

Towards Stakeholder-Grounded Requirements Engineering in AI System Specification

Danniell Hu, Westley Weimer, Elizabeth Bondi-Kelly

University of Michigan, Ann Arbor
{dannihu, weimerw, ecbk}@umich.edu

Abstract

AI systems often fail when their designs do not fit the needs, values, capacities, or constraints of the people they are meant to serve. Prior work has identified many reasons why this alignment can break down, including limited stakeholder authority, organizational incentives, deployment constraints, and power imbalances. Much of this work emphasizes eliciting the right stakeholder needs from the right people. We examine what happens after elicitation, when developers must translate stakeholder needs into concrete requirements and downstream AI design choices.

We conducted a controlled human-subjects study ($n = 49$) using Requirements Engineering (RE) artifacts to study this stakeholder-to-design translation process. Participants worked from stakeholder scenarios to produce requirements artifacts, then selected AI system components to “build” systems they believed would best fit each context. We evaluated these design choices against expert judgments of best-fit, aligned AI designs for each scenario.

We find that conventionally well-formed requirements were not significantly associated with AI design choices that more closely matched expert judgments. Neither self-reported AI knowledge nor prior AI development experience significantly predicted design performance. In contrast, participants whose requirements identified plausible stakeholders, articulated contextually grounded goals, foregrounded human needs or constraints, and acknowledged relevant values were more likely to produce expert-aligned AI designs. We identify these dimensions as *stakeholder-grounded reasoning*. Our findings suggest that preserving stakeholder context through early design depends not only on eliciting stakeholder needs, but also on how developers translate those needs into written design guidance. AI development workflows should more explicitly teach, support, and evaluate this translation step so that stakeholder context is preserved as teams move from stakeholder needs to technical design.

1 Introduction

Artificial Intelligence (AI) systems are increasingly used to address societal challenges across domains such as health-care, education, and social welfare. In these settings, AI systems may shape high-stakes decisions, mediate access to resources, and influence how benefits and harms are distributed across populations (Shi, Wang, and Fang 2020).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

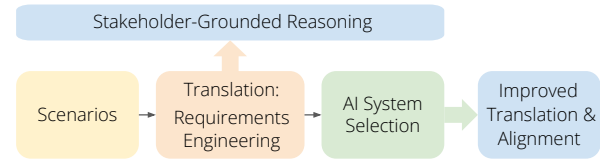


Figure 1: Stakeholder-grounded requirements engineering to support translation from stakeholder scenarios to AI systems.

Motivation. Despite their promise, AI systems frequently fail in real-world settings because of misalignments between system designs and deployment realities (Lin et al. 2024; Hu et al. 2025). In this paper, we use *alignment* to refer to the extent to which an AI system’s objectives, assumptions, technical capabilities, and operational constraints fit the needs, values, capacities, and constraints of the stakeholders it is intended to serve or affect.

Prior work has identified many reasons why alignment can break down during AI development. Stakeholder goals may be displaced by funder priorities, technology-team incentives, deployment constraints, power dynamics, or limited stakeholder authority over technical decisions (Lin et al. 2024; Erete et al. 2025; Majumdar et al. 2025). In response, researchers have emphasized participation, co-design, and more equitable governance as ways to better surface stakeholder needs and ensure they shape AI development (Zytka et al. 2022; Delgado et al. 2023; Birhane et al. 2022).

These approaches are critical, but surfacing stakeholder needs does not guarantee that those needs are addressed by technical design. For example, stakeholders may express a need for timely support, but developers must decide whether that need is best modeled as an automated prediction task, a decision-support tool, a human review process, or some combination. Each technical implementation choice encodes different assumptions about what errors are acceptable, what data are appropriate, how much human judgment should be involved, and beyond.

This points to a difficult “translation” problem. Previous literature recognizes that moving from stakeholder needs to technical specification is challenging because needs are often broad or ambiguous, while specifications must be con-

crete and specific enough to guide design decisions (Van de Poel 2013; Li et al. 2015). This difficulty is also recognized in literature focusing specifically on AI formulations (Saxena et al. 2025; Passi and Barocas 2019; Kross and Guo 2021). However, we have limited empirical guidance on what developers should preserve during specification to maintain alignment in AI systems.

Approach. We examine this translation problem through Requirements Engineering (RE), a common software engineering process for converting stakeholder needs into actionable design guidance (Pacheco and Garcia 2012; Liywalii 2024; Ahmad et al. 2023). In RE, techniques for specification produce written outputs that express how stakeholder needs have been interpreted as requirements, constraints, success criteria, and design priorities. We use these written requirements outputs as an inspectable site for studying stakeholder-to-design translation. Our central question is whether particular properties of these outputs are associated with downstream AI designs that better fit stakeholder needs, values, capacities, and constraints.

To answer this question, we conducted a controlled human-subjects study. Rather than eliciting needs from live stakeholders, participants worked from the same fixed stakeholder scenarios, allowing us to hold constant problem context, stakeholder narrative, stated goals, contextual constraints, and design information. For each scenario, participants produced requirements outputs, then selected AI system components to “build” a system they believed would best fit the stakeholder context. We evaluated these selections against judgments from AI experts (Section 4.1).

We first test whether participants who produce high-quality requirements, as measured by conventional RE rubrics, also produce AI designs that are significantly associated with expert-aligned AI design choices. We find they do not. Self-reported AI knowledge and prior AI development experience also do not significantly predict design performance. Instead, we find that stronger alignment is associated with what we call *stakeholder-grounded reasoning*—the extent to which a requirements artifact preserves stakeholder roles, contextually realistic goals, human needs and constraints, and relevant values (Section 4.3). Participants whose requirements reflected these dimensions produced AI designs that were significantly closer to expert judgments.

Human-Centered Design (HCD) and Value Sensitive Design (VSD) already emphasize attention to human needs, contextual constraints, and values (Margetis et al. 2021; Van Der Hoven and Manders-Huits 2020). Our results provide empirical evidence that applying these principles during the translation from stakeholder context to technical design is associated with stronger downstream alignment, motivating more explicit support and evaluation of this translation step in AI development.

In this paper, we make the following contributions:

1. We present a controlled human-subjects study of stakeholder-to-design translation in AI development, using RE artifacts to investigate how stakeholder needs are converted into design guidance.
2. We identify and operationalize *stakeholder-grounded*

reasoning as a measurable construct for evaluating whether translation artifacts preserve stakeholder roles, contextual goals, human needs, and relevant values.

3. We provide empirical evidence that stakeholder-grounded reasoning in requirements artifacts is associated with AI design choices that more closely match expert judgments of aligned AI systems, while conventional requirements quality, self-reported AI knowledge, and prior AI development experience are not significant predictors in our study.

2 Related Work

We investigate how stakeholder needs are translated into AI design choices. We first review work on alignment in AI systems, then discuss socio-technical design approaches that seek to keep stakeholder needs and values central to system development. Finally, we describe the RE artifacts used in our study and explain why they provide a useful site for examining stakeholder-to-design translation.

2.1 Stakeholder Alignment in AI Systems

Misalignment between AI systems and the needs, values, and constraints of stakeholders is well documented. Prior work shows that AI failures often arise from mismatches between problem framing, system objectives, evaluation metrics, and real-world definitions of success (Varshney and Mojsilovic 2019; Shi, Wang, and Fang 2020; Lin et al. 2024; Saxena et al. 2025; Passi and Barocas 2019). Such systems may be technically functional while remaining unusable or unsustainable for the populations affected by them.

These failures can emerge throughout the AI development pipeline. Projects may face limited stakeholder participation, recruitment challenges, or power imbalances that constrain whose needs shape the system (Kim et al. 2025; Birhane et al. 2022). Later, systems may fail because they lack deployment support, maintenance capacity, or evaluation metrics that capture stakeholder-defined outcomes beyond technical performance (Varshney and Mojsilovic 2019; Lin et al. 2024; Bondi et al. 2021). This literature motivates our focus on translation and alignment.

2.2 Participation, Values, and Stakeholder Needs

Participatory AI, HCD, and VSD offer approaches for ensuring that stakeholder needs, constraints, and values shape system development. Participatory approaches involve affected stakeholders in problem formulation, design, and evaluation, often in response to concerns that technical teams may otherwise reproduce institutional priorities or power asymmetries (Zytka et al. 2022; Delgado et al. 2023; Birhane et al. 2022). HCD emphasizes human needs and contexts of use, while VSD foregrounds the values and ethical commitments implicated by technical systems (Margetis et al. 2021; Van Der Hoven and Manders-Huits 2020).

These approaches surface stakeholder needs but do not necessarily focus on including those needs during translation to technical design. Once input has been elicited, developers must convert it into requirements, objectives, data assumptions, model choices, and evaluation criteria—a translation

in which stakeholder needs may be simplified into generic user goals, converted into model-centered objectives, or separated from the constraints and values that made them meaningful. While some technical methods aim to allow stakeholders to shape modeling choices like objective functions directly (e.g., (Behari et al. 2024)), we broadly examine this translation step, including high-level model choices.

2.3 Requirements Engineering

Requirements Engineering (RE) is a software engineering process concerned with identifying, analyzing, documenting, and prioritizing system goals, constraints, and assumptions (Pacheco and Garcia 2012; Ahmad et al. 2023). Because RE is intended to translate stakeholder needs into actionable guidance for development, it provides a useful lens for studying whether stakeholder intent is preserved as technical design decisions are made.

In this paper, we focus on user stories, acceptance criteria, and prioritization methods, which are commonly used RE techniques in agile software development contexts (Cohn 2004; Franch et al. 2024). These artifacts usually fall within requirements specification, where stakeholder interactions are converted into statements that guide design and implementation. We use classic RE approaches because relatively little work has empirically examined whether common RE artifacts support early AI design decisions. This allows us to study foundational RE concepts before evaluating more complex or domain-specific methods.

User Stories

Definition. User stories are short, structured statements that describe functionality from the user’s perspective. They help keep development centered on stakeholder goals and needs.

Format:

As a [type of user], **I want** [to do something], **so that** [I achieve this goal].

Example.

As a patient, I want to control how my medical images are used in training, so that I can protect my privacy.

Acceptance Criteria

Definition. Acceptance criteria define the conditions under which a user story is considered complete and successful. They clarify ambiguity and specify how a requirement can be tested against stakeholder expectations.

Format.

Given some initial context, **When** a specific action is taken, **Then** a desired outcome should occur.

Example.

Given that I am completing hospital consent forms, when I am asked about the use of my medical data for educational purposes, then I must have a clear option to either consent or decline.

MoSCoW Method

The MoSCoW method is a prioritization technique that categorizes requirements by importance (Clegg and Barker 1994):

- **Must have:** required for the system to work.
- **Should have:** important, but not essential.
- **Could have:** useful, but lower priority.
- **Won’t have:** explicitly excluded from current scope.

We evaluate conventional user story quality using a rubric adapted from the INVEST framework (Buglione and Abran 2013) (Section 4.2), which provides a structured lens for assessing whether user stories are well formed. We observe that conventional RE rubrics assess whether requirements are procedurally well formed, but they may not capture whether requirements preserve the stakeholder context needed for AI design. We therefore compare conventional requirements quality with stakeholder-grounded reasoning as predictors of stakeholder-aligned AI design.

3 Study Design

We conducted a controlled human-subjects study to examine how developers translate stakeholder needs into requirements artifacts and AI design choices. The study had four phases, administered via a Qualtrics survey: (1) RE training, (2) requirements artifact creation, (3) AI system design, and (4) a post-task questionnaire.¹

Our study is guided by three research questions:

- **RQ1:** Is requirements quality, as measured by conventional RE rubrics, associated with more expert-aligned AI design choices?
- **RQ2:** Which properties of requirements artifacts are associated with more expert-aligned AI design choices?
- **RQ3:** Is lightweight RE training associated with higher-quality requirements artifacts during stakeholder-to-design translation?

3.1 Participant Recruitment

Participants were recruited via email outreach, social media, and the Prolific crowdsourcing platform (Prolific 2024). Eligibility criteria included being at least 18 years old, fluent in English, and having prior programming experience. Participants also self-reported completion of or current enrollment in an introductory data structures and algorithms course.

We recruited participants with programming experience because our goal was to study people who plausibly could be, or soon become, developers of AI systems. Such developers may be asked to translate stakeholder needs into design choices even if they have not received formal RE training. Participants who completed the study received \$30 in compensation. The study protocol was approved by our Institutional Review Board (protocol HUM00272282).

¹Study materials, including stakeholder scenarios, survey, rubrics, etc. are available at <https://github.com/realize-lab/RE-AI>.

Group	Category	Total	Prolific	Other
Age	18–24	31	6	25
	25–39	15	13	2
	40+	3	3	0
Gender	Man	38	14	24
	Woman	11	8	3
Race / ethnicity	Black or African American	17	16	1
	Asian	14	0	14
	White	15	5	10
	Mixed race	3	1	2
Current education	UG CS	14	7	6
	MS CS	8	7	1
	PhD CS	8	7	1
	UG non-CS	10	2	8
	MS non-CS	6	3	3
	PhD non-CS	1	1	0
	None	4	2	2
Current job	CS industry	19	13	6
	CS academia	7	1	1
	Other industry	3	1	2
	Other academia	7	7	0
	Self-employed	1	1	0
	Not working	13	0	13

* Participants could select multiple education or job categories.

Table 1: Participant demographics and recruitment.

3.2 Participant Demographics

We report demographics in Table 1. Participants were primarily young, with 63% aged 18–24, 31% aged 25–39, and 6% aged 40 or older. The sample skewed male, with 78% identifying as men and 22% identifying as women. Participants reported diverse racial and ethnic backgrounds, including Black or African American (35%), Asian (29%), White (31%), and multiracial (6%) identities. Participants varied in educational and professional background, including undergraduate students, master’s students, PhD students, industry practitioners, and participants outside computer science.

3.3 Training Conditions

Participants were randomly assigned to one of three RE training conditions. All conditions introduced the same core RE concepts: user stories, acceptance criteria, and MoSCoW prioritization (Section 2.3). The conditions varied in how much structured instruction and practice participants received before completing the specification task.

Minimal Training. Participants received basic definitions of user stories, acceptance criteria, and MoSCoW prioritization. This condition did not include examples, a training scenario, reflective prompts, or interactive exercises.

Generic RE Training. Participants received detailed, standard RE instruction grounded in a stakeholder scenario (Section 3.4) about a hospital residency program that could no longer use real patient MRI scans for diagnostic training because of privacy concerns. The stakeholder requested a generic (non-AI) software platform to support resident training. With the stakeholder scenario as context, the training

introduced user stories, acceptance criteria, and MoSCoW prioritization through examples, interactive exercises, and practice applying these techniques.

RE Training with AI Context. Participants received the same instructional structure and underlying hospital MRI training scenario as the generic RE condition, but the stakeholder explicitly requested an AI-based solution, such as generating realistic synthetic MRI images for training. This condition allowed us to test whether adapting the training context to AI development, without otherwise changing the RE instruction, affected later artifacts or AI design choices.

3.4 Stakeholder Scenarios

To contextualize the RE techniques, we use a fictional *stakeholder scenario*, a structured narrative simulating an early-stage scoping conversation between a developer and a stakeholder. Scenarios have been shown to be an effective template for requirements elicitation (Cirqueira et al. 2020). Each scenario introduces a focal stakeholder, their context, and a challenge they face. Following established scenario design practices, scenarios specify stakeholder goals, constraints, and relevant contextual factors. Each scenario was constructed with an intended best-fit AI design in mind. Scenarios also include *distractor cues*, or details that make certain design options appear reasonable even though they are ultimately misaligned with stakeholder goals or constraints. These cues prompt participants to distinguish viable from unsuitable options, reinforcing trade-off analysis central to effective RE (Laplante and Kassab 2022).

3.5 Participant Tasks

After training, all participants completed tasks for four new stakeholder scenarios. Each participant saw the same four scenarios, and scenario order was randomized to reduce order effects. The scenarios spanned multiple application domains, including social care, urban planning, education, and environmental sustainability.

For each stakeholder scenario, participants produced requirements artifacts intended to capture the stakeholder needs and constraints they believed should guide system development. Specifically, participants:

1. Identified the number of key stakeholders they believed were most relevant to the scenario,
2. Wrote user stories for those stakeholders,
3. Wrote acceptance criteria for each user story, and
4. Prioritized acceptance criteria using MoSCoW.

After creating requirements artifacts for a scenario, participants selected AI system components to “build” an AI system they believed would best fit the stakeholder context. The task included five curated, non-exhaustive design categories: data type, learning paradigm, model type or technique, training regime, and system behavior or setting. Participants selected one option from each category and could consult a reference document defining the available options. Table 2 summarizes the categories and example options.

The component set was constrained to support comparison across participants. Our goal was not to simulate full

Category	Options
Learning paradigm	Supervised; unsupervised; reinforcement; semi-supervised; imitation; human-in-the-loop learning
Model type / technique	Graph neural networks; transformers; bandits; clustering; rule-based systems; diffusion models
Data type	Numeric; tabular; images; audio; time series
Training regime	Transfer learning; few-shot/meta learning; pre-trained models; trained from scratch
System behavior / setting	Complementary; cooperative; fully interactive; decision support; autonomous

Note. System behavior/setting options vary in human involvement, from decision support and deferral to cooperative, continuous, or autonomous operation.

Table 2: AI system component categories used in the design task.

AI system design, but to evaluate whether participants’ design choices were appropriate for the stakeholder goals and constraints represented in the scenarios.

3.6 Post-Task Questionnaire

To mitigate stereotype threat, participants completed a post-task questionnaire after completing the training, requirements task, and AI design tasks. We collected demographic information, self-reported theoretical AI knowledge, and self-reported practical AI development experience.

4 Analysis Methodology

Our analysis examines whether different properties of participant-authored requirements artifacts are associated with stakeholder-aligned AI design. We compare requirements artifacts to participants’ downstream AI component selections. Our analysis proceeds in four steps. First, we define an expert reference set for AI systems with good alignment. Second, we score participant AI design choices against that reference set. Third, we evaluate participant-authored requirements artifacts using both conventional RE rubrics and our stakeholder-grounded reasoning rubrics. Fourth, we use regression models to examine which measures are associated with stakeholder-aligned AI design.

Data Cleaning Participants were instructed not to use generative AI tools during the study and consented to this condition. Responses were excluded if they contained explicit ChatGPT or similar prompting stems, if they later self-reported using generative AI tools during the study in a self-report item, or if their responses were unintelligible (gibberish, or the same copy pasted response for each question). From an initial sample of 61 respondents, the final analysis includes 49 participants.

4.1 Assessing AI-Stakeholder Alignment

To evaluate participants’ AI design choices, we established an expert reference set for each stakeholder scenario. We recruited four AI experts spanning multiple subfields, including multi-agent systems and game theory, ML for health and

sustainability, AI for science, and human–AI interaction. All had post-graduate AI development and research experience and demonstrated involvement in designing, deploying, or sustaining real-world AI-for-social-impact systems across academic, industry, and nonprofit settings. Experts reviewed each stakeholder scenario individually, without coordination or knowledge of the intended best-fit design, selecting the AI components they believed best fit the stakeholder’s needs, constraints, and context. Experts did not author RE artifacts.

Experts showed strong convergence across the 20 scenario-by-component decisions (five component categories across four scenarios): 8 decisions were unanimous, 9 had a 3-of-4 majority, and only 3 lacked a clear majority. Because multiple AI designs could reasonably fit a given scenario, we used the union of expert selections as the acceptable reference set rather than requiring full consensus: any component selected by at least one expert was treated as acceptable for that scenario. Of the 104 options across the 20 decisions, 33 appeared in the reference set.

As described in Section 3.5, participants selected one option from each of five AI component categories to “build” an AI system for each of the four scenarios. We computed an *expert-alignment score* by counting the number of participant-selected components that appeared in the expert reference set for the corresponding scenario and summing across all four scenarios. Higher scores indicate that a participant’s design choices more closely matched expert judgments of which components fit the stakeholder context. The union-based reference set remains discriminative while accommodating expert disagreement: a participant selecting randomly from each category would score approximately 6.4 of 20, whereas participants averaged 11.96 (SD = 3.35).

We treat expert agreement as a proxy for stakeholder alignment, not a direct measure of stakeholder satisfaction or deployment fit. Expert judgments are limited to the controlled translation task: experts reviewed the same scenarios as participants and judged which component choices best fit those scenarios. We therefore interpret the expert-alignment score as convergence with experienced practitioners’ judgments under matched conditions. Experts may share assumptions with our HCD- and VSD-motivated rubrics, so this measure should not be read as evidence that a design would be endorsed by stakeholders in deployment.

4.2 Conventional Requirements Quality

We evaluated participant-authored requirements artifacts using conventional RE quality criteria to test whether procedurally well-formed requirements were associated with expert-aligned AI design.

User story quality. We evaluated user stories using criteria adapted from the conventional RE rubric INVEST, which stands for *Independent, Negotiable, Valuable, Estimable, Small, and Testable*. We assessed three INVEST criteria: *Negotiable, Valuable, and Testable* (NVT). These criteria capture whether a user story expresses a flexible need rather than a fixed solution, articulates clear value for the stakeholder, and can be plausibly verified to pass or fail. The full grading rubric appears in Table 4.

Indicator	Score = 1	Score = 0
Verifiable	Describes an outcome that could translate into one or more clear tests.	Describes an outcome that is subjective, vague, or not testable.
Atomic	Describes a single condition and a single expected outcome.	Contains multiple independent behaviors or bundled requirements.
Implementation-Independent	Avoids specifying UI elements, algorithms, technologies, or internal logic.	References specific implementation details, such as UI layouts, database operations, or algorithms.

Table 3: RE acceptance criteria evaluation rubric.

Indicator	Score = 1	Score = 0
Negotiable	Describes a goal, preference, or problem rather than a fixed solution.	Locks in a specific method, constraint, or structure.
Valuable	Clearly describes how it helps the identified stakeholder.	Has an unclear or generic benefit.
Testable	Could plausibly be tested to determine whether the goal was achieved.	Has an abstract or subjective outcome that is not realistically testable.

Table 4: Conventional RE user story evaluation rubric.

We did not assess *Independent*, *Estimable*, or *Small* because these criteria assume planning and implementation contexts that were not present in our survey-based translation task, such as managing dependencies across requirements, estimating development effort, or decomposing work for sprint-based implementation. Each evaluated criterion was scored binarily using the predefined rubric, and each user story received a total score ranging from 0 to 3.

Acceptance criteria quality. To evaluate acceptance criteria, we applied a rubric grounded in widely cited RE guidance rather than a single standard, as acceptance criteria quality is context-dependent and often discussed across practitioner and methodological literature rather than formalized into a unified framework (Schön, Thomaschewski, and Escalona 2017). Each acceptance criterion was evaluated in relation to its associated user story using three properties (Table 3): *Verifiable*, *Atomic*, and *Implementation-Independent*. *Verifiable* assesses whether the criterion specifies an objectively testable outcome; *Atomic* whether it expresses a single condition; and *Implementation-Independent* whether it avoids prescribing specific technical solutions or design choices. Similarly, each evaluated acceptance criterion received a total score from 0 to 3.

4.3 Stakeholder-Grounded Reasoning

Standard RE rubrics assess whether requirements are procedurally well formed, but we observe that they may not capture whether requirements preserve the stakeholder context needed for AI design. We define *stakeholder-grounded reasoning* as the extent to which a requirements artifact carries

forward the stakeholder roles, goals, needs, constraints, and values that make a design problem meaningful in context. Motivated by HCD, VSD, and recent work integrating these perspectives into RE (Margetis et al. 2021; Van Der Hoven and Manders-Huits 2020), we developed additional rubrics to assess stakeholder-grounded reasoning. Rubrics were pre-defined before any data analysis.

For user stories, we evaluated four dimensions: *Stakeholder Plausibility*, *Goal Plausibility*, *Human-Centered Framing*, and *Value-Oriented Framing*. These dimensions are important because a user story may be negotiable, valuable, and testable while still being misaligned with the stakeholder context. Together, they capture whether participants identified appropriate stakeholders, articulated realistic goals, foregrounded human needs or constraints, and acknowledged relevant values or ethical considerations. Each dimension was scored on a three-level scale (0.0, 0.5, 1.0), with higher scores indicating stronger stakeholder-grounded reasoning. The full user story rubric appears in Table 5.

For acceptance criteria, we evaluated two dimensions: *User Story Relevance* and *Scenario Scope Validity* (Table 6). These dimensions assess whether success conditions are logically tied to both the associated user story and the broader stakeholder scenario. They capture whether acceptance criteria specify realistic, in-scope indicators of success rather than vague or misaligned outcomes. Each dimension was scored on the same three-level scale. We use a three-level ordinal scale for stakeholder-grounded reasoning because these dimensions admit meaningful partial credit. These dimensions can be present strongly, weakly, or not at all. In contrast, the conventional RE dimensions in our study function as more direct procedural checks, such as whether a user story is negotiable or whether an acceptance criterion is verifiable. We therefore scored conventional RE quality binarily and stakeholder-grounded reasoning ordinally.

4.4 Coding of Requirements Artifacts

We evaluated participant-authored requirements artifacts using a combination of manual coding and LLM-assisted rubric-guided coding. Manual coding was used for conventional user story quality, while LLM-assisted coding was used for conventional acceptance criteria quality and stakeholder-grounded reasoning.

Manual Coding We used human-only annotation to evaluate user stories with the traditional INVEST framework. Two researchers jointly annotated an initial subset of user

Indicator	Score = 1.0	Score = 0.5	Score = 0.0
Stakeholder Plausibility	Role is clearly a plausible system user or directly affected party.	Role is adjacent, indirectly affected, plausible but weakly justified, or ambiguous.	Role does not make sense for the scenario or is too generic or unrealistic.
Goal Plausibility	Goal reflects a realistic responsibility or motivation for the role.	Goal is reasonable for the role but not central or clearly justified.	Goal does not match the responsibilities or impacts implied for the role.
Human-Centered Framing	Story centers a concrete human need, limitation, or contextual risk.	Human need or constraint is vague, implied, or underspecified.	Story is system- or technology-centered with no human need.
Value-Oriented Framing	Story expresses a clear stakeholder value or ethical objective.	Story expresses general improvement goals without clear value framing.	Story omits stakeholder values or ethical objectives.

Table 5: Stakeholder-grounded user story evaluation rubric.

Indicator	Score = 1.0	Score = 0.5	Score = 0.0
User Story Relevance	Directly evaluates a success condition required by the user story.	Related to the user story, but loosely linked or partially scoped.	Not meaningfully linked or introduces unrelated success conditions.
Scenario Scope Validity	Plausible, in scope, and consistent with the scenario and stakeholder role.	Plausible but slightly out of scope, underspecified, or weakly grounded.	Unrealistic, contradictory, or outside the scenario context or role.

Table 6: Stakeholder-grounded acceptance criteria evaluation rubric.

Rubric item	κ	n
<i>Acceptance criteria quality</i>		
Verifiable	0.72	76
Atomic	0.72	78
Implementation-independent	0.39	77
<i>Stakeholder-grounded reasoning</i>		
Stakeholder plausibility	0.74	89
Goal plausibility	0.71	86
Human-centered framing	0.71	81
Value-oriented framing	0.73	85

Table 7: Human–LLM agreement for LLM-coded rubric items.

stories to calibrate their interpretations of the rubric before splitting the data and annotating individually. Inter-rater reliability (IRR) was assessed on $n = 46$ double-coded user stories, yielding substantial agreement for *Negotiable* ($\kappa = 0.78$) and *Testable* ($\kappa = 0.79$), and moderate agreement for *Valuable* ($\kappa = 0.60$).

Coding using LLMs To enable scalable evaluation across these tasks, we additionally employ a large language model (LLM) as an evaluator. We note that we are not using LLMs to generate user stories (Pasquale et al. 2025), but instead to help score them (validated by human checking).

Prior work shows that LLMs can perform rubric-based annotation with substantial agreement to human annotators when criteria are explicitly defined and outputs are con-

strained (Chew et al. 2023; De Paoli 2024; Tai et al. 2024). The evaluation tasks in this paper are well suited to this approach: our rubrics are narrowly scoped and designed to measure defined properties of participant artifacts. Consistent with best practices, we treat the LLM as a measurement tool rather than an authoritative source of ground truth.

We follow a rubric-guided evaluation process adapted from prior work (Chew et al. 2023). For each task, we defined a task-specific rubric specifying explicit scoring criteria and developed structured prompts instructing the model to apply the rubric and assign scores for each dimension. The model was prompted to generate brief justifications to expose how it interpreted and applied each rubric dimension, enabling inspection and refinement of the prompt.

Before annotating the full dataset, we validated the LLM’s application of each rubric item on a subset of artifacts. Human annotators independently coded the same subset, and we assessed agreement between the human-coded labels and LLM-coded labels using unweighted Cohen’s κ . We interpreted κ values using established guidelines (Landis and Koch 1977). After prompt refinement, agreement was moderate to substantial across rubric items, suggesting that the LLM applied the rubrics in a manner broadly consistent with human judgment. After validation, the remaining artifacts were evaluated using the LLM.

All evaluations used Gemini 2.5 Pro with temperature 0.0 and fixed prompt templates, with system instructions constraining the model’s role and enforcing strict rubric adherence. Agreement with human annotators is not interpreted as

evidence that the LLM produces objectively “correct” judgments. Rather, these analyses support the use of the LLM as a consistent and scalable rubric evaluator.

Table 7 reports agreement for each LLM-coded rubric item. Agreement was substantial for all stakeholder-grounded user story dimensions and for two of the three traditional acceptance criteria dimensions. Agreement for *Implementation-Independent* was high in absolute terms, with raw agreement greater than 95%, but Cohen’s κ was lower because of class imbalance: most participants did not specify implementation details in their acceptance criteria.

4.5 Statistical Analysis

We used regression models to examine whether requirements quality and stakeholder-grounded reasoning were associated with the *expert-alignment score* described in Section 4.1. This score was the primary dependent variable for analyses of AI design performance. Models controlled for self-reported theoretical AI knowledge and practical AI development experience, current job, and current education.

All predictors were aggregated to the participant level before modeling. Conventional user story quality was measured as each participant’s average NVT score across user stories from all stakeholder scenarios. Conventional acceptance criteria quality was measured analogously as each participant’s average score across stakeholder scenarios.

Stakeholder-grounded reasoning was measured using four participant-level summary indices: Stakeholder Plausibility, Goal Plausibility, Human-Centered Framing, and Value-Oriented Framing. Requirements artifacts were scored on a three-level ordinal scale (0.0, 0.5, 1.0) for each dimension. We then averaged each dimension across a participant’s requirements artifacts from all stakeholder scenarios. Because these indices are averages over multiple ordinal ratings rather than single artifact-level scores, we modeled them as approximately continuous summary measures. We interpret coefficients as associations with higher average stakeholder-grounded reasoning, not as precise interval-scale effects of moving between individual rubric levels.

We fit Ordinary Least Squares (OLS) regressions with HC3 robust standard errors. Each stakeholder-grounded reasoning index was tested in a separate regression model with the same control variables. We therefore interpret these models as estimating associations between each stakeholder-grounding dimension and expert-alignment score.

To examine whether training condition affected requirements artifact quality, we treated training condition as a categorical predictor and tested whether it was associated with conventional requirements quality or stakeholder-grounded reasoning outcomes. Models were fit using `statsmodels` in Python. Throughout, we report Benjamini–Hochberg-adjusted p-values (q-values) to correct for false discovery across multiple comparisons ($q < 0.05$).

5 Results

We organize the results around the three research questions introduced in Section 3.

5.1 RQ1: Conventional Requirements Quality Was Not Associated with Expert-Aligned AI Design Choices

We first examine whether requirements quality, as measured by conventional RE rubrics, is associated with expert-aligned AI design choices. We use expert alignment as a proxy for stakeholder alignment in the design task. If conventional RE quality captures the aspects of requirements artifacts that matter for stakeholder-to-design translation, then higher scores on user story and acceptance criteria rubrics should be associated with higher expert alignment scores.

We did not find this pattern. User stories evaluated using the INVEST criteria of Negotiable, Valuable, and Testable did not significantly predict expert-alignment ($\beta = 1.65$, 95% CI $[-1.10, 4.40]$, partial $R^2 = .03$, $q > 0.95$). Acceptance criteria quality, measured using Verifiable, Atomic, and Implementation-Independent, was also not significantly associated with expert-alignment ($\beta = 0.48$, 95% CI $[-3.08, 4.04]$, partial $R^2 < .01$, $q > 0.95$).

Thus, procedurally well-formed requirements were not necessarily associated with higher expert-alignment scores. A user story can be negotiable, valuable, and testable while still naming an implausible stakeholder, stating a generic goal, or omitting relevant human constraints. Similarly, an acceptance criterion can be verifiable and atomic while remaining irrelevant to the stakeholder scenario.

5.2 RQ2: Stakeholder-Grounded Reasoning Was Associated with Expert-Aligned AI Design Choices

We found significant positive associations for four dimensions of stakeholder-grounded reasoning (Section 4.3): *Stakeholder Plausibility*, *Goal Plausibility*, *Human-Centered Framing*, and *Value-Oriented Framing*. As shown in Table 8, each dimension was associated with higher expert-alignment scores after controlling for self-reported AI knowledge, prior AI development experience, current job, and current education. The coefficients represent the expected change in expert-alignment score for a full one-point increase in the corresponding stakeholder-grounded reasoning index. Interpreted more practically, a 0.10 increase in the stakeholder-grounded reasoning score corresponds to approximately one additional expert-aligned component choice. Overall, these results suggest that participants whose requirements artifacts better preserved stakeholder roles, contextual goals, human needs or constraints, and relevant values tended to make AI design choices that more closely matched expert judgments of scenario fit.

Predictor	β	95% CI	q	Partial R^2
Stakeholder Plausibility	14.65	[7.97, 21.32]	< .001	.301
Goal Plausibility	9.72	[3.33, 16.10]	.009	.171
Human-centered framing	13.10	[5.24, 20.96]	.004	.199
Value-oriented framing	10.55	[4.08, 17.01]	.004	.192

Table 8: Stakeholder-grounded reasoning indices predicting the expert-alignment score.

Score	Example User Story
1.0	As a local emergency dispatch center worker, I want to accurately prioritize incoming calls, so that we can get services to those who need them most in a timely manner.
0.5	As a worker, I want an automated system, so that information can be processed fast enough.
0.0	As a company, I want to accord engineers necessary tools so they can build the system without struggle.

Table 9: Ex. of user stories scored for Stakeholder Plausibility.

Score	Example User Story
1.0	As an electrical worker, I want to check whether a certain area has power, so that I can fix outages as soon as possible.
0.5	As a policy maker, I want cost/benefit projections for smart grid funding decisions.
0.0	As a resident, I want to have access to my media, so that I can be able to post about the weather.

Table 10: Ex. of user stories scored for Goal Plausibility.

Stakeholder Plausibility User stories that identified a plausible system user or directly affected stakeholder were associated with higher performance on the AI component selection task. The examples in this subsection are drawn from the emergency response scenario. As shown in Table 9, the 1.0-scored user story explicitly names a local emergency dispatch center worker, the role responsible for prioritizing calls, coordinating responders, and making time-critical decisions. In contrast, the 0.0-scored example names “a company,” which is not relevant to the scenario as a system user.

Goal Plausibility User stories that articulated goals plausibly aligned with the identified actor’s role were associated with higher AI component selection performance. The examples in this subsection are drawn from the urban planning and sustainability scenario. From Table 10, the 1.0-scored example aligns with an electrical field technician role whose responsibilities include monitoring infrastructure and rapidly repairing outages. In contrast, the 0.0-scored example identifies a resident, which is reasonable, but assigns a goal unrelated to the scenario’s design problem.

Human-Centered Framing User stories that stated a concrete human need, limitation, or risk relevant to the scenario were associated with higher performance on the AI component selection task. The examples in this subsection are drawn from the social benefits and social care scenario. As shown in Table 11, the 1.0-scored example foregrounds the case worker’s need to triage applications efficiently, reflecting workload management and decision support in a resource-constrained setting. In contrast, the 0.0-scored example focuses on technical reuse and implementation without referencing a human need or contextual risk.

Score	Example User Story
1.0	As a case worker, I want the system to categorize applications into “likely eligible,” “unclear,” or “likely ineligible,” so that I can prioritize my review efforts efficiently.
0.5	As a policy maker, I want to track processing times before and after AI implementation, so that I can measure efficiency gains.
0.0	As a system engineer, I want to extend the legacy system without starting from scratch.

Table 11: Ex. of user stories scored for Human-Centered Framing.

Score	Example User Story
1.0	As a student, I want to anonymously share my struggles outside of school, so that I can get help without feeling embarrassed or judged.
0.5	As a counselor, I want better understanding of student survey data, so that I can better understand what kinds of problems the students are facing.
0.0	As a counselor, I want to analyze students’ feedback, so that I figure out the biggest struggles for students.

Table 12: Ex. of user stories scored for Value-Oriented Framing.

Value-Oriented Framing User stories that articulated a relevant human value or ethical objective were associated with higher AI component selection performance. The examples in this subsection are drawn from the education-focused student risk detection and early intervention scenario. As shown in Table 12, the 1.0-scored example foregrounds student vulnerability and reluctance to disclose sensitive struggles, emphasizing values such as privacy, dignity, and harm avoidance. In contrast, the 0.0-scored example focuses on gaining understanding, which is an analytical task rather than a value or normative objective.

5.3 RQ3: Lightweight RE Training Was Not Associated with Higher Requirements Quality

We did not find evidence that the brief training improved either conventional requirements quality or stakeholder-grounded reasoning. Training condition was not significantly associated with conventional requirements quality for user stories ($q = 0.51$) and acceptance criteria ($q = 0.51$), or stakeholder-grounded reasoning scores for user stories and acceptance criteria (all $q > 0.62$). In other words, participants who received more structured RE instruction did not produce measurably stronger requirements artifacts than those in the minimal training condition. This result should be interpreted cautiously. The interventions were intentionally lightweight, so the finding does not suggest that RE training is unimportant. Rather, it suggests that brief exposure to RE concepts may be insufficient to improve how participants preserve stakeholder context in requirements artifacts.

5.4 Additional Analysis

For acceptance criteria, *User Story Relevance* did not significantly predict component selection performance ($q = 0.09$), while *Scenario Scope Validity* did ($q = 0.003$). We also examined whether performance improved when high-quality acceptance criteria, defined as the average of *User Story Relevance* and *Scenario Scope Validity*, were prioritized as *Must have* requirements. This association was positive but not significant ($q = 0.10$). These results suggest that stakeholder-grounded reasoning may matter when developers prioritize success conditions, though this finding requires replication.

We also examined whether expert alignment could be explained by prior AI expertise. Neither self-reported theoretical AI knowledge nor practical AI development experience significantly predicted expert-alignment, and controlling for these measures did not eliminate the observed stakeholder-grounded reasoning associations.

6 Discussion

Our results indicate that stakeholder-to-design translation is not an automatic outcome of stakeholder elicitation. Even when the same stakeholder information is available, developers may preserve, omit, or simplify different aspects of that context as they move from stakeholder needs to requirements and technical design choices.

These results should not be interpreted as evidence that stakeholder-grounded requirements causally produce better AI design choices. Participants who understood the scenarios more deeply, or had stronger reading and writing skills, may have both written more stakeholder-grounded requirements and selected AI components that better matched expert judgments. In this interpretation, requirements artifacts are not necessarily the mechanism that improves design; instead, they are observable traces of participant comprehension and reasoning. We therefore interpret stakeholder-grounded requirements as evidence of how participants carried stakeholder context forward during the task, not as proof that the artifacts themselves caused stronger performance.

This distinction bounds the scope of our claims. Our study shows that, during specification, requirements preserving plausible stakeholders, contextual goals, human needs or constraints, and relevant values were associated with AI design choices that more closely matched expert judgments of scenario fit. Requirements artifacts can therefore help diagnose whether stakeholder context is maintained as developers move toward technical design. Finally, we do not claim to identify the sole cause of real-world AI misalignment, which prior work attributes to many factors across the development pipeline. Rather, we identify translation as one specific site where targeted intervention may be possible.

6.1 Implications for AI Development and Training

RE is one technique for supporting stakeholder-to-design translation. We do not claim that RE is the only way to support specification, nor that RE artifacts should substitute for participatory design, stakeholder governance, or domain-specific evaluation. Rather, if teams use RE, it may be ben-

eficial for their artifacts to be evaluated both for procedural quality and for the preservation of stakeholder context.

More broadly, stakeholder-to-design translation could be treated as a core design skill. Lightweight RE training was not associated with higher requirements quality in our study, and self-reported AI knowledge and prior AI development experience did not significantly predict performance on the AI component selection task. These null results should be interpreted cautiously given our sample size and brief interventions. They do not show that RE training is unimportant or that technical expertise does not matter for AI design. Instead, they indicate that brief exposure to standard RE formats, even when adapted to an AI context, may be insufficient to help developers preserve stakeholder context as they move from problem understanding to system design.

This interpretation is consistent with work in human-centered AI requirements and responsible AI practice which shows that socially appropriate AI systems require complex reasoning about users, contexts of use, organizational constraints, and stakeholder values (Ahmad et al. 2023; Madaio et al. 2022). Existing supports such as training materials, review checklists, structured prompts, and tool-assisted feedback often emphasize procedural correctness, completeness, or consistency (Lucassen et al. 2016; Buglione and Abran 2013; Wong, Madaio, and Merrill 2023). Our findings indicate that these supports should also help developers practice identifying affected parties, distinguishing stakeholder goals from technical objectives, recognizing contextual constraints, and tracing values across design decisions.

6.2 Limitations

Our study has several limitations. First, the study's fictional scenarios, simplified design tasks, sample of early-career developers, and selection of aligned designs by multiple experts may not fully capture the complexity of real-world AI development or the expertise of seasoned practitioners. Second, stakeholder reasoning was approximated via written rubrics, meaning the results might reflect baseline scenario understanding. Finally, because the study was administered as a survey, the undetected use of generative AI tools may influence results despite strict filtering.

These limitations reflect trade-offs in the study design. The controlled task enabled comparison across participants while preserving the core translation activity. We encourage mixed methods approaches and further study of practical questions through real-world case studies in future work.

6.3 Conclusion

This study examined the translation step between stakeholder elicitation and AI design. Using RE artifacts as an inspectable record of this step, we found that conventional requirements quality was not significantly associated with expert-aligned AI design choices, while requirements preserving plausible stakeholders, contextual goals, human constraints, and relevant values were. These findings point to stakeholder-grounded reasoning as a useful lens for evaluating whether early design artifacts carry stakeholder context forward into technical design.

Acknowledgments

We thank our study participants for their time and thoughtful contributions, and the four AI experts who provided reference judgments for the stakeholder scenarios. We also thank Sarah Snay for helpful discussions and feedback on manual coding. This work was supported by the University of Michigan.

References

- Ahmad, K.; Abdelrazek, M.; Arora, C.; Bano, M.; and Grundy, J. 2023. Requirements engineering for artificial intelligence systems: A systematic mapping study. *Information and Software Technology*, 158: 107176.
- Behari, N.; Zhang, E.; Zhao, Y.; Taneja, A.; Nagaraj, D.; and Tambe, M. 2024. A decision-language model (dlm) for dynamic restless multi-armed bandit tasks in public health. *Advances in Neural Information Processing Systems*, 37: 3964–4002.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022. Power to the people? Opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.
- Bondi, E.; Xu, L.; Acosta-Navas, D.; and Killian, J. A. 2021. Envisioning communities: a participatory approach towards AI for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 425–436.
- Buglione, L.; and Abran, A. 2013. Improving the user story agile technique using the invest criteria. In *2013 joint conference of the 23rd International Workshop on Software Measurement and the 8th International Conference on Software Process and Product Measurement*, 49–53. IEEE.
- Chew, R.; Bollenbacher, J.; Wenger, M.; Speer, J.; and Kim, A. 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*.
- Cirqueira, D.; Nedbal, D.; Helfert, M.; and Bezbradica, M. 2020. Scenario-based requirements elicitation for user-centric explainable AI: A case in fraud detection. In *International cross-domain conference for machine learning and knowledge extraction*, 321–341. Springer.
- Clegg, D.; and Barker, R. 1994. *Case Method Fast-Track: A RAD Approach*. Addison-Wesley.
- Cohn, M. 2004. *User stories applied: For agile software development*. Addison-Wesley Professional.
- De Paoli, S. 2024. Performing an inductive thematic analysis of semi-structured interviews with a large language model: An exploration and provocation on the limits of the approach. *Social Science Computer Review*, 42(4): 997–1019.
- Delgado, F.; Yang, S.; Madaio, M.; and Yang, Q. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–23.
- Erete, S.; Corbett, E.; Smith-Walker, N.; Cunningham, J. L.; Gatz, E.; Park, T.; Perry, T.; Wilcox, L.; and Denton, R. 2025. Towards Equitable Community-Industry Collaborations: Understanding the Experiences of Nonprofits’ Collaborations with Tech Companies. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–31.
- Franch, X.; Steffe, H.-J.; Bühne, S.; López, L.; and Sturm, S. 2024. Stories vs. user stories: a terminological clarification. In *International Conference on Agile Software Development*, 236–241. Springer.
- Hu, D.; Navas, D. A.; Gaube, S.; Mozannar, H.; Taylor, M. E.; Dvijotham, K.; and Bondi-Kelly, E. 2025. Human at the Center: A Framework for Human-Driven AI Development. *AI Magazine*, 46(4): e70043.
- Kim, E.; Balloli, V.; Karimian, B.; Bondi-Kelly, E.; and Fish, B. 2025. Relationship Building: Analyzing Recruitment Strategies for Participatory AI. In *AAAI/ACM Conference on AI, Ethics, and Society*.
- Kross, S.; and Guo, P. 2021. Orienting, framing, bridging, magic, and counseling: How data scientists navigate the outer loop of client collaborations in industry and academia. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–28.
- Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.
- Laplante, P. A.; and Kassab, M. 2022. *Requirements engineering for software and systems*. Auerbach Publications.
- Li, F.-L.; Horkoff, J.; Borgida, A.; Guizzardi, G.; Liu, L.; and Mylopoulos, J. 2015. From stakeholder requirements to formal specifications through refinement. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*, 164–180. Springer.
- Lin, H.; Karusala, N.; Okolo, C. T.; D’Ignazio, C.; and Gajos, K. Z. 2024. Come to us first: Centering Community Organizations in Artificial Intelligence for Social Good Partnerships. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2): 1–28.
- Liywalii, E. 2024. Requirements Engineering (RE) in Artificial Intelligence (AI) Systems Implementation: The need to emphasize non-functional requirements (nfrs) for ethical ai. In *Artificial Intelligence Research: 5th Southern African Conference, SACAIR 2024, Bloemfontein, South Africa, December 2–6, 2024, Proceedings*, volume 2326, 423. Springer Nature.
- Lucassen, G.; Dalpiaz, F.; van der Werf, J. M. E.; and Brinkkemper, S. 2016. Improving agile requirements: the quality user story framework and tool. *Requirements engineering*, 21(3): 383–403.
- Madaio, M.; Egede, L.; Subramonyam, H.; Wortman Vaughan, J.; and Wallach, H. 2022. Assessing the fairness of AI systems: AI practitioners’ processes, challenges, and needs for support. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1): 1–26.
- Majumdar, A.; Zhang, W.; Prawal, K.; and Yadav, A. 2025. The Hardness of Achieving Impact in AI for Social Impact Research: A Ground-Level View of Challenges & Opportunities. *arXiv preprint arXiv:2506.14829*.

- Margetis, G.; Ntoa, S.; Antona, M.; and Stephanidis, C. 2021. Human-centered design of artificial intelligence. *Handbook of human factors and ergonomics*, 1085–1106.
- Pacheco, C.; and Garcia, I. 2012. A systematic literature review of stakeholder identification methods in requirements elicitation. *Journal of Systems and Software*, 85(9): 2171–2181.
- Pasquale, L.; Ragone, A.; Piemontese, E.; and Darban, A. A. 2025. Exploring the Use of LLMs for Requirements Specification in an IT Consulting Company. In *Requirements Engineering Conference*, 389–399.
- Passi, S.; and Barocas, S. 2019. Problem formulation and fairness. In *Proceedings of the conference on fairness, accountability, and transparency*, 39–48.
- Prolific. 2024. What is Prolific and how does it work. <https://participant-help.prolific.com/en/article/dc132c>.
- Saxena, D.; Jung, J.-Y.; Forlizzi, J.; Holstein, K.; and Zimmerman, J. 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.
- Schön, E.-M.; Thomaschewski, J.; and Escalona, M. J. 2017. Agile Requirements Engineering: A systematic literature review. *Computer standards & interfaces*, 49: 79–91.
- Shi, Z. R.; Wang, C.; and Fang, F. 2020. Artificial intelligence for social good: A survey. *arXiv preprint arXiv:2001.01818*.
- Tai, R. H.; Bentley, L. R.; Xia, X.; Sitt, J. M.; Fankhauser, S. C.; Chicas-Mosier, A. M.; and Monteith, B. G. 2024. An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods*, 23: 16094069241231168.
- Van de Poel, I. 2013. Translating values into design requirements. In *Philosophy and engineering: Reflections on practice, principles and process*, 253–266. Springer.
- Van Der Hoven, J.; and Manders-Huits, N. 2020. Value-sensitive design. In *The ethics of information technologies*, 329–332. Routledge.
- Varshney, K. R.; and Mojsilovic, A. 2019. Open platforms for artificial intelligence for social good: common patterns as a pathway to true impact. *arXiv preprint arXiv:1905.11519*.
- Wong, R. Y.; Madaio, M. A.; and Merrill, N. 2023. Seeing like a toolkit: How toolkits envision the work of AI ethics. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–27.
- Zytka, D.; J. Wisniewski, P.; Guha, S.; PS Baumer, E.; and Lee, M. K. 2022. Participatory design of AI systems: opportunities and challenges across diverse users, relationships, and application domains. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 1–4.