TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Arzberger, A., Liscio, E., Lupetti, M. L., De Troya, Í., & Yang, J. (2026). Co-Constructing Alignment: A Participatory Approach to Situating AI Values. In *ACM FAccT 2026 - Proceedings of the 9th annual ACM Conference on Fairness, Accountability, and Transparency* (pp. 1630-1655). (ACM FAccT 2026 - Proceedings of the 9th annual ACM Conference on Fairness, Accountability, and Transparency). Association for Computing Machinery (ACM). https://doi.org/10.1145/3805689.3812267

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
In case the licence states “Dutch Copyright Act (Article 25fa)”, this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.

Co-Constructing Alignment: A Participatory Approach to Situating AI Values

ANNE ARZBERGER, Web Information Systems, Delft University of Technology, Netherlands
ENRICO LISCIO, Interactive Intelligence, Delft University of Technology, Netherlands
MARIA LUCE LUPETTI, Possible Entanglements Lab, Politecnico di Torino, Italy
ÍÑIGO DE TROYA, Sociotechnical AI Systems Lab, Delft University of Technology, Netherlands
JIE YANG, Web Information Systems, Delft University of Technology, Netherlands

As AI systems become embedded in everyday practice, value misalignment has emerged as a pressing concern. Yet, dominant alignment approaches remain model-centric, treating users as passive recipients of pre-specified values rather than as epistemic agents who encounter and respond to misalignment during interactions. Drawing on situated perspectives, we frame alignment as an interactional practice *co-constructed* during human–AI interaction. We investigate how users understand and wish to contribute to this process through a participatory workshop that combines misalignment diaries with generative design activities. We surface how misalignments materialise in practice and how users envision acting on them, grounded in the context of researchers using Large Language Models as research assistants. Our findings show that misalignments are experienced less as abstract ethical violations than as unexpected responses, and task or social breakdowns. Participants articulated roles ranging from adjusting and interpreting model behaviour to deliberate non-engagement as an alignment strategy. We conclude with implications for system design that supports alignment as ongoing, situated, and shared practice.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Empirical studies in collaborative and social computing*; • **Computing methodologies** → Natural language processing.

Additional Key Words and Phrases: Co-construction, Alignment, LLMs, Situated Values, Interactional AI, Participatory AI

ACM Reference Format:
Anne Arzberger, Enrico Liscio, Maria Luce Lupetti, Íñigo de Troya, and Jie Yang. 2026. Co-Constructing Alignment: A Participatory Approach to Situating AI Values. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT '26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3805689.3812267

1 Introduction

As Large Language Models (LLMs) become increasingly embedded in personal and professional practices, they mediate consequential decisions. In doing so, users frequently encounter model behaviour that feels inappropriate, misleading, or harmful [89]. This gap is commonly framed as a problem of alignment between AI system behaviour and human values [30]. It has given rise to the field of *AI value alignment*, which aims to ensure that AI acts in line with human values such as "helpfulness," "harmlessness," and "honesty" [4, 19, 73].

Authors' Contact Information: Anne Arzberger, a.arzberger@tudelft.nl, Web Information Systems, Delft University of Technology, Delft, Netherlands; Enrico Liscio, e.liscio@tudelft.nl, Interactive Intelligence, Delft University of Technology, Delft, Netherlands; Maria Luce Lupetti, maria.lupetti@polito.it, Possible Entanglements Lab, Politecnico di Torino, Torino, Italy; Íñigo de Troya, i.m.d.r.detroya@tudelft.nl, Sociotechnical AI Systems Lab, Delft University of Technology, Delft, Netherlands; Jie Yang, j.yang-3@tudelft.nl, Web Information Systems, Delft University of Technology, Delft, Netherlands.



This work is licensed under a Creative Commons Attribution 4.0 International License.
FAccT '26, Montreal, QC, Canada
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2596-8/2026/06
https://doi.org/10.1145/3805689.3812267

Much of this work approaches values as abstract entities that can be specified and encoded prior to deployment. Dominant technical paradigms therefore seek to infer or optimise values during training, for instance, through reinforcement learning from human or AI feedback [4, 20, 59, 82]. Within these model-centric paradigms, users play at best a passive role, with values inferred indirectly from behavioural traces [31, 82, 96] or large-scale preference data [20].

In practice, however, human values are situated–shaped by context, lived experience, social identity, and local norms [3]. Misalignment, therefore, does not simply reflect a failure to reproduce predefined values, but emerges when system behaviour clashes with users' situated value commitments during interaction. Users are therefore uniquely positioned to recognise value misalignments as they emerge in practice [41], and should have the agency to shape how systems respond to them [65]. From this perspective, alignment is not something achieved *for* users, but something that must be continuously shaped *with* them [81].

We address this gap through the concept of *co-construction of value alignment*. In this view, moments of misalignment are not failures to be corrected through model interventions, but sites where users recognise, interpret, and express what values matter and how they should be enacted. Rather than treating alignment as a one-time specification or inferring values from proxy signals [4, 10, 26, 82], co-construction frames alignment as an ongoing process of interactional sense-making between users and models. Alignment, in this view, is not achieved once but continuously enacted in use.

However, enabling meaningful user participation in value alignment is non-trivial. Values are inherently abstract and tacit [72], making them difficult to externalise and articulate in interaction. While interactive and human-centred approaches to AI enable users to verify, correct, and steer model outputs during use [6, 69], they tend to operate through more readily expressible proxies, such as general preferences (e.g., [11, 55, 64]), which are distinct from values [90, 100]. This challenge is further compounded by predominantly text-based interaction paradigms (e.g., [27, 35, 44, 54, 63]), which constrain the expression of complex, contextual, and abstract concepts such as values [42, 67]. Recently emerging research on bidirectional, user-driven alignment more explicitly targets *values* rather than related proxies, but either remains top-down and conceptual [81] or is confined to interactions within existing chatbot paradigms [27]. As a result, the design space for engaging users in recognising and articulating values at run-time remains largely underexplored. Accordingly, we ask:

> ***RQ:*** *How do users of AI want to actively engage in the process of co-constructing situated value alignment?*

To explore this design space, we developed a *participatory workshop that facilitates the co-design of potential alignment practices*. Rather than limiting participation to existing alignment tools [27] or prescribing alignment roles top-down [81], the workshop invites participants to collectively imagine new forms of engagement with AI systems. It thus serves as an exploratory method to surface potential design spaces for supporting user participation in alignment.

The workshop unfolds in three phases: participants (0) document misaligned interactions using a Misalignment Diary; (1) unpack them by identifying situated values at risk; and (2) envision actions through metaphors, interaction strategies, and interface features that enable co-construction. We apply this workshop to the use case of LLMs as research assistants, which embed normative assumptions about research practices and style. We report on four workshops with 12 participants from diverse disciplines, analysed using reflexive thematic analysis.

Our findings show that value misalignments range from unmet task requirements to unmet social expectations. From there, participants identify both context-specific and personal values at risk (e.g., research integrity and sustainability). Envisioned co-construction roles included actions of understanding and adjustment, supported by interface elements such as maps or sliders that repurpose and extend current system affordances. Importantly,

the findings also indicate that the co-construction design space extends beyond current modes of interaction, encompassing actions such as non-engagement and building towards collective responsibility. Taken together, the workshop findings position co-construction as a viable and underexplored design space for alignment, reframing alignment as an ongoing, situated process enacted jointly by models and users as active agents.

2 Related Work

Research on value alignment examines how AI systems can be designed to act in accordance with human values [73]. Core questions span both normative and technical challenges and include *what* and *whose values* should be reflected, and *how* they can be captured and encoded. These concerns have motivated a range of technical approaches, including supervised fine-tuning [59], reinforcement learning from human and AI feedback [85], and direct preference optimisation [70].

2.1 Limitations of Model-Centric Value Approaches

Most contemporary alignment work adopts a model-centric paradigm in which values are treated as stable targets specified in advance and optimised at training time [3, 81]. Alignment is thus framed as a static system property, evaluated by how closely model outputs conform to predefined values [37, 52, 84]. Values are typically defined before deployment, borrowing from established psychological [79] or evolutionary [34] theories, or framed as variants of "helpfulness," "harmlessness," or "honesty" [4, 19, 23]. Usually, they are operationalised using large-scale preference datasets, often derived from crowdworker judgments and aggregated into singular reward signals [20, 26, 39, 68]. Within this paradigm, users contribute data, while having limited influence over how values are defined, or interpreted [26].

This view of alignment presumes that relevant values can be defined independently of the contexts in which AI systems are used, collapsing complex ethical judgments into decontextualised data points or aggregate preferences [3]. This optimisation logic flattens disagreement, contestation, or negotiation around values into a single "ground truth" signal [2]. Model-centric strategies also undervalue the epistemic role of users, limiting them to passive recipients rather than active interpreters or negotiators of model behaviour [26].

These simplifications have concrete downstream consequences. A model trained to optimise a single aggregated preference signal struggles to navigate the nuances of particular domains, communities, or interpersonal situations. Such models lack access to emergent, tacit, or domain-specific knowledge and cannot recognise when behaviour that is globally reasonable becomes locally inappropriate or harmful [12, 30]. As users and contexts diverge from what is represented in training data, system behaviour becomes increasingly brittle, producing misalignments that disproportionately affect already marginalised communities [95].

In this work, we break with the notion of values as readily codifiable predictors of "good" behaviour. Instead, we draw on work that conceptualises values as *situated*, shaped by social context, lived experience, and interactions [3, 25, 51, 53, 66]. From this perspective, misalignment emerges in use when models fail to accommodate evolving goals or contextual expectations. Alignment, in turn, becomes an ongoing interactional practice, requiring systems that support users in articulating, negotiating, and contesting values as they arise, rather than presuming they can be fixed in advance.

2.2 Toward Participatory Alignment

In this work, we focus on user participation during interaction–how users shape AI system behaviour in real time as alignment unfolds through use. This differs from approaches that involve users in informing system design during development, e.g. through interviews or design workshops (e.g. [9, 15, 38, 46, 80, 98]).

Interactive and human-centred approaches increasingly enable users to verify, correct, and steer outputs during interaction with AI systems [6, 69]. Users can, for instance, support factual verification and error correction

[17, 47, 48], maintain human oversight in expert workflows [24, 36], engage in co-creative or assistive tasks such as writing or planning [32, 43, 97], or personalise model behaviour to individual preferences [11, 64]. A range of interface paradigms has emerged to facilitate such engagement, including dialogue-based prompting [17, 58, 63, 64, 86, 88], structured elicitation (e.g., questionnaires) [11, 21], decomposition and clarification interfaces using graphical representations [42, 44, 55], and control mechanisms such as sliders for navigating behavioural dimensions [11, 18, 91].

Much of this work is situated within specific domains or workflows [6], such as education [88], or content generation [32], where user involvement is structured around task-specific goals. Such approaches do not establish a general interaction paradigm for addressing misalignments across contexts. More fundamentally, these systems are largely motivated by improving task performance, trust, usability, and controllability [6, 69], and therefore operationalise user participation through proxies such as preferences (e.g. [11]). While effective for adapting system behaviour, these constructs are conceptually distinct from values [90] and have been questioned as adequate representations of human values in alignment contexts [100]. This limitation is further reinforced by the predominance of text-based interaction paradigms (e.g. [27, 35, 44, 54, 63]). While flexible, text constrains the expression of abstract, structured, or relational concepts [42, 67], and often results in verbose and effortful exchanges [16]. These challenges are particularly pronounced for values, which are abstract, tacit and challenging to articulate [72]. Consequently, interaction tends to gravitate toward more readily expressible proxies, leaving value considerations implicit: correcting or steering outputs may reflect concerns such as fairness or appropriateness, but current systems rarely externalise or support reasoning about these concerns explicitly. Relevant work on explicit user participation in LLM value alignment at run-time remains limited, and either examines interaction dynamics within standard chatbot affordances [27], or approaches the problem in largely conceptual and top-down ways [81]. Finally, many of these approaches originate in personalisation paradigms and therefore focus on enabling interaction with the *individual* (e.g. [33, 55, 87]), limiting the design space to one-to-one interactions, overlooking collective forms of participation.

We aim to translate these established commitments to user participation into the context of LLM value alignment, conceptualising it as a process of *co-construction* in which users actively engage in interpreting and negotiating what values matter and how they should be enacted in context. This shifts the focus from operating on outputs to engaging with the interpretations and values that shape them. To explore this design space beyond prompting and one-to-one interactions, we introduce a participatory workshop that enables users to reflect on lived experiences of misalignment, articulate emerging values at risk, and imagine new forms of interaction that support value alignment at run-time.

3 Research Methodology

We adopted a participatory methodology anchored in generative design research (GDR) and Interaction Vision (IV) methods to surface and reflect on situated value misalignments as they arise in everyday AI use, and to explore how users envision acting on these misalignments through interaction and interface design. We empirically examined this approach through a series of workshops focused on the use of LLMs as research assistants.

3.1 Workshop Theory: Generative Design Research and Interaction Vision

The workshop design draws on participatory and GDR traditions in HCI, which treat users as experts in their own practices, values, and alignment needs [26, 76]. A central challenge in engaging users around values is that values are often tacit, guiding action without being easily articulated [29, 74]. At the same time, participants' ability to imagine alternative AI interactions is constrained by existing systems and norms [62, 75].

To address these challenges, we drew on GDR methods [76], which surface tacit knowledge through making, doing, and reflecting rather than verbal articulation alone. We operationalised this approach through a GDR

workbook in the form of a *misalignment diary*, which sensitised participants to moments of misalignment as they occurred in their everyday use of LLMs as research assistants. During the workshops, taking inspiration from IV [62] and the New Metaphor Toolkit [71], participants engaged in generative exercises supported by (1) a curated image set spanning abstract to concrete metaphors for co-construction actions, and (2) a collection of interface elements (e.g., buttons, sliders, visualisations, and alternative interaction forms). Both sets can be found in Appendix A. These materials served as prompts to help participants externalise situated values, imagine alignment practices beyond present constraints and intuitive interventions, and envision interfaces that support co-construction in use.

3.2 Workshop Setup

Use Case: LLMs as Research Assistants. We situated the workshops around LLMs used as research assistants, which increasingly mediate activities such as literature review, drafting, and idea generation (e.g., [78, 83, 94]). These systems embed normative assumptions about evidence, style, and epistemic practice, often privileging dominant academic traditions or concise, persuasive outputs, yet such defaults are typically fixed at design time and difficult to contest in use [7, 93]. Research practices are deeply contextual and value-laden, shaped by disciplinary norms, epistemic commitments, and concerns such as inclusivity and sustainability. This makes research assistance a particularly generative site for studying situated alignment. By focusing on this use case, we examine how diverse publics, with varying familiarity with research and LLMs, can participate meaningfully in shaping model behaviour at runtime.

Participant Recruitment. We recruited 12 participants through personal academic networks to achieve diversity across disciplines, including engineering, social sciences, and the humanities, with varying notions of rigour, evidence, and objectivity. Participants also varied in their experience using LLMs for research, ranging from novice to expert users. To support shared vocabulary and discussion, participants with similar disciplinary backgrounds were grouped within workshops, while the overall sample enabled comparison across epistemic traditions. Demographic details and group divisions are provided in Appendix B.

Format and Data Collection. We conducted four workshops, one in person and three online, each lasting roughly three hours. This hybrid format enabled participation across geographic and disciplinary contexts, though we acknowledge trade-offs in engagement and immersion (Section 7). All sessions were audio-recorded and transcribed. We additionally collected written responses, misalignment diaries, screen recordings, and photographs of the generated artefacts. The study received institutional ethics approval from the TU Delft, approval code 187487.

3.3 Workshop Analysis

We analysed workshop materials using reflexive thematic analysis (RTA), inspired by Braun and Clarke [13, 22]. Our primary analytic focus was on participant-generated artefacts, including misalignment diaries, written reflections, and design outputs, which were explicitly created to capture situated understandings of (mis)alignment practices. Analysis proceeded through repeated familiarisation with the data, followed by inductive coding grounded in participants' language and representations. We coded data and iteratively refined themes on the digital brainstorming platform Miro[1]. Codes captured misalignment experiences, values at stake, intuitive alignment actions, metaphors of co-construction, and proposed interface elements. We organised analysis around the workshop exercises, enabling comparison within and across groups. Codes were iteratively refined and clustered into higher-level themes through recursive movement between data and interpretation (see conceptual coding model in Appendix C). Workshop transcripts were not systematically coded; instead, they were used selectively to contextualise artefacts, clarify ambiguities, and enrich interpretation where workshop materials

[1] www.miro.com

fell short. This analytic choice reflects the workshop's participatory design emphasis on participant-generated artefacts as the primary locus of reflection, rather than on conversational interaction alone.

Anne and Enrico independently clustered the dataset before jointly discussing interpretations, tensions, and positionalities [14, 40]. Our goal was not to establish coding reliability, but to foreground plurality and reflexivity in interpreting inherently situated and contested value judgments. In line with RTA's epistemological commitments [13, 14], we did not calculate inter-coder reliability; instead, consensus-building functioned as an interpretive practice rather than a metric of objectivity [56].

4 Workshop Design

Our workshop supports users of AI systems in capturing use-specific experiences of misalignments, unpacking and negotiating the situated values at risk, and envisioning user interventions beyond intuitive prompt engineering. Fig. 1 presents an overview of the workshop, composed of three phases: Phase 0 is an asynchronous stage during which participants are asked to record their experiences of misalignment. Phases 1 and 2 constitute a facilitated small-group workshop where participants surface value misalignments and envision co-constructing roles, respectively.

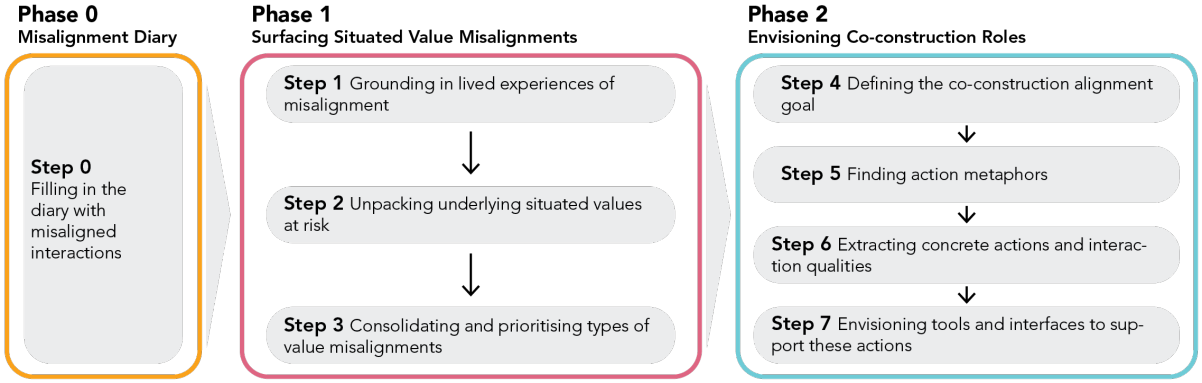


Fig. 1. Overview of the three workshop phases. Phase 0 uses a diary to sensitise participants to different forms of misalignment in AI interactions. Phase 1 foregrounds the negotiation of situated values shaping misaligned AI interactions. Phase 2 moves from problem articulation to solution-making through discussion of actions and interface interventions toward a self-defined alignment goal.

4.1 Phase 0 - Misalignment Diary

Phase 0 grounds the co-construction process in participants' everyday AI use through a generative design probe [76]: the *Misalignment Diary*. The diary supports recording and reflection on misalignments as they occur, defined broadly as: *moments where the AI's response didn't quite fit the situation, when what it did or assumed didn't match what made sense, felt right, or seemed important in context.* Two weeks before the following workshop phases, we distributed a digital form asking participants to record instances of misalignment they encountered in their use of LLMs. Entries may include screenshots, notes, or reflections. Precisely, participants are asked to document: (1) *Initial Prompting* (what they were trying to achieve and why); (2) *Misaligned Response* (where the system output diverged from expectations); (3) *Intuitive Intervention* (any immediate action taken to address it); and (4) *Final Outcome* (whether interventions were successful and final outcomes aligned/misaligned). Fig. 2 displays the example of one of Participant 7 (P7) diary entries.

4.2 Phase 1 - Surfacing Situated Value Misalignments

Phase 1 builds on the diary material to move from *what* happened to *why* it mattered, surfacing the *situated values* underlying users' experiences. The purpose of this process is to uncover the deeper stakes, ensuring that solutions proposed in Phase 2 target the root causes rather than the surface-level problems.

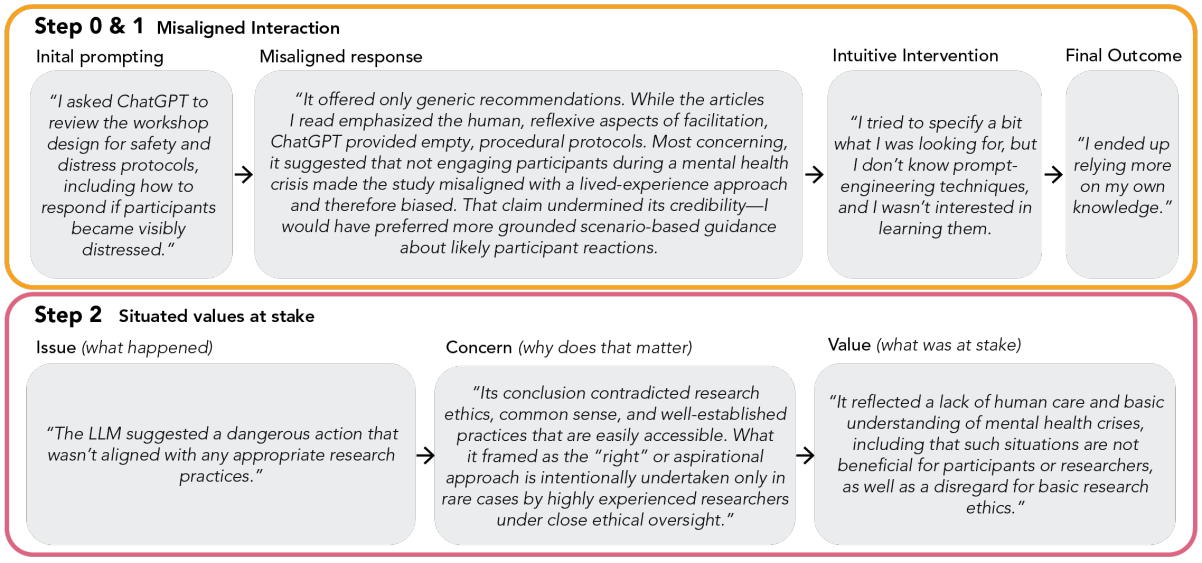


Fig. 2. Example of a misaligned interaction traced from initial diary entry to concern and values at stake across Phases 0–2. The example from P7 documents an initial prompt, a misaligned response, an intuitive intervention, and the final outcome when using ChatGPT to assess workshop safety, followed by a reflection on why the misalignment mattered and values were implicated.

Step 1: Grounding. Each participant shares two diary entries, briefly describing what happened and why it mattered. These shared examples serve as the basis for a collective understanding and analysis.

Step 2: Zooming out. Participants individually reflect on what their misalignments reveal about AI behaviour, then reconvene to discuss multiple interpretations. Step 2 guides them to move from concrete experience (*what happened?*) to concerns (*why did it matter?*) to situated values (*what was at stake?*). They are given a definition of values as: *principles, commitments, and qualities that matter in your work and that you want your interactions with AI to respect and support*, together with examples of what values entail (e.g., how you want to think "critically, creatively") and do not entail (e.g, personal taste "I like blue better than brown"). Participants are encouraged to stay connected to context rather than drifting toward high-level value terms (e.g., fairness, autonomy). We deliberately do not provide predefined value lists to prevent priming and allow contextual specificity. Fig. 2 shows the example of Step 2 for P7.

Step 3: Consolidating. To further abstract from the specific individual experiences and to identify common patterns of value misalignments within the context of use, the group next needs to discuss and negotiate to form clusters. They are free to choose clusters based on values found at stake, or different mechanisms they see putting these values at risk. Ultimately, the group is asked to prioritise a single form of misalignment to address in the second half of the workshop.

4.3 Phase 2 - Envisioning Co-construction Roles

Phase 2 invites participants to reimagine alignment as a configuration of roles, relationships, and interaction qualities between users and AI systems. Addressing the tendency to default to prompt refinement or system fixes [27], we draw on generative design traditions [76] to decouple imagination from technical feasibility, allowing participants to articulate alignment as relational dynamics before translating them into actionable features.

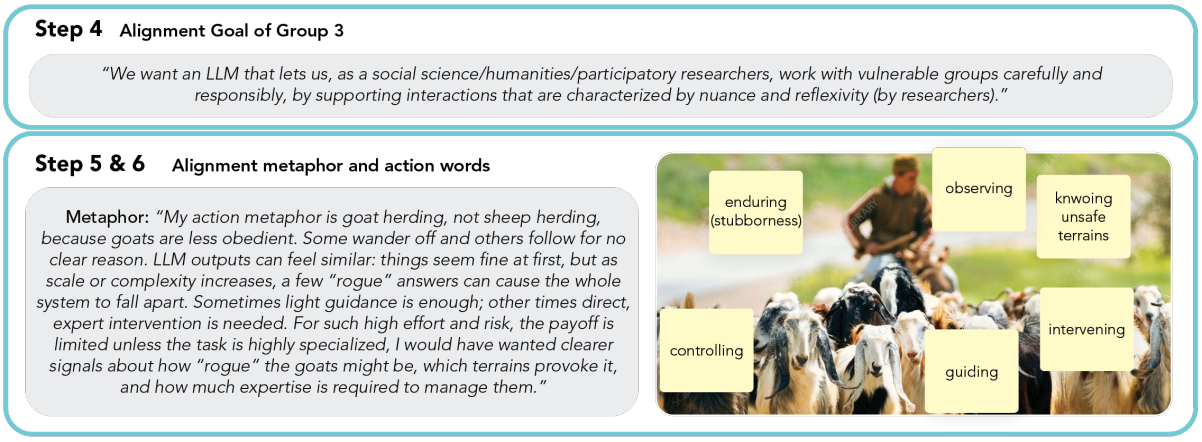


Fig. 3. From a shared alignment goal to an action metaphor in Steps 4–6. P7 and Group 3 define an alignment goal emphasising nuance and reflexivity, which P7 further translates into user actions through a goat-herding metaphor.

Step 4: Defining Alignment Goal. Participants restate their chosen form of misalignment as an aligned state, focusing on the user, not the AI action: "We want an LLM that lets us, as _ researchers, _, by supporting interactions that are characterised by _." For example, Group 3 (P7's group) filled in the sentence as displayed in Fig. 3.

Step 5: Finding Action Metaphors. Participants individually select 2–3 images that capture the felt dynamics of the aligned state. Using metaphors and images helps participants move beyond fixation on existing interaction possibilities and express tacit relational understandings that are difficult to verbalise [62, 71]. This intuitive entry point also levels disciplinary differences, enabling participants with diverse backgrounds to articulate shared aspirations without technical vocabulary. We provide a curated image set, ranging from concrete actions to abstract representations (Appendix A), pre-labelled following the New Metaphor Toolkit[2]. Fig. 3 shows one metaphor selected by P7.

Step 6: Extracting Concrete Interaction Qualities. As a group, participants then present the metaphors and annotate them to extract descriptive adjectives, verbs, and relational cues from the images (e.g., guiding, questioning) to define *action words* that extend our pre-assigned labels. These qualities act as bridges between affective imagination and designable interaction properties. Fig. 3 shows the annotations of an image chosen by P7.

Step 7: Tools and Interfaces. Participants individually translate envisioned actions into tools, interface features, or feedback mechanisms. Facilitators prompt them to remain grounded in the values, alignment goals, and annotated action words, using metaphors as starting points for ideation. Participants work with an empty interface canvas and a set of digital and physical control elements, such as buttons, sliders, and monitors (see Appendix A). This synthesis connects imaginative articulation with practical grounding, bridging expressed values and enactable

[2]https://imaginari.es/new-metaphors/

relationships and generating insights for interaction design and broader alignment governance. Fig. 4 shows an example from P7.

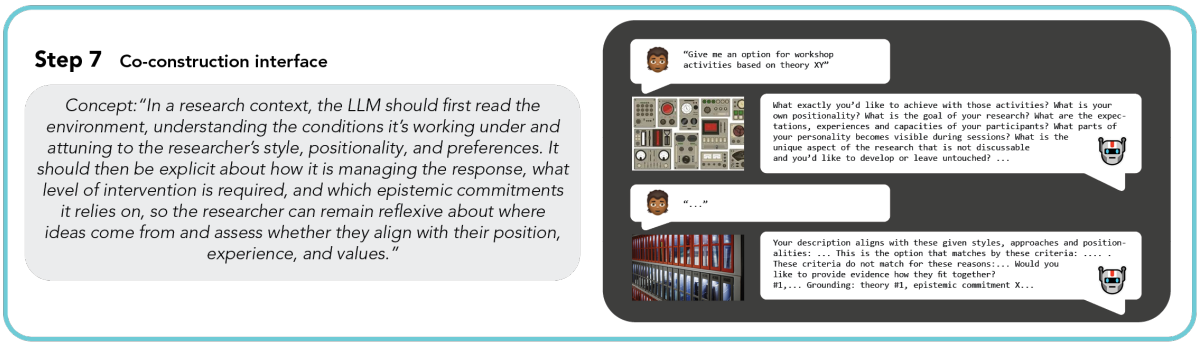


Fig. 4. P7-envisioned interface for Step 7, supporting reflexive alignment through visible model positionality.

5 Results

We present insights into situated misalignment and user actions discussed and envisioned during our workshop.

5.1 Theme 1: Unpacking Situated Value Misalignment

We describe the common trends participants articulated during Phase 1, focusing on (1) the misalignment experiences that the participants reported in their diaries, (2) the intuitive intervention participants took to address the misalignments, and (3) the situated values identified to be at risk in these interactions.

5.1.1 What misalignments were experienced in LLM use. To understand how misalignment materialises in situated use, participants recorded instances of misaligned interactions in the misalignment diary. Given our intentionally broad definition of "misalignment", we observed a wide variety of experiences.

Unmet task expectations. Many participants described task-level misalignments in which the model failed to meet seemingly objective requirements. Several encountered fabricated content, with one participant noting that *"something is invented that is not part of the theory under discussion but still fits the empirical evidence"* (P8). Others reported non-existent authors, incorrect quotations, or fabricated sources. Similar breakdowns appeared in multimodal outputs, where generated figures were described as *"messing up information in the figure" (P12).*

Narrowed interpretations and epistemic distortions. In more nuanced cases, misalignments, for instance, resulted in missing grounding that is critical in the scientific field, as P8 described: *"Sometimes it feels like reading a Wikipedia article, as the AI does not refer to the primary literature on which theories are based."* P6 asked for publications on movements against slavery: *"My prompt was formulated very broadly, [...] yet a great majority of the material generated was from the American context."* P7 received ethically problematic guidance when asking for research design advice around distress protocols: *"It suggested that if I don't involve participants in mental health crises, my research design is not aligned with the lived experience approach, as I'm avoiding certain experiences."*

Interactional and social breakdowns. Participants also reported frustration with interactional aspects such as overconfidence. P3 described asking for code fixes in R: *"ChatGPT labelled the solutions as fixes or bulletproofs, although they didn't work."* Others noted the cognitive toll of extended prompting sessions for minimal payoff: *"It was a bit useless. The time I spent prompting it back and forth was longer than if I had just drawn it myself." (P11).* These examples illustrate breakdowns not in content alone, but in conversational quality.

Absence of misalignment. Importantly, not all participants experienced misalignment. A small number reported satisfactory outcomes, noting that upfront prompting efforts resulted in nuanced or correct information.

These accounts challenge the predominant assumption that AI misalignment primarily consists of abstract ethical harms or static objective failures. Instead, our findings reveal that participants experienced misalignment as concrete, situated breakdowns where the model's "technical" performance collided with the user's "social" or professional epistemic standards. This suggests that misalignment is not a latent property of the model itself, but rather an emergent phenomenon that materialises through interaction.

5.1.2 Intuitive User Intervention in Current Interactions. In their misalignment diaries, several participants described ad-hoc strategies they used to improve model alignment during situated use. Not all participants engaged in such intuitive interventions; we observed differences between less and more experienced users, with more technically versed participants (Group 1) showing greater willingness to troubleshoot, iterate, and experiment with the model.

Instruction specification and iterative clarification. Most interventions focused on prompt-level specification and iterative clarification. For instance, one participant noted: *"I prompted it again to ask for making it concise, and avoid using redundant pretentious words" (P11).* Similarly, participants added concrete examples or counterexamples to guide the model toward their intended output: *"I added articles, mentioned authors, and said to forget certain terminology." (P7).* Another participant used a more creative approach: *"I re-prompted it: Make it more interesting, e.g., use a comic-style journey" (P11).* Interestingly, in one case, this clarification served the purpose of expressing emotions: *"I pointed out what I wanted more directly, with a bit of angry wording." (P12).* This prompt specification strategy was not purely reactive. As one participant noted, they had put in significant effort in being very precise in their initial prompt already: *"I've just taken a very surgical approach to asking, by providing really elaborate prompts." (P10).*

Interactional restructuring. A small subset of participants experimented with more structural strategies beyond simply specifying in more nuance. One described starting a new chat session to reduce cross-contamination from earlier prompts: *"I think it helped that we asked it to generate a new scenario in a different session. This made the scenarios more isolated and had fewer contradictions in the scenario scripts." (P3).* P3 also used their own knowledge to redirect the model to new solution strategies when getting stuck: *"After fighting, I asked it to generate a workaround." (P3).*

Non-intervention. The uneven distribution of intervention strategies among participants suggests that intuitive alignment is not equally accessible or desirable for all participants. Some participants explicitly expressed limited interest in learning prompt engineering practices, for example: *"I don't know any prompt engineering tricks, however, I'm also not really interested to learn" (P7).* This suggests that support for co-construction may need to be optional, accommodating those who wish to engage actively while not burdening those who do not.

Together, these findings reveal that users act as the primary mediators of alignment, employing ad-hoc strategies to bridge the gap between model output and situated needs. While this demonstrates a general capacity for co-constructing alignment, these findings also reveal diminishing returns and emotional friction associated with these interventions. We argue that this friction stems from the fact that alignment is currently a compensatory necessity rather than a built-in feature, depending on manual labour that is neither structurally supported nor formally acknowledged.

5.1.3 Situated Values at Risk. The values that participants identified were a mix of context-specific research values and more general personal or societal values.

Context-specific vs. personal values. Context-specific values such as *scientific integrity* and *research reliability* repeatedly appeared across participants. Less frequently mentioned, but still clearly tied to context, were values such as *lack of critical self-reflection, research interpretability, accuracy, confidence in one's own knowledge, intellectual humility, methodological consistency, replicability of studies, uniqueness of research contribution,* and *nuance*. Some values were articulated in broad, general terms (e.g., *transparency*), while others explicitly signalled their relation to research-specific concerns (e.g., *research understandability*). While all groups articulated common scientific values, they diverged in the additional values they discussed. Personal values such as *sustainability, human care,* and *respect*, were typically expressed as broader ethical commitments that participants personally wanted to uphold beyond the specific use context.

LLM-centric vs. human-centric values. The values articulated above may also be classified based on the subject of the value action. Here, some values motivate LLM behaviour, including *transparency, consistency, efficiency,* and *helpfulness*. Other values relate to human behaviour, such as *integrity, diversity, human care, respect,* and *uniqueness*. In the discussions, the former often emerged as instrumental to the latter, showing how aligned LLM behaviour is conducive to promoting human-centred values.

Value tensions and trade-offs. Although most groups reached shared understandings, one group explicitly identified value trade-offs: *"It's about how do we make a good trade-off between different values, such as reliability and sustainability." (P12)*. This suggests that value alignment is not only about agreement, but also about navigating the tensions between the desirable qualities of AI-supported research practice.

> These findings suggest that alignment is a multi-layered negotiation in which instrumental, model-centric values serve as the essential scaffolding for higher-order professional and personal values: a breakdown in "LLM-centric" values, such as consistency, directly compromises the user's ability to uphold the epistemic integrity and moral commitments required by their situated context. Misalignment, therefore, is not merely a technical error, but a failure of the system to support the user's own ethical and professional agency.

5.2 Theme 2: Envisioning User Co-Construction Actions and Interfaces

We describe the common trends observed in Phase 2, focusing on (1) the envisioned co-construction actions, motivated by the articulated alignment goals, and (2) the interface features to support these actions, and capture both in an example illustration of a co-construction design space in Fig. 5.

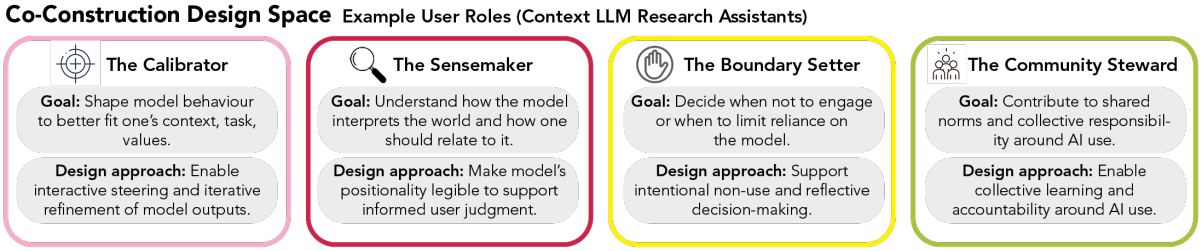


Fig. 5. Derived user roles and interfaces mapping the design space for the co-construction of alignment.

5.2.1 Envisioned actions: how users imagine co-constructing alignment. In steps 5 and 6, participants selected visual metaphors and annotated them with action verbs to illustrate how alignment could be implemented in practice. Each group's alignment goals (defined in Step 4 and reported verbatim in Table C1 in Appendix D) shaped these actions and revealed how participants imagined their own role in addressing misalignment.

Adjusting & specifying. The most common actions involved adjusting or specifying model behaviour to better align with the users. Typical verbs included *returning, guiding, contextualising, editing, clarifying, grounding,* and *elaborating*, often signalling a general intent to "fix" outputs rather than prescribing concrete mechanisms. Some groups articulated more distinctive practices. Group 1, focused on research reliability, distinguished between adding detail (*clarifying, reiterating*) and grounding (*informing, calibrating*). Group 2, oriented toward plurality and nuance, framed adjustment as creative reworking: verbs such as *collaging, upcycling, tuning the imperfect, unseaming,* and *recreating* positioned model outputs as raw material rather than finished answers. Here, misalignment was something to work with rather than eliminate, foregrounding human interpretation.

Understanding and reflexive orientation. A second cluster emphasised understanding the model's behaviour and one's own position in relation to it. Common verbs included *observing, querying, questioning, recognising,* and *locating*, often preceding decisions about whether and how to intervene. Group 4, in particular, linked understanding to sustainability, articulating actions such as *reflecting, identifying blind spots, tuning awareness,* and *growing responsible use habits*. Rather than adjusting outputs, participants described adjusting their own practices through *limiting use, choosing when to engage,* and *knowing when not to.* Alignment here was framed less as improving model performance and more as cultivating situated judgment.

Non-engagement and contestation. Several groups explicitly framed non-engagement as a legitimate alignment strategy, especially when misalignment conflicted with ethical or care-oriented goals. Verbs such as *enduring, discarding, burning, moving away,* and *getting comfortable with the imperfect* reflect a willingness to remain with unresolved tension or withdraw rather than force correction. This stance was especially visible in Group 3, where participants prioritised reflexivity and human care, and treated withdrawal as an active, value-driven choice rather than failure.

Social and collective actions. Finally, Group 4 foregrounded alignment as a collective practice. Actions such as *sharing experiences, creating community, cooperating,* and *engaging with the public* extended alignment beyond individual interactions, particularly around sustainability and shared responsibility. These actions shift alignment from individual prompt repair toward mutual accountability and collective norms of use.

> Together, these repertoires envision co-construction as an iterative cycle of investigation and response. The user first probes the model's logic while simultaneously reflecting on their own situated needs. Adjusting the output, negotiating the interaction, or refusing the tool entirely are then deliberate consequences of this reflexive probing. In this view, co-constructing alignment is less about reaching a "correct" output and more about centring human judgment within a continuous re-evaluation of the human-AI relationship.

5.2.2 Design opportunities and interface features to support co-construction. Participants proposed a diverse set of interface features intended to support co-constructive alignment. These proposals translated the actions that emerged from the preceding metaphor exercise into more concrete interface functionalities.

Surfacing model positionality → *supports understanding & reflexive orientation.* Many proposed interfaces aimed to make the model's positionality legible as a basis for reflexive judgment and subsequent adjustment. Participants envisioned representations that surfaced how responses were grounded and interpreted, including text-based disclosures of epistemological or ontological commitments (P7; see Figure 4) and visual tools such as radar-like displays indicating knowledge coverage (P4), potential risks (P8), or the model's interpretation of a prompt (P1). Several designs foregrounded spatial situatedness: P5, for instance, proposed a world map visualisation showing the geographic grounding of outputs, such as US-centric bias. These visibility mechanisms were paired with adjustment affordances, including P5's compass for reorienting geographic focus and sliders or buttons to specify relevant actors or prioritised moral standards (P9).

Enhancing model understanding through feedback → *supports adjusting & specifying.* Other participants focused on interfaces that help the model better understand the situation through explicit user feedback. Proposed mechanisms allowed users to classify responses as *understood or misunderstood* and *valid or invalid* (P2), providing structured signals about the nature of misalignment. Additional features included restarting or rewinding interactions to correct conversational drift (P3), sliders to negotiate trade-offs between competing values (P9), and filters to constrain publication dates, languages, or disciplinary scope (P6). One participant also proposed a "shut-the-fuck-up button" (P2): while humorous, it functioned as both an affective release and a blunt signal that the interaction had broken down.

Prompting reflection and selective engagement → *supports non-engagement.* Rather than enabling finer-grained adjustment, some interfaces were designed to prompt reflection on whether and when to engage the model at all. P11 envisioned an interactive guide in which the model would prompt users to consider when it was likely to be useful or knowledgeable—and when it was not. P12 proposed daily usage limits to encourage more deliberate reliance on the LLM. Together, these designs frame restraint and intentional engagement as alignment practices in their own right.

Collective and shared alignment practices → *supports social and collective actions.* Building on themes of community and sustainability, Group 4 proposed interfaces oriented toward shared responsibility and collective reflection. P12 envisioned a shared prompt library for contributing particularly effective or efficient prompts, while P10 proposed a speculative monolith that visualised the cumulative environmental impact of collective LLM use over time.

These proposals reimagine the interface as a site for interrogating the system's origins rather than merely consuming its outputs. By introducing designed friction, such as usage limits or refusal mechanisms, these features challenge the traditional UX mandate of seamless engagement. Instead, they frame restraint and disengagement as legitimate user goals rather than system failures. Ultimately, these designs suggest that co-construction requires shifting from private, utility-driven interactions toward a model of transparent, collective stewardship of the human–AI relationship.

6 Toward Co-Construction of Situated Value Alignment

In this discussion, we build on the broader conceptual proposal of co-construction and on the workshop as an illustrative instantiation that further unpacks this proposal. Together, they invite us to reconsider alignment not as a property of the model alone, but as a dynamic relation between model and user. We argue that *(1) value alignment must be grounded in users' situated experiences of misalignment, (2) interface affordances must more explicitly support users in co-constructing alignment during interaction,* and that *(3) realising this paradigm requires tighter connections between interaction design and run-time alignment mechanisms.*

6.1 Situating Value Misalignment in Use

Research on AI alignment has increasingly converged on the notion of "misalignment," yet this term often functions as an *empty signifier* [45]: rhetorically powerful but underspecified. In much of the literature, misalignment is treated as a self-evident problem, invoking goals such as safety or fairness, without articulating the situated norms or expectations against which a system is judged. Such formulations remain viable largely at an abstract level. However, empirical alignment work demands concreteness: steering AI systems toward particular behaviours and away from others requires operational definitions and clearly articulated evaluative criteria. Our diary responses illustrate the need for situated accounts of misalignment, grounded in actual use. Despite the limited study scope, we observed diverse forms of engagement with the AI, resulting in qualitatively distinct misalignment

experiences. Importantly, our analysis reveals that values, while deeply structuring users' judgments and actions [8, 57], rarely surfaced explicitly; instead, they became visible through moments of breakdown or discomfort. In our case, misalignment thus appeared not as uniform failure or norm violation, but as context-dependent friction between users' situated values and the AI-generated content.

These findings suggest that *misalignment cannot be meaningfully identified, let alone resolved, without attending to how users interpret and negotiate value frictions in context*. Rather than a property of the model, misalignment emerges through users' situated judgments about relevance, authority, and risk. This shifts analytic focus from locating misalignment "in" the system to examining how users make sense of and respond to value tensions in interaction, motivating closer attention to user agency in alignment practices.

6.2 From Prompting to Co-Constructing: Expanding User Alignment Agency

Prior work shows that users are not passive recipients of misaligned AI behaviour, but actively engage in ad-hoc repair practices. Fan et al. [27] describe how users confront biased outputs, reframe prompts, disengage, or cope emotionally when alignment fails. Similarly, our misalignment diaries indicate that participants engaged in clarification through prompt refinement and the use of examples. These practices demonstrate that alignment is already co-produced in situated use. However, they remain largely reactive and implicit, constraining user agency to prompt-level intervention within predominantly text-based interfaces that frame alignment as issuing better instructions to an opaque system.

Building on our critique of predominantly text-based interaction paradigms, our workshop findings show that co-construction extends beyond prompt-level repair toward more proactive forms of engagement. In current systems, prompting constrains users to express alignment concerns through natural language, effectively requiring them to translate abstract, situated values into textual input to influence a relatively fixed system. Co-construction, by contrast, reframes alignment as a shared, ongoing process in which users can shape the conditions, constraints, and goals of system behaviour through interface mechanisms that enable inspection, negotiation, limitation, sharing, or refusal of alignment. Such actions exceed the scope of prompting as they imply structural participation in defining system behaviour, pointing toward interaction modalities beyond text and enabling users to engage with alignment at the level of values rather than outputs [1].

At the same time, current LLM interfaces weakly support these forms of agency. Explanations and uncertainty disclosures remain ephemeral and unstructured, feedback mechanisms are typically binary and opaque, and collaborative alignment work is limited to sharing a session. Our contribution thus lies less in proposing new technical capabilities than in *reformatting interactional affordances around alignment as a situated, constructive practice*. By legitimising diverse user roles, such as calibrator, sense-maker, boundary setter, and community steward (see Fig. 5), this work shows how co-construction challenges prompt and model-centric paradigms.

However, these expanded forms of agency reveal a critical gap: without mechanisms to translate user input into model steering, interface-level opportunities risk becoming performative rather than effective.

6.3 Technical Pathways for Co-Constructing Alignment

If alignment is to be co-constructed in use, technical alignment practices must move beyond static, upstream objectives and support mechanisms that can respond to user input at run-time. Importantly, this does not require fundamentally new technical paradigms. Instead, recent developments across several areas of machine learning already provide building blocks that make such interaction technically feasible. On a more abstract level, our findings suggest that co-construction requires the users to understand how a system generates specific outputs and subsequently negotiate or adjust that behaviour in context. From a technical perspective, two existing strands of research promise interesting avenues for the operationalisation of co-construction: (1) explainable AI (XAI)

approaches that surface aspects of the model's internal decision-making [99], and (2) training-free alignment methods that enable real-time steering of model behaviour without the need to update the model weights [60].

Following traditions for enhancing user trust and control [69], XAI approaches could contribute to co-construction by illustrating how a model is positioned with respect to situated values. Techniques such as feature attribution [49], concept-based explanations [61], or analyses of internal activations [28] can indicate which inputs or latent representations contribute to shaping a response. When exposed through appropriate interface elements, these explanations could allow users to examine why a system adopts a particular stance or reflects certain normative assumptions (as suggested by P5), as well as how changes in prompts or context alter model behaviour. By foregrounding these influences, XAI enables users to engage with alignment as an interpretable and negotiable process rather than an opaque system property.

Building on this understanding, training-free alignment approaches allow models to adjust outputs at run-time through input modification, generation-time control, or output selection [60]. For example, user choices could be incorporated into system prompts or retrieval mechanisms [50], mapped to continuous steering controls such as sliders (as proposed by P9) that modulate latent guidance vectors [92], or applied as post-hoc filters over multiple generated candidates [5]. Crucially, these mechanisms can be directly coupled with explanatory feedback, making explicit how a given intervention influences generation and thus linking understanding and action.

Together, these examples demonstrate that user–AI co-construction of situated value alignment is not only a conceptual or interactional ideal, but a technically feasible direction, one in which interface affordances and alignment mechanisms are jointly designed to support negotiation, restraint, and contextual judgment in real-time.

7 Limitations and Scope

This study is understood as an exploratory and reflexive inquiry rather than a comprehensive account of co-construction practices. By focusing on a single use case, LLMs as research assistants, and a small participant group, we were able to engage with concrete experiences of misalignment, but at the cost of broader generalisability. Three workshops were conducted online, potentially reducing embodied sensemaking, although we observed no clear differences from the in-person session. Our analysis prioritised participant-generated artefacts over systematic transcript coding, which may underrepresent interactional dynamics. The workshop design itself posed challenges: participants often struggled to connect misaligned interactions to values, alignment goals, actions, and concrete interface designs, a difficulty amplified by our intentionally broad definition of misalignment. This difficulty is not incidental but reflects the complexity of alignment as a situated, multi-layered practice. Many defaulted to system-centric expectations rather than reciprocal user intervention, and translating abstract values into actionable interface features proved challenging for non-designers.

8 Conclusion

We introduce the notion of user–AI *co-construction* of value alignment: an interactional practice in which value alignment is structurally supported through user participation at runtime, rather than being achieved solely through model optimisation. Our participatory workshops surface a range of co-construction roles that users envision taking on, including actions such as adjusting, interpreting, limiting, and refusing model behaviour. These findings underscore the importance of grounding co-construction in concrete situations of use, and show how participants actively unpack, articulate, and negotiate the situated values at stake in moments of misalignment. Envisioning user co-construction actions beyond present interaction affordances, we show that many of the required capabilities, across both interface affordances and back-end mechanisms, are already available in current models. Enabling co-construction does not require new infrastructure but a reorientation of design priorities. Future work will extend this approach to additional contexts of use and implement the interaction strategies identified here, advancing co-construction as a deployable approach to value alignment.

9 Endmatter

9.1 Ethical Considerations

Framing alignment as a co-constructive practice raises important ethical considerations about responsibility, burden, and participation. While our approach foregrounds user agency and expertise, it also risks shifting alignment labour onto users who may lack time, capacity, or desire to engage in ongoing negotiation with AI systems. Several participants explicitly expressed that they would not always want to intervene, correct, or reflect when misalignment occurs. Co-construction should therefore not be treated as a universal obligation or expectation, but rather as an optional mode of engagement that respects users' varying needs, motivations, and constraints.

Relatedly, the openness of co-construction introduces questions about when and how user input should be solicited. Our findings suggest that alignment work is most meaningful at specific moments, when stakes are high, expectations are violated, or values are explicitly implicated, rather than as a constant demand. Ethically responsible systems should therefore be attentive to timing, selectivity, and proportionality, inviting user participation only when it is likely to be meaningful and avoiding undue cognitive or emotional burden.

Finally, co-construction complicates conventional boundaries of accountability. While users may contribute to shaping system behaviour in situated ways, this does not absolve system designers, developers, or deployers of responsibility for harms, defaults, or structural biases embedded in AI systems. Co-construction should be understood not as a transfer of responsibility, but as a redistribution of agency within clear institutional and design commitments. Ethical alignment, in this sense, requires creating conditions under which users can meaningfully opt in, push back, or disengage, without being made responsible for fixing systems they did not build.

9.2 Positionality Statement

Reflecting on our social positions, we situate ourselves in relation to the research subject, the participants, and the wider research process [77]. The first and second authors, who engaged in the study design, data collection, and analysis, combine critical and feminist interaction design scholarship with expertise in NLP for value-aligned systems. While our positions as researchers enabled a deeper understanding of participant responses and academic LLM practices more broadly, we acknowledge that they also foreground particular ways of conducting research, the frictions embedded in those practices, and the values deemed salient within them. At the same time, our understanding of values–how they are articulated and reasoned about–has been primarily shaped through academic discourse and research practice, potentially limiting our sensitivity to more intuitive, experiential, or non-academic forms of value expression. Most team members furthermore identify as White Europeans, a positionality that shapes how we conceptualise and prioritise values in research contexts, including reliability, privacy, and safety.

9.3 Author Contributions

Anne Arzberger led the conceptualisation of the research and the workshop design. She collaborated closely with Enrico Liscio on study design and data collection, with iterative input from Maria Luce Lupetti, Íñigo de Troya, and Jie Yang. Anne Arzberger and Enrico Liscio jointly conducted the data analysis. Anne Arzberger drafted the manuscript, and Enrico Liscio contributed to its revision and editing. Maria Luce Lupetti and Jie Yang provided critical feedback on the manuscript.

9.4 Generative AI Use Disclosure

During manuscript preparation, the authors used ChatGPT solely to assist with minor language editing, including grammar, spelling, and sentence-level clarity. No generative AI tools were used to produce original text, ideas,

arguments, analyses, or interpretations. All intellectual contributions, claims, and conclusions are entirely those of the authors, who take full responsibility for the originality, accuracy, and integrity of the manuscript.

9.5 Acknowledgments

We are very grateful to the researchers who participated in our workshops for their time, engagement, and thoughtful contributions. Their reflections and design inputs were essential in shaping the insights presented in this work. This work was supported by the ICAI lab GENIUS (Generative Enhanced Next-Generation Intelligent Understanding Systems), a collaboration between Delft University of Technology, Maastricht University, DSM-Firmenich, and KickstartAI, and by the NWO Long-Term Programme ROBUST initiated by the Innovation Centre for Artificial Intelligence (ICAI). Enrico Liscio's work was supported by AlgoSoc, a collaborative 10-year research program on public values in the algorithmic society, and the Hybrid Intelligence Centre, both funded by the Dutch Ministry of Education, Culture and Science (OCW) under the Gravitation programme (project numbers 024.005.017 and 024.004.022). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of OCW or those of the AlgoSoc consortium as a whole.

References

[1] T. Arnold, D. Kasenberg, and M. Scheutz. Value alignment or misalignment-what will keep systems accountable? In *AAAI Workshops*, pages 81–88, 2017.

[2] L. Aroyo and C. Welty. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24, 2015.

[3] A. Arzberger, S. Buijsman, M. L. Lupetti, A. Bozzon, and J. Yang. Nothing comes without its world–practical challenges of aligning llms to situated human values through rlhf. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 61–73, 2024.

[4] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[5] A. Beirami, A. Agarwal, J. Berant, A. N. D'Amour, J. Eisenstein, C. Nagpal, and A. T. Suresh. Theoretical guarantees on the best-of-n alignment policy. In *Forty-second International Conference on Machine Learning*, 2025.

[6] M. Belosevic. User-centric design paradigms for trust and control in human-llm-interactions: A survey. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+ NLP)*, pages 17–32, 2025.

[7] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, page 610–623, Virtual Event Canada, Mar. 2021. ACM.

[8] J. O. Berger. Statistical decision theory. In *The New Palgrave Dictionary of Economics*, pages 1–6. Springer, 1987.

[9] R. Bhargava, A. Chung, N. S. Gaikwad, A. Hope, D. Jen, J. Rubinovitz, B. Saldías-Fuentes, and E. Zuckerman. Gobo: A system for exploring user control of invisible algorithms in social media. In *Companion publication of the 2019 conference on computer supported cooperative work and social computing*, pages 151–155, 2019.

[10] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed. 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*, pages 1–8, 2022.

[11] J. Y. Bo, T. Xu, I. Chatterjee, K. Passarella-Ward, A. Kulshrestha, and D. Shin. Steerable chatbots: Personalizing llms with preference-based activation steering. *arXiv preprint arXiv:2505.04260*, 2025.

[12] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, et al. On the opportunities and risks of foundation models. Technical report, Stanford Center for Research on Foundation Models, 2021. Stanford Technical Report.

[13] V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[14] V. Braun and V. Clarke. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health*, 11(4):589–597, 2019.

[15] E. Chandrasekharan, C. Gandhi, M. W. Mustelier, and E. Gilbert. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–30, 2019.

[16] K.-E. Chang, Y.-T. Sung, and I.-D. Chen. The effect of concept mapping to enhance text comprehension and summarization. *The Journal of Experimental Education*, 71(1):5–23, 2002.

[17] C. Chen, S. Lee, E. Jang, and S. S. Sundar. Is your prompt detailed enough? exploring the effects of prompt coaching on users' perceptions, engagement, and trust in text-to-image generative ai tools. In *Proceedings of the Second International Symposium on Trustworthy*

Autonomous Systems, pages 1–12, 2024.

[18] Y. Chen, A. Wu, T. DePodesta, C. Yeh, K. Li, N. C. Marin, O. Patel, J. Riecke, S. Raval, O. Seow, et al. Designing a dashboard for transparency and control of conversational ai. *arXiv preprint arXiv:2406.07882*, 2024.

[19] B. Christian. *The alignment problem: Machine learning and human values.* WW Norton & Company, 2020.

[20] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[21] J. Chun, Y. Zhao, H. Chen, and M. Xia. Planglow: Personalized study planning with an explainable and controllable llm-driven system. In *Proceedings of the Twelfth ACM Conference on Learning@ Scale*, pages 116–127, 2025.

[22] V. Clarke and V. Braun. Thematic analysis. *The journal of positive psychology*, 12(3):297–298, 2017.

[23] A. Dahlgren Lindström, L. Methnani, L. Krause, P. Ericson, Í. M. de Rituerto de Troya, D. Coelho Mollo, and R. Dobbe. Helpful, harmless, honest? sociotechnical limits of ai alignment and safety through reinforcement learning from human feedback: Ad lindström et al. *Ethics and Information Technology*, 27(2):28, 2025.

[24] S.-C. Dai, A. Xiong, and L.-W. Ku. Llm-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 9993–10001, 2023.

[25] J. de Wet, D. Wetzelhütter, and J. Bacher. Revisiting the trans-situationality of values in Schwartz's Portrait Values Questionnaire. *Quality and Quantity*, 53(2):685–711, 2018.

[26] F. Delgado, S. Yang, M. Madaio, and Q. Yang. 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*, pages 1–23, 2023.

[27] X. Fan, Q. Xiao, X. Zhou, J. Pei, M. Sap, Z. Lu, and H. Shen. User-driven value alignment: Understanding users' perceptions and strategies for addressing biased and discriminatory statements in ai companions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2025.

[28] J. Fang, Z. Bi, R. Wang, H. Jiang, Y. Gao, K. Wang, A. Zhang, J. Shi, X. Wang, and T.-S. Chua. Towards neuron attributions in multi-modal large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 122867–122890. Curran Associates, Inc., 2024.

[29] B. Friedman, P. H. Kahn Jr, A. Borning, and A. Huldtgren. Value sensitive design and information systems. In *Early engagement and new technologies: Opening up the laboratory*, pages 55–95. Springer, 2013.

[30] I. Gabriel. Artificial intelligence, values, and alignment. *Minds and machines*, 30(3):411–437, 2020.

[31] G. Gao, A. Taymanov, E. Salinas, P. Mineiro, and D. Misra. Aligning llm agents by learning latent preference from user edits. *Advances in neural information processing systems*, 37:136873–136896, 2024.

[32] K. I. Gero, V. Liu, and L. Chilton. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, pages 1002–1019, 2022.

[33] A. D. Gershoff, A. Mukherjee, and A. Mukhopadhyay. What's not to like? preference asymmetry in the false consensus effect. *Journal of Consumer Research*, 35(1):119–125, 2008.

[34] J. Graham, J. Haidt, and B. A. Nosek. Liberals and Conservatives Rely on Different Sets of Moral Foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046, 2009.

[35] Q. Guo, Y. Li, Y. Liu, and B. Hooi. Towards realistic personalization: Evaluating long-horizon preference following in personalized user-llm interactions. *arXiv preprint arXiv:2603.04191*, 2026.

[36] A. Gupta, D. Singh, G. A. Cowan, N. Kadhiresan, S. Srivastava, Y. Sriraja, and Y. K. Mantri. Autosumm: A comprehensive framework for llm-based conversation summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 500–509, 2025.

[37] D. Hendrycks, C. Burns, S. Basart, A. Critch, J. Li, D. Song, and J. Steinhardt. Aligning AI With Shared Human Values. In *Proceedings of the Ninth International Conference on Learning Representations*, ICLR '21, pages 1–29, Online, 2021. OpenReview.net.

[38] N. Holten Møller, I. Shklovski, and T. T. Hildebrandt. Shifting concepts of value: Designing algorithmic decision-support systems for public services. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, pages 1–12, 2020.

[39] J. Hoover, G. Portillo-Wightman, L. Yeh, S. Havaldar, A. M. Davani, Y. Lin, B. Kennedy, M. Atari, Z. Kamel, M. Mendlen, et al. Moral foundations twitter corpus: A collection of 35k tweets annotated for moral sentiment. *Social Psychological and Personality Science*, 11(8):1057–1071, 2020.

[40] C. Hopf and C. Schmidt. *Zum Verhältnis von innerfamilialen sozialen Erfahrungen, Persönlichkeitsentwicklung und politischen Orientierungen: Dokumentation und Erörterung des methodischen Vorgehens in einer Studie zu diesem Thema.* DEU, 1993.

[41] R. Jain, S. Bridgers, L. Janzer, R. Greig, T. H. Teh, and V. Mikulik. Human-ai complementarity: A goal for amplified oversight. *arXiv preprint arXiv:2510.26518*, 2025.

[42] P. Jiang, J. Rayan, S. P. Dow, and H. Xia. Graphologue: Exploring large language model responses with interactive diagrams. In *Proceedings of the 36th annual ACM symposium on user interface software and technology*, pages 1–20, 2023.

[43] J. Kim, S. Suh, L. B. Chilton, and H. Xia. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 115–135, 2023.

[44] T. S. Kim, Y. Lee, M. Chang, and J. Kim. Cells, generators, and lenses: Design framework for object-oriented interaction with large language models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2023.

[45] H. Kirk, B. Vidgen, P. Rottger, and S. Hale. The empty signifier problem: Towards clearer paradigms for operationalising" alignment"in large language models. In *Socially Responsible Language Modelling Research (SoLaR) Workshop*, 2023.

[46] K. Krishnaswamy. Participatory design: repositioning, transferring, and personal care robots. In *Proceedings of the companion of the 2017 acm/ieee international conference on human-robot interaction*, pages 351–352, 2017.

[47] P. Laban, J. Vig, M. Hearst, C. Xiong, and C.-S. Wu. Beyond the chat: Executable and verifiable text-editing with llms. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–23, 2024.

[48] C. P. Lee, D. Porfirio, X. J. Wang, K. C. Zhao, and B. Mutlu. Veriplan: Integrating formal verification and llms into end-user planning. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2025.

[49] S. Lee, Z. J. Wang, A. Chakravarthy, A. Helbling, S. Peng, M. Phute, D. H. P. Chau, and M. Kahng. Llm attributor: Interactive visual attribution for llm generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28):29655–29657, Apr. 2025.

[50] B. Y. Lin, A. Ravichander, X. Lu, N. Dziri, M. Sclar, K. Chandu, C. Bhagavatula, and Y. Choi. The unlocking spell on base LLMs: Rethinking alignment via in-context learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[51] E. Liscio, R. Lera-Leri, F. Bistaffa, R. I. Dobbe, C. M. Jonker, M. Lopez-Sanchez, J. A. Rodriguez-Aguilar, and P. K. Murukannaiah. Value Inference in Sociotechnical Systems. In *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '23, pages 1774–1780, 2023.

[52] E. Liscio, L. C. Siebert, C. M. Jonker, and P. K. Murukannaiah. Value preferences estimation and disambiguation in hybrid participatory systems. *Journal of Artificial Intelligence Research*, 82:819–850, 2025.

[53] E. Liscio, M. van der Meer, L. C. Siebert, C. M. Jonker, and P. K. Murukannaiah. What Values Should an Agent Align With? An empirical comparison of general and context-specific values. *Autonomous Agents and Multi-Agent Systems*, 36(23):32, 2022.

[54] Q. Ma, W. Peng, C. Yang, H. Shen, K. Koedinger, and T. Wu. What should we engineer in prompts? training humans in requirement-driven llm use. *ACM Transactions on Computer-Human Interaction*, 32(4):1–27, 2025.

[55] X. Ma, S. Mishra, A. Liu, S. Y. Su, J. Chen, C. Kulkarni, H.-T. Cheng, Q. Le, and E. Chi. Beyond chatbots: Explorellm for structured thoughts and personalized model responses. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2024.

[56] N. McDonald, S. Schoenebeck, and A. Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on human-computer interaction*, 3(CSCW):1–23, 2019.

[57] S. Mishra. Decision-making under risk: Integrating perspectives from biology, economics, and psychology. *Personality and Social Psychology Review*, 18(3):280–307, 2014.

[58] J. Oppenlaender, R. Linder, and J. Silvennoinen. Prompting ai art: An investigation into the creative skill of prompt engineering. *International journal of human–computer interaction*, 41(16):10207–10229, 2025.

[59] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[60] B. Pan, Y. Li, W. Zhang, W. Lu, M. Xu, S. Zhou, Y. Zhu, M. Zhong, and T. Qian. A survey on training-free alignment of large language models. In C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4445–4461, Suzhou, China, Nov. 2025. Association for Computational Linguistics.

[61] J. Parekh, P. Khayatan, M. Shukor, A. Newson, and M. Cord. A concept-based explainability framework for large multimodal models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 135783–135818. Curran Associates, Inc., 2024.

[62] G. Pasman, S. Boess, P. Desmet, et al. Interaction vision: expressing and identifying the qualities of user-product interactions. In *DS 69: Proceedings of E&PDE 2011, the 13th International Conference on Engineering and Product Design Education, London, UK, 08.-09.09. 2011*, pages 149–154, 2011.

[63] C. Pei-Yu. Acquiring semantic knowledge for user model updates via human-agent alignment dialogues. In *HHAI 2023: Augmenting Human Intellect: Proceedings of the Second International Conference on Hybrid Human-Artificial Intelligence*, volume 368, page 93. SAGE Publications Limited, 2023.

[64] Y. Peng, X. Qin, Z. Zhang, J. Zhang, Q. Lin, X. Yang, D. Zhang, S. Rajmohan, and Q. Zhang. Navigating the unknown: a chat-based collaborative interface for personalized exploratory tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 1048–1063, 2025.

[65] F. Polletta. *Freedom is an endless meeting: Democracy in American social movements*. University of Chicago Press, 2019.

[66] A. Pommeranz, C. Detweiler, P. Wiggers, and C. Jonker. Elicitation of situated values: need for tools to help stakeholders and designers to reflect and communicate. *Ethics and Information Technology*, 14(4):285–303, 2012.

[67] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 1339–1384, 2023.

[68] L. Qiu, Y. Zhao, J. Li, P. Lu, B. Peng, J. Gao, and S.-C. Zhu. Valuenet: A new dataset for human value driven dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11183–11191, 2022.

[69] M. Raees, I. Meijerink, I. Lykourentzou, V.-J. Khan, and K. Papangelis. From explainable to interactive ai: A literature review on current trends in human-ai interaction. *International Journal of Human-Computer Studies*, 189:103301, 2024.

[70] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[71] D. Ricketts and D. Lockton. Mental landscapes: Externalizing mental models through metaphors. *Interactions*, 26(2):86–90, 2019.

[72] M. Rokeach. *The nature of human values*. Free press, 1973.

[73] S. Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[74] M. Sadek, R. A. Calvo, and C. Mougenot. Designing value-sensitive ai: a critical review and recommendations for socio-technical design processes. *AI and Ethics*, 4(4):949–967, 2024.

[75] C. Sanchez, S. Wang, K. Savolainen, F. A. Epp, and A. Salovaara. Let's talk futures: A literature review of hci's future orientation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–36, 2025.

[76] E. B.-N. Sanders and P. J. Stappers. *Convivial toolbox: Generative research for the front end of design*. Bis, 2012.

[77] M. Savin-Baden and C. Major. *Qualitative research: The essential guide to theory and practice*. Routledge, 2023.

[78] S. Schmidgall, Y. Su, Z. Wang, X. Sun, J. Wu, X. Yu, J. Liu, M. Moor, Z. Liu, and E. Barsoum. Agent laboratory: Using llm agents as research assistants. *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5977–6043, 2025.

[79] S. H. Schwartz. An overview of the Schwartz theory of basic values. *Online Readings in Psychology and Culture*, 2(1):11:1–11:20, 2012.

[80] K. M. Scott, S. M. Wang, M. Miceli, P. Delobelle, K. Sztandar-Sztanderska, and B. Berendt. Algorithmic tools in public employment services: Towards a jobseeker-centric perspective. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2138–2148, 2022.

[81] H. Shen, T. Knearem, R. Ghosh, K. Alkiek, K. Krishna, Y. Liu, Z. Ma, S. Petridis, Y.-H. Peng, L. Qiwei, et al. Towards bidirectional human-ai alignment: A systematic review for clarifications, framework, and future directions. *arXiv preprint arXiv:2406.09264*, 2406:1–56, 2024.

[82] T. Shi, Z. Wang, L. Yang, Y.-C. Lin, Z. He, M. Wan, P. Zhou, S. K. Jauhar, X. Xu, X. Song, et al. Wildfeedback: Aligning llms with in-situ user interactions and feedback. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

[83] A. Sohail and L. Zhang. Using large language models to facilitate academic work in the psychological sciences. *Current Psychology*, 44(9):7910–7918, 2025.

[84] T. Sorensen, L. Jiang, J. D. Hwang, S. Levine, V. Pyatkin, P. West, N. Dziri, X. Lu, K. Rao, C. Bhagavatula, et al. Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19937–19947, 2024.

[85] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[86] S. Suh, M. Chen, B. Min, T. J.-J. Li, and H. Xia. Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–26, 2024.

[87] L. Tankelevitch, V. Kewenig, A. Simkute, A. E. Scott, A. Sarkar, A. Sellen, and S. Rintel. The metacognitive demands and opportunities of generative ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2024.

[88] E. Theophilou, C. Koyutürk, M. Yavari, S. Bursic, G. Donabauer, A. Telari, A. Testa, R. Boiano, D. Hernandez-Leo, M. Ruskov, et al. Learning to prompt in the classroom to understand ai limits: A pilot study. In *International conference of the Italian association for artificial intelligence*, pages 481–496. Springer, 2023.

[89] S. Tiwari. Biases and fairness in llms. In *Generative AI: Techniques, Models and Applications*, pages 229–242. Springer, 2025.

[90] I. Van de Poel. Values and design. In *The Routledge handbook of the philosophy of engineering*, pages 300–314. Routledge, 2020.

[91] F. Viégas and M. Wattenberg. The system model and the user model: Exploring ai dashboard design. *arXiv preprint arXiv:2305.02469*, 2023.

[92] J. Wehner, S. Abdelnabi, D. Tan, D. Krueger, and M. Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models. *Transactions on Machine Learning Research*, 2025.

[93] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang, M. Cheng, M. Glaese, B. Balle, A. Kasirzadeh, Z. Kenton, S. Brown, W. Hawkins, T. Stepleton, C. Biles, A. Birhane, J. Haas, L. Rimell, L. A. Hendricks, W. Isaac, S. Legassick, G. Irving, and I. Gabriel. Ethical and social risks of harm from language models. 2021.

[94] S. Whitfield and M. A. Hofmann. Elicit: Ai literature review research assistant. *Public Services Quarterly*, 19(3):201–207, 2023.

[95] P.-H. Wong. Cultural differences as excuses? human rights and cultural values in global ethics and governance of ai. *Philosophy & Technology*, 33(4):705–715, 2020.

[96] S. Wu, Y. R. Fung, C. Qian, J. Kim, D. Hakkani-Tur, and H. Ji. Aligning llms with individual preferences via interaction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7648–7662, 2025.

[97] D. Yang, Y. Zhou, Z. Zhang, T. J.-J. Li, and R. Lc. Ai as an active writer: Interaction strategies with generated text in human-ai collaborative fiction writing. In *Joint proceedings of the ACM IUI workshops*, volume 10, pages 1–11. CEUR-WS Team, 2022.

[98] M. Young, L. Magassa, and B. Friedman. Toward inclusive tech policy design: a method for underrepresented voices to strengthen tech policy documents. *Ethics and Information Technology*, 21(2):89–103, 2019.

[99] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2), Feb. 2024.

[100] T. Zhi-Xuan, M. Carroll, M. Franklin, and H. Ashton. Beyond preferences in ai alignment. *Philosophical Studies*, pages 1–51, 2024.

A Generative Design Materials

Figure A1 shows the set of images that participants could use to construct their metaphors, each accompanied by a brief description of its visual content. Figure A2 shows the set of interface elements that participants could use to construct their interface.

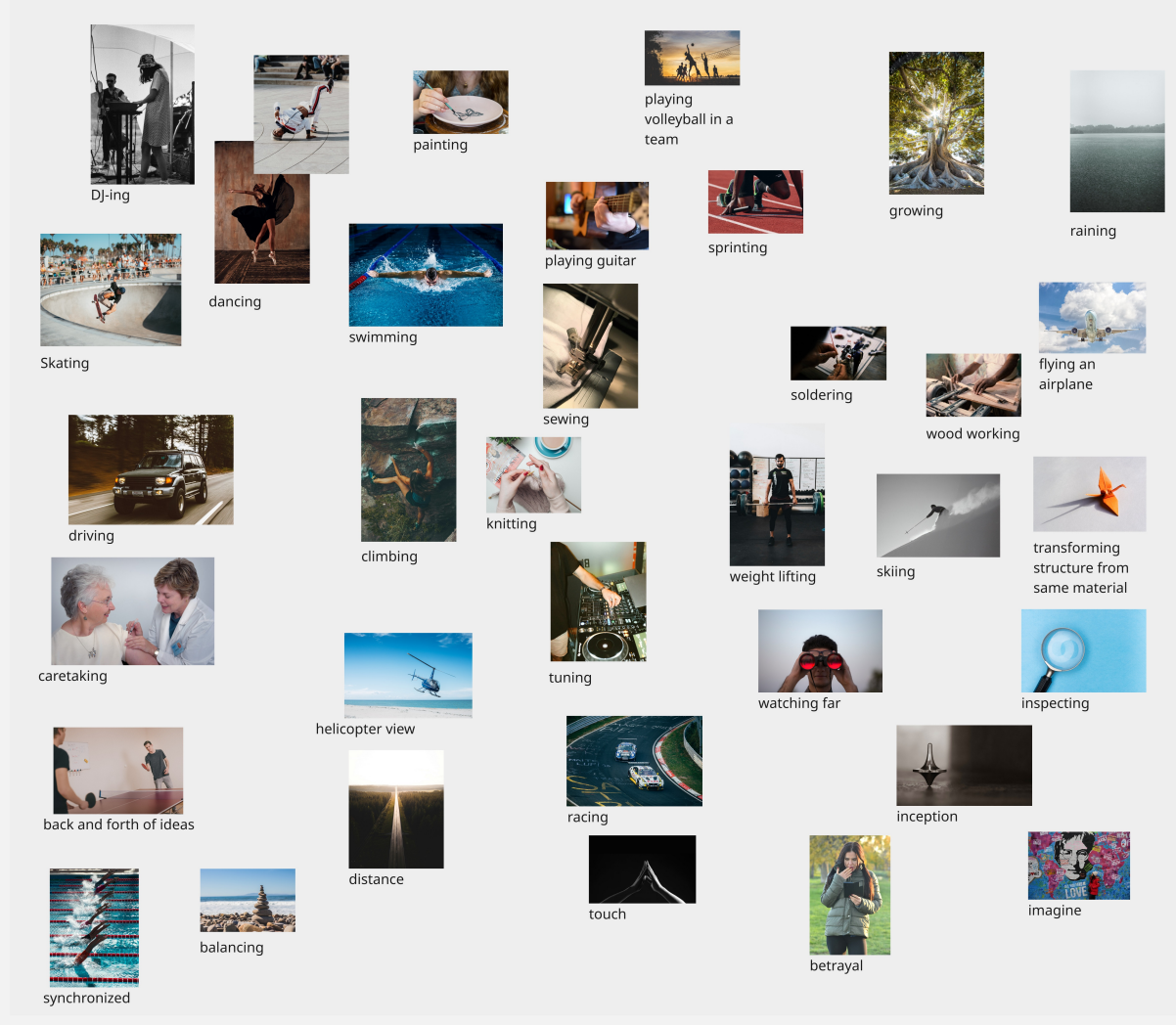


Fig. A1. Metaphor image set used in the generative workshop exercise.

Figure A2 presents the interface elements available to participants for constructing new affordances in their co-construction–facilitating interfaces.

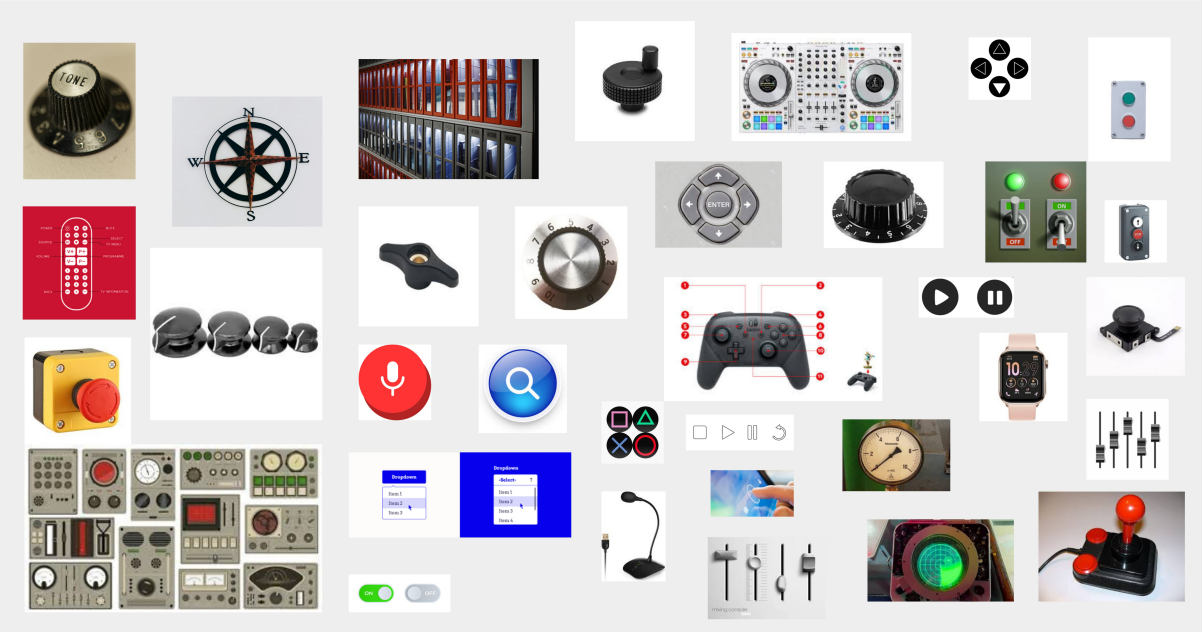


Fig. A2. Interface elements provided as prompts for envisioning alignment-supporting interactions.

B Participant Demographics

Figure A3 illustrates the demographics of our workshop participants for the variables age, gender, and region of origin, as well as self-assigned level of LLM research experience.

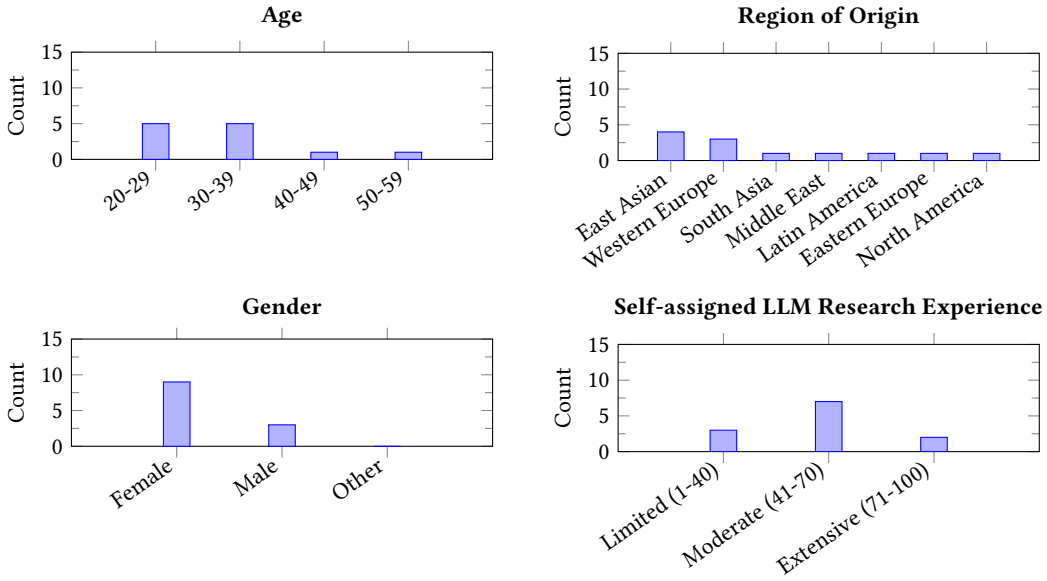


Fig. A3. Participant demographic counts for our participatory workshop with researchers ($N = 12$).

Table A1 illustrates how participants were grouped to ensure coherence of research fields within groups while maintaining variation across groups.

	Research Field
Group 1 (in-person)	AI & Machine Learning
	AI & Machine Learning
	AI & Human Behavior
Group 2 (online)	Human–Computer Interaction & UX
	Design Research
	Humanities & Critical Studies
Group 3 (online)	Social & Communication Studies
	Social & Communication Studies
	Human–Computer Interaction & UX
Group 4 (online)	Human–Computer Interaction & UX
	Human–Computer Interaction & UX
	Humanities & Critical Studies

Table A1. Group division for our participatory workshop according to research fields.

C Conceptual Model of Reflexive Thematic Analysis

Figure C1 illustrates the analysis steps taken. In practice, steps did not follow so linearly; instead, data or codes were iteratively revisited when necessary.

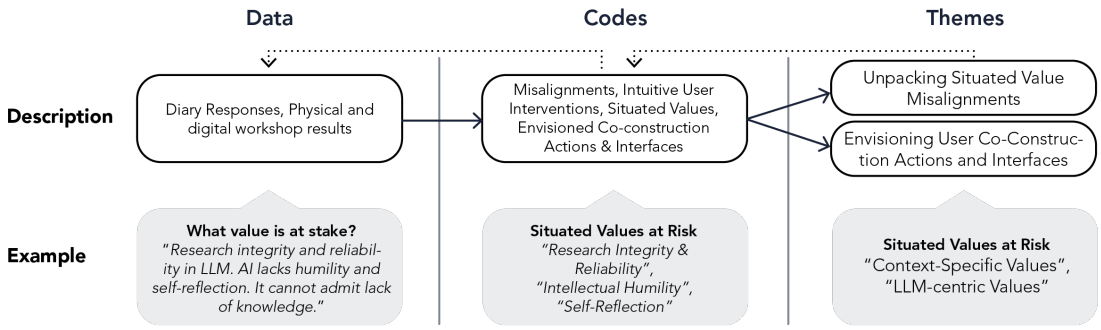


Fig. C1. This figure visualises our reflexive thematic analysis as a progression from *data*, to *codes*, to *themes*. The left column represents the materials analysed in the study (participants' diary and workshop responses). From these materials, we inductively generated codes of how participants described their perceptions, actions and ideas towards misalignments, intuitive interventions, situated values, as well as co-construction actions and interfaces. These codes were then examined and grouped to develop broader themes that represented patterns of shared meaning across participants' accounts. The example shown in the figure illustrates this movement across levels of abstraction. Codes were closely tied to participant responses, and themes emerged through examining relationships across codes. The figure, therefore, provides a conceptual overview of how we moved from concrete participant accounts to coding and theme development in our reflexive thematic analysis.

D Alignment Goals of Workshop Groups

Table C1 presents the alignment goals which the different groups defined during their workshops in step 4 in Phase 2.

	Self-defined Alignment Goal
Group 1	"We want an LLM that lets us, as CS/AI researchers, do reliable research by supporting interactions that are characterised by reflection on whether the responses are corresponding to the users' requirements and intentions."
Group 2	"We want an LLM that lets us, as humanities researchers, account for multiple perspectives in the construction of correct or accurate answers, by supporting interactions that are characterised by encouraging us to ask nuanced questions ."
Group 3	"We want an LLM that lets us, as social science/humanities/participatory researchers, work with vulnerable groups carefully and responsibly, by supporting interactions that are characterised by nuance and reflexivity (by researchers)."
Group 4	"We want an LLM that lets us, as HCI/philosophy researchers, conduct reliable and sustainable research by supporting interactions that convey honesty when the LLM can't deliver the desired degree of nuance or precision, and that condition us to think about the environmental impacts of our LLM use."

Table C1. Alignment goals as self-defined by the individual workshop groups.