestions for improving this stage of AI innovation.

Data Collection & Analysis. We recorded and transcribed each workshop, and collected all artifacts, including individual rating sheets and the collective Risk-Benefit matrices. For analysis, three researchers conducted affinity diagramming across transcripts, artifacts, and session notes to identify key themes and gain insights into how the design activities shaped participants' reflections and outcomes.

6.2 Findings

All participants seemed eager to collaborate and to share their views. We found the team discussions to be dynamic and engaging. Disagreements about the risks and benefits were common; however, all teams consistently reached a consensus. As shown in Figure 8, all teams placed Concepts #1 and #5 in the high-risk, low-benefit quadrant. (Note that although we developed our overall set of concepts to span a spectrum of risks and benefits, we did not have a specific ordering in mind.) They generally placed Concepts #2 and #3 in the low-risk, high-benefit quadrant. This provides some preliminary evidence that a more formal approach to assessing AI concepts can lead to consistent results across different teams.

6.2.1 Acceptance of RAI in Concept Selection. Participants uniformly recognized RAI issues. The process allowed them to tie their RAI concerns to commercial goals and better justify why they considered an AI concept to be high-risk or low-benefit. Far from being an afterthought, ethical concerns sometimes became the center of the discussion. For example, Concept #1 (Teacher Evaluation) was rated high-risk by all four teams. They questioned its desirability, citing likely teacher opposition. They questioned its feasibility, noting potential harmful biases and a likely failure to meet performance standards. They also questioned its viability, arguing that the operational costs would surpass the benefits. While participants expected that the concept might offer medium to high-benefits to administrators (i.e. payer), all agreed it posed too great a risk to teachers, who were neither the end-users nor the payers. This led to an explicit and deliberate decision by participants to prioritize the ethical concerns for teachers, concluding that the ethical risks to impacted stakeholders far outweighed the potential financial benefits.

Ethical risks prompted rich, multidimensional discussions that considered intersections between technical feasibility, user acceptance, ethical considerations, and financial viability. Discussing one dimension (e.g., technical feasibility) almost always prompted realizations in other dimensions (e.g., user acceptance or ethical considerations). For example, when evaluating Concept #5 (AI Proctor), participants noted that collecting bio-privacy data to detect cheating behavior was unethical, technically infeasible, and likely expensive. While some participants recognized the potential high benefit of the system due to its scalability and human-in-the-loop design—which flags student behavior rather than imposing direct penalties—many raised concerns about its operation in a morally ambiguous area, such as defining what constitutes cheating. They highlighted the technical challenges in reliably detecting cheating, the difficulties in accurately labeling data affected by factors like 'nervousness,' which can vary significantly due to a student's cultural background. Moreover, they stressed the high costs associated with errors in mislabeling students, which could severely damage the relationships between students and teachers, as well as between students and the school.

The Risk-Benefit Matrix facilitated nuanced discussions, emphasizing that the realization of potential benefits was contingent upon addressing identified risks. For example, in the case of Concept #2 (Career Counseling), although it was positively assessed by all four teams, participants raised concerns about potential biases in job recommendations based on gender, race, and socioeconomic status

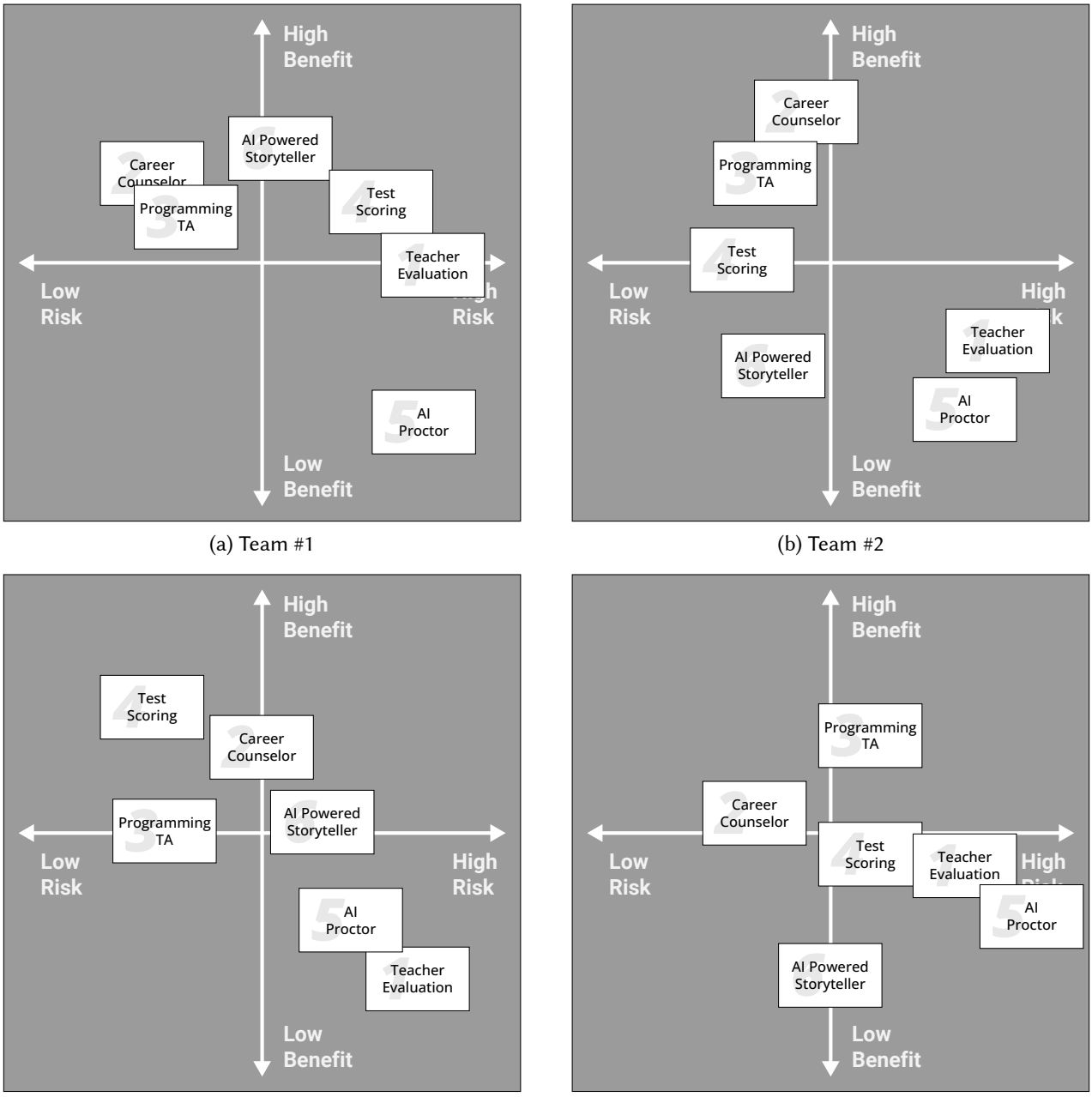


Figure 8: Risk-Benefit analysis of AI concepts across Teams 1–4. While all teams rated Concepts 1 and 5 as high-risk and low-benefit (bottom-right), Concepts 2 and 3 were often placed in the low-risk, high-benefit quadrant (top-left), with some teams placing other concepts there as well.

present in historical data. They underscored that the concept's high potential benefits could only be achieved if these significant risks were effectively mitigated. This nuanced dialogue highlighted the critical need to resolve these issues before considering the implementation of the concept.

This approach contrasts with the Impact-Effort matrix, which generally assumes the system is buildable and focuses primarily on the effort required for development. We noticed that discussions using the Risk-Benefit Matrix often involved questioning whether

the application was even possible to build and examining the problem statement in a more fundamental manner. For instance, in discussing Concept #4 (Test Scoring), participants examined how the system could potentially improve fairness and grading quality by flagging overlooked rubric components. However, they emphasized that such benefits were unrealizable if the potential technical risk of achieving consistency remained unaddressed.

Participants strongly emphasized the critical importance of considering policy and regulatory restrictions, such as HIPAA [2] and GDPR [3], early in the development of AI products. They highlighted these concerns as "very real" due to the additional costs and feasibility challenges they introduce. This entails making preliminary plans for data handling and addressing security issues, which they identified as key factors. Furthermore, they noted that in their practice, they not only assess the difficulty of building the system but also strive to quantify the extent of these challenges. They viewed policy risk as equally important to the four risks our probe explored, and they seemed almost puzzled that it was not included. Our probe seems to have effectively prompted participants to bring up these rich ideas and suggestions on early-stage considerations.

6.2.2 Selection Behavior. We observed that discussions about risks quickly led to consensus among the team, allowing for swift agreement about where a concept fit on the low to high-risk dimension of the Risk-Benefit matrix. However, reaching agreement on benefits was more challenging. It often sparked extensive discussions and careful evaluations of trade-offs among stakeholders. Participants frequently questioned, "but benefits for whom?" For example, during the discussion of Concept #3 (Programming TA) there was significant debate about whether the university's potential financial gain justified possibly reducing learning opportunities for students to be trained as tutors, alongside affecting their income opportunities. Concept #2 (Career Counseling) sparked debates on whether the benefits to end-users outweighed those to the universities themselves. With Concept #4 (Test Scoring), participants struggled to assess benefits, recognizing that while teachers would significantly benefit, they were not the ones financing the service. Participants noted the need to exercise caution, recognizing that the benefits valued by school districts might not necessarily align with those of the teachers, as evidenced by the current system's failure to compensate teachers for unpaid grading tasks performed outside of school hours.

Participants often based initial judgments on the status quo when assessing the benefits and risks. However, the team discussion when using the matrix helped them to more deeply interrogate concepts. For example, in discussions about Concept #2 (Career Counseling), some participants expressed skepticism about AI's effectiveness in offering career suggestions to students, citing their limited capabilities. Other participants expressed that current school career counselors often provide limited insights to students, despite the evident needs and justified school spending in this area. Some participants felt that Concept #2 could provide 'good enough' suggestions and had the potential to reach a broader range of students more effectively, potentially motivating students to reflect more on their goals. Similarly, the perceived benefits of Concept #3 (Programming TA) were diminished by the existence of ChatGPT, as participants questioned the additional value it could provide to students who already use ChatGPT to seek help and complete their assignments.

We observed that participants frequently referenced information from the Concept Cards as they articulated and voiced their concerns. For example, in discussions about Concept #1 (Teacher Evaluation), they highlighted the subjectivity of listed data components (i.e., classroom observation, teacher commitment) and the impossibility of establishing a definitive measure of teacher performance, questioning the validity of the outcome itself, which is a key ethical consideration [46]. Similarly, for Concept #5 (AI Proctor), participants highlighted the 'video footage' data component, expressing concerns about the financial burden of data collection and training. Additionally, they raised issues regarding the complexity and privacy concerns of managing biodata, questioning the value of such efforts given the minimal benefits. In discussing Concept #2 (Career Counseling), participants mentioned concerns about the costs associated with consistently updating the model with rapidly-changing job market data. They also noted the large volume of data and the significant effort required to train and fine-tune the model to ensure accuracy and relevance.

Participants noted that the collective ranking and discussion activities were critical for evaluating concepts. Collective discussion allowed team members to highlight strengths and weaknesses together. Observing agreements and disagreements helped participants better understand different perspectives, often arising from their disciplinary perspectives. They all seemed open to changing their minds about issues that are outside of their area of expertise. For example, many shared how initial uncertainties were resolved after hearing their teammates articulate their reasoning, prompting them to adjust their initial assessments.

Participants felt that assessing concepts using individual Rating Sheets before proceeding to collective discussions gave them clarity and confidence in forming their opinions. However, they pointed out that individual ratings would have limited use without contrasting them with others' viewpoints. When assessing the concept alone, participants recognized they were making numerous judgments and often lacked confidence in these. Through team dialogue, these assumptions were clarified, leading to a consensus and more confident judgment about the concept. Additionally, the prompts from Individual Rating Sheet encouraged critical thinking and stepping outside comfort zones, which enhanced their understanding of the concepts. This preparatory phase ensured that collective discussions were informed and productive.

Unexpectedly, participants demonstrated a desire to refine and reframe concepts, rather than simply selecting among the existing concepts as-is. They often redescribed concepts in ways that increased benefits and lowered risks. For instance, during discussions about Concept #4 (Test Scoring), one participant suggested changing the customer and reframing the concept as an AP prep tool to increase its appeal to payers. Another proposed positioning it as a teacher companion tool, offering example essays for comparison rather than direct scoring. In almost all cases, they made changes to the business aspects, not the technical aspects of the concept.

6.2.3 Reflections on Current Practices. Majority of the participants shared that they did not participate in deciding what new products

[2]The Health Insurance Portability and Accountability Act (HIPAA)
[3]The General Data Protection Regulation (GDPR)

or services their organization should make. In reflecting on their practice, most noted that they did participate in decision-making for incremental improvements to a current application. Participants from larger organizations also shared that introducing a new product or service impacts existing product ecosystems, a risk our probe did not address. Almost all felt our probe pushed them into doing work and making decisions outside of their normal responsibilities. Several participants noted that they participated in design sprints as part of an Agile development process. They shared that sprints could lead to a pivot, a change to a new application or a new target market. This serial process is similar to our probe, but it lacks the ideation of many possible concepts proceeding assessment and selection.

The few participants who said they were actively involved in concept selection (e.g., product managers) talked about following a structured process by using tools such as RICE (Reach, Impact, Confidence, and Effort) scoring and Impact-Effort matrices. They described decision-making on AI innovation within their organizations as hierarchical. They noted that "final" decisions were often overridden by board members or investors, who prioritized ambitious, high-risk concepts driven by market trends. They explained that senior executives and the tech culture often dismissed concerns about feasibility risks, believing that all technical challenges could be resolved with sheer effort. These statements align with prior work suggesting managers and executives overestimate benefits and underestimate the technical risks associated with AI innovation. This sentiment was shared across all four workshop teams. Some participants described it as engaging in "watercooler shit talk." They shared that they frequently complained and questioned their colleagues about the real value of the AI concept that they were working on. Others expressed a feeling that the corporation pushes for AI merely for the sake of it.

They shared that decision-making varied significantly based on organization size and their title. Participants from larger organizations described a top-down approach, where key decisions were made at senior levels. Smaller organizations employed decision-making that heavily prioritized technical feasibility and speed. Participants from startups, for example, shared that their decisions were driven by the need to quickly develop a Minimum Viable Product (MVP), often within just a few days. Concepts that could not meet short timelines were deprioritized or ignored.

Participants noted differences between their current practices and the workflow our probe suggested. They described their approach as more 'iterative,' involving building MVPs and pivoting based on feedback. They noted that in their current practice, they did not typically examine all risks and benefits simultaneously or comprehensively. They describe the issues of technical feasibility, financial viability, user acceptance, and ethical risks as showing up individually, at different stages of the development process. The fast pace and earlier commitment to model development is consistent with previous studies suggesting there might be a lack of ideation in AI innovation [102, 103]. The lack of rigorous ideation and focused attention on MVPs and development partially explains why there are many AI innovation tools and methods for prototyping and almost nothing for ideation.

6.2.4 Concept Card. When asked what aspects of the probe stood out as particularly useful, participants unexpectedly highlighted the Concept Card. Initially, we viewed it as a secondary part of our probe, assuming it was tangential to the overall process. However, participants emphasized that its structured template was instrumental in providing the necessary information for sound reasoning about the concepts. Some even expressed interest in adopting the concept card for their own organizations.

Participants noted that the concept card filled a significant gap in their current decision-making processes. They particularly liked how it clarified feasibility by listing relevant data and outlining what the system could realistically achieve. This level of detail not only grounded their evaluations, but also ensured that discussions focused on concepts presented with comparable degrees of specificity.

The differentiation between stakeholders—end-users, payers, and impacted stakeholders—was also highlighted. Participants found this helpful for identifying the root causes of a concept's challenges. For instance, in discussions about Concept #4 (Test Scoring), participants noted it was beneficial for end-users (teachers) but less so for payers (schools) and impacted stakeholders (students). The clarity around stakeholders enriched the conversations, enabling teams to assess benefits and risks with greater precision and insight.

7 Discussion

Our study explored a more structured approach to considering risks and benefits at the point of project selection. Through three design experiments, we investigated how RAI considerations could be integrated into commercial innovation practices at an early stage and how the process might scaffold teams to avoid pursuing problematic concepts that could lead to failure. These experiments highlighted key gaps in current workflows and provided actionable insights for improving future methods.

Below, we delve into the challenges and opportunities associated with executing early-stage ideation and selection processes in AI innovation. We explore the implications of integrating RAI into commercial frameworks and examine how service design can aid this integration. Additionally, we reflect on the implications of our HCI research methods, identify areas for improvement, and propose open questions for further research.

7.1 Challenges and Opportunities of Ideation in AI Innovation

Our findings suggest that the AI innovation process might be negatively dominated by top-down management decision-making. Our participants did not regularly participate in brainstorming new applications or in the selection of what to make. They only participated in the ideation of small, incremental improvements. Their reactions raise questions. Who is doing ideation? Is anyone doing the ideation? Is AI innovation driven by one-off ideas from managers? Design and HCI research shows the importance of ideation to develop more effective and successful products and services [30]. Furthermore, previous studies have shown that people tend to give higher ratings to a design when presented with a single option and are more hesitant to criticize it, compared to when they are

shown multiple concepts simultaneously [87]. Yet we saw little evidence that ideation was happening. Prior HCI suggested a lack of ideation for AI innovation might be due to the fact that data scientists were not trained to ideate [104]; however, our probe hints that organizational leaders don't support innovation teams in ideating.

Our observation that managers might be envisioning and selecting concepts raises additional concerns about a problematic innovation process. Previous research shows low AI literacy among executives with only 3% of executives at S&P 500 companies viewed as AI literate [70]. Prior RAI work raised this concern, noting that ethical assessments in AI are primarily considered by senior management [8]. It seems like the people who know the least about AI's technical challenges and about how it might create unintended harm are playing a very large role in determining the best way to get value from AI's opportunity. If this really is the current state of the industry, it begins to explain the tension participants shared about effectively communicating technical risks to executives with decision-making power.

Despite being largely excluded from decision-making processes around what to build, our interdisciplinary teams showed strong capabilities for doing this work and doing it in a way that tapped into their collective intelligence. Participants engaged in thoughtful discussions about risks and benefits, achieving consistency across teams in promoting and demoting concepts based on benefits and risks. Their frustrations when talking about their leaders indicate their desire to participate in ideation and selection. With both the desire and the skills, the question now becomes one of changing organizational culture to unleash the benefit of their participation. Future research should study the implications and potential benefits of actively including product-facing roles in the decision-making process. This includes examining current workflows and practices to identify who is truly driving decision-making and where ideation originates. By exploring these dynamics, we can move towards a deeper understanding of AI innovation.

Our research reveals a potential gap in how HCI and UX design are taught and the needs and processes in the industry. Practitioners working at Youtube have spoken about the need to differentiate three types of user-centered innovation, calling these "Versioning, Visioning, and Venturing" [19, 26, 62]. Versioning involves small, incremental improvements to a current product. Visioning focuses on creating a new product or major new feature and has a time scale running to three years. Our probe of assessing and selecting concepts fit visioning, not the versioning activities our participants were familiar with. We don't see this distinction of three types of innovation as particularly new. Versioning, venturing, and visioning seem like an almost perfect mapping to McKinsey's "three horizons framework" [11, 74]. Interestingly, in HCI research and education, we seem to talk about innovation and create methods for doing HCI practice as if these three different types of making did not exist. We sort of teach a single approach to user-centered design, ignoring that the scale of the innovation under consideration matters. It might be time to advance our own research and curriculums to better fit what seem to be very real distinctions being used in practice.

7.2 Supporting Early-Stage RAI Practices in Industry

In response to the high failure rates of AI projects, several technology and consulting firms have started using centralized AI governance teams as one way of addressing risks [4, 45, 52, 58]. These teams (e.g., PwC AI Factory [58]) are generally staffed with AI practitioners and domain experts who decide which AI concepts the company should pursue. They are tasked with identifying and assessing AI use cases, staffing AI projects, adopting RAI practices, ensuring compliance with regulation, and supporting the development and deployment processes; thereby creating a one-stop-shop for the governance of AI products and services.

These innovation teams suggest adopting RAI practices from the start, however, research in RAI shows that there is insufficient support in operationalizing RAI principles in practice [44, 49, 60, 76]. RAI tools have supported innovation practice in various ways (e.g., data and model documentation [25, 34, 66], detecting and mitigation bias [9, 41], etc.) but there is a misalignment. RAI work has often framed ethical risks as standalone considerations that must be addressed on top of the practitioners' usual tasks. Consequently, ethical concerns are often sidestepped or deprioritized due to tight project timelines and organizational pressures [76].

In practice, innovation teams must assess multiple AI concepts together and scaffold this assessment within commercial goals. Prior RAI research suggests that ethical concerns often surface informally and are vulnerable to project constraints [76]. Yet, our findings reveal three key insights for RAI – 1) ethical risks are intertwined with commercial goals and impact the desirability, feasibility, and viability of AI concepts, and 2) embedding ethical considerations into commercial criteria can prompt more nuanced discussions and holistic evaluations. Far from being sidelined, ethical risks often took precedence during discussions, with some concepts being rejected primarily due to their ethical implications, 3) assessing multiple concepts together allowed the practitioners to identify high-risk AI concepts as well as lower-risk AI concepts that could be refined further.

This suggests that structured tools can elevate ethical considerations to a central role in decision-making without displacing other business priorities. This integrated approach can help innovation teams discover critical flaws before resources are invested, systematically avoid early missteps, and prioritize higher-value and lower-risk concepts.

7.3 Centering Responsible AI Work in Service Design

Our approach embraces more of a service design as opposed to a UX design perspective. We look at both the financial aspects and the larger ecology of stakeholders involved and impacted by the AI system [82]. We grounded our tools in service design to emphasize a holistic view of the entire service ecosystem, including stakeholders, organizational resources, and processes.

Prior work in RAI shows how ethical concerns are often overshadowed by commercial goals [44, 76]. However, by focusing on the broader ecosystem that will be impacted by an AI concept, we wove ethical questions directly into business-focused discussions, allowing participants to see ethical concerns as a complimentary

priority and not a competing one. For example, Concept #1 (Teacher Evaluation), even though financially appealing, was deemed too risky due to potential harm to educators. This more holistic framing shows how balancing commercial and ethical factors from the beginning can foster more nuanced conversations and inform responsible decisions.

The concept card explicitly maps out payers, end-users, and impacted stakeholders, a core service design principle that asks designers to differentiate between different user groups and consider each group's unique needs, risks, and benefits. This breakdown appeared to have a huge impact on participants' ability to understand a concept, compare it with other concepts, and reason about its risks and benefits. Discussing varying benefits and risks based on different stakeholders almost felt natural for them. We observed that deliberately listing it out for the assessment process helped facilitate and make this process of negotiating trade-offs more transparent. The key sticking point revolved around who should be receiving the most benefits and how this might make a concept better or worse.

Iterative learning and a co-creative environment are central to a service design process. Our participants shared their expertise when comparing AI concepts side-by-side and identified shared challenges and risks. This ability to compare multiple concepts simultaneously fostered richer discussions and subsequently led to the participants' seeking to refine AI concepts to lower their risk. Participants proposed actionable refinements to weaker concepts, such as repositioning the Test Scoring tool to better serve teachers rather than automating grading. Comparative evaluation, supported by tools like the Risk-Benefit Matrix, can encourage more nuanced decision-making and creative problem-solving.

By focusing on the broader AI ecosystem and not just technical factors, our probe materials collectively functioned as a boundary object, facilitating effective cross-disciplinary collaboration. Participants from diverse roles—data scientists, designers, and domain experts—were able to contribute effectively to discussions by identifying ethical risks that others might have missed. These tools ensured that all voices were heard and helped bridge gaps in knowledge and expertise.

HCI does not hold a strong concern for the financial aspects of innovation. Most methods attend to user needs, not to the financial needs of service providers. HCI research rarely explores the financial aspects of design concepts, which are critical to a product or service's commercial success. Our paper contributes to HCI research by integrating this perspective, offering a more holistic view of the potential impact a design concept can have on society. Our integrations of financial, technical, user acceptance, and ethical considerations offer some evidence that HCI, RAI, and business considerations and priorities can effectively be brought together.

7.4 Lessons from Piloting with Students vs. Professionals

We conducted pilot tests, largely with students, and then ran probe sessions with professionals. Their strikingly different behaviors, concerns, and reactions raise issues with respect to how our research community will often use students as a proxy for professionals. While students provided valuable perspectives, they often struggled with unfamiliar workflows and hesitated to confidently share clear opinions. Their assessments were often very slow, and we had concerns that professionals would not be able to quickly assess the concepts. In contrast, the professional teams adopted the probe with ease, leveraging their industry experience to evaluate concepts quickly and effectively while considering market and organizational factors. This contrast highlighted the critical importance of seeking feedback from real-world audiences—such as AI innovation teams—for generating meaningful and actionable insights in HCI studies.

7.5 Limitations

One limitation of our study is that participants evaluated concepts they had not envisioned. We realize that sorting behavior may differ when participants evaluate concepts they have ideated themselves. We made this choice due to time constraints and the challenge of recruiting enough professional AI innovators who all work in the same domain. Future research should explore how self-generated concepts may affect sorting outcomes and team dynamics. Furthermore, our study involved ad-hoc teams with no prior rapport, and who had little if any expertise in educational technology. While participants quickly adapted to the process and showed little struggle in evaluating the AI concepts, future studies should examine how the probe may perform within established teams in real organizational settings, where shared histories and domain expertise may influence results. Additionally, as noted in the Discussion, our probe focused on visioning-scale concepts; had we used more versioning-oriented concepts with smaller variances, participants' engagement with the tool might have differed. Finally, the probe was designed for use within a single organization, but its applicability to multi-stakeholder collaborations or cross-organizational projects remains untested. Future work should investigate how the tool can be scaled or adapted for broader, more complex contexts.

8 Conclusion

This paper addresses a critical challenge in early-stage AI innovation: how to effectively ideate, sort, and select AI concepts while integrating RAI considerations into the process. By designing and deploying a probe, we demonstrated how interdisciplinary teams can navigate complex trade-offs, integrating RAI considerations seamlessly with technical, financial, and user-acceptance factors. Our findings suggest that we might have an overly top-down innovation process, coupled with a lack of thorough ideation that restricts full exploration of potential solution space. This work bridges the gap between RAI principles and real-world practice, offers the HCI and design community an empirical insight for embedding RAI concerns into AI systems, and suggests a mechanism for participation from a broader number of stakeholders on a team. By advancing our understanding of early-stage AI innovation practices, we position design as a driving force in developing AI systems that are more beneficial and less risky in terms of financial, technical, user acceptance, and ethical considerations.

Acknowledgments

This research is supported by the National Science Foundation under Grant No. (2007501), the Presidential Postdoctoral Fellowship

Program (PPFP) at Carnegie Mellon University, the Digital Transformation and Innovation Center at Carnegie Mellon University sponsored by PwC, the UL Research Institutes through the Center for Advancing Safety of Machine Intelligence, and the Block Center for Technology and Society at Carnegie Mellon University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors or research partners. We would also like to thank Dan Saffer, Raelin Musuraca, Laura Vinchesi, and Skip Shelly for their valuable feedback, as well as our anonymous reviewers for their insightful comments and suggestions.

References

[1] [n. d.]. AIAAIC - AIAAIC Repository. https://www.aiaaic.org/aiaaic-repository

[2] 2022. Introduction | Machine Learning. https://developers.google.com/machine-learning/guides/text-classification

[3] admin. 2022. How to Use the Impact Effort Matrix to Prioritize Projects. https://www.sixsigmadaily.com/how-to-use-the-impact-effort-matrix/

[4] Deloitte AI. 2024. Deloitte AI Institute: Connecting enterprises through perspectives and analysis to the entire AI ecosystem. *https://www2.deloitte.com/us/en/pages/deloitte-analytics/articles/advancing-human-ai-collaboration.html* (2024).

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[6] Louis Anthony (Tony)Cox Jr. 2008. What's Wrong with Risk Matrices? *Risk Analysis* 28, 2 (2008), 497–512. doi:10.1111/j.1539-6924.2008.01030.x _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2008.01030.x.

[7] Apple. 2023. Human Interface Guidelines: Machine learning. https://developer.apple.com/design/human-interface-guidelines/machine-learning

[8] Jacqui Ayling and Adriane Chapman. 2022. Putting AI ethics to work: are the tools fit for purpose? *AI and Ethics* 2, 3 (Aug. 2022), 405–429. doi:10.1007/s43681-021-00084-x

[9] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. doi:10.48550/arXiv.1810.01943 arXiv:1810.01943 [cs].

[10] William Robert Bitman and Nawaz Sharif. 2008. A conceptual framework for ranking R&D projects. *IEEE Transactions on Engineering Management* 55, 2 (2008), 267–278.

[11] Steve Blank. 2019. McKinsey's Three Horizons Model Defined Innovation for Years. Here's Why It No Longer Applies. https://hbr.org/2019/02/mckinseys-three-horizons-model-defined-innovation-for-years-heres-why-it-no-longer-applies

[12] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. doi:10.48550/arXiv.2005.14050 arXiv:2005.14050 [cs].

[13] Sara Bly and Elizabeth F. Churchill. 1999. Design through matchmaking: technology in search of users. *interactions* 6, 2 (March 1999), 23–31. doi:10.1145/296165.296174

[14] Kirsten Boehner, William Gaver, and Andy Boucher. 2012. Probes. In *Inventive Methods*. Routledge. Num Pages: 17.

[15] Margarita Boyarskaya, Alexandra Olteanu, and Kate Crawford. 2020. Overcoming Failures of Imagination in AI Infused System Development and Deployment. doi:10.48550/arXiv.2011.13416 arXiv:2011.13416 [cs].

[16] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, S. J. Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, and Dario Amodei. 2024. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation. doi:10.48550/arXiv.1802.07228 arXiv:1802.07228 [cs].

[17] Marion Buchenau and Jane Fulton Suri. 2000. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques (DIS '00)*. Association for Computing Machinery, New York, NY, USA, 424–433. doi:10.1145/347642.347802

[18] William Buxton. 2007. *Sketching user experiences: getting the design right and the right design*. Elsevier/Morgan Kaufmann, Amsterdam Boston.

[19] Conversions. 2019. Product Visioning at Google. https://www.youtube.com/watch?v=fMCc89kO5BI

[20] Alan Cooper. 1999. The Inmates are Running the Asylum. In *Software-Ergonomie '99: Design von Informationswelten*, Udo Arend, Edmund Eberleh, and Knut Pitschke (Eds.). Vieweg+Teubner Verlag, Wiesbaden, 17–17. doi:10.1007/978-3-322-99786-9_1

[21] Alan Cooper, Robert Reimann, David Cronin, and Christopher Noessel. 2014. *About Face: The Essentials of Interaction Design*. John Wiley & Sons. Google-Books-ID: w9Q5BAAAQBAJ.

[22] A. Feder Cooper, Ellen Abrams, and NA NA. 2021. Emergent Unfairness in Algorithmic Fairness-Accuracy Trade-Off Research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. Association for Computing Machinery, New York, NY, USA, 46–54. doi:10.1145/3461702.3462519

[23] Robert G. Cooper. 1981. An empirically derived new product project selection model. *IEEE Transactions on Engineering Management* EM-28, 3 (Aug. 1981), 54–61. doi:10.1109/TEM.1981.6448587 Conference Name: IEEE Transactions on Engineering Management.

[24] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. 2023. A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. 690–704. doi:10.1109/SaTML54575.2023.00050

[25] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. 2022. Interactive Model Cards: A Human-Centered Approach to Model Documentation. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 427–439. doi:10.1145/3531146.3533108

[26] Kevin Dame. 2019. The Power of Visioning: A six-step guide to supercharging product innovation and thinking big at YouTube. https://design.google/library/youtube-visioning

[27] 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3544548.3581026

[28] 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 (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 705–716. doi:10.1145/3593013.3594037

[29] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 278–288. doi:10.1145/3025453.3025739

[30] Steven P. Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L. Schwartz, and Scott R. Klemmer. 2011. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Trans. Comput.-Hum. Interact.* 17, 4 (Dec. 2011), 18:1–18:24. doi:10.1145/1879831.1879836

[31] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. 2020. Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2803–2813. https://proceedings.mlr.press/v119/dutta20a.html ISSN: 2640-3498.

[32] Tatiana Ermakova, Julia Blume, Benjamin Fabian, Elena Fomenko, Marcus Berlin, and Manfred Hauswirth. 2021. *Beyond the Hype: Why Do Data-Driven Projects Fail?* http://hdl.handle.net/10125/71237

[33] Bill Gaver, Tony Dunne, and Elena Pacenti. 1999. Design: Cultural probes. *Interactions* 6, 1 (Jan. 1999), 21–29. doi:10.1145/291224.291235

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

[35] Francesco Gentile, Jean Charle Yaacoub, James Gleave, Michael Fernandez, Anh-Tien Ton, Fuqiang Ban, Abraham Stern, and Artem Cherkasov. 2022. Artificial intelligence–enabled virtual screening of ultra-large chemical libraries with deep docking. *Nature Protocols* 17, 3 (2022), 672–697.

[36] Sarah Gibbons. 2018. Using Prioritization Matrices to Inform UX Decisions. https://www.nngroup.com/articles/prioritization-matrices/

[37] Sarah Gibbons. 2019. Dot Voting: A Simple Decision-Making and Prioritizing Technique in UX. https://www.nngroup.com/articles/dot-voting/

[38] Google. 2019. People + AI Guidebook. https://pair.withgoogle.com/guidebook

[39] Connor Graham, Mark Rouncefield, Martin Gibbs, Frank Vetere, and Keith Cheverst. 2007. How probes work. In *Proceedings of the 19th Australasian conference on Computer-Human Interaction: Entertaining User Interfaces (OZCHI '07)*. Association for Computing Machinery, New York, NY, USA, 29–37. doi:10.

1145/1324892.1324899
[40] Dave Gray, Sunni Brown, and James Macanufo. 2010. *Gamestorming: A Playbook for Innovators, Rulebreakers, and Changemakers*. "O'Reilly Media, Inc.".
[41] Michaela Hardt, Xiaoguang Chen, Xiaoyi Cheng, Michele Donini, Jason Gelman, Satish Gollaprolu, John He, Pedro Larroy, Xinyu Liu, Nick McCarthy, Ashish Rathi, Scott Rees, Ankit Siva, ErhYuan Tsai, Keerthan Vasist, Pinar Yilmaz, Muhammad Bilal Zafar, Sanjiv Das, Kevin Haas, Tyler Hill, and Krishnaram Kenthapadi. 2021. Amazon SageMaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM, Virtual Event Singapore, 2974–2983. doi:10.1145/3447548.3467177
[42] John S. Harris. 1982. New Product Profile Chart. *IEEE Engineering Management Review* 10, 3 (Sept. 1982), 17–25. doi:10.1109/EMR.1982.4305929 Conference Name: IEEE Engineering Management Review.
[43] Anne D Henriksen and Ann Jensen Traynor. 1999. A practical R&D project-selection scoring tool. *IEEE transactions on engineering management* 46, 2 (1999), 158–170.
[44] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé, 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 (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3290605.3300830
[45] IBM. 2024. IBM Garage: Let's create an approach that turns ideas into outcomes. *https://www.ibm.com/garage* (2024).
[46] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. doi:10.1145/3442188.3445901
[47] Mayur P. Joshi, Ning Su, Robert D. Austin, and Anand K. Sundaram. 2021. Why So Many Data Science Projects Fail to Deliver. *MIT Sloan Management Review* 62, 3 (March 2021). https://sloanreview.mit.edu/article/why-so-many-data-science-projects-fail-to-deliver/
[48] Nathan Kallus and Angela Zhou. 2018. Residual Unfairness in Fair Machine Learning from Prejudiced Data. In *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2439–2448. https://proceedings.mlr.press/v80/kallus18a.html ISSN: 2640-3498.
[49] Anna Kawakami, Amanda Coston, Haiyi Zhu, Hoda Heidari, and Kenneth Holstein. 2024. The Situate AI Guidebook: Co-Designing a Toolkit to Support Multi-Stakeholder, Early-stage Deliberations Around Public Sector AI Proposals. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3613904.3642849
[50] Chrissy Kidd. 2018. Why does Gartner predict up to 85% of AI projects will "not deliver" for CIOs? https://www.bmc.com/blogs/cio-ai-artificial-intelligence/
[51] Beth Kindig. 2024. AI Spending To Exceed A Quarter Trillion Next Year. https://www.forbes.com/sites/bethkindig/2024/11/14/ai-spending-to-exceed-a-quarter-trillion-next-year/
[52] KPMG. 2024. KPMG Ignition: Let's build a brighter future together. *https://kpmg.com/ca/en/home/services/ignition.html* (2024).
[53] 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2 (Oct. 2021), 311:1–311:28. doi:10.1145/3476052
[54] 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–24. doi:10.1145/3544548.3581290
[55] Hao-Ping Hank Lee, Lan Gao, Stephanie Yang, Jodi Forlizzi, and Sauvik Das. 2024. "I Don't Know If We're Doing Good. I Don't Know If We're Doing Bad": Investigating How Practitioners Scope, Motivate, and Conduct Privacy Work When Developing AI Products. In *Proceeding of the 33rd USENIX Security Symposium*.
[56] Michelle Seng Ah Lee and Jat Singh. 2021. The Landscape and Gaps in Open Source Fairness Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. doi:10.1145/3411764.3445261
[57] Houjiang Liu, Anubrata Das, Alexander Boltz, Didi Zhou, Daisy Pinaroc, Matthew Lease, and Min Kyung Lee. 2024. Human-centered NLP Fact-checking: Co-Designing with Fact-checkers using Matchmaking for AI. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2 (Nov. 2024), 423:1–423:44. doi:10.1145/3686962
[58] PricewaterhouseCoopers LLP. 2024. Why you need an AI factory: A CIO's guide to generative AI. *https://www.pwc.com/us/en/tech-effect/ai-analytics/guide-to-generative-ai-for-the-cio.html* (2024).
[59] Chade Lv, Xin Zhou, Lixiang Zhong, Chunshuang Yan, Madhavi Srinivasan, Zhi Wei Seh, Chuntai Liu, Hongge Pan, Shuzhou Li, Yonggang Wen, et al. 2022. Machine learning: an advanced platform for materials development and state prediction in lithium-ion batteries. *Advanced Materials* 34, 25 (2022), 2101474.
[60] 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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376445
[61] Kit-Kay Mak, Yi-Hang Wong, and Mallikarjuna Rao Pichika. 2024. Artificial intelligence in drug discovery and development. *Drug discovery and evaluation: safety and pharmacokinetic assays* (2024), 1461–1498.
[62] David Mannheim. 2021. Experimenting for versioning, visioning and venturing forms of innovation. https://davidleemannheim.medium.com/experimenting-for-versioning-visioning-and-venturing-forms-of-innovation-331d1cbb149b
[63] Tuuli Mattelmäki. 2006. *Design probes*. Aalto University.
[64] Aaron Mcdade. 2024. More Than 40% of CEOs Expect To Boost AI Spend To Gain Competitive Edge—KPMG Survey. https://www.investopedia.com/more-than-40-of-ceos-expect-to-boost-ai-spend-to-gain-competitive-edge-survey-8630199
[65] Scarlett R Miller and Brian P Bailey. 2014. Searching for inspiration: An in-depth look at designers example finding practices. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, Vol. 46407. American Society of Mechanical Engineers, V007T07A035.
[66] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, Atlanta GA USA, 220–229. doi:10.1145/3287560.3287596
[67] 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 (ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 413–425. doi:10.1145/3510003.3510209
[68] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 39–48. doi:10.1145/3287560.3287567
[69] Dorian Peters, Karina Vold, Diana Robinson, and Rafael A. Calvo. 2020. Responsible AI—Two Frameworks for Ethical Design Practice. *IEEE Transactions on Technology and Society* 1, 1 (March 2020), 34–47. doi:10.1109/TTS.2020.2974991 Conference Name: IEEE Transactions on Technology and Society.
[70] Marc Pinski, Monideepa Tarafdar, and Alexander Benlian. 2024. Why Executives Can't Get Comfortable with AI. https://sloanreview.mit.edu/article/why-executives-cant-get-comfortable-with-ai/
[71] Erich Prem. 2023. From ethical AI frameworks to tools: a review of approaches. *AI and Ethics* 3, 3 (Aug. 2023), 699–716. doi:10.1007/s43681-023-00258-9
[72] Stuart Pugh. 1981. Concept Selection: A Method that Works. In *Proceedings of the International conference on Engineering Design*. Heurista, Zürich, 497–506.
[73] Edward O Pyzer-Knapp, Jed W Pitera, Peter WJ Staar, Seiji Takeda, Teodoro Laino, Daniel P Sanders, James Sexton, John R Smith, and Alessandro Curioni. 2022. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Computational Materials* 8, 1 (2022), 84.
[74] McKinsey Quarterly. 2009. Enduring Ideas: The three horizons of growth | McKinsey. https://www.mckinsey.com/capabilities/strategy-and-corporate-finance/our-insights/enduring-ideas-the-three-horizons-of-growth
[75] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 959–972. doi:10.1145/3531146.3533158
[76] 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 (April 2021), 1–23. doi:10.1145/3449081
[77] Goldman Sachs. 2024. Will the $1 trillion of generative AI investment pay off? https://www.goldmansachs.com/insights/articles/will-the-1-trillion-of-generative-ai-investment-pay-off
[78] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Ken Holstein, and John Zimmerman. 2024. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. New York, NY, USA.
[79] Devansh Saxena, Ji-Youn Jung, Jodi Forlizzi, Kenneth Holstein, and John Zimmerman. 2025. AI Mismatches: Identifying Potential Algorithmic Harms Before AI Development. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.
[80] Daniel Schiff, Bogdana Rakova, Aladdin Ayesh, Anat Fanti, and Michael Lennon. 2021. Explaining the Principles to Practices Gap in AI. *IEEE Technology and Society Magazine* 40, 2 (June 2021), 81–94. doi:10.1109/MTS.2021.3056286 Conference Name: IEEE Technology and Society Magazine.
[81] Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*

(FAT '19)*. Association for Computing Machinery, New York, NY, USA, 59–68. doi:10.1145/3287560.3287598

[82] Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N'Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, et al. 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*. 723–741.

[83] Gregory Smith, Jenkins Richardson, Joshua D. Summers, and Gregory M. Mocko. 2012. Concept Exploration Through Morphological Charts: An Experimental Study. *Journal of Mechanical Design* 134, 051004 (April 2012). doi:10.1115/1.4006261

[84] Brian Spisak, Louis Rosenberg, and Max Beilby. 2023. 13 Principles for Using AI Responsibly. https://hbr.org/2023/06/13-principles-for-using-ai-responsibly

[85] JR Steele, RA Babione, LA Shikashio, AJ Wacaster, and AD Paterson. 1994. *Commercial integration and partnering at Savannah River Site*. Technical Report. Savannah River Site (SRS), Aiken, SC (United States).

[86] Erik Stolterman. 2008. The Nature of Design Practice and Implications for Interaction Design Research. *International Journal of Design* 2 (2008). https://ijdesign.org/index.php/IJDesign/article/view/240

[87] Maryam Tohidi, William Buxton, Ronald Baecker, and Abigail Sellen. 2006. Getting the right design and the design right. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Association for Computing Machinery, New York, NY, USA, 1243–1252. doi:10.1145/1124772.1124960

[88] Gerald G Udell and Kenneth G Baker. 1982. Evaluating new product ideas.... systematically. *Technovation* 1, 3 (1982), 191–202.

[89] Annemiek G.C. van Boeijen, J.J. Daalhuizen, and Jelle Zijlstra. 2020. *Delft Design Guide: Perspectives, models, approaches, methods*. BIS Publishers.

[90] Angelina Wang, Sayash Kapoor, Solon Barocas, and Arvind Narayanan. 2024. Against Predictive Optimization: On the Legitimacy of Decision-making Algorithms That Optimize Predictive Accuracy. *ACM J. Responsib. Comput.* 1, 1 (March 2024), 9:1–9:45. doi:10.1145/3636509

[91] 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3544548.3581278

[92] Joyce Weiner. 2020. *Why AI/Data Science Projects Fail - Google Books*. Morgan & Claypool Publishers. https://www.google.com/books/edition/Why_AI_Data_Science_Projects_Fail/jYMSEAAAQBAJ?hl=en&gbpv=0

[93] David Gray Widder, Derrick Zhen, Laura Dabbish, and James Herbsleb. 2023. It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 467–479. doi:10.1145/3593013.3594012

[94] Ian H. Witten and Eibe Frank. 2002. Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record* 31, 1 (March 2002), 76–77. doi:10.1145/507338.507355

[95] 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 (April 2023), 1–27. doi:10.1145/3579621

[96] 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 (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300415

[97] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating How Experienced UX Designers Effectively Work with Machine Learning. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. Association for Computing Machinery, New York, NY, USA, 585–596. doi:10.1145/3196709.3196730

[98] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376301

[99] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.

[100] Qian Yang, John Zimmerman, Aaron Steinfeld, and Anthony Tomasic. 2016. Planning Adaptive Mobile Experiences When Wireframing. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems (DIS '16)*. Association for Computing Machinery, New York, NY, USA, 565–576. doi:10.1145/2901790.2901858

[101] Nur Yildirim, Alex Kass, Teresa Tung, Connor Upton, Donnacha Costello, Robert Giusti, Sinem Lacin, Sara Lovic, James M O'Neill, Rudi O'Reilly Meehan, Eoin Ó Loideáin, Azzurra Pini, Medb Corcoran, Jeremiah Hayes, Diarmuid J Cahalane, Gaurav Shivhare, Luigi Castoro, Giovanni Caruso, Changhoon Oh, James McCann, Jodi Forlizzi, and John Zimmerman. 2022. How Experienced Designers of Enterprise Applications Engage AI as a Design Material. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–13. doi:10.1145/3491102.3517491

[102] Nur Yildirim, Changhoon Oh, Deniz Sayar, Kayla Brand, Supritha Challa, Violet Turri, Nina Crosby Walton, Anna Elise Wong, Jodi Forlizzi, James McCann, and John Zimmerman. 2023. Creating Design Resources to Scaffold the Ideation of AI Concepts. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. ACM, Pittsburgh PA USA, 2326–2346. doi:10.1145/3563657.3596058

[103] 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 (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3544548.3580900

[104] 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, Dan Ricketts, Lu Tang, Venkatesh Sivaraman, Adam Perer, Sarah M. Preum, James McCann, and John Zimmerman. 2024. Sketching AI Concepts with Capabilities and Examples: AI Innovation in the Intensive Care Unit. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3613904.3641896

[105] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (May 2020), 22:1–22:23. doi:10.1145/3392826

[106] John Zimmerman, Jodi Forlizzi, and Shelley Evenson. 2007. Research through design as a method for interaction design research in HCI. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 493–502. doi:10.1145/1240624.1240704

[107] John Zimmerman, Aaron Steinfeld, Anthony Tomasic, and Oscar J. Romero. 2022. Recentering Reframing as an RtD Contribution: The Case of Pivoting from Accessible Web Tables to a Conversational Internet. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3491102.3517789

A Study Material: 6 AI Concepts

A.1 Concept 1: Teacher Evaluation System

- **Description:** AI system that reviews and evaluates teachers' performance through (i) automated classroom observations, (ii) student test scores, (iii) student evaluations, and (iv) assessment of teacher's commitment to school community.
- **Stakeholder** (*Who are the...*)
  - **Payer:** School District (K-12 & College)
  - **End-User:** School administrators
  - **Servicing Party:** EduTech Company (EduNova)
  - **Impacted Stakeholder:** Teachers
- **System description** (*What should the system...*)
  - **Do/Act:** Present a report with grades of teacher performance in different categories (classroom instruction, student achievement, commitment to community)
  - **Infer/Reason:** Streamline and standardize the teacher evaluation process and help school districts make necessary cuts by laying off under-performing teachers.
- **Datasets** (*What do we know?*)
  - **Data 1:** Classroom Footage (Collected through newly installed cameras in classrooms to gauge student engagement)
  - **Data 2:** Student test scores and student evaluations
  - **Data 3:** Teacher's "Commitment to the Community" Score

A.2 Concept 2: AI-Driven Career Counselling

- **Description:** A career advisory service that uses AI to analyze a student's skills, interests, and job market trends to suggest suitable career paths and necessary courses or skills development. It takes a special account of future workforce forecasts.
- **Stakeholder** (*Who are the...*)
  - **Payer:** Highschool Parents, College Students
  - **End-User:** High School and College Students
  - **Servicing Party:** EduTech Company (EduNova)
  - **Impacted Stakeholder:** Students
- **System description** (*What should the system...*)
  - **Do/Act:** Present a list of potential career paths using student academic records, extra-curricular activities, and interests and explain why the jobs were recommended to students.
  - **Infer/Reason:** From the datasets, find a good career match for the student.
- **Datasets** (*What do we know?*)
  - **Data 1:** Job Market Data (current job listings and descriptions, previous job market trends and tendencies, industry trends and future job market predictions).
  - **Data 2:** Educational Data (Course catalogs from various educational institutions)
  - **Data 3:** Student Data. (1) Academic records and transcripts. (2) Extracurricular activities and achievements. (3) Personality assessments and career interest surveys.
  - **Data 4:** Alumni Data. (Career paths of graduates from different programs)

A.3 Concept 3: LLM-based Programming TA

- **Description:** LLM-powered programming assistant in "Introduction to Programming" courses that respond to students' conceptual questions about coding, without revealing direct code solutions. It is expected to provide ongoing and immediate support to students struggling with coding assignments, just like a TA during office hours.
- **Stakeholder** (*Who are the...*)
  - **Payer:** Universities
  - **End-User:** College students in "Introduction to Programming" course
  - **Servicing Party:** EduTech Company (EduNova)
- **System description** (*What should the system...*)
  - **Do/Act:** Generate conversational answers to students' technical questions
  - **Infer/Reason:** Understand students' questions and generate conversational replies without revealing direct code solutions. It tells students when you have arrived at the correct answer.
- **Datasets** (*What do we know?*)
  - **Data 1:** Few-Shot Learning Examples. (Current job listings and descriptions, Industry trends and future job market predictions)
  - **Data 2:** Programming Language Document (i.e., Python, javascript, etc)
  - **Data 3:** Code Execution Traces. (Data on how code is executed, including variable values, function calls, and control flow.)

A.4 Concept 4: Test Scoring for & Open-ended Essay Questions for High School Students

- **Description:** An AI system that automates the grading of essays and provides the reason/rubric behind the grades and open-ended questions for highschool assignments and exams.
- **Stakeholder** (*Who are the...*)
  - **Payer:** School District (High School)
  - **End-User:** Teachers
  - **Servicing Party:** EduTech Company (EduNova)
- **System description** (*What should the system...*)
  - **Do/Act:** Grade the quality of student responses and provide percentage points.
  - **Infer/Reason:** Compare student answers to the right / "high-quality" answers.
- **Datasets** (*What do we know?*)
  - **Data 1:** Sample Answers and Essays (Annotated datasets of sample answers for open-ended questions, including different levels of quality and correctness.)
  - **Data 2:** Scoring Rubrics (Rubrics and guidelines for grading open-ended questions and essays, specifying criteria for different levels of performance.)

A.5 Concept 5: AI Proctor in Classrooms

- **Description:** During exams in the physical classroom, assess each student's behavior by tracking their facial expressions,

eye gaze, posture, and lip movement in real-time and abnormal sounds in the room to flag the possibility of cheating to the proctor. The video footage will be stored and accessible to teachers for further review to determine if cheating occurred.

- **Stakeholder** (*Who are the...*)
  - **Payer:** School District (K-12 & College)
  - **End-User:** High School and College Students
  - **Servicing Party:** EduTech Company (EduNova)
  - **Impacted Stakeholder:** Students
- **System description** (*What should the system...*)
  - **Do/Act:** Flag suspicious cheating behavior by analyzing each students' behavior
  - **Infer/Reason:** Identify suspicious cheating behavior by analyzing real-time video and audio (i.e., eye gaze, facial expressions, posture, movement of lips, abnormal sounds in the room).
- **Datasets** (*What do we know?*)
  - **Data 1:** Video Footage of Exams (Large amounts of labeled video footage from classroom exams showing both typical and suspicious behaviors).
  - **Data 2:** Facial Expressions (Dataset of various facial expressions and head movements during exams, labeled as normal or suspicious).
  - **Data 3:** Body Language (Annotated data of different body postures and movements, such as turning their heads often, which might indicate cheating).
  - **Data 4:** Audio Recordings of Exams (Labeled audio data from exam settings, identifying normal sounds (e.g., rustling of papers) and suspicious sounds (e.g., whispering)).

A.6 Concept 6: AI-Powered Storyteller

- **Description:** An AI-driven application where middle-school kids can generate their own stories, read them, and share them with classmates and friends. The tool will read the story aloud and also generate illustrations for the stories. The stories will be stored in the system to allow users to continually update them.
- **Stakeholder** (*Who are the...*)
  - **Payer:** School District (Middle School)
  - **End-User:** Middle School Students
  - **Servicing Party:** EduTech Company (EduNova)
- **System description** (*What should the system...*)
  - **Do/Act:** Generate young adult (middle school level) stories. ; Generate illustrations for stories. ; Read these stories in different voices.
  - **Infer/Reason:** Generate stories based on characters and plots provided by students. Update the story based on new information (e.g., plot twists).
- **Datasets** (*What do we know?*)
  - **Data 1:** Young-Adult(middle school) Story to fine-tune LLM (A large dataset of stories, including fairy tales, fables, and modern literature, to fine-tune an LLM like GPT-3).
  - **Data 2:** Story Templates (Datasets of various story templates and outlines that can serve as the foundation for story generation).
  - **Data 3:** Vocabulary Lists (Age-appropriate vocabulary lists to ensure the generated stories are suitable for the target audience).
  - **Data 4:** Voice Data (Audio recordings of stories being read aloud, including different voices and intonations, to train text-to-speech models).