TU Delft

Delft University of Technology

Document Version
Final published version

Licence
CC BY

Citation (APA)
Arzberger, A., Offerman, C., Gadiraju, U., Bozzon, A., & Yang, J. (2026). "label from Somewhere": Reflexive Annotating for Situated AI Alignment. In *ACM FAccT 2026 - Proceedings of the 9th annual ACM Conference on Fairness, Accountability, and Transparency* (pp. 1656-1680). (ACM FAccT 2026 - Proceedings of the 9th annual ACM Conference on Fairness, Accountability, and Transparency). Association for Computing Machinery (ACM). https://doi.org/10.1145/3805689.3812268

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.

“Label from Somewhere”: Reflexive Annotating for Situated AI Alignment

ANNE ARZBERGER, Web Information Systems, Delft University of Technology, Netherlands
CÉLINE OFFERMAN, Knowledge and Intelligence Design, Delft University of Technology, Netherlands
UJWAL GADIRAJU, Web Information Systems, Delft University of Technology, Netherlands
ALESSANDRO BOZZON, Knowledge and Intelligence Design, Delft University of Technology, Netherlands
JIE YANG, Web Information Systems, Delft University of Technology, Netherlands

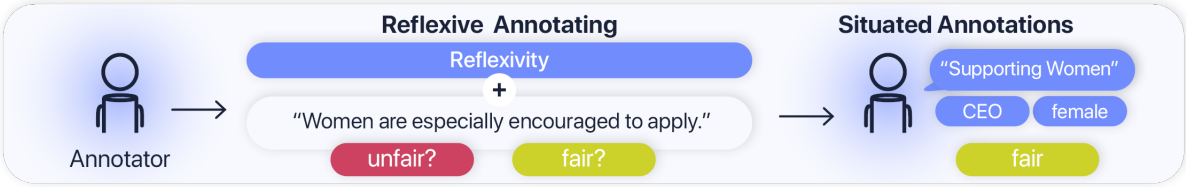


Fig. 1. Overview of our contributions: extending annotation through reflexivity to produce situated fairness annotations.

AI alignment relies on annotator judgments, yet annotation pipelines often treat annotators as interchangeable, obscuring how their social position shapes annotation. We introduce *reflexive annotating* as a probe that invites crowd workers to reflect on how their positionality informs subjective annotation judgments in a language model alignment context. Through a qualitative study with crowd workers ($N = 30$), including follow-up interviews ($N = 5$), we examine how our probe shapes annotators’ behaviour, experience, and the situated metadata it elicits. We find that reflexive annotating captures epistemic metadata beyond static demographics by eliciting intersectional reasoning, surfacing positional humility, and nudging viewpoint change. Crucially, we also denote tensions between reflexive engagement and affective demands such as emotional exposure. We discuss the implications of our work for richer value elicitation and alignment practices that treat annotator judgments as situated and selectively integrate positional metadata.

CCS Concepts: • **Human-centered computing** → *Empirical studies in collaborative and social computing*.

Additional Key Words and Phrases: Reflexive Annotating, Situated Annotations, Reflexivity, Value Alignment, AI

ACM Reference Format:
Anne Arzberger, Céline Offerman, Ujwal Gadiraju, Alessandro Bozzon, and Jie Yang. 2026. “Label from Somewhere”: Reflexive Annotating for Situated AI Alignment. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’26), June 25–28, 2026, Montreal, QC, Canada.* ACM, New York, NY, USA, 25 pages. https://doi.org/10.1145/3805689.3812268

Authors’ Contact Information: Anne Arzberger, a.arzberger@tudelft.nl, Web Information Systems, Delft University of Technology, Delft, Netherlands; Céline Offerman, c.e.offerman-1@tudelft.nl, Knowledge and Intelligence Design, Delft University of Technology, Delft, Netherlands; Ujwal Gadiraju, u.k.gadiraju@tudelft.nl, Web Information Systems, Delft University of Technology, Delft, Netherlands; Alessandro Bozzon, a.bozzon@tudelft.nl, Knowledge and Intelligence Design, 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/10.1145/3805689.3812268

1 Introduction

Alignment aims to ensure that artificial intelligence (AI) acts in line with human values, such as fairness [24]. It is driven by the value judgments of annotators, whose preferences shape the model's moral backbone through large-scale annotation pipelines. Here, annotators are routinely treated as interchangeable instruments for collecting "objective" labels [4, 37, 68, 70]. Yet, despite these aspirations to objectivity, value judgments remain inherently subjective and situated: whether a text is deemed fair or unfair depends not only on semantic content, but also on the annotator's positionality —the lived experiences, cultural norms, social identities, and personal histories that shape their interpretation [11, 34, 111]. Positionality does not neatly translocate across contexts; which of its facets matter shift with the scenario at hand [27, 41, 54, 94]. Without accounting for this fluid positioning, models might inherit false universals, flattening diverse perspectives into averages that mirror the norms of the most represented or audible annotators [10, 11, 35, 111].

Challenging this "view from nowhere" [73], prior work has argued for centring annotator perspectives and documenting annotation context, treating crowd workers as knowers, foregrounding how positionality shapes subjective judgements [34, 72, 111]. With respect to value alignment, such positional metadata could provide avenues for making value judgements traceable, surfacing overlooked or marginalised viewpoints [34, 88], supporting engagement with plural perspectives [72, 107], and enabling personalised alignment [63, 64]. However, related work either (1) attaches context-transcending demographic metadata (e.g., [9, 64]), which overlook how positionality is salient to a particular value judgment, or (2) infers annotator characteristics from behavioural traces or labels (e.g., [21]), reducing workers to objects of prediction, and sidelining their interpretive agency.

We argue for a third path: *reflexivity*. Rooted in feminist theory and social science, reflexivity is a practice of self-interrogation traditionally used by researchers to articulate how their social position shapes knowledge production and interpretation [94]. We adapt reflexivity to the annotation process, inviting crowd workers, at the moment of judgement, to articulate how they arrived at it, so as to surface which facets of their positionality shape their value judgements. We refer to this process as *reflexive annotating*, and to its output as *situated annotation*, to resist the broader tendency in the field to collapse process and product under the single term "annotation", and to emphasise that annotations are constituted through situated human judgement. Designing for *reflexive annotating* is non-trivial: individuals hold multiple, intersecting identities that are fluid and context-dependent, rendering reflexive engagement cognitively demanding even for trained practitioners [26, 27, 54]. These demands sit in tension with the efficiency-driven logics of large-scale crowdsourcing, where time and cognitive effort are tightly constrained. Understanding how value judgments become socially grounded, therefore, requires attending to how annotators recognise and articulate positionality during annotation. This motivates our research questions:

> *(RQ1) How do annotators reason about value judgments when provided with a mechanism for reflexive engagement? (RQ2) What are the implications of reflexive engagement for annotators?*

To advance this agenda, we aim to characterise what *reflexive annotating* looks like in practice and how it reshapes both the content and conduct of value annotation. In doing so, we aim to move beyond critiques of "simple alignment" toward empirically grounded accounts of how situated human judgment can be meaningfully engaged within alignment pipelines. Drawing from design research traditions, we adopt a qualitative experimental approach that iteratively designs and tests a reflexive annotating probe within a crowdsourcing alignment task. Workers assess the fairness of short text excerpts and are prompted to (i) indicate which aspects of their identity or background felt relevant in making their assessment, (ii) highlight textual cues shaping their interpretation, and (iii) provide a brief rationale linking experience, social positioning, and reading of the scenario.

Through an empirical study with crowd workers ($N = 30$) and follow-up interviews ($N = 5$), we demonstrate that *reflexive annotating* can elicit rich rationales, fluid and task-dependent identity salience, and context-specific

framings that surpass the limitations of static demographic attributes. Reflexive engagement makes visible how value judgments unfold through personal experience as well as imaginative or empathetic engagement with others' perspectives, often accompanied by explicit uncertainty. We further find that reflexive engagement alters how workers approach annotation, encouraging slower deliberation, perspective-taking, and recognition of epistemic limits. We position *reflexive annotating* as a distinct epistemic mode of alignment work, oriented toward *situated annotations* rather than decontextualised labels. We argue that such an orientation is essential for *situated alignment*, in which subjectivity, uncertainty, and social positioning are treated as epistemic resources rather than obstacles.

2 Background and Related Work

We trace how annotation for value alignment[1] has been shaped by the ideal of objectivity, introduce a feminist lens on situatedness and reflexivity, and highlight gaps in related work on unpacking the social grounding of annotator labels.

2.1 The Fiction of Objectivity

Common alignment approaches, such as supervised fine-tuning (SFT) [76], reinforcement learning from human feedback (RLHF) [97], and direct preference optimisation (DPO) [83], depend heavily on large-scale human annotation, typically asking annotators to compare outputs and indicate preferences. Platforms such as Prolific [3] and MTurk [1] enable this scaling [47, 103], but incentivise throughput and consensus over depth or reflection. Here, annotators often complete dozens of microtasks [22, 42, 47, 103]. Agreement is enforced as alignment pipelines are built around loss functions that assume a single, commensurable target for optimisation. When annotations disagree, they are typically resolved through methods such as majority voting or averaging [87], or by deferring to an "expert" adjudicator [101] to produce a single ground-truth or gold label.

These alignment practices inherit epistemologies of neutrality and objectivity, rooted in ideals of subject-independent and reproducible knowledge [37]. Adams' critique in Deleting the Subject [4], extending Nagel's "view from nowhere" [73], shows how systems like Cyc enforced a singular worldview under the guise of objectivity, echoing alignment work that assumes a generalisable notion of what is "preferred" or "aligned" [10, 11, 71, 100]. This epistemic stance shapes annotation pipelines, which optimise for consensus and scale, thereby erasing the very subjectivity they rely on [59]. Approaches such as filtering out workers with "strong opinions" [56], counterfactual debiasing [46], and Bayesian bias mitigation [109] reinforce what Kapania et al. describe as "representationalist thinking" [59]: the assumption that labels transparently reflect reality. Critical data studies show how this erasure privileges dominant viewpoints, whether in fairness definitions that marginalise communities in India [88], annotation labour shaped by power hierarchies [108], or the broader situated contexts of data production [102, 106]. Ekbia and Nardi's notion of "heteromation" underscores how the labour and perspectives of those who make data are routinely obscured [33].

This flattening has epistemic and ethical consequences. Divergent perspectives are collapsed into a single training signal, leaving no trace of who held which views or why. As a result, models learn statistical conformity rather than sensitivity to situated values. At deployment, this creates an asymmetry: systems approximate what "most people" prefer, not what particular users or communities value. This asymmetry gives rise to epistemic injustice, where marginalised perspectives are excluded or distorted [60], and systems default to the norms

[1]Alignment is the primary focus of this study and serves as a particularly salient site for examining these dynamics, as it explicitly foregrounds normative judgments about fairness, harm, and acceptability. At the same time, alignment is not exceptional in its reliance on subjective interpretation: many widely used NLP annotation tasks, such as sentiment analysis or toxicity and abuse detection, similarly depend on annotators' situated judgments. We therefore treat alignment as a focal case through which to examine broader questions about subjectivity, objectivity, and social grounding in annotation practices.

of dominant populations [8]. Alignment thus risks homogenising plural, situated perspectives, suppressing disagreement and reproducing epistemic harms [34].

2.2 Situated Judgments and Reflexive Interrogation

If objectivity cannot serve as a viable epistemic ideal for value-laden judgments, subjectivity must be understood not as a flaw but as constitutive of how judgment is formed. Feminist epistemology advances this view by arguing that knowledge is never disembodied or universal, but always situated [7, 48, 98]. Central here is *positionality*, the social coordinates through which individuals interpret and engage with the world, such as race, gender, class, political ideology, as well as personal and professional experiences, political and ideological stances, and other aspects of our social biography, or "lifeworld" [16, 17, 23, 92, 113], which structure what is seen, valued, or ignored [93].

These insights call for a reconsideration of objectivity in AI alignment. Harding's *strong objectivity* [49, 50] reframes objectivity as obstructive, advocating explicit engagement with positionality and the power relations embedded in annotation. Here, objectivity often masks normative assumptions [4, 64, 82], reinforcing dominant worldviews. For alignment, this means interrogating not just who annotates, but how values are encoded, whose perspectives are centred, and what is lost in pursuit of coherence. This commitment leads to the imperative of reflexivity: the interrogation of how one's own position shapes task design, prompts, and interpretation, central to feminist and critical HCI [13, 14, 16, 57, 91, 112]. Yet, operationalising reflexivity remains a challenge. As Soedirgo and Glass [94] observe, positionality is too often reduced to identity checklists or static declarations. As intersectional scholars emphasise [26, 27, 54], individuals hold multiple, intersecting identities that interact in fluid and contextual ways. Our positions do not neatly translocate from one setting to another; they shift in relation to time, space, power, and task. What is salient in one annotation judgment may be irrelevant in another [26, 41, 52]. Additionally, in feminist and reflexive HCI, reflexivity remains inward-facing, leaving its insights inaccessible to alignment pipelines at scale [12, 15, 38]. Consequently, how positionality shapes epistemic judgment remains underexamined, calling for tools and methods that surface and disclose the social positions shaping AI value judgments.

2.3 Beyond Ground Truth: Aiming for Positionality in Annotation

A growing body of work in NLP, ML, and HCI has challenged the ideal of a single, objective ground truth in annotation, particularly for subjective and value-laden tasks. Zhang et al. [114] show that human preferences vary far more than LLM outputs reflect, partly because common annotation methods rely on homogeneous candidate responses. Other studies link annotators' lived experiences, social positions, and beliefs to systematic differences in labelling political stance [67], sentiment [30], and toxic language [78, 89, 99]. Enforcing consensus through aggregation can therefore erase meaningful contextual nuance in human judgments [40]. In response, several approaches aim to capture annotator positionality explicitly or implicitly. Systems such as DICES [9], PRISM [64], Community Alignment [114], and POPQUORN [79] attach static demographic metadata (e.g., age, gender, nationality), often accompanied by explanations to analyse disagreement patterns, while related HCI scholarship calls for documenting annotation context and treating crowdworkers as epistemic contributors rather than interchangeable labour [70, 82]. Parallel computational work instead infers positionality from annotation behaviour, using techniques such as annotator fingerprinting, position mining, or annotator embeddings to recover latent preference structures from disagreement patterns [21, 28, 29]. Recent work has further shown the value of capturing annotator perspectives explicitly in subjective learning tasks, demonstrating that such perspectives improve both interpretability and downstream performance [72].

Despite these advances, existing approaches tend to treat positionality either as a fixed, task-agnostic attribute encoded through demographic metadata or as a latent statistical pattern inferred post hoc from disagreement.

While some systems collect natural language rationales, these explanations are typically task-specific and do not explicitly connect judgments to annotators' social positionalities. As a result, links between identity and judgment are generally reconstructed analytically rather than made explicit within the annotation process. Here, annotators remain largely passive: their positionality is either predefined through externally imposed categories or inferred from their behaviour, with limited opportunity to actively interpret, negotiate, or express how their social position shapes a given judgment. This risks essentialising intersectional experience and obscuring which aspects of identity become salient in context [26, 27, 54]. Such approaches rarely capture when, why, or how positional factors matter in situ, nor how annotators reason through value judgments as they unfold [52, 94]. Taken together, these limitations point to a methodological gap: while annotator positionality is widely acknowledged as relevant, it is seldom captured as a dynamic, per-judgment, and explicitly articulated form of reasoning produced by annotators themselves. This paper addresses this gap by introducing reflexive annotating as an active epistemic practice, in which annotators are not treated as sources of signals to be extracted, but as agents who articulate how their situated perspectives shape each value judgment at the moment it is made.

3 Methodology

To inform situated alignment, we empirically characterise reflexive annotating through a design probe and follow-up interviews with crowd workers. The probe elicited situated fairness judgments in an RLHF-style alignment task [97]. We adopt a Big Q qualitative approach [62] and analysed both data sources through reflexive thematic analysis [19, 20].

3.1 Iterative Probe Design Process

We introduced a design probe to create conditions for reflexivity in annotation. Probes are design-oriented tools often using visual and tangible kits to trigger new insights or to reveal hidden ideas [43]. In our study, the probe functioned as a generative device to elicit perspectives obscured in conventional annotation workflows. The probe was iteratively developed through low-fidelity prototypes on the collaborative online platform Miro, and piloted with PhD candidates from design and computer science. These pilots helped balance cognitive load, time demands, and the depth of introspection. Our aim was to create an interface legible to crowd workers while minimising fatigue.

To enable *reflexive annotating*, we turned to situational mapping practices that embed reflexive moments within empirical action. Jacobson & Mustafa's [58] *Social Identity Map* proved particularly well-suited: a concise,

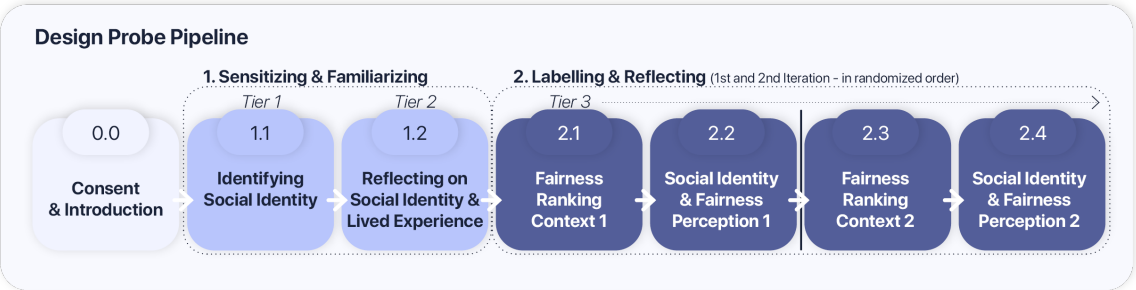


Fig. 2. The anatomy of the design probe used in our crowd computing study. The reflexive annotating process consists of two core stages: (1) *Sensitising & Familiarising*, where annotators identify and reflect on social identity facets and facts; (2) *Labelling & Reflecting*, consisting of two randomised iterations where annotators rank fairness in context and reflect on social identity. This structure was inspired by tiers of the social identity map [58].

three-tiered tool that prompts participants to (1) name their social identities, (2) reflect on their influence more generally, and (3) more specifically, given a particular context at hand. This structure lends form to inherently abstract, fluid, and evolving positionality, enabling participants to recognise and disclose how social identity facets shape their value judgments. For our purposes, the Social Identity Map offered the right balance of guided structure to support non-experts, flexibility to foreground identity facets relevant to particular judgments, and scalability for lightweight, iterative reflexivity in crowdsourcing contexts.

We translated the Social Identity Map into a design probe (see Figure 2) that facilitated reflexive annotating by enabling participants to record their reasoning, and interface elements that allow for explicit and intersectional expression of salient social identity facets. We also balanced open-ended reflection with task feasibility and the degree of structure imposed on annotators' reflections. The resulting probe was implemented as a custom web application (React frontend with a minimal server-side component for task delivery and data storage), integrating annotation and reflection within a single interface and including attention checks for quality control (see Figure 5).

3.2 Data Collection

3.2.1 Crowdsourced Reflections. We deployed our design probe in a qualitative crowdsourcing study with ($N = 30$) participants. The study received approval from the ethics review committee of Delft University of Technology under code 5829. Participants were recruited via Prolific[2] using English fluency and approval-rate filters and provided demographics in Tier 1. Two participants who nevertheless responded in Spanish were excluded because their responses could not be reliably interpreted. Recruitment proceeded iteratively across several rounds while maintaining eligibility, with participation filters adjusted to broaden intersectional representation across gender, ethnicity, and ability (see Appendix A). Participants ranged in age from 19 to 67 years ($M = 34.4$, $SD = 12.9$). In terms of gender identity, 14 participants identified as male, 13 as female, and 3 as non-binary. Ethnicity was self-reported as White ($n = 17$), Black ($n = 10$), Asian ($n = 1$), Mixed ($n = 1$), and Other ($n = 1$). On average, participants spent 32 minutes ($SD = 15.0$, range = 10–67) completing the task and were paid £9.00/hr, deemed a "good wage" on the platform. The probe scaffolded two layers of reflection (see Figure 2 and Section 4): (1) sensitising participants to positionality and reflexivity, (2) labelling and reflecting on a fairness-related annotation task. Each layer combined structured items with open-ended prompts, encouraging participants to articulate how their social positions shaped their fairness judgments, which enabled us to collect both judgments and meta-level reflections. This structure positioned the probe as both a site of annotation and of reflexive knowledge production. While fairness ratings were collected on Likert scales, our primary data consisted of participants' written responses to the reflective prompts. These texts provided a window into how workers navigate identity, power, and fairness within the constraints of high-throughput annotation work.

3.2.2 Follow-Up Interviews with Annotators. To complement our understanding of how participants experienced and justified their reflexive engagement, we conducted one-hour semi-structured follow-up interviews. All previous ($N = 30$) participants were invited to the optional follow-up interviews in response to the substantial variation in annotators' positional accounts. Five of those participated. The interview protocol drew on frameworks from meta-cognition [32], self-directed learning [95], and reflexive interviewing [80] (see the interview protocol in Appendix B). We asked participants to walk us through their experience with the tool and revisit moments of tension, clarity, or uncertainty in their responses. Our aim was to move beyond the procedural accounts of annotation and surface the emotions, assumptions, and social contexts shaping them. This dialogic, exploratory approach allowed us to probe both the cognitive and affective dimensions of situated annotation and to validate or add nuance to themes emerging from the initial probe data.

[2]https://www.prolific.com

3.3 Data Analysis: Reflexive Thematic Analysis

We conducted a reflexive thematic analysis, following Braun and Clarke's six-phase approach [19, 25]. We began with repeated reading of participant responses and interview transcripts to build familiarity, followed by inductive coding grounded in participant reflections. Codes were iteratively refined and clustered into themes through recursive discussion. Anne and Céline independently coded the full dataset before meeting to compare interpretations, discuss tensions, and reflect on their own positionalities (which can be found in Section 9) [20, 55]. We used multiple coders because questions of fairness, identity, and reflexivity are best approached from different standpoints. This plural perspective helped surface blind spots, interrogate assumptions, and enrich our analysis. Consistent with Braun and Clarke's epistemology [19, 20], we did not calculate inter-coder reliability; instead, consensus-building was treated as an interpretive act rather than a metric of objectivity [69]. To support transparency, we include a conceptual model of our analytic process in Appendix D. We applied the same analytic approach to both survey and interview data. The supplementary quantitative data (e.g., Likert ratings and identity selections) were analysed descriptively to contextualise and support patterns emerging from the qualitative findings.

4 The Final Design Probe: Designing with and for Positionality in Crowd Sourcing

Through several design iterations and pilot tests, we developed a design probe inspired by contemporary LLM alignment annotation workflows. This experimental probe, as illustrated in Figure 2, included a concise introduction to minimise cognitive load while ensuring that participants understood key concepts such as *social identity* and *social identity facets*, preparing them for the reflective exercises ahead. Following this, participants were asked to describe, in their own words, what they understood by social identity. This served both as a quality control measure and as a source of rich data for our thematic analysis. Participants were then guided through *Tier 1 & 2 reflection exercises*, to familiarise and sensitise them for the key reflection exercise in *Tier 3*, where they were asked to rate text pieces in terms of fairness and reflect on their social identity facets that influenced these fairness judgments.

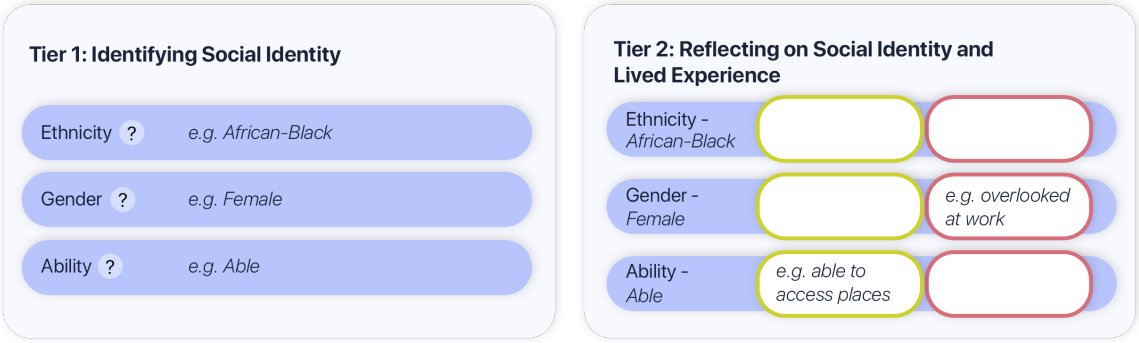


Fig. 3. Tier 1 and Tier 2 reflection activities in the design probe. In Tier 1, annotators identified facets of their social identity by completing each category. In Tier 2, they reflected on how these facets shaped their lived experiences, noting both positive and negative impacts. Instructions asked annotators to (1) consider whether facets such as ethnicity, gender, or ability affected their education, work, relationships, or other life domains, and (2) record keywords describing those positive or negative impacts.

Tier 1 – Familiarising. In the first Tier (Fig. 3), participants listed three facets of their social identity (ethnicity, gender, and ability) to become more familiar with key concepts such as positionality and social identity, laying

the groundwork for the subsequent reflections. Recognising that novice annotators might struggle to articulate their social identities while being wary of the time constraints of crowd work, we chose three social identity facets as an accessible yet meaningful subset of the eight dimensions in Jacobson & Mustafa's original map [58], balancing familiarity (ethnicity, gender) with a less commonly discussed dimension (ability). Definitions of key terms were available via clickable icons, and one example was provided to guide participants.

Tier 2 – Sensitising. The second Tier (Fig. 3) prompted participants to consider how these identity facets shaped everyday experiences, using short guiding questions (e.g., "Did my gender affect my work or education?"). Examples illustrated the expected format. By connecting identities to lived experiences, this Tier aims to deepen awareness of positionality and prime workers to link identity to value judgments in Tier 3.

Tier 3 – Labelling & Reflecting. Tier 3 formed the core of the probe. Participants first rated the fairness of two texts offered in randomised order (see Fig. 4). These texts were strategically designed to highlight contrasting dynamics related to gender discrimination—a job vacancy for a secretary position to reflect biases disadvantaging women, and an advertisement on beauty products to reverse this dynamic (see Appendix C). These were intentionally loaded with fairness-relevant cues to support reliable, reflexive evaluations. We used Likert-scale feedback to further promote subjectivity and permit preference intensity to be expressed beyond "chosen:rejected" pairings. This choice reflects our focus on capturing the texture and grounding of individual value judgments, rather than producing directly optimizable preference data. Fairness was chosen as the focal value for its intuitive resonance and defined as "treating individuals justly, without bias or discrimination, and ensuring equal opportunities while recognising diverse needs" [2]. A clickable icon revealed this definition, safeguarding clarity without disrupting flow. Participants were encouraged to rely on their intuitive judgments when rating fairness.

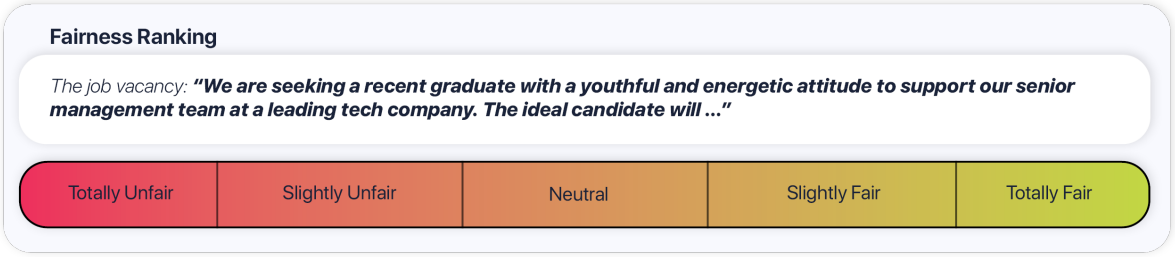


Fig. 4. Annotation task for capturing annotators' fairness perceptions of the job vacancy text sample. Annotators are asked to read the job description and rate its fairness on a Likert scale according to their judgment. The full text pieces can be found in the Appendix C.

Participants then revisited the texts in an interactive interface (Fig. 5), highlighted text passages they perceived as (un)fair, and tagged them with one or more identity facets. One or more facets could be used to label a text passage to capture intersectionality intuitively. Based on the Social Identity example [58], participants were given a non-exhaustive list of optional identity facets (Class, Citizenship, Ability, Age/Generation, Ethnicity, Sexual Orientation, Cis/Trans, Native Language(s), Gender, Religion, Socio-economic Background), which they could extend if needed, and provided a short rationale for each tag. This process enabled situated, intersectional reflections by tying fairness judgments directly to identity. Due to its complexity, Tier 3 was supported by a brief tutorial video available throughout the task.

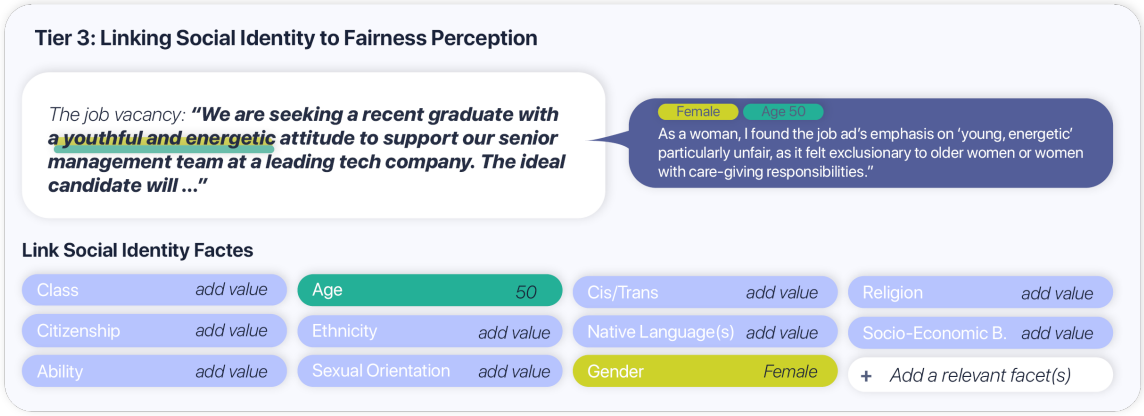


Fig. 5. Tier 3 of the reflection interface allows annotators to highlight passages they perceive as fair or unfair, tag them with one or more predefined (or custom) social identity facets, and explain how these facets shape their judgment, supporting intersectional labelling. Annotators were instructed to: (1) re-read the text and highlight passages relevant to their fairness perception; (2) reflect on "How did my social identity shape this perception?"; (3) attach one or more identity facets to each highlighted passage by dragging them from the right-hand panel; and (4) briefly describe how these facets positively or negatively influenced their view. They were encouraged to add new facets when needed and to select only those most salient to their interpretation.

5 Results

Our findings draw on situated annotations and follow-up interviews to examine how value judgments become socially grounded through reflexive engagement, and how this reshapes annotators' experiences of alignment work. Using reflexive thematic analysis, we identify inductively constructed patterns of shared meaning across accounts. We organise the findings around our two research questions, focusing on (RQ1) how annotators reason about fairness through positional reflection, and (RQ2) how reflexive annotating shapes annotators' conduct and epistemic orientation.

5.1 Reflexive Annotating in Practice (RQ1)

Reflexive prompts shifted annotators from providing decontextualised labels to engaging in situated forms of reasoning. When asked to indicate relevant identity facets, highlight textual cues, and articulate how their experiences shaped their judgment, workers responded with layered, often deeply personal reflections that reveal how value-laden judgments unfold in practice. Figure 6 illustrates typical situated annotations, showing how identity salience, cue interpretation, and rationale-building intertwine within a single response. The figure highlights the kinds of epistemic materials generated by the probe and grounds the analysis that follows. Addressing RQ1, we cover three subthemes: (1) Positioning the Self, (2) Intersectional Reasoning, and (3) Reflexivity Under Tension.

5.1.1 Positioning the Self. A central pattern in annotators' reasoning was the deliberate act of positioning themselves in relation to the scenario. Workers often grounded their evaluations in who they were, what they had lived, and how they located themselves relative to the subjects involved. Reflexive prompts surfaced two recurring modes of anchoring, lived experience and imaginative solidarity, as well as distinct narrative strategies for claiming (or withholding) epistemic authority. This positioning was central to their interpretive process. Many workers grounded their judgments in first-hand experience. For example, **P17 (White, Female, Able)** linked her

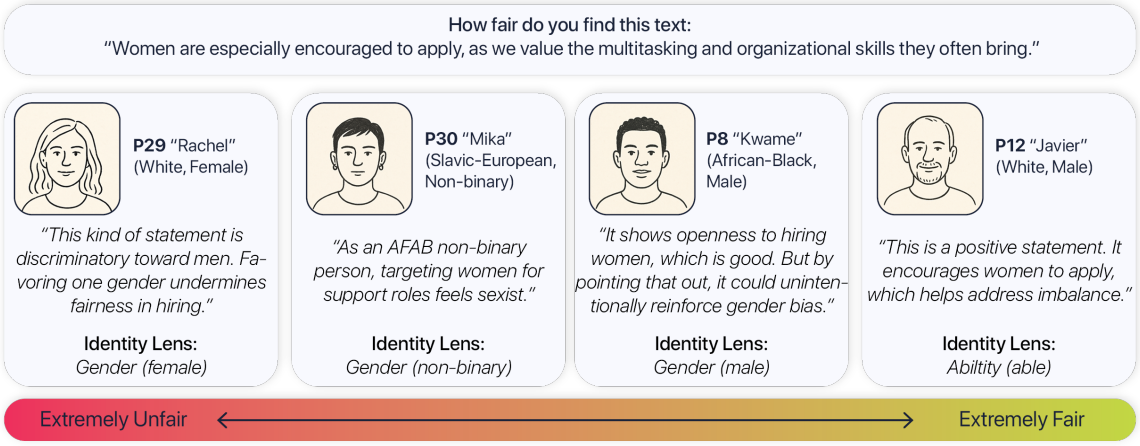


Fig. 6. Exemplary situated annotations from our study that introduce our annotators as situated individuals with varying valid perceptions of fairness. The names have been assigned freely, and their rationales have been paraphrased to ensure anonymity.

own psychological challenges to a broader understanding of disability, explaining: *"Given my personal experiences with psychological issues, I have a broader understanding of how people with mental disabilities must see the world."* Similarly, **P10 - Interview (African-Black, Female)** described how the task resonated with experiences of discrimination she had repeatedly encountered: *"It was not difficult for me to reflect because these are the things that we've been experiencing as we grow up."* Other annotators reasoned through imaginative solidarity, acknowledging a lack of personal experience while still articulating empathic moral evaluations. As **P25 - Interview (White and Black Caribbean, Male)** noted: *"I don't think I've ever experienced hiring discrimination myself where I've applied for a job, and people were only accepting women, but I can easily put myself in the shoes of someone in that situation. I'd be upset."*

Across these variants, positioning the self, whether through lived experience or imagined solidarity, was a key part of how annotators reasoned through value-laden scenarios. These strategies show that reflexive annotating surfaces how people locate themselves within moral judgments, not just what judgments they make.

Annotators also didn't express positionality uniformly; rather, they selectively foregrounded it, with different facets becoming salient across contexts. **P2 (African-Black, Female, Able)**, for instance, foregrounded gender in a hiring scenario (*Context 1: "It's giving a slightly misogynistic vibe with the male-dominated executive team reference." Assigned Identity: Gender (Female)*) but ethnicity or class in a consumer product context (*Context 2: "This presupposes black or yellow or other skin types that are not white to be inferior." Assigned Identity: Ethnicity (African-Black)*). In line with our theoretical stance, this underpins the argument that positionality should be understood and approached as a fluid, context-sensitive construct.

5.1.2 Intersectional Reasoning. Annotators frequently reasoned through multiple, co-constitutive identity facets when explaining their fairness judgments. Many participants explicitly linked gender, race, class, or ability, showing how overlapping identities shaped their interpretations of the task. Sometimes this appeared as distributed reasoning, with different facets applied to different text segments; in others, annotators intertwined facets into a single, compound standpoint. For example, **P28's (Caucasian, Non-Binary)** connected feminist commitments, gendered social pressure, and assigned sex at birth in a critique of a job advertisement: *"Woah, what a misogynist*

job vacancy! Women and people who present as women already have a lot of pressure on them put by society; it's better if we don't include confidence in this too. Confidence comes from the inside and self-worth." Assigned Identity: Militantism (Feminist), Assigned sex at birth (Female). This kind of intersectional anchoring enriched the interpretive space of the task, as annotators contextualised them within broader patterns of compounded discrimination. Intersectional reasoning thus contributed directly to epistemic richness, surfacing how fairness evaluations emerge from lived social complexity.

5.1.3 Reflexivity Under Tension. While the design probe helped annotators surface rich perspectives, they also revealed moments where reasoning was ambivalent or incomplete. These tensions exposed the limits and frictions inherent in articulating value judgments within constrained annotation workflows. Annotators grappled with how to express moral evaluations, map them onto fixed categories, and separate fairness concerns from personal reactions.

One form of tension appeared in under-articulated reflections, where annotators pointed to fairness-relevant cues but left their evaluative stance implicit. **P2 (African-Black, Female)**, for instance, highlighted "exclusive" in a luxury skincare advertisement but wrote only: *"The use of the word exclusive."* This left unclear whether she viewed the phrasing as aspirational, exclusionary, or simply noteworthy. Interview data showed that such brevity often reflected an assumption of shared understanding. **P10 - Interview (African-Black, Female)** explained that she omitted explicit approval because it seemed obvious: *"It's a good thing that it considers women's input. I just didn't write that it's a good thing because I think, as you can see from the text, it's good that in a job vacancy they consider women's input."* These moments illustrate how annotators rely on implicit normativity rather than explicating their reasoning, creating ambiguity for systems that depend on explicit signals and in light of plural moral perspectives.

A second form of tension emerged through labelling dissonance, where annotators' selected identity facets diverged from the nuance of their reasoning. **P21 (African-Black, Female)** critiqued exclusion rooted in class and income, yet selected "Gender (Female)" as the identity category: *"The text suggests that perfection is not only desirable but also attainable via luxury products. Individuals from lower-income backgrounds may feel excluded for not affording such products, while gender norms often push perfectionist beauty ideals more heavily."* Such mismatches suggest that annotators sometimes defaulted to familiar categories rather than those most aligned with their reflection, highlighting the difficulty of mapping contextualised reasoning onto rigid taxonomies.

A third tension arose in emotionally charged or beauty-oriented content, eliciting personal reactions that overshadowed fairness evaluation. Responding to the marketing phrase *"the flawless fairness you truly deserve,"* **P3 (African-Black, Female)** disclosed: *"This makes me feel good because I am not happy with the way my skin looks and feels."* In contrast, **P17 (White, Female)** interpreted the same line as a normative judgment about appearance: *"It's making an assumption that I am imperfect, it flat out says that I am not pretty enough for society."* Together, these tensions reveal the complex terrain that annotators navigate when interpreting texts that resonate with personal insecurities or aspirations. Annotators moved between clarity and ambiguity, precision and approximation, moral critique and personal response. These frictions illuminate the messy nature of value judgments and underscore why reflexive annotating should be understood as revealing the lived complexity behind them.

5.2 Experiencing Reflexive Annotating (RQ2)

The design probe shaped how annotators anchored their value judgment in their social position and experienced the work. Asking workers to name relevant identities, revisit personal memories, or articulate interpretive cues introduced new expressive, emotional, and moral demands that diverged from efficiency-driven platform labour. These experiences influenced whether annotators engaged deeply, withdrew, or revised their own understanding

of fairness. Related to RQ2, we describe three subthemes: (1) Emotional Exposure, (2) Strategic Withdrawal, and (3) Perspective Awareness.

5.2.1 Emotional Exposure. For many annotators, reflexive annotating elicited feelings of vulnerability by prompting them to revisit painful experiences or confront privilege. **P25 - Interview (White and Black Caribbean, Male)** described the emotional weight of returning to memories of assimilation: *"When I do think about and I do remember about that stuff as a child, it can be a little bit emotional."* Others experienced discomfort when acknowledging their own structural advantages. **P16 - Interview (White, Female)** reflected on this unease: *"That's quite an awkward thing to think about because you feel a bit guilty. [...] It's difficult to admit that you've had more privilege based just on colour. It's not fair, is it?"* For some, emotional discomfort was compounded by concerns about privacy and data use. **P22 - Interview (White, Female)** described initial hesitation: *"Yeah, it was uncomfortable [to share intimate information], until I was reassured that Prolific would not share my data."* Yet anonymity could also function as a protective buffer, enabling deeper engagement for certain workers. As **P25 - Interview (White and Black Caribbean, Male)** noted: *"It makes me a lot more comfortable just being able to sit at my computer and enter the details myself… it doesn't make me worried about anything like that."* Together, these accounts show that reflexive annotating involves affective labour: annotators must navigate the emotional consequences of revealing, revisiting, or acknowledging aspects of their lives that are normally irrelevant in data work.

5.2.2 Strategic Withdrawal. Some annotators responded to the cognitive and emotional demands of reflexive engagement by scaling back their participation. Some reflections were minimal or disconnected from the fairness task, such as **P1's (White, Male)** terse description: *"Women"* or **P5's (African-Black, Female)** unrelated comment: *"They are willing to work extra hard to gain more knowledge."* These instances of minimal engagement do not necessarily reflect misunderstanding. We interpret these as moments where annotators chose not to invest further emotional or cognitive effort, rather than as misunderstandings. Reasons included discomfort with the task's personal nature, uncertainty about whether identity should matter, or the pressure of crowd work that rewards speed over introspection. Strategic withdrawal thus shows how reflexivity can become an additional burden that annotators may selectively refuse.

5.2.3 Perspective Awareness. Reflexive annotating made several annotators more aware of the limits of their own perspective. They explicitly noted what they could not know, particularly when encountering forms of discrimination or marginalisation outside their lived experience. Some openly noted the partiality: **P13 (White, Male)** admitted: *"Occasional difficulty relating to experiences of ethnic minorities."* Similarly, **P2 (African-Black, Female)** reflected: *"I don't have first-hand understanding of disability challenges, and I tend to overlook those needs."* Others recognised the structural advantages shaping their own viewpoints. **P16 (White, UK)** stated this explicitly: *"I have a lot of privilege being white. I have had more opportunities than other non white people."*

This often took the form of reflexive approximation, with annotators thinking through uncertainty. **P16 - Interview (White, Female)**, for instance, questioned her own categorisation out loud: *"Well, I don't know. I presume it's gender. Would age come into it?"* Such expressions of uncertainty signalled an awareness that fairness judgments are necessarily partial and provisional. For some annotators, this awareness deepened over time and even led them to revise their judgments. We also observed cases where workers' initial fairness ratings diverged from their later reflections. **P21 (African-Black, Female)** described a job vacancy as "very fair," yet upon reflection, reconsidered its implications: *"Requiring native English speakers will possibly exclude individuals who are bilingual but with perfect fluency. This may discriminate against non native speakers."* These moments show how reflexive annotating cultivates perspective awareness and epistemic humility, as annotators engage not only with the text but also with the boundaries of their own standpoint.

6 Discussion and Implications

In this discussion, we move from empirical findings to a conceptual account of *reflexive annotating* as an epistemic practice. Concluding our empirical account, and building on Gentles et al. [44] and Olmos-Vega et al. [74], we define *reflexive annotating* as a *set of continuous and multifaceted practices through which annotators disclose and interrogate how their social position and lived experience shape their judgments*. The goal is not to strip annotation of its subjectivity, a task both impossible and undesirable [84], but to foreground subjectivity as a legitimate locus of knowledge production in alignment. We discuss our findings by identifying two main contributions of reflexive annotating: **enhancing epistemic richness** by surfacing how diverse value perspectives are grounded in experience, and **enhancing epistemic accountability** by making visible how annotators understand their perspectives as partial and revisable. We discuss both aims below and consider **how reflexive annotating could enhance value alignment**.

6.1 Enhancing Epistemic Richness

By surfacing the positionality that informs annotators' decisions, reflexive annotating reveals the social conditions under which labels are produced. We showed that resulting situated annotations capture dynamic and intersectional positional information that static metadata and inference-based approaches often overlook. Annotators expressed positionality in multiple narrative stances, from explicit lived experience *("as a non-binary person")*, to relational or community framings *("women in my family")*, to solidarity-based empathy *("I can put myself in the shoes of someone in that situation")*. Such variation expands the interpretive space of annotations: judgments articulated through self-, community-, or other-as-subject framings all carry epistemic weight. This empirically observed variation demonstrates that positional engagement is not uniform but contextually enacted.

These findings substantiate feminist accounts of situated knowledge (e.g. [27, 54, 94]) by providing insight into how annotators selectively foreground different aspects of themselves *(e.g., gender in one context, ethnicity in another)*. While prior work has similarly shown that annotation behaviour depends on factors beyond static demographics [6, 18, 39, 75, 85], our results offer direct empirical documentation of how such variability becomes explicitly expressed when reflexivity is supported by design. Without reflexive annotating, these layered positional articulations would have remained invisible.

Our findings suggest that this framing expands the interpretive space of annotation by treating socially derived variation as epistemically meaningful structure for alignment research. In doing so, they provide empirical grounding for rethinking alignment as a practice that can account for the broader social, political, historical, and personal contexts in which value judgments are situated.

6.2 Enhancing Epistemic Accountability

Reflexive annotating also helped annotators recognise assumptions, biases, and blind spots that might otherwise remain implicit, fostering more critical engagement with the task. Workers frequently stated what they could not know *("difficulty relating to experiences of ethnic minorities")*, what they tended to overlook *("I don't have first-hand understanding of disability challenges")*, or where they needed to approximate *("I don't know, I presume it's gender")*. Such statements reflect epistemic humility [77, 94]. In Harding's terms, this strengthens objectivity not by masking subjectivity but by making it visible and open to scrutiny [49]. At the same time, reflexive engagement was affectively and expressively demanding. Some annotators struggled to articulate intuitive judgments, experienced discomfort in acknowledging privilege, or revisited emotionally charged memories. Others withdrew strategically. These findings echo prior work on the affective costs of data work [5, 31, 47], suggesting that reflexive annotating must balance epistemic benefits with labour implications. Annotators also noted when their perspective shifted, showing reflexivity as an unfolding process in which structured prompts allow judgments to deepen or change over time. Here, epistemic humility and uncertainty are not signs of

disengagement but markers of a provisional and accountable stance toward knowledge. Designing for epistemic accountability thus means embedding support for articulation (not just prompts), preserving anonymity, and acknowledging that reflexivity is a laborious process.

6.3 Toward Situated Alignment: Opportunities and Future Work

Building on our empirical account of reflexive annotating, we outline several contributions this practice enables for alignment research, system design, and implementation. We show how reflexive annotating opens up new ways of reasoning about where value judgments come from, whose perspectives are represented, how disagreement is interpreted, and how alignment data might be used in context-sensitive and scalable ways.

Traceable Value Judgments. Our findings highlight the importance of understanding how a value judgment arises and from which perspective. Rigid prompts and predefined identity labels sometimes obscured what annotators themselves considered salient, flattening contextual signals such as professional background or lived experience. Supporting traceability, therefore, requires designs that allow annotators to indicate which aspects of their perspective matter, for example, by separating affective reactions from evaluative claims or enabling brief rationales and uncertainty markers. Such rationales have been shown to improve data quality and inform the design and evaluation of model-generated rationales, supporting explainability in AI systems [53]. Technically, such reflexive signals provide structured metadata that link judgments to their conditions of production—e.g., from personal experience or imaginative solidarity, with levels of uncertainty due to awareness of personal standpoint—enabling alignment processes to reason over situated provenance rather than decontextualised labels.

Surfacing Marginalised Perspectives. Our study suggests that marginalisation in alignment data can occur not only through under-representation but through representational mismatch, when systems recognise only certain forms of positionality as legitimate. Reflexive annotating allows annotators to articulate values and experiences in their own terms, helping surface perspectives that might otherwise remain invisible. This aligns with work showing that minority or dissenting judgments often raise concerns overlooked by consensus-driven approaches [34], as well as standpoint-theoretic arguments that such perspectives may be most epistemically valuable for alignment [50, 51].

Making Disagreement Legible. Following recent calls to value pluralism [96], situated annotations render disagreement and value pluralism interpretable. Prior work has shown that disagreement often reflects subjectivity or ambiguity rather than error and should be preserved [36, 40, 66, 104]. Existing technical approaches, such as modelling annotators individually [29, 45], learning from label distributions [86, 110], or using disagreement for uncertainty-aware prediction [61], could be extended by situated annotations that clarify why judgments diverge. In this way, disagreement becomes legible as a relational phenomenon between annotator, input, and context.

Context-Sensitive and Personalised Alignment. Reflexive annotating also has implications for context-sensitive or personalised alignment, where models adapt to user-specific values while maintaining transparency about whose values are enacted [63]. Datasets such as PRISM [64] demonstrate how linking judgments to richer participant profiles enables exploration of attribution and personalisation. By retaining explicit links between judgments and situated perspectives, reflexive annotating may further support alignment mechanisms that respond to context at deployment time rather than assuming a single, static value configuration.

Scalability vs. Reflexivity. Lastly, we would like to acknowledge that, beyond questions of utility, future work will need to further engage with how intentional, slow, and small situated approaches speak to or intersect with prevailing practices in data annotation. Resolving such tensions is beyond the scope of this paper, but we believe that reflexivity need not scale uniformly. It could be selectively deployed in high-stakes or normatively complex contexts or integrated into specific training pathways, such as active learning workflows [105]. While we have already tried to strike a balance between cost and reflexivity, we also believe future design refinements (e.g., merging Tier 1 & 2) may further minimise cost while preserving epistemic richness.

Taken together, these contributions motivate a move beyond "simple alignment" and point toward what we term *situated alignment*: approaches that treat value judgments as grounded in social position and context, and that seek to improve alignment by preserving the provenance, plurality, and contextual relevance of human judgments. Rather than replacing existing alignment techniques, situated alignment reframes their assumptions, shifting attention from producing a single "correct" behaviour to supporting alignment processes that can reason over whose values are represented, under what conditions, and with what limits.

7 Limitations, Caveats, and Ethical Considerations

Several methodological limitations merit consideration. First, despite our intent to foster reflexivity, prompts elicited avoidance in some annotators, discomfort in others, and, at times, performative engagement, suggesting that reflexive annotating may not scale reliably without careful design adaptations. Second, our focus on fairness leaves open how these practices apply to other alignment-relevant values such as harm, dignity, or truthfulness. Third, our partly categorical approach to capturing positionality, adapted from the Social Identity Map [58], supported structured reflection but necessarily constrained the dynamic and narrative character of positionality [16, 17, 23, 92, 94, 113]. While more open-ended formats could better represent its fluidity, early pilots showed that fully self-defined identities risked overwhelming participants, suggesting that greater complexity does not automatically yield deeper reflexivity. Future work might therefore explore designs that move beyond fixed demographic labels while avoiding cognitive overload (e.g., predefined yet more nuanced, bottom-up formed categories) to better balance expressiveness, usability, and situated nuance. Finally, our largely English-speaking sample limits cultural diversity and calls for replication in multilingual and cross-cultural settings.

Ethical and political risks also warrant attention. Collecting positionality data, even pseudonymised, carries re-identification and privacy risks: combined data points may reveal sensitive traits via the mosaic effect [81]. Furthermore, while personalisation can enhance cultural relevance and user experience, it also raises the spectre of large-scale profiling, bias entrenchment, privacy infringements, and targeting of vulnerable populations [63]. Prior work shows that personalisation can produce filter bubbles and affective polarisation, amplifying harms familiar from social media and search [65].

Together, these considerations underscore that reflexive annotating sits at the intersection of technical design, epistemic theory, and ethical labour practice. Its promise lies in foregrounding annotator positionality and accountability while promoting situated alignment, but its viability depends on whether we, as a community, can create infrastructures that protect workers' privacy, recognise their emotional and cognitive labour, and resist reproducing the very asymmetries that reflexivity seeks to expose.

8 Conclusion

This paper introduced *reflexive annotating* as a design probe that foregrounds annotators' social positions, experiences, and interpretive cues in producing alignment data. Through qualitative crowdsourcing and follow-up interviews, we showed how reflexive prompts elicit intersectional reasoning, surface uncertainty and perspective shifts, and generate epistemic metadata inaccessible through static demographics or behavioural inference. Our findings also underscore the affective demands of reflexivity, highlighting the need for designs that protect anonymity and respect limits on disclosure. We argue that alignment pipelines should treat annotations as situated, provenance-rich claims rather than neutral ground truths, and selectively incorporate reflexive metadata to make model behaviour more plural, interpretable, and accountable.

9 Endmatter Section

9.1 Author Positionality

We unpack our social position according to Savin-Baden & Major's [90] and locate ourselves in relation to the research subject, the participants, and the wider research process. Our team combines critical/feminist design scholarship with computer science expertise in crowd work and LLMs/NLP, shaping our approach to annotation and its role in AI alignment. We acknowledge approaching crowd work primarily from an academic perspective. None of us relies on crowd work for income, though some of us have previously participated in sessions to better understand the experience. We engage with these issues largely as researchers with institutional privilege, able to purchase time on platforms rather than depending on them for our livelihood. This stance reproduces the power imbalances inherent in crowd work, where anonymity and platform design obscure the people behind the labour. Our interpretative stance is informed by feminist commitments and Western European academic contexts. Several team members identify as female, most as white, shaping how we consider fairness, reflexivity, and pluralism in annotation. We also adopt a strongly qualitative approach, which, while less common in crowd work research, is essential for addressing our research questions.

9.2 Author Contributions

Anne Arzberger led the conceptualisation of the research, designed the study, and collected the data, with iterative input from Alessandro Bozzon, Jie Yang, and Ujwal Gadiraju. Anne Arzberger and Céline Offerman jointly analysed the data. Anne Arzberger drafted the manuscript, while Céline Offerman contributed to writing the results section, as well as to revising and editing the paper. Alessandro Bozzon, Jie Yang, and Ujwal Gadiraju provided iterative feedback on the manuscript.

9.3 Generative AI Use Disclosure

During manuscript preparation, the authors used ChatGPT (OpenAI; GPT-4–class model) 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.

9.4 Acknowledgments

We are very grateful for all the crowd workers who participated in our study and who trusted us with their often intimate and sometimes painful experiences. Without them, this work would not have been possible. We want to thank Shreyan Biswas for his great support with setting up and running the web application, and Catalina Lagos Rojas for her thoughtful contributions to early-stage study design. We also thank our colleagues who provided peer feedback along the way, in particular Fabio Figoli, Himanshu Verma, Ariane Lucchini, and Karin Bogdanova. 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).

References

[1] Amazon Mechanical Turk — mturk.com. https://www.mturk.com/. [Accessed 30-12-2025].
[2] Fairness. Cambridge Dictionary, Cambridge University Press. Accessed: 2025-09-05.
[3] Prolific | Easily collect high-quality data from real people — prolific.com. https://www.prolific.com/. [Accessed 30-12-2025].
[4] A. Adam. Deleting the subject: A feminist reading of epistemology in artificial intelligence. *Minds and Machines*, 10(2):231–253, 2000.

[5] M. M. AlEmadi and W. Zaghouani. Emotional toll and coping strategies: Navigating the effects of annotating hate speech data. In *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies@ LREC-COLING 2024*, pages 66–72, 2024.

[6] S. Alipour, I. Sen, M. Samory, and T. Mitra. Robustness and confounders in the demographic alignment of llms with human perceptions of offensiveness. *arXiv preprint arXiv:2411.08977*, 2024.

[7] E. Anderson. Feminist Epistemology and Philosophy of Science. In E. N. Zalta and U. Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition, 2024.

[8] C. Apicella, A. Norenzayan, and J. Henrich. Beyond weird: A review of the last decade and a look ahead to the global laboratory of the future, 2020.

[9] L. Aroyo, A. Taylor, M. Diaz, C. Homan, A. Parrish, G. Serapio-García, V. Prabhakaran, and D. Wang. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342, 2023.

[10] 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.

[11] 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.

[12] A. Arzberger, M. L. Lupetti, and E. Giaccardi. Reflexive data curation: Opportunities and challenges for embracing uncertainty in human–ai collaboration. *ACM Transactions on Computer-Human Interaction*, 31(6):1–33, 2024.

[13] M. Asad. Prefigurative design as a method for research justice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–18, 2019.

[14] S. Bardzell. Feminist hci: taking stock and outlining an agenda for design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1301–1310, 2010.

[15] E. P. Baumer. Reflective informatics: conceptual dimensions for designing technologies of reflection. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 585–594, 2015.

[16] R. Berger. Now i see it, now i don't: Researcher's position and reflexivity in qualitative research. *Qualitative research*, 15(2):219–234, 2015.

[17] R. J. Bernstein. *The restructuring of social and political theory*. University of Pennsylvania Press, 1978.

[18] L. Biester, V. Sharma, A. Kazemi, N. Deng, S. Wilson, and R. Mihalcea. Analyzing the effects of annotator gender across nlp tasks. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, pages 10–19, 2022.

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

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

[21] S. A. Cambo and D. Gergle. Model positionality and computational reflexivity: Promoting reflexivity in data science. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2022.

[22] C. Cant, J. Muldoon, and M. Graham. *Feeding the machine: The hidden human labor powering AI*. Bloomsbury Publishing USA, 2024.

[23] D. W. Carbado. Colorblind intersectionality. *Signs: Journal of Women in Culture and Society*, 38(4):811–845, 2013.

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

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

[26] P. H. Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*. routledge, 2022.

[27] K. W. Crenshaw. Mapping the margins: Intersectionality, identity politics, and violence against women of color. In *The public nature of private violence*, pages 93–118. Routledge, 2013.

[28] A. M. Davani, M. Díaz, and V. Prabhakaran. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110, 2022.

[29] N. Deng, X. F. Zhang, S. Liu, W. Wu, L. Wang, and R. Mihalcea. You are what you annotate: Towards better models through annotator representations. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[30] M. Díaz, I. Johnson, A. Lazar, A. M. Piper, and D. Gergle. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018.

[31] C. D'ignazio and L. F. Klein. *Data feminism*. MIT press, 2023.

[32] J. Dunlosky and J. Metcalfe. *Metacognition*. Sage Publications, 2008.

[33] H. Ekbia and B. Nardi. Heteromation and its (dis) contents: The invisible division of labor between humans and machines. *First Monday*, 2014.

[34] S. Fazelpour and W. Fleisher. The value of disagreement in ai design, evaluation, and alignment. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 2138–2150, 2025.

[35] E. Fleisig, R. Abebe, and D. Klein. When the majority is wrong: Modeling annotator disagreement for subjective tasks. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6715–6726, Singapore, Dec. 2023. Association for Computational Linguistics.

[36] E. Fleisig, S. L. Blodgett, D. Klein, and Z. Talat. The perspectivist paradigm shift: Assumptions and challenges of capturing human labels. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies (Volume 1: Long Papers), pages 2279–2292, 2024.

[37] D. E. Forsythe. Engineering knowledge: The construction of knowledge in artificial intelligence. *Social studies of science*, 23(3):445–477, 1993.

[38] C. Frauenberger. Critical realist hci. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 341–351, 2016.

[39] D. A. Freedman. Ecological inference and the ecological fallacy. *International Encyclopedia of the social & Behavioral sciences*, 6(4027-4030):1–7, 1999.

[40] S. Frenda, G. Abercrombie, V. Basile, A. Pedrani, R. Panizzon, A. T. Cignarella, C. Marco, and D. Bernardi. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746, 2025.

[41] L. A. Fujii. *Killing neighbors: Webs of violence in Rwanda*. Cornell University Press, 2017.

[42] U. Gadiraju, A. Checco, N. Gupta, and G. Demartini. Modus operandi of crowd workers: The invisible role of microtask work environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):1–29, 2017.

[43] B. Gaver, T. Dunne, and E. Pacenti. Design: cultural probes. *interactions*, 6(1):21–29, 1999.

[44] S. J. Gentles, S. M. Jack, D. B. Nicholas, and K. A. McKibbon. Critical approach to reflexivity in grounded theory. *The Qualitative Report*, 19(44):1–14, 2014.

[45] M. Geva, Y. Goldberg, and J. Berant. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 1161–1166. Association for Computational Linguistics, 2019.

[46] B. Ghai, Q. V. Liao, Y. Zhang, and K. Mueller. Measuring social biases of crowd workers using counterfactual queries. *arXiv preprint arXiv:2004.02028*, 2020.

[47] M. L. Gray and S. Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Harper Business, 2019.

[48] D. Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective 1. In *Women, science, and technology*, pages 455–472. Routledge, 2013.

[49] S. Harding. "strong objectivity": A response to the new objectivity question. *Synthese*, 104:331–349, 1995.

[50] S. Harding. Rethinking standpoint epistemology: What is "strong objectivity"? In *Feminist epistemologies*, pages 49–82. Routledge, 2013.

[51] S. G. Harding. *The feminist standpoint theory reader: Intellectual and political controversies*. Psychology Press, 2004.

[52] M. Henry, P. Higate, and G. Sanghera. Positionality and power: The politics of peacekeeping research. *International Peacekeeping*, 16(4):467–482, 2009.

[53] E. Herrewijnen, D. Nguyen, F. Bex, and K. van Deemter. Human-annotated rationales and explainable text classification: a survey. *Frontiers in Artificial Intelligence*, 7:1260952, 2024.

[54] B. Hooks. *Feminist theory: From margin to center*. Pluto Press, 2000.

[55] 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. 1993.

[56] C. Hube, B. Fetahu, and U. Gadiraju. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[57] L. Irani, J. Vertesi, P. Dourish, K. Philip, and R. E. Grinter. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1311–1320, 2010.

[58] D. Jacobson and N. Mustafa. Social identity map: A reflexivity tool for practicing explicit positionality in critical qualitative research. *International Journal of Qualitative Methods*, 18:1609406919870075, 2019.

[59] S. Kapania, A. S. Taylor, and D. Wang. A hunt for the snark: Annotator diversity in data practices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2023.

[60] J. Kay, A. Kasirzadeh, and S. Mohamed. Epistemic injustice in generative ai. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 684–697, 2024.

[61] U. Khurana, E. Nalisnick, A. Fokkens, and S. Swayamdipta. Crowd-calibrator: Can annotator disagreement inform calibration in subjective tasks? *arXiv preprint arXiv:2408.14141*, 2024.

[62] L. H. Kidder and M. Fine. Qualitative and quantitative methods: When stories converge. *New directions for program evaluation*, 1987(35):57–75, 1987.

[63] H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, 6(4):383–392, 2024.

[64] H. R. Kirk, A. Whitefield, P. Röttger, A. Bean, K. Margatina, J. Ciro, R. Mosquera, M. Bartolo, A. Williams, H. He, et al. The prism alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37:105236–105344, 2024.

[65] T. Lazovich. Filter bubbles and affective polarization in user-personalized large language model outputs. In *Proceedings on*, pages 29–37. PMLR, 2023.

[66] E. Leonardelli, S. Menini, A. Palmero Aprosio, M. Guerini, S. Tonelli, et al. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539. Association for Computational Linguistics, 2021.

[67] Y. Luo, D. Card, and D. Jurafsky. Detecting stance in media on global warming. *arXiv preprint arXiv:2010.15149*, 2020.

[68] A. Mateescu and M. Elish. Ai in context: the labor of integrating new technologies. 2019.

[69] 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.

[70] M. Miceli, M. Schuessler, and T. Yang. Between subjectivity and imposition: Power dynamics in data annotation for computer vision. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–25, 2020.

[71] S. Mohamed, M.-T. Png, and W. Isaac. Decolonial ai: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33:659–684, 2020.

[72] N. Mokhberian, M. Marmarelis, F. Hopp, V. Basile, F. Morstatter, and K. Lerman. Capturing perspectives of crowdsourced annotators in subjective learning tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7337–7349, 2024.

[73] T. Nagel. *The view from nowhere*. oxford university press, 1989.

[74] F. M. Olmos-Vega, R. E. Stalmeijer, L. Varpio, and R. Kahlke. A practical guide to reflexivity in qualitative research: Amee guide no. 149. *Medical teacher*, 45(3):241–251, 2023.

[75] M. Orlikowski, P. Röttger, P. Cimiano, and D. Hovy. The ecological fallacy in annotation: Modeling human label variation goes beyond sociodemographics. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023.

[76] 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.

[77] T. Pachirat. The tyranny of light. *Qualitative & Multi-Method Research*, 13(1), 2015.

[78] D. Patton, P. Blandfort, W. Frey, M. Gaskell, and S. Karaman. Annotating social media data from vulnerable populations: Evaluating disagreement between domain experts and graduate student annotators. 2019.

[79] J. Pei and D. Jurgens. When do annotator demographics matter? measuring the influence of annotator demographics with the popquorn dataset. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, 2023.

[80] A. S. G. Pessoa, E. Harper, I. S. Santos, and M. C. D. S. Gracino. Using reflexive interviewing to foster deep understanding of research participants' perspectives. *International journal of qualitative methods*, 18:1609406918825026, 2019.

[81] D. E. Pozen. The mosaic theory, national security, and the freedom of information act. *Yale LJ*, 115:628, 2005.

[82] V. Prabhakaran, A. M. Davani, and M. Diaz. On releasing annotator-level labels and information in datasets. In *Proceedings of the joint 15th linguistic annotation workshop (LAW) and 3rd designing meaning representations (DMR) workshop*, pages 133–138, 2021.

[83] 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.

[84] C. E. Rees, P. E. Crampton, and L. V. Monrouxe. Re-visioning academic medicine through a constructionist lens. *Academic Medicine*, 95(6):846–850, 2020.

[85] W. S. Robinson. Ecological correlations and the behavior of individuals. *International journal of epidemiology*, 38(2):337–341, 2009.

[86] F. Rodrigues and F. Pereira. Deep learning from crowds. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[87] M. Sabou, K. Bontcheva, L. Derczynski, and A. Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *LREC*, pages 859–866, 2014.

[88] N. Sambasivan, E. Arnesen, B. Hutchinson, T. Doshi, and V. Prabhakaran. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 315–328, 2021.

[89] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. Smith. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. 2022.

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

[91] S. Shehata. Ethnography, identity, and the production of knowledge. In *Interpretation and Method*, pages 209–227. Routledge, 2015.

[92] C. R. Showden and S. Majic. *Youth who trade sex in the US: Intersectionality, agency, and vulnerability*. Temple University Press, 2018.

[93] D. Simandan. Revisiting positionality and the thesis of situated knowledge. *Dialogues in human geography*, 9(2):129–149, 2019.

[94] J. Soedirgo and A. Glas. Toward active reflexivity: Positionality and practice in the production of knowledge. *PS: Political Science & Politics*, 53(3):527–531, 2020.

[95] L. Song and J. R. Hill. A conceptual model for understanding self-directed learning in online environments. *Journal of interactive online learning*, 6(1):27–42, 2007.

[96] T. Sorensen, J. Moore, J. Fisher, M. Gordon, N. Mireshghallah, C. M. Rytting, A. Ye, L. Jiang, X. Lu, N. Dziri, et al. Position: a roadmap to pluralistic alignment. pages 46280–46302, 2024.

[97] 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.

[98] L. A. Suchman. *Plans and situated actions: The problem of human-machine communication*. Cambridge university press, 1987.
[99] Z. Talat. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142, 2016.
[100] Z. Talat, H. Blix, J. Valvoda, M. I. Ganesh, R. Cotterell, and A. Williams. On the machine learning of ethical judgments from natural language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 769–779. Association for Computational Linguistics, 2022.
[101] Z. Talat and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.
[102] A. S. Taylor, S. Lindley, T. Regan, D. Sweeney, V. Vlachokyriakos, L. Grainger, and J. Lingel. Data-in-place: Thinking through the relations between data and community. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2863–2872, 2015.
[103] C. Toxtli, S. Suri, and S. Savage. Quantifying the invisible labor in crowd work. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–26, 2021.
[104] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021.
[105] M. van der Meer, N. Falk, P. K. Murukannaiah, and E. Liscio. Annotator-centric active learning for subjective NLP tasks. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18537–18555, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.
[106] J. Vertesi and P. Dourish. The value of data: considering the context of production in data economies. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pages 533–542, 2011.
[107] R. Wan, H. Wang, T.-H. Huang, and J. Gao. From noise to nuance: Enriching subjective data annotation through qualitative analysis. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+ NLP)*, pages 240–254, 2025.
[108] D. Wang, S. Prabhat, and N. Sambasivan. Whose ai dream? in search of the aspiration in data annotation. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–16, 2022.
[109] F. L. Wauthier and M. Jordan. Bayesian bias mitigation for crowdsourcing. *Advances in neural information processing systems*, 24, 2011.
[110] T. C. Weerasooriya, A. Ororbia, R. Bhensadadia, A. KhudaBukhsh, and C. Homan. Disagreement matters: Preserving label diversity by jointly modeling item and annotator label distributions with disco. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4679–4695, 2023.
[111] 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.
[112] D. Yanow. Interpretive analysis and comparative research. In *Comparative policy studies: Conceptual and methodological challenges*, pages 131–159. Springer, 2014.
[113] D. Yanow. Thinking interpretively: Philosophical presuppositions and the human sciences. In *Interpretation and method*, pages 5–26. Routledge, 2015.
[114] L. H. Zhang, S. Milli, K. L. Jusko, J. Smith, B. Amos, W. Bouaziz, J. Kussman, M. Revel, L. Titus, B. Radharapu, et al. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *2nd Workshop on Models of Human Feedback for AI Alignment*.

A Participant Demographics

Table 1 presents the demographics of the participants who engaged in our prolific main study.

Age (years)	$M = 34.4$, $SD = 12.9$, range 19–67
Gender	Male: 14 (47%) Female: 13 (43%) Non-binary: 3 (10%)
Ethnicity	White: 17 (57%) Black: 10 (33%) Asian: 1 (3%) Mixed: 1 (3%) Other: 1 (3%)
Ability	Abled: 21 (70%) Disabled: 9 (30%)

Table 1. Participant demographics for the main study ($N = 30$). Percentages are of N.

Table 2 presents the demographics of the participants who were interviewed after their participation in our main prolific study.

Age (years)	$M = 31$, $SD = 10.35$, range 20–44
Gender	Female: 4 (80%) Male: 1 (20%)
Ethnicity	White: 2 (40%) Black: 2 (40%) Asian: 1 (20%)
Ability	Abled: 3 (60%) Disabled: 2 (40%)

Table 2. Participant demographics for the follow-up interview, with a subsample from the Prolific participants ($N = 5$). Percentages are of N.

B Follow-up Interview Protocol

In what follows, we present the interview protocol for our semi-structured interviews with the crowd workers.

B.1 1. Introduction

A few questions to get to know the participant/ease them into the interview:

- Could you tell me a bit about yourself?
- How did you first get into crowd work?
- What kind of tasks do you typically do, and how long have you been doing this work?

B.2 2. Before the Task

These questions explore the participant's thoughts before starting the annotation:

- Before you started annotating, what did you expect from the task?
- Did you consider how aspects of your identity—like gender, ethnicity, or personal experiences—might shape how annotate? [follow-up] What reflections did that bring up for you?
- How would you define "social identity"? [follow-up] (If needed, clarify that in this context it refers to aspects like background, gender, or ability—things that shape how we see the world to create a shared understanding of the word for the remaining interview.)

B.3 3. During the Task – Reflections

These questions allow us to revisit some of the annotation tasks the participant performed during the prolific study, including examples from their responses.

- Can you describe all the tasks you did (from start to finish) in your own words?
- In comparison to usual tasks you perform on prolific, was this exceptionally time intensive?
- And how was the cognitive load?
- How did you feel overall while doing the reflection exercises?

With respect to Tier 1& 2 annotation tasks: [example selected from participant responses]

- Was anything especially easy or difficult? Helpful or confusing?
- Did you feel like you could express your identity comfortably here?
- What did you learn about yourself?
- Was it emotionally difficult?
- Was it uncomfortable to share such intimate information?

With respect to Tier 3 annotation task: [example selected from participant responses]

- Can you describe this experience in three words? Why those?
- What helped you connect this fairness reflection to your social identity?
- Were there any personal memories or feelings that came up?
- Did this feel meaningful? Challenging?
- (Was anything surprising?)
- How do you know that you provided a complete list of relevant identity facets? That there is none that would be better fitting to describe your perception here?
- If you could redesign this tool or process, what would you change?

B.4 4. After the Task - Reflections

The following question relates to the participants' thoughts and feelings after the annotation tasks.

- Would you have liked to adjust your fairness ranking after you analysed the text in more detail? [follow-up] If so, how?
- Has this reflection process changed anything in how you view your annotation work?
- Are there any personal assumptions and biases that you became aware of?
- Will any of the insights you gained potentially affect future annotation tasks you perform? [follow-up] How?

B.5 5. Reflecting on Specific Moments

These questions aim to foster reflection on specific moments from the participants' annotations where something stood out—either because it was unclear, surprising, or where we saw a bit of tension or complexity. 1–3 relevant moments from below are selected, related to the participants' annotation responses. The most insightful or ambiguous ones are chosen.

Misalignment / Unclear Reasoning [add example]

- Here's a moment where your answer and explanation seemed a bit at odds. Can you walk me through what you were thinking?
- Looking back, do you think you'd approach it differently now?

Lower Engagement or Ambiguity [add example]

- Here, your response was brief or detached. Was that intentional, or was something unclear or unengaging at that moment?
- What could've helped keep your focus or interest?

Misunderstanding or Instructional Confusion [add example]

- Did any parts of the task or wording feel confusing or ambiguous?
- How did that affect what you wrote or thought?

Struggling with Expression / Misarticulation [add example]

- Sometimes it's hard to put our thoughts into words—did you experience that at any point?
- Were there moments where you felt your reflection didn't fully capture what you meant?
- What support or changes could help make it easier to express yourself?

Multiple Identity Facets / Intersectionality [add example]

- Here you mentioned more than one aspect of your identity. Can you tell me what you were thinking when adding those two facets?

Detached Framing [add example]

- In this part, you pointed out an issue, but didn't link it to your own perception or personal identity. Did you see it as personally relevant? Why or why not?

Avoiding Reflexivity [add example]

- At this moment, it seemed like the task didn't fully resonate with you. Did the reflection feel meaningful to you? Why or why not?
- What could make such a task feel more relevant or engaging for you?

B.6 Closing

- Is there anything else I didn't ask that feels important to share?

C Text Pieces for Fairness Ranking Task

The following text samples were ranked by the crowd workers in terms of fairness. They were presented in randomised order. Text sample a) resembles a job vacancy, while sample b) relates to an advertisement of a skin care product.

C.1 Text Sample a) Job Vacancy

The job vacancy: *"We are seeking a recent graduate with a youthful and energetic attitude to support our senior management team at a leading tech company. The ideal candidate will excel in managing executives' demanding schedules, preparing polished, high-level documents and presentations, coordinating meetings seamlessly, and maintaining absolute discretion. Candidates must have native English proficiency and demonstrate impeccable grooming and a professional demeanour at all times. Proficiency in office software is not necessary, but women are especially encouraged to apply, as we value the multitasking and organisational skills they often bring. This role suits someone willing to work long hours, thrives in fast-paced environments, and is ready to align themselves with the vision of a high-achieving, male-dominated executive team."*

C.2 Text Sample b) Product Advertisement

The advertisement: *"Experience the ultimate self-care with our new line of luxury skincare products, specially designed to meet the unique needs of soft, sensitive skin. From rejuvenating night creams to hydrating serums, our products are made for those who understand the importance of embracing their natural beauty and taking the time to prioritise themselves. Join the countless users who trust us to support their journey toward healthier, glowing skin. Because everyone deserves to feel confident in their own skin."*

D Conceptual Model of Reflexive Thematic Analysis

Figure 7 illustrates the analysis steps taken. In practice, steps did not follow so linearly; instead, data or codes were iteratively revisited when necessary.

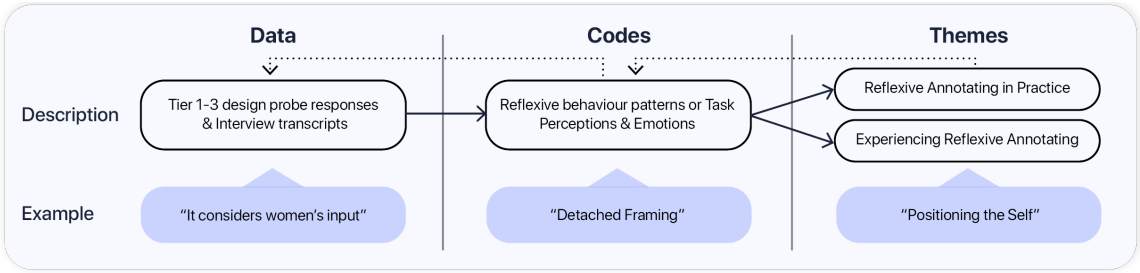


Fig. 7. This figure visualises our reflexive thematic analysis as a progression from *data*, to *codes*, to *themes*. The left column represents the materials analysed in the study (participants’ probe responses and interview transcripts). From these materials, we inductively generated codes of recurring patterns in how annotators described their reasoning and feelings towards reflexive annotating. These interpretive 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 developed and revised through repeated engagement with the dataset, and themes emerged through examining relationships across codes. The figure therefore provides a conceptual overview of how we moved from concrete participant accounts to interpretive coding and theme development in our reflexive thematic analysis.