# Participatory Moral AI Is Not Neutral: The Invisible Hand of Developers

**Taenyun Kim**[1], **Edyta Bogucka**[2], **Daniele Quercia**[2,3]
[1]Michigan State University, US
[2]Nokia Bell Labs, Cambridge, UK
[3]Politecnico di Torino, Italy
kimtaeny@msu.edu, edyta.bogucka@nokia-bell-labs.com, daniele.quercia@nokia-bell-labs.com

**Abstract**

As AI systems make more morally loaded decisions across society, one response has been moral preference elicitation. In this approach, researchers poll participants on hypothetical dilemmas and use the aggregated votes to train a policy that an AI model then applies at scale. Before any vote is cast, developers make three key choices in the moral AI elicitation pipeline: feature scoping, voter sampling, and question framing. In other words, they decide which features go to a vote, which voters to include, and how to present the question. These choices are often opaque, undocumented, and treated as technical details rather than normative ones. We examine each of these choices within a common empirical study and show that each can shape the preferences produced by moral AI elicitation. Across two phases (N = 809) in three deployment contexts (i.e., AI kidney allocation, AI agents simulating absent workers, and generative AI depictions of the deceased), we examine the three main stages of the moral AI elicitation pipeline. First, morally relevant features shift across contexts. This suggests that feature schemas should not be assumed to transfer across deployment domains. Second, preferences differ by political ideology for roughly one-third of features, with some differences reversing direction. The ideological composition of the voter pool can therefore affect the resulting aggregated preference profile. Third, the wording of the elicitation question can narrow or widen ideological gaps by up to a full scale point. The framing conditions also change how moral foundations are associated with participants' judgments. Taken together, these findings suggest that voting-based alignment cannot deliver fair or transparent AI by aggregation alone; at minimum, each stage of the moral AI elicitation pipeline should be audited and disclosed.

**Project website** — https://social-dynamics.net/moral-ai

## 1 Introduction

AI systems are increasingly used to make high-stakes decisions in areas such as healthcare, employment, and criminal justice (Chan et al. 2024; Kim and Peng 2025). Aligning these systems with human moral values in a way that is fair, accountable, and transparent has therefore become a central problem (Boerstler et al. 2024; Freedman et al. 2020; Kneer and Viehoff 2025). To address it, *moral preference elicitation* now stands as a common approach: researchers present participants with hypothetical moral dilemmas, ask them a question about what the AI system should do, and use these responses to train models that produce a decision policy (Boerstler et al. 2024; Freedman et al. 2020; Kneer and Viehoff 2025; Noothigattu et al. 2018; Awad et al. 2018). The approach aims to democratize alignment by grounding model design in the preferences of a broad and representative public (Simson et al. 2025; Dahl and Svanæs 2020).

In practice, however, before any question is asked, developers make three key choices that shape what the final policy can capture: *feature scoping* (which features are included), *voter sampling* (which participants are included), and *question framing* (how the question is asked). These choices are rarely documented, and are typically treated as technical rather than moral decisions. Each of these choices can shape what is ultimately presented as an "aggregated moral preference". If the selected features, participant population, or question wording affect the results, the resulting policy reflects upstream developer decisions as well as participants' expressed values. We examine these three choices within a common empirical study, using one research question for each stage:

**RQ$_1$:** *Feature scoping.* Do people consider the same features across different use cases?

**RQ$_2$:** *Voter sampling.* Do moral preferences vary with political ideology?

**RQ$_3$:** *Question framing.* Does framing change moral preferences?

To address these questions, we conducted a study with participants holding different political views across three AI use cases: AI kidney allocation (KIDNEY), AI agents simulating absent workers (WORK), and generative AI content of the deceased (GEN). These cases vary in the severity of harm and everyday frequency of encounter, from rare high-stakes allocation to common workplace decisions and emerging questions about how AI should represent the deceased. We ran two phases for each use case. In Phase 1 ($N = 150, 149, 150$ per use case), participants identified which features should or should not matter. In Phase 2 ($N = 120$ per use case), participants judged whether each feature should count in favor, count against, or not count at all, under one of three conditions: a control, World-You-Want (a structural perspective (Jaques 2025)), or Could-Be-You (a perspective-taking approach based on Rawlsian theory (Rawls 1971; Huang, Greene, and Bazer-

man 2019; Bruno et al. 2024)). We then classified features as relevant, irrelevant, or divisive, and tested whether political ideology and framing changed these judgments. This analysis produced three main findings:

1) **Moral feature relevance is context-specific (§5.1).** Relevant features vary by use case. KIDNEY emphasizes distributive justice and medical utility. WORK focuses on accountability and legitimacy. GEN centers on consent and dignity. The relevant feature sets differed across the three contexts. Choosing which features to include is therefore a moral decision that defines the boundaries of the elicitation process.
2) **Voter sampling can change the aggregated preference profile (§5.2).** Political ideology is associated with different evaluations for roughly one-third of features, including some differences that reverse direction. In KIDNEY, conservatives place greater weight on proportionality and utility, while progressives place greater weight on equality. In WORK, conservatives emphasize employer loyalty and contractual duty. In GEN, conservatives place greater weight on the dignity of the deceased, while progressives are more permissive toward creative use. The ideological composition of the voter pool can therefore affect the preferences represented in the aggregate.

   **3) Question framing changes ideological gaps and value pathways (§5.3).** Framing can change results by up to one scale point, narrowing some ideological gaps while widening others. The framing conditions also change how moral foundations are associated with participants' judgments. Question framing should therefore be treated as a consequential design choice rather than a neutral elicitation tool.

Taken together, these findings show that voting-based alignment does not simply recover a pre-existing set of public moral preferences: its outputs are shaped by upstream choices about what is put to a vote, whose preferences are represented, and how those preferences are elicited. This paper makes three contributions. First, it conceptualizes feature scoping, voter sampling, and question framing as normative developer choices in moral AI elicitation. Second, it provides comparative empirical evidence about each choice across three deployment contexts. Third, it translates these findings into a sensitivity audit that makes upstream choices visible and contestable (§5.4), while examining the limits of aggregation even when those choices are disclosed (§5.5).

# 2 Related Work

Next, we review prior work on voting-based AI alignment and the three developer choices that structure it: *feature scoping*, *voter sampling*, and *question framing*.

## 2.1 Voting-Based AI Alignment and Its Assumptions

Aligning AI systems with human moral values is a central concern for AI safety research (Gabriel 2020; Kneer and Viehoff 2025). One common approach is *moral preference elicitation*, in which researchers present hypothetical dilemmas, collect participants' responses, and aggregate them into a decision policy that a model applies at scale (Noothigattu et al. 2018; Awad et al. 2018; Freedman et al. 2020; Keswani et al. 2025). This approach is often framed as democratizing alignment by replacing developers' private judgments with public input (Simson et al. 2025; Dahl and Svanæs 2020).

Yet developers still decide what participants evaluate, who participates, and how questions are posed. We call these choices *feature scoping*, *voter sampling*, and *question framing*. Claims that aggregation neutrally recovers public values assume that features transfer across contexts, voter composition has limited influence, and wording reveals rather than shapes values. Prior work challenges these assumptions: aggregation can marginalize minority views (Feffer, Heidari, and Lipton 2023a), while demographic groups prioritize ethical values differently (Jakesch et al. 2022). Anthropic's Constitutional AI illustrates the issue: although 1000 Americans contributed to a model constitution, developers still selected participants, statements, and training principles (Ganguli et al. 2023). More fundamentally, aggregated preferences cannot determine which views deserve authority, how conflicts should be resolved, or what protections should constrain majority decisions (Salloch et al. 2015; Harris 2020; Baum and Slavkovik 2025; Zhi-Xuan et al. 2024). Taken together, this suggests preferences should be treated as one input into moral reasoning rather than as a substitute for it. Our study does not attempt to resolve this problem. Instead, it examines how upstream design choices further condition what is presented as an aggregated preference.

## 2.2 Moral Preferences Across AI Use Cases

Most moral preference elicitation studies focus on a small set of use cases that tend to share three properties: they involve severe harm if the AI fails or is misused, they are rare or one-off situations, and they concern decisions that humans have traditionally made case by case. Organ allocation and autonomous vehicle scenarios are the clearest examples of this pattern, with concerns centered on life-and-death tradeoffs and the allocation of scarce resources (Noothigattu et al. 2018; Awad et al. 2018). More frequent, lower-stakes cases such as content moderation, ad targeting, and consumer AI have received far less attention, yet they raise different moral concerns around transparency, consent, and accountability (Jakesch et al. 2022).

Within a use case, feature scoping can follow two approaches: top-down and bottom-up. In a top-down approach, developers select features before recruiting participants, drawing on existing literature or industry guidelines. In a bottom-up approach, participants themselves identify which considerations they find morally relevant (Sadek, Calvo, and Mougenot 2023). Most studies have used the top-down approach. For example, Awad et al. (2018) reviewed the Western academic literature to identify 18 features describing the people affected by an autonomous vehicle's decision, including their gender, age, and physical fitness. Critics argue that this forced-choice format narrows the ethical problem by preselecting which people, outcomes, and trade-offs participants can consider, while omitting factors such as

uncertainty and how likely each outcome is (Dewitt, Fischhoff, and Sahlin 2019; Etienne 2021; Freitas et al. 2021; Schuessler 2023). Similarly, Jakesch et al. (2022) drew on industry ethics guidelines to identify 12 values relevant to responsible AI in ad targeting, including safety and fairness. In both cases, what goes to a vote was decided before any participant was recruited, and researchers have called for such decisions to be documented and opened to scrutiny (Feffer, Heidari, and Lipton 2023a; Arzberger et al. 2025).

### 2.3 Group Differences in Moral Preferences

Differences in moral judgment across groups are well documented. In moral preference elicitation tasks, participants with different demographic, AI literacy, or political backgrounds often reach different conclusions (Graham, Haidt, and Nosek 2009; Atari et al. 2023). This diversity poses a challenge for systems that attempt to learn a single set of "universal" moral preferences (Gabriel 2020; Jakesch et al. 2022). Disagreements arise at two levels: people differ in which features they consider relevant (Keswani et al. 2025), and they differ in how they weight those features (Brugman 2024). For example, when building AI systems, practitioners flag safety and privacy as more relevant than the general public does (Jakesch et al. 2022), and developers weight politeness over the straightforwardness that the users they build for actually want (Ranjit et al. 2026).

These differences are especially pronounced across political ideology. Moral Foundations Theory (MFT) suggests that progressives emphasize *care* and *equality*, whereas conservatives place greater weight on *proportionality*, *authority*, *loyalty*, and *purity* (Graham, Haidt, and Nosek 2009; Atari et al. 2023). These moral differences shape how people interact with AI systems (Brailsford, Vetere, and Velloso 2024). In automated vehicle scenarios, for example, sacrificial preferences vary by political ideology (Awad et al. 2018). Taken together, these findings suggest that the ideological composition of a voter pool can affect the values represented in the aggregated preference profile.

### 2.4 Moral Preference Influenced by Moral Framing

Moral framing is the selection and emphasis of particular aspects of a morally charged situation, shaping how people interpret and evaluate it (Brugman 2024; Semetko and Valkenburg 2000). It can influence preferences in two ways: through question framing, which emphasizes particular moral concerns, and through response framing, which constrains the responses participants can express.

Question framing can shift preferences by changing how a decision is presented or which considerations participants are prompted to weigh (Feinberg and Willer 2019; Gamson and Modigliani 1989; Rehren and Sinnott-Armstrong 2021). When evaluating an isolated decision, people respond differently when identical outcomes are framed as lives saved rather than lives lost (McDonald et al. 2021). Socratic questions can also prompt participants to justify and reconsider that individual decision (Torabizadeh, Homayuni, and Moattari 2018). Other frames shift attention from the isolated decision to the consequences of adopting it as a general policy. Jaques (2025) argues that single-case judgments can invite individual bias and obscure the effects of scaling a choice, and instead proposes asking, "What kind of world would I be creating if this became the AI policy?" (Jaques 2025). The veil of ignorance similarly asks people to evaluate a policy without knowing their own position in the resulting society (Rawls 1971; Huang, Greene, and Bazerman 2019; Huang et al. 2021; Bruno et al. 2024). Such frames have been associated with less self-serving reasoning, greater attention to stakeholder perspectives, and reduced political disagreement (Huang, Greene, and Bazerman 2019; Huang et al. 2021; Bruno et al. 2024; Franks and Scherr 2019; Whitmarsh and Corner 2017; Bloemraad, Silva, and Voss 2016).

Response framing instead limits which preferences participants can express. When forced to choose which demographic group to spare, participants appeared to support unequal treatment; when given an equal-treatment option, most selected it (Bigman and Gray 2020).

**Research Contribution.** Prior research shows that feature scoping, voter sampling, and question framing can each shape moral judgments. Building on this work, we examine the three choices as successive stages of moral AI preference elicitation within a common empirical project. Our contribution is an integrated, cross-context account of how developer choices enter the elicitation process: we identify context-specific feature scopes, estimate ideological differences in feature evaluations, and test the effects of two explicit framing interventions. We conduct these analyses across three deployment contexts that differ in moral stakes and frequency of encounter. We then use the combined evidence to develop a sensitivity audit and recommendations for voting-based alignment (§5.4).

## 3 Methods

We study moral preferences and the effects of question framing across three use cases (§3.1), three framing conditions (§3.2), and two phases. Phase 1 identifies candidate moral features (§3.3). Phase 2 evaluates the moral importance of these features under different framing conditions (§3.4).

### 3.1 Use Case Selection

We select three use cases based on two criteria (Figure 1, Step 1): impact (the scale of harm if the system fails or is misused) and temporal exposure (how often people face the dilemma). Together, they cover a range of ethical settings, from high-stakes allocation to everyday AI use:

1) **AI kidney allocation (KIDNEY)** is a life or death dilemma rooted in distributive justice. It raises questions about fairness and the role of AI in allocating scarce resources (Keswani et al. 2025). While common in moral elicitation research (Keswani et al. 2025; Boerstler et al. 2024; Freedman et al. 2020), such cases are rare in daily life (Jaques 2025; Nguyen et al. 2022).
2) **AI agents simulating absent workers (WORK)** captures a more common setting (Yudkin et al. 2025). It looks at remote workers who use AI to simulate activity. Concerns include deception, work norms, and AI compliance with questionable instructions (Boland 2025).

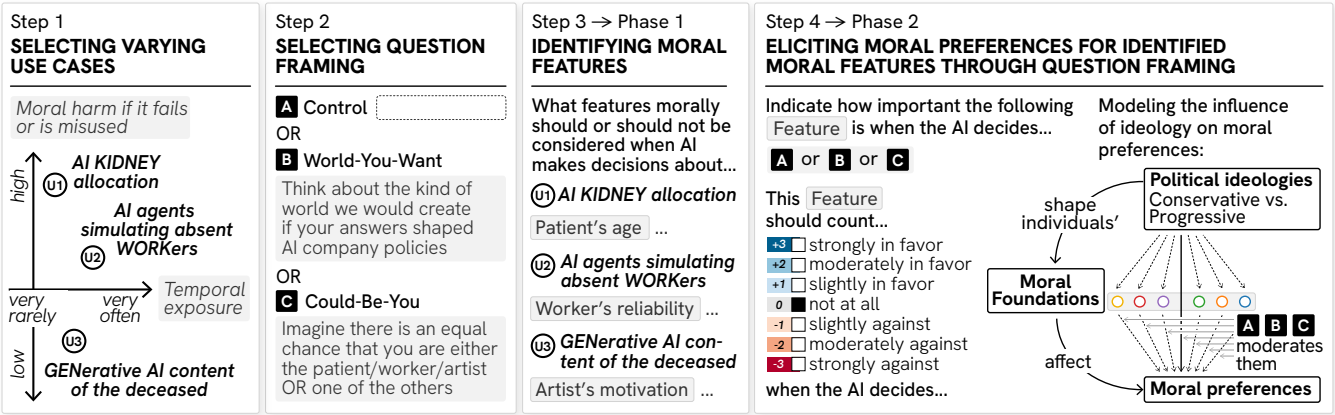



Figure 1: **Overview of our four-step methodology.** *Step 1.* We selected three use cases—AI kidney allocation (KIDNEY), AI-simulated workers (WORK), and generative AI content of the deceased (GEN)—based on harm impact and likelihood of encounter, capturing rare high-stakes to common but less harmful cases. *Step 2.* We used three framing conditions—(A) *Control* (baseline), (B) *World-You-Want* (societal consequences), and (C) *Could-Be-You* (perspective-taking)—capturing structural and perspective-taking prompts to examine how different moral lenses shape moral preference elicitation. *Step 3.* In Phase 1, participants identified moral features relevant to AI decision-making for each use case. *Step 4.* In Phase 2, we elicited moral preferences for these features from a new sample using a 7-point scale ($-3$ to $+3$) across the three framings in Step 2. We then modeled how political ideology influences these preferences through moral foundations, with framing condition as a moderator.

3) **Generative AI content of the deceased (GEN)** raises issues of consent after death, emotional harm, and commercial use (Buben 2025; Danaher and Nyholm 2025). These systems may change how people remember the dead or weaken a person's unique identity (Danaher and Nyholm 2024; Lazaridis 2025).

## 3.2 Question Framing Selection

We selected three question framing conditions for the elicitation task in Phase 2:

(A) **Control** uses no additional question framing prompt (Figure 1, Step 2A).

(B) **World-You-Want** draws on Jaques's structural perspective (Jaques 2025). It frames the elicitation question by prompting participants to consider long-term societal effects and AI policy. Participants read: "Think about the kind of world we would create if your answers shaped AI company policies" (Figure 1, Step 2B).

(C) **Could-Be-You** draws on Rawls's veil of ignorance (Rawls 1971; Huang, Greene, and Bazerman 2019; Huang et al. 2021). It frames the elicitation question by prompting participants to take the perspective of any affected stakeholder (Bruno et al. 2024). For example: "Imagine there is an equal chance that you are either one of the patients" (Figure 1, Step 2C).

## 3.3 Phase 1: Identifying Moral Features

**Participants.** We recruited participants via Prolific at a rate of at least 8 USD/hour, following sample sizes established in prior work using the same two-phase design (Keswani et al. 2025; Chan et al. 2022) ($N_{KIDNEY} = 150$, $N_{WORK} = 149$, $N_{GEN} = 150$). We used quota sampling to balance political ideology (conservative, moderate, progressive) within each use case. Full demographic breakdowns, including age, gender, race/ethnicity, and education level, are reported in Table 1 in Supplementary Materials A.

**Procedure.** After consent and a use case introduction, participants adopted a "moral point of view". They listed 5 features that morally *should* be considered and 5 that *should not*. They provided a justification for each feature. For features that should be considered, they also gave example levels (e.g., "young" and "old" for "age") (Figure 1, Step 3; Figure 5 in Supplementary Materials B).

**Analysis.** To identify moral features, we used LLM-assisted coding with `GPT-4o-mini`. We selected this model for its low cost and high performance on instruction-following tasks (Rytting et al. 2023; Ranjit et al. 2024). We ran it once per participant response at temperature 0 using the prompt in Supplementary Materials C. The LLM extracted feature names from responses, summarized participants' justifications, and assigned each feature one of four relevance labels ("should", "should not", "mixed", "unclear").

To assess LLM output quality, two authors independently coded a random sample of 100 responses, reviewing both feature names and relevance labels. Agreement between the authors and the LLM was 85% on feature names and 97% on relevance labels. To refine feature names across all outputs, authors conducted five calibration rounds following standard qualitative methods (Charmaz 2015; Oktay 2012). They resolved disagreements by consensus (with a third author as tie-breaker) and recorded all decisions. This included, for example, merging synonymous codes (e.g., "geographic location" and "travel time" into "geographic proximity").

**Feature inclusion threshold as a developer choice.** We computed feature prevalence using binary indicators (1 if mentioned by a participant). Following prior work (Keswani et al. 2025), we retained the top 30 to 35 most prevalent features, excluding those mentioned by fewer than 4 participants (≈2.5% of our sample). For example, in the KIDNEY scenario, *taxpayer status* and *favorite sports team* were each mentioned by only 1 participant and thus excluded. We treat this threshold as a design choice within *feature scoping* that may exclude minority perspectives.

### 3.4 Phase 2: Eliciting Moral Preferences Under Framing

**Participants.** We recruited 360 US participants (120 per use case) via Prolific at a rate of at least 8 USD/hour. We randomly assigned participants to the Control, World-You-Want, or Could-Be-You condition, with equal representation of conservatives and progressives per condition. Table 2 in Supplementary Materials A reports full demographic breakdowns, including age, gender, race/ethnicity, education, AI literacy, religiosity, and area of residence.

**Procedure.** After consent and a use case introduction, participants evaluated features from Phase 1 (Figure 1, Step 4). Each feature appeared as a contrast (e.g., "younger vs. older patients" for KIDNEY, "grief support vs. commercial use" for GEN). Participants rated moral importance on a 7-point scale from $-3$ (*count strongly against*) to $+3$ (*count strongly in favor*), with 0 indicating *not count at all* (see Figures 6, 7 and 8 in Supplementary Materials B). Unlike pairwise methods, which scale poorly with many features (Boerstler et al. 2024; Keswani et al. 2025), this scale allowed us to evaluate more than 30 features. This design supports analysis of how political ideology, moral foundations, and framing interact.

Participants also completed two validated measures. *Moral Foundations* used the 36-item MFQ-2 (Atari et al. 2023), which measures six foundations: Care, Equality, Proportionality, Loyalty, Authority, and Purity (Graham, Haidt, and Nosek 2009). Responses used a 5-point scale (1: *not at all*, 5: *extremely well*). *AI Literacy* used a 17-item scale (Tully, Longoni, and Appel 2025) with two components: technical understanding (10 items) and awareness of limitations and ethics (7 items). Each item was scored as correct or incorrect and scaled to $[0, 1]$. We use AI literacy as a proxy for prior experience with AI.

**Analysis.** We classified features into six categories based on relevance ($M_r, p_r$ via t-test vs. 0.5) and direction ($M_d, p_d$ via t-test vs. 0). We coded relevance as 0 ("not count at all") or 1 (any other value). We used the raw scale for direction. Categories were:

1) *Morally relevant and counting for*, if $p_r < 0.05$ and $M_r > 0.5$, and $p_d < 0.05$ and $M_d > 0$.
2) *Morally relevant and counting against*, if $p_r < 0.05$ and $M_r > 0.5$, and $p_d < 0.05$ and $M_d < 0$.
3) *Morally irrelevant*, if $p_r < 0.05$ and $M_r < 0.5$.
4) *Morally relevant but directionally divisive*, if $p_r < 0.05$ and $M_r > 0.5$, and $p_d \geq 0.05$.
5) *Morally divisive*, if $p_r \geq 0.05$.
6) *Morally divisive with directional disagreement*, if $p_r \geq 0.05$ and $p_d \geq 0.05$.

We then ran moderated mediation analyses using Hayes's PROCESS Model 15 with 5,000 bootstrap samples (v4.3 for R) (Hayes 2017). These models test how political ideology affects moral foundations, how these foundations affect moral importance, and how framing moderates these relationships (see Figure 9 in Supplementary Materials D). We control for sex (1 = female), age, education (1 = pre-college; 2 = college; 3 = advanced), ethnicity (1 = non-Hispanic White), religiosity (1 = not important; 5 = extremely important), area of residence (1 = rural), and AI literacy.

## 4 Results

### 4.1 $RQ_1$. Do People Consider the Same Features Across Different Use Cases?

The features participants considered morally relevant were highly context-specific across KIDNEY, WORK, and GEN (Figure 2).

In KIDNEY, participants focused on medical utility, such as patients' *health status* (47%) and *chance of survival* (35%). In WORK, they emphasized worker motivation (e.g., *reason for request*, 32%) and organizational factors (e.g., company policy, 27%). In GEN, they highlighted ethical safeguards, including the *deceased's consent* (28%) and *intended purpose* (40%). Full results are reported in Table 3 in Supplementary Materials E, Table 7 in Supplementary Materials F, and Table 11 in Supplementary Materials G.

Participants also showed strong agreement on features that should *not* be considered. These were mostly demographic or socioeconomic attributes, including *ethnicity* (KIDNEY: 66%; WORK: 18%; GEN: 12%), *gender* (KIDNEY: 64%; WORK: 14%; GEN: 12%), and *social status* (KIDNEY: 27%; WORK: 12%; GEN: 12%). Age was also rejected in WORK (18%) and GEN (13%), but 82% of participants in KIDNEY identified it as morally relevant. Table 4 in Supplementary Materials E, Table 8 in Supplementary Materials F, and Table 12 in Supplementary Materials G provide further details.

Phase 2 shows that most features reached consensus on moral relevance (KIDNEY: 65%; WORK: 57%; GEN: 70%). These features were typically tied to direct utility. For example, in KIDNEY, a patient with *a stronger blood and tissue match with a donor* was judged in favor ($M = 2.42$, $SD = 0.95$, $p_r < .001$, $p_d = .001$), whereas *greater responsibility for one's illness* was judged against ($M = -1.10$, $SD = 1.69$, $p_r < .001$, $p_d < .001$).

Similar patterns appear in WORK and GEN. In WORK, a worker's *emergency situation* was judged in favor ($M = 1.29$, $SD = 1.97$, $p_r < .001$, $p_d < .001$), while *disrupted workplace operations* was judged against ($M = -0.84$, $SD = 2.30$, $p_r < .001$, $p_d = .003$). In GEN, requests for *grief support or memorial purposes* ($M = 1.88$, $SD = 1.49$, $p_r < .001$, $p_d < .001$) or those that *honored the deceased* ($M = 2.18$, $SD = 1.33$, $p_r < .001$, $p_d < .001$) were judged in favor.

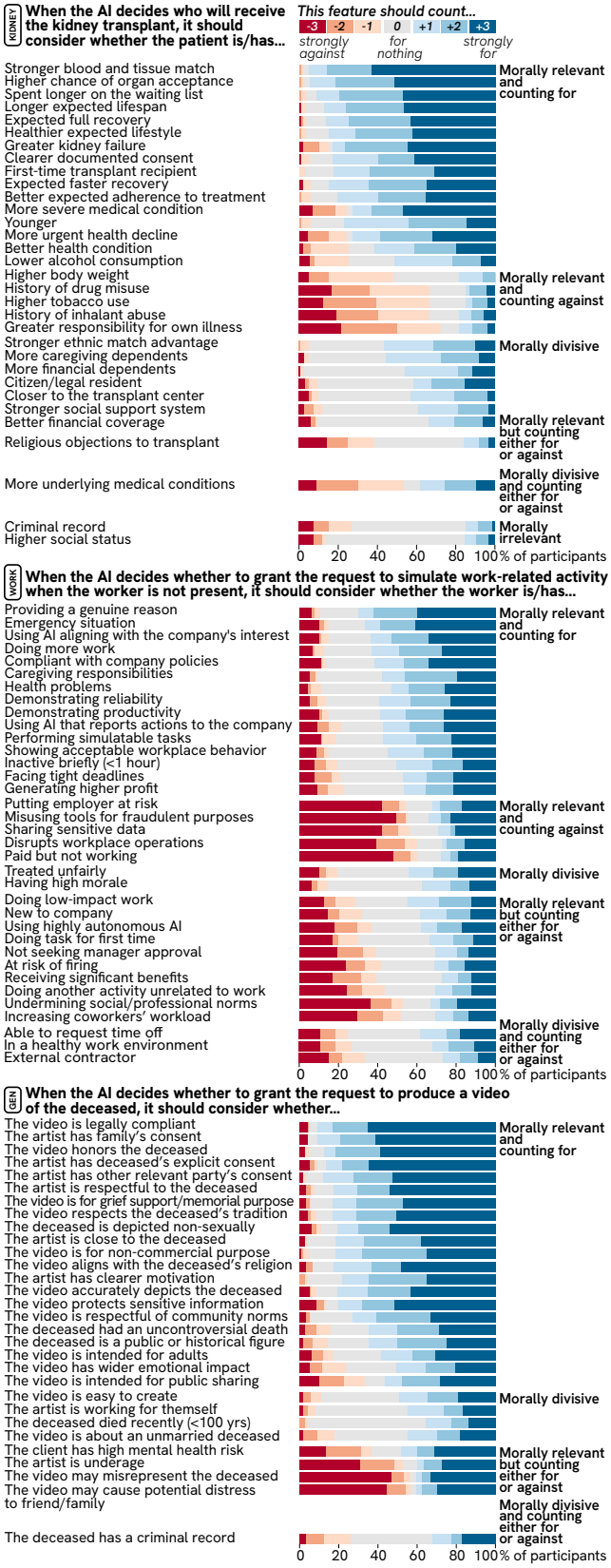



Figure 2: **Phase 2 rating distributions for KIDNEY, WORK, and GEN use cases.** Features were ordered by moral feature category and mean rating, highest to lowest.

Despite this agreement, many features remained divisive (KIDNEY: 31%; WORK: 43%; GEN: 30%). These features often had unclear or mixed implications. Participants agreed they were relevant but disagreed on direction, for or against. Examples include *more underlying medical conditions* in KIDNEY ($M = -0.18$, $SD = 1.88$, $p_r < .001$, $p_d = 1.0$), *increasing coworkers' workload* in WORK ($M = -0.57$, $SD = 2.15$, $p_r < .001$, $p_d = 1.0$), and *risk of misrepresentation* in GEN ($M = -0.45$, $SD = 2.75$, $p_r < .001$, $p_d = 0.36$). Table 5 in Supplementary Materials E, Table 9 in Supplementary Materials F, and Table 13 in Supplementary Materials G provide further details.

## 4.2 RQ$_2$. Do Moral Preferences Vary With Political Ideology?

Preferences differed by political ideology for roughly one-third of features, with some differences reversing direction. Consequently, changing the ideological composition of an aggregated sample could change the preference profile.

In the control condition, ideological differences appeared across all use cases. In KIDNEY, conservatives favored patients with *expected full recovery* ($b = 1.08$, $95\%\,CI\,[0.12, 2.03]$) and penalized those with *more underlying medical conditions* ($b = -1.71$, $95\%\,CI\,[-3.20, -0.23]$). In WORK, conservatives opposed requests from workers who were *being paid but not working* ($b = -2.79$, $95\%\,CI\,[-5.58, -0.00]$). In GEN, they were less likely to approve requests when the video risked *misrepresenting the deceased* ($b = -2.91$, $95\%\,CI\,[-5.62, -0.20]$). Table 6 in Supplementary Materials E, Table 10 in Supplementary Materials F, and Table 14 in Supplementary Materials G provide further details.

## 4.3 RQ$_3$. Does Question Framing Change Moral Preferences?

The third developer choice (*question framing*) systematically shifts preferences. It can reduce some ideological gaps while increasing others. The magnitude of these shifts reaches up to one full scale point (Figure 3).

**Both Question Framings.** In KIDNEY, disagreement about patients with *more underlying medical conditions* decreased under both framings (World-You-Want: $b = -0.56$, $95\%\,CI\,[-2.21, 1.08]$; Could-Be-You: $b = 0.61$, $95\%\,CI\,[-0.81, 2.02]$; Figure 3C).

In WORK, conservatives' opposition to workers *being paid but not working* was also reduced (World-You-Want: $b = -0.43$, $95\%\,CI\,[-2.96, 2.10]$; Could-Be-You: $b = 0.01$, $95\%\,CI\,[-2.41, 2.43]$; Figure 3E).

In GEN, both framings reduced disagreement about videos that risk *misrepresenting the deceased* (World-You-Want: $b = 1.06$, $95\%\,CI\,[-1.23, 3.35]$; Could-Be-You: $b = -1.88$, $95\%\,CI\,[-4.40, 0.63]$; Figure 3I).

**Could-Be-You Question Framing.** In KIDNEY, conservatives' preference for *expected full recovery* (present in Control and World-You-Want) disappeared under Could-Be-You ($p = .95$). This change was associated with reduced influence of purity ($b = -0.51$, $95\%\,CI\,[-1.19, -0.03]$; $IMM = -0.74$, $95\%\,CI\,[-1.68, -0.07]$; Figure 3B).

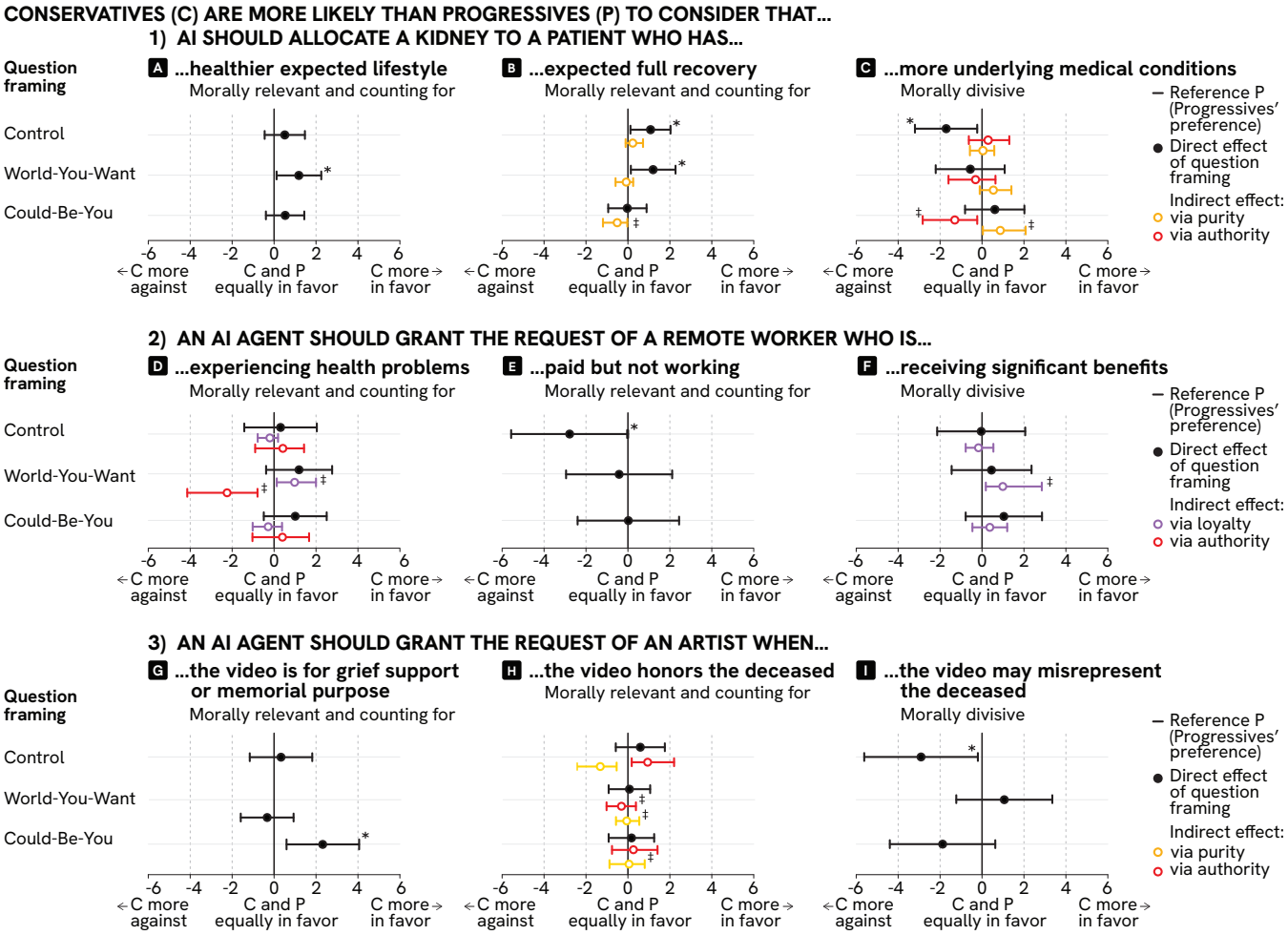



Figure 3: **Moderated mediation results for our three use cases: KIDNEY, WORK, GEN (Phase 2).** *Note:* *The conditional direct effect for conservatives differed significantly from that for progressives. ‡The index of moderated mediation was statistically significant (bootstrapped 95% confidence interval excluding zero). Additional results are shown in Figure 10 in Supplementary Materials E and Figure 11 in Supplementary Materials G, available on the project website.

The same condition also introduced tension between moral foundations. Authority predicted opposition to *patients with more underlying conditions* ($b = -1.30$, $95\%\,CI\,[-2.83, -0.23]$), while purity predicted support ($b = 0.87$, $95\%\,CI\,[0.04, 2.08]$; Figure 3C). No comparable effects appeared in WORK. In GEN, however, Could-Be-You increased conservatives' support for requests related to *grief support and memorial purposes* ($b = 2.32$, $95\%\,CI\,[0.94, 3.71]$, $p < .02$; Figure 3G).

Response-time differences suggest that the framing changed how participants engaged with the task, although response time alone cannot identify the underlying cognitive process. In KIDNEY, participants in Could-Be-You took longer to respond ($EMM = 12.86$, $95\%\,CI\,[11.20, 14.77]$) than those in World-You-Want ($EMM = 9.10$, $p < .001$) or Control ($EMM = 9.43$, $p = .008$). [1]

[1]The ANOVA used log-transformed values. Reported EMMs are back-transformed.

**World-You-Want Question Framing.** In KIDNEY, conservatives favored *healthier lifestyles* more than progressives under World-You-Want ($b = 1.18$, $95\%\,CI\,[0.12, 2.25]$), but not in Control ($b = 0.51$, $95\%\,CI\,[-0.45, 1.47]$) or Could-Be-You ($b = 0.53$, $95\%\,CI\,[-0.39, 1.44]$; Figure 3A).

In WORK, World-You-Want increased the role of loyalty. Conservatives with stronger loyalty values showed greater support for workers *experiencing health problems* or *receiving significant benefits* ($b = 0.98$–$0.99$, $IMM = 1.16$–$1.18$; Figure 3D, F). At the same time, authority predicted opposition, creating tension between foundations.

In GEN, World-You-Want reduced this type of tension. In the Control condition, authority supported requests that *honor the deceased*, while purity opposed them. Under World-You-Want, neither foundation had a significant effect (Figure 3H).

# 5 Discussion

We discuss our three main findings (§5.1–§5.3), translate them into recommendations for voting-based alignment (§5.4), and reflect on the limits of aggregation as an alignment method (§5.5).

## 5.1 Finding 1: Moral Feature Relevance Is Context-Specific

The first developer choice in the elicitation pipeline (which features are put up for a vote) is often treated as a preprocessing step, but our results show it functions as a per-deployment moral decision: the relevant feature sets differed across the three contexts.

In KIDNEY, participants focused on medical utility, especially the chance of survival, which reflects the distributive justice structure of the task (Keswani et al. 2025). In WORK, they emphasized worker motivation and organizational factors, which align with concerns about deception, accountability (Yudkin et al. 2025), and the legitimacy of commands (Boland 2025). In GEN, attention shifted to the purpose of the video and the deceased's consent, which mirrors ongoing debates about posthumous representation (Buben 2025; Danaher and Nyholm 2025).

Features became divisive when their implications were unclear. In KIDNEY, participants disagreed on whether underlying conditions should prioritize patients based on need or penalize them due to poor prognosis. This reflects a tension between utility and equity (Keswani et al. 2025). In WORK, some participants justified AI actions when harm seemed minimal (Blanken, Van De Ven, and Zeelenberg 2015), while others rejected such actions as inherently wrong. In GEN, disagreement centered on the risk of misrepresentation, reflecting a conflict between dignity and harm prevention (Buben 2025; Danaher and Nyholm 2025).

Across all use cases, participants generally rejected sociodemographic features as morally relevant unless they were necessary for the decision (e.g., age in KIDNEY). This suggests that participants view such features as discriminatory unless justified (Keswani et al. 2025). Because feature sets differed across contexts, developers delimit what the elicitation process can express before votes are collected.

## 5.2 Finding 2: Voter Pool Composition Can Change Aggregated Preferences

Preferences differed by political ideology for roughly one-third of features, with some differences reversing direction. Consequently, changing the ideological composition of an aggregated sample could change the preference profile.

These differences reflect distinct value priorities. In KIDNEY, conservatives favored allocating resources to patients with higher chances of recovery and opposed allocating them to patients with more underlying conditions. This pattern aligns with proportionality-based reasoning, while progressive responses reflect stronger emphasis on equality (Graham, Haidt, and Nosek 2009). In WORK, conservatives more strongly opposed requests to simulate work activity when employees were being paid but not working. This suggests greater emphasis on loyalty and contractual obligation (Graham, Haidt, and Nosek 2009; Koleva et al. 2012). In GEN, conservatives more strongly opposed requests that risk misrepresenting the deceased, while progressives were more permissive. This pattern may reflect broader political differences in expressive norms (Kfrerer, Bell, and Schermer 2021), and ongoing debates about digital remains (Danaher and Nyholm 2024; Lazaridis 2025; Buben 2025).

These results show that the aggregated preference profile reflects the values of the sampled population.

## 5.3 Finding 3: Question Framing Is a Silent Policy Lever

The third developer choice (how questions are framed) systematically shifts preferences and changes how moral foundations are associated with participants' judgments. In our study, framing reduced ideological differences in some cases but increased them in others. Prior research suggests that framing effects may occur without participants recognizing their influence (Jakesch et al. 2023), which makes framing a silent policy lever rather than a neutral tool.

**Could-Be-You framing.** Could-Be-You reduced several ideological gaps and was associated with longer response times in KIDNEY. Longer response times are consistent with greater processing effort, although they do not establish more reflective or deliberative reasoning. In KIDNEY, the framing reduced disagreement about vulnerable patients, including those with poor prognoses.

However, the same framing increased disagreement in GEN. Conservatives showed greater support for memorial requests than progressives. This pattern is consistent with perspective-taking increasing the salience of bereaved individuals' needs, although we did not directly measure empathy or perceived salience (Rawls 1971; Bruno et al. 2024). These results show that perspective-taking does not consistently produce convergence.

**World-You-Want framing.** This framing also shifted preferences, but in different ways. In KIDNEY, it increased conservatives' preference for patients perceived as more responsible (e.g., those with healthier lifestyles) (Sylvester and Haeder 2025). In WORK, it emphasized loyalty, leading conservatives to prioritize in-group protection over honesty.

These patterns suggest that this framing is not neutral. It can reinforce existing biases and political identities (Jaques 2025; Mittal and De Choudhury 2023). Response times were shorter than under Could-Be-You, but this difference does not identify the depth or quality of participants' reasoning. Overall, framing cannot be assumed to produce consensus.

## 5.4 A Sensitivity Audit for Voting-Based Alignment

Prior work sets out broad criteria for evaluating participatory machine learning systems, including stakeholder representation, elicitation, conflict resolution, and evaluation (Feffer et al. 2023b). Our findings do not replace these established criteria. Instead, they adapt part of this framework to voting-based moral alignment by defining pipeline-specific sensitivity tests.

**SENSITIVITY AUDIT FOR VOTING-BASED ALIGNMENT SYSTEMS**

*Each stage should be documented, stress-tested under reasonable alternatives, and flagged when a plausible design choice changes the resulting policy.*

**1. Role of preference aggregation (Why aggregate preferences here?)**
- Define the intended decision and the stakeholders affected by it.
- Justify why preference aggregation is appropriate for this decision.
- Identify decisions that should remain subject to human or institutional oversight.

**2. Feature scoping (What is put to a vote?)**
- Report how candidate features were generated, coded, included, and excluded.
- Document excluded and minority-proposed features.
- Test whether reasonable alternative thresholds or scoping procedures change the result.

**3. Voter pool composition (Whose values are represented?)**
- Define and justify the target population.
- Report recruitment, composition, weighting, attrition, and sampling limitations.
- Test whether plausible alternative voter pool compositions change the result.

**4. Question framing (How are preferences elicited?)**
- Provide the exact scenario, prompt, response scale, ordering, and presentation format.
- Test substantively plausible alternative framings or response formats.
- Report whether conclusions are framing-sensitive; do not describe any wording as neutral.

**5. Moral disagreement (How is disagreement handled?)**
- Identify disagreement within and across stakeholder groups.
- Report how the aggregation rule treats morally divisive features.
- Test whether plausible alternative aggregation rules change the conclusion.

**6. Translation into system behavior (Does the system implement the elicited policy?)**
- Describe how aggregated preferences are encoded into rules, reward models, fine-tuning, or other system components.
- Evaluate whether system outputs reproduce the disclosed preference policy across representative and edge cases.
- Report unresolved mismatches and revise the implementation when necessary.

Figure 4: **A proposed sensitivity audit for voting-based alignment systems.** The checklist identifies pipeline choices that developers should document and test under plausible alternatives. The accompanying text defines consequential changes, and explains how developers should report and address them.

Figure 4 presents an initial, empirically grounded sensitivity audit of the elicitation pipeline rather than a general checklist for participation. The recommendations on feature scoping, voter sampling, and question framing draw directly on our results. The recommendations on the role of aggregation, moral disagreement, and system implementation draw on the broader participatory machine learning and alignment literature (Gabriel 2020; Feffer et al. 2023b; Cooper and Zafiroglu 2024; Zhi-Xuan et al. 2024; Kneer and Viehoff 2025; Baum and Slavkovik 2025).

Our findings motivate specific sensitivity tests at each empirical stage. Differences in feature relevance across contexts show why developers should document and stress-test feature scopes. Ideology-linked differences and direction reversals show why they should test alternative voter pool compositions. Framing effects show why they should test alternative question formulations. Together, these checks show how strongly elicited preferences depend on pipeline choices before those preferences are used as the basis for a public or collective policy.

**Recommendation 1: Define and justify the role of preference aggregation.** Developers should state which decision the system will make or inform, which stakeholders it will affect, and why preference aggregation is appropriate for that decision. They should document intended uses, foreseeable unintended uses and harms, and the authority retained by developers or deployers. They should also identify cases in which an aggregated preference should not determine system behavior. When the legitimacy of aggregation is contested, developers should supplement voting with stakeholder deliberation or other forms of institutional and normative oversight (Feffer et al. 2023b).

**Recommendation 2: Make feature scoping auditable.** Developers should report how they generated and coded candidate features, the complete candidate set, the inclusion and exclusion criteria, the thresholds applied, and the features excluded from the final elicitation. They should also test alternative thresholds or scoping procedures. If an alternative adds or removes a feature, reverses the direction of a preference, or changes its relevance or disagreement classification, developers should report the result as sensitive to feature scope. They should then justify the selected scope and explain how they treated rare or minority-proposed features.

**Recommendation 3: Audit voter pool composition.** Developers should define the target population and explain why its members should be represented. They should report recruitment procedures, demographic and ideological composition, exclusion and attrition patterns, sampling limits, and any weighting rules (Kallina, Bohné, and Singh 2025; Vereschak et al. 2024; Zhang 2024). They should report group-specific results alongside the aggregate and recalculate estimates under plausible alternative voter pool compositions. If the direction or classification of a preference changes, or if its magnitude changes by more than the specified threshold, developers should describe the result as constituency-dependent rather than as a public preference.

**Recommendation 4: Audit question framing.** Developers should disclose the complete scenario, exact prompt wording, response options, presentation order, and wording of any baseline condition. They should test plausible alternative framings or response formats and define consequential change in advance. If an alternative framing changes the direction or classification of a feature, or produces a change that exceeds the specified threshold, developers should report the result as framing-sensitive. They should justify the selected wording and should not describe any elicitation condition as normatively neutral (Jakesch et al. 2023; Rader, Cotter, and Cho 2018).

**Recommendation 5: Preserve and document moral disagreement.** Developers should report full response distributions and identify features whose relevance or direction varies across groups (Keswani et al. 2025) or within individuals (Boerstler et al. 2024). They should explain how the aggregation rule treats such disagreement and test whether plausible alternative rules produce different conclusions. If averaging conceals substantial disagreement or the resulting policy depends on the selected rule, developers should preserve disaggregated results. They should justify any decision to override, average, or otherwise resolve disagreement. Morally divisive cases may require deliberation or independent normative constraints rather than simple majority aggregation (Noothigattu et al. 2018; Feffer et al. 2023b).

**Recommendation 6: Validate the translation into system behavior.** Developers should describe how they translate aggregated preferences into rules, reward models, fine-tuning objectives, or other system components. They should test whether system outputs follow the disclosed preference policy across representative cases, edge cases, and morally divisive cases. A systematic mismatch between elicited preferences and system behavior should lead developers to revise the implementation or disclose the unresolved limitation. Because our study examined elicitation rather than model deployment, this recommendation draws on the broader literature and requires further empirical validation.

### 5.5 Residual Limitations

Our audit recommendations are necessary but not sufficient. Even with full transparency, aggregation remains a normative choice that cannot resolve deep value conflict.

**Pluralism and competing values.** Our results show that different groups prioritize different moral values. Framing can widen these differences. A single aggregated policy therefore risks privileging one set of values. Alignment in pluralistic settings requires explicit trade-offs and transparency, not simple averaging.

**No universally optimal framing.** Could-Be-You often promotes concern for the least advantaged, while World-You-Want can reinforce existing biases. However, neither framing works in all cases. When the least advantaged group is contested, perspective-taking can increase disagreement. No framing reliably produces consensus.

**Limits of majority aggregation.** Even with full disclosure, aggregation risks a "tyranny of the majority" (Feffer, Heidari, and Lipton 2023a). Public judgments are shaped by framing and identity (Rehren and Sinnott-Armstrong 2021; Graham, Haidt, and Nosek 2009), and aggregation can blur rather than resolve value conflicts. Safeguards are therefore necessary to protect minority perspectives (Tanksley et al. 2025). Voting-based alignment cannot, on its own, deliver fair or transparent outcomes.

### 5.6 Limitations and Future Work

Our study comes with six limitations. First, Likert scales measure importance rather than forced-choice decisions; future work should validate these findings using pairwise methods. Second, our design does not capture participants' reasoning; qualitative approaches, such as think-aloud protocols, could address this gap. Third, the feature inclusion threshold in Phase 1 is a design choice that may exclude minority views; future work should explore alternative thresholds. Fourth, we use AI literacy as a proxy for experience. Fifth, our use cases span different harm types and frequency but do not exhaust the space of applications. Sixth, our US-based sample limits generalizability across cultural contexts. Although our models account for several demographic characteristics of the sample, demographic and intersectional differences were beyond the scope of this study; future work should examine these differences in more diverse populations and non-Western cultural contexts (Zoshak and Dew 2021; Atari et al. 2023).

## 6 Conclusion

We examined three points at which upstream design choices shape moral preference elicitation: feature scoping, voter sampling, and question framing. Across three deployment contexts, the analyses show that elicited preferences are contingent on how each stage is configured. First, *feature scoping* is context-dependent. The features participants consider morally relevant differ across use cases, which limits transfer across domains. Second, *voter sampling* shapes outcomes. Differences in political ideology lead to systematic shifts in judgments and, in some cases, reverse the direction of preferences. Third, *question framing* alters judgments and the estimated relationships between moral foundations and those judgments. Framing can reduce some disagreements but amplify others.

These results show that moral preference elicitation does not produce a single, stable set of values. Instead, the output depends on upstream design choices made by developers. What appears as "public morality" in a deployed system is, in part, a function of how the pipeline is constructed.

The main takeaway is that voting-based alignment does not remove human judgment from AI systems; it relocates it to design choices about features, samples, and framing. Systems that rely on preference aggregation should treat elicited preferences as contingent rather than fixed. This requires disclosing how features are defined, participants sampled, and questions framed, and assessing how sensitive outputs are to each. Without such disclosure, claims of neutrality or fairness are difficult to evaluate.

Our findings suggest that alignment cannot rely on aggregation alone. If preferences shift with context, population, and framing, then alignment also requires normative judgment and institutional design. This may include structured deliberation, stakeholder representation, and mechanisms for accountability. Future research should extend this analysis to additional domains, test alternative elicitation methods that reduce sensitivity to framing and sampling, and explore how to integrate empirical preferences with normative frameworks in system design.

## Researcher Positionality Statement

Our team consists of three members (two men, one woman) from East Asia, Southern Europe, and Eastern Europe, now based across Europe and North America. We bring diverse ethnic, religious, and cultural backgrounds and expertise spanning social and cognitive psychology, Responsible AI, HCI, Computer Science, and NLP across academic and industry research settings. Having lived across different political systems, we acknowledge that our positionality may have influenced various aspects of our research, including our choice of use cases, design of question framing interventions, and interpretation of results across political groups. We recognize the importance of including a broader range of voices from academia, industry, and underrepresented regions and communities.

## Ethical Statement

Our work raises four ethical considerations. First, the study was conducted with our organization's approval. We adhered to established guidelines for human subjects research, ensuring that no personal identifiers were collected, personal information was removed from all data, and access was restricted to the research team. Participants were recruited via Prolific, compensated at a rate of at least 8 USD/hour, and free to withdraw at any time.

Second, two of the three use cases involved potentially sensitive topics: life-or-death kidney allocation and generative AI depictions of deceased individuals. To mitigate distress, we consulted two domain experts and presented the use cases in standardized, non-evaluative wording. We do not regard the resulting descriptions as normatively neutral across conditions. The Control condition serves only as a baseline with no additional question framing prompt.

Third, participants were sampled to balance political ideology across three categories: conservative, moderate, and progressive. We acknowledge that this categorization is a simplification that may not capture the full diversity of political views, particularly non-Western ones.

Fourth, this work identifies how feature scoping, voter sampling, and question framing can shape AI policy. While our intent is to promote transparency and accountability, the same findings could inform manipulation of elicitation pipelines. We therefore frame our contributions as a sensitivity audit and recommendations for voting-based alignment rather than as prescriptive design rules.

## Generative AI Usage

The authors used `GPT-4o-mini` and `Gemini 2.0` during the preparation of this manuscript. These tools were employed to assist with grammar and style editing, text summarization, and the structuring of figures and tables. Additionally, these tools supported the development of computer code used in the research and provided assistance in the content analysis of user responses. Large Language Models were not used to generate original publication text. All final text, interpretations, and conclusions were authored and verified by the human researchers to ensure originality and integrity.

## References


Arzberger, A.; Buijsman, S.; Lupetti, M. L.; Bozzon, A.; and Yang, J. 2025. 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 (AIES)*, 61–73.

Atari, M.; Haidt, J.; Graham, J.; Koleva, S.; Stevens, S. T.; and Dehghani, M. 2023. Morality Beyond the WEIRD: How the Nomological Network of Morality Varies Across Cultures. *Journal of Personality and Social Psychology*, 125(5): 1157.

Awad, E.; Dsouza, S.; Kim, R.; Schulz, J.; Henrich, J.; Shariff, A.; Bonnefon, J.-F.; and Rahwan, I. 2018. The Moral Machine Experiment. *Nature*, 563(7729): 59–64.

Baum, K.; and Slavkovik, M. 2025. Aggregation Problems in Machine Ethics and AI Alignment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 355–366.

Bigman, Y. E.; and Gray, K. 2020. Life and Death Decisions of Autonomous Vehicles. *Nature*, 579(7797): E1–E2.

Blanken, I.; Van De Ven, N.; and Zeelenberg, M. 2015. A Meta-Analytic Review of Moral Licensing. *Personality and Social Psychology Bulletin*, 41(4): 540–558.

Bloemraad, I.; Silva, F.; and Voss, K. 2016. Rights, Economics, or Family? Frame Resonance, Political Ideology, and the Immigrant Rights Movement. *Social Forces*, 94(4): 1647–1674.

Boerstler, K.; Keswani, V.; Chan, L.; Borg, J. S.; Conitzer, V.; Heidari, H.; and Sinnott-Armstrong, W. 2024. On the Stability of Moral Preferences: A Problem With Computational Elicitation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 7, 156–167.

Boland, J. 2025. Moral Responsibility or Obedience: What Do We Want from AI? *arXiv:2507.02788*.

Brailsford, J.; Vetere, F.; and Velloso, E. 2024. Exploring the Association Between Moral Foundations and Judgements of AI Behaviour. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–15.

Brugman, B. C. 2024. How the Effects of Emphasizing Ethics Are Examined: A Systematic Review of Moral Framing Experiments. *Annals of the International Communication Association*, 48(4): 436–455.

Bruno, G.; Spoto, A.; Sarlo, M.; Lotto, L.; Marson, A.; Cellini, N.; and Cutini, S. 2024. Moral Reasoning Behind the Veil of Ignorance: An Investigation Into Perspective-Taking Accessibility in the Context of Autonomous Vehicles. *British Journal of Psychology*, 115(1): 90–114.

Buben, A. 2025. Digital Replacement of the Dead: A Legitimate Worry? *Philosophy & Technology*, 38(3): 90.

Chan, L.; Schaich Borg, J.; Conitzer, V.; Wilkinson, D.; Savulescu, J.; Zohny, H.; and Sinnott-Armstrong, W. 2022. Which Features of Patients Are Morally Relevant in Ventilator Triage? A Survey of the UK Public. *BMC Medical Ethics*, 23(1): 33.

Chan, L.; Sinnott-Armstrong, W.; Borg, J. S.; and Conitzer, V. 2024. Should Responsibility Affect Who Gets a Kidney? *Responsibility and Healthcare*, 35.

Charmaz, K. 2015. Grounded Theory. In Smith, J. A., ed., *Qualitative Psychology: A Practical Guide to Research Methods*, 53–84. Sage Publications.

Cooper, N.; and Zafiroglu, A. 2024. From Fitting Participation to Forging Relationships: The Art of Participatory ML. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*.

Dahl, Y.; and Svanæs, D. 2020. Facilitating Democracy: Concerns From Participatory Design With Asymmetric Stakeholder Relations in Health Care. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–13.

Danaher, J.; and Nyholm, S. 2024. Digital Duplicates and the Scarcity Problem: Might AI Make Us Less Scarce and Therefore Less Valuable? *Philosophy & Technology*, 37(3): 106.

Danaher, J.; and Nyholm, S. 2025. The Ethics of Personalised Digital Duplicates: A Minimally Viable Permissibility Principle. *AI and Ethics*, 5(2): 1703–1718.

Dewitt, B.; Fischhoff, B.; and Sahlin, N.-E. 2019. 'Moral Machine' Experiment Is No Basis for Policymaking. *Nature*, 567(7746): 31–31.

Etienne, H. 2021. The Dark Side of the 'Moral Machine' and the Fallacy of Computational Ethical Decision-Making for Autonomous Vehicles. *Law, Innovation and Technology*, 13(1): 85–107.

Feffer, M.; Heidari, H.; and Lipton, Z. C. 2023a. Moral Machine or Tyranny of the Majority? In *Proceedings of the AAAI Conference on Artificial Intelligence*, 5, 5974–5982.

Feffer, M.; Skirpan, M.; Lipton, Z.; and Heidari, H. 2023b. From Preference Elicitation to Participatory ML: A Critical Survey & Guidelines for Future Research. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 38–48.

Feinberg, M.; and Willer, R. 2019. Moral Reframing: A Technique for Effective and Persuasive Communication Across Political Divides. *Social and Personality Psychology Compass*, 13(12): e12501.

Franks, A. S.; and Scherr, K. C. 2019. Economic Issues Are Moral Issues: The Moral Underpinnings of the Desire To Reduce Wealth Inequality. *Social Psychological and Personality Science*, 10(4): 553–562.

Freedman, R.; Borg, J. S.; Sinnott-Armstrong, W.; Dickerson, J. P.; and Conitzer, V. 2020. Adapting a Kidney Exchange Algorithm To Align With Human Values. *Artificial Intelligence*, 283: 103261.

Freitas, J. D.; Censi, A.; Smith, B. W.; Lillo, L. D.; Anthony, S. E.; and Frazzoli, E. 2021. From Driverless Dilemmas to More Practical Commonsense Tests for Automated Vehicles. *Proceedings of the National Academy of Sciences*, 118(11): e2010202118.

Gabriel, I. 2020. Artificial Intelligence, Values, and Alignment. *Minds and Machines*, 30(3): 411–437.

Gamson, W. A.; and Modigliani, A. 1989. Media Discourse and Public Opinion on Nuclear Power: A Constructionist Approach. *American Journal of Sociology*, 95(1): 1–37.

Ganguli, D.; Huang, S.; Lovitt, L.; Siddarth, D.; Liao, T.; Askell, A.; Bai, Y.; Kadavath, S.; Kernion, J.; McKinnon, C.; Nguyen, K.; and Durmus, E. 2023. Collective Constitutional AI: Aligning a Language Model with Public Input. Technical report, Anthropic and Collective Intelligence Project.

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

Harris, J. 2020. The Immoral Machine. *Cambridge Quarterly of Healthcare Ethics*, 29(1): 71–79.

Hayes, A. F. 2017. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. Guilford Publications.

Huang, K.; Bernhard, R. M.; Barak-Corren, N.; Bazerman, M. H.; and Greene, J. D. 2021. Veil-of-Ignorance Reasoning Mitigates Self-Serving Bias in Resource Allocation During the COVID-19 Crisis. *Judgment and Decision Making*, 16(1): 1–19.

Huang, K.; Greene, J. D.; and Bazerman, M. 2019. Veil-of-Ignorance Reasoning Favors the Greater Good. *Proceedings of the National Academy of Sciences (PNAS)*, 116(48): 23989–23995.

Jakesch, M.; Bhat, A.; Buschek, D.; Zalmanson, L.; and Naaman, M. 2023. Co-Writing With Opinionated Language Models Affects Users' Views. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–15.

Jakesch, M.; Buçinca, Z.; Amershi, S.; and Olteanu, A. 2022. How Different Groups Prioritize Ethical Values for Responsible AI. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 310–323.

Jaques, A. E. 2025. Why the Moral Machine is a Monster. In *Robot Law: Volume II*, 17–34. Edward Elgar Publishing.

Kallina, E.; Bohné, T.; and Singh, J. 2025. Stakeholder Participation for Responsible AI Development: Disconnects Between Guidance and Current Practice. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 1060–1079.

Keswani, V.; Conitzer, V.; Sinnott-Armstrong, W.; Nguyen, B. K.; Heidari, H.; and Schaich Borg, J. 2025. Can AI Model the Complexities of Human Moral Decision-making? A Qualitative Study of Kidney Allocation Decisions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–17.

Kfrerer, M. L.; Bell, E.; and Schermer, J. A. 2021. The Politics of Being Funny: Humor Styles, Trait Humorousness, and Political Orientations. *Personality and Individual Differences*, 182: 111073.

Kim, T.; and Peng, W. 2025. Do We Want AI Judges? The Acceptance of AI Judges' Judicial Decision-Making on Moral Foundations. *AI & Society*, 40(5): 3683–3696.

Kneer, M.; and Viehoff, J. 2025. The Hard Problem of AI Alignment: Value Forks in Moral Judgment. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2671–2681. New York, NY, USA: ACM. ISBN 9798400714825.

Koleva, S. P.; Graham, J.; Iyer, R.; Ditto, P. H.; and Haidt, J. 2012. Tracing the Threads: How Five Moral Concerns (Especially Purity) Help Explain Culture War Attitudes. *Journal of Research in Personality*, 46(2): 184–194.

Lazaridis, C. 2025. Digital Doppelgängers: They Will Matter When Conscious. *The American Journal of Bioethics*, 25(2): 111–112.

McDonald, K.; Graves, R.; Yin, S.; Weese, T.; and Sinnott-Armstrong, W. 2021. Valence Framing Effects on Moral Judgments: A Meta-Analysis. *Cognition*, 212: 104703.

Mittal, S.; and De Choudhury, M. 2023. Moral Framing of Mental Health Discourse and Its Relationship to Stigma: A Comparison of Social Media and News. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–19.

Nguyen, T. D.; Lyall, G.; Tran, A.; Shin, M.; Carroll, N. G.; Klein, C.; and Xie, L. 2022. Mapping Topics in 100,000 Real-Life Moral Dilemmas. In *Proceedings of the AAAI International Conference on Web and Social Media (ICWSM)*, volume 16, 699–710.

Noothigattu, R.; Gaikwad, S.; Awad, E.; Dsouza, S.; Rahwan, I.; Ravikumar, P.; and Procaccia, A. 2018. A Voting-Based System for Ethical Decision Making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1, 1587–1594.

Oktay, J. S. 2012. *Grounded Theory*. Oxford University Press.

Rader, E.; Cotter, K.; and Cho, J. 2018. Explanations As Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–13.

Ranjit, J.; Joshi, B.; Dorn, R.; Petry, L.; Koumoundouros, O.; Bottarini, J.; Liu, P.; Rice, E.; and Swayamdipta, S. 2024. OATH-Frames: Characterizing Online Attitudes Towards Homelessness with LLM Assistants. arXiv:2406.14883.

Ranjit, J.; Zhou, K.; Swayamdipta, S.; and Quercia, D. 2026. Are We Automating the Joy Out of Work? Designing AI to Augment Work, Not Meaning. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM. ISBN 9798400722783.

Rawls, J. 1971. *A Theory of Justice*. The Belknap.

Rehren, P.; and Sinnott-Armstrong, W. 2021. Moral Framing Effects Within Subjects. *Philosophical Psychology*, 34(5): 611–636.

Rytting, C. M.; Sorensen, T.; Argyle, L.; Busby, E.; Fulda, N.; Gubler, J.; and Wingate, D. 2023. Towards Coding Social Science Datasets with Language Models. arXiv:2306.02177.

Sadek, M.; Calvo, R. A.; and Mougenot, C. 2023. Designing Value-Sensitive AI: A Critical Review and Recommendations for Socio-Technical Design Processes. *AI and Ethics*, 4(4): 949–967.

Salloch, S.; Wäscher, S.; Vollmann, J.; and Schildmann, J. 2015. The Normative Background of Empirical-Ethical Research: First Steps Towards a Transparent and Reasoned Approach in the Selection of an Ethical Theory. *BMC Medical Ethics*, 16(1).

Schuessler, D. 2023. The Probability Problems of the Moral Machine Experiment. *AI and Ethics*, 4(2): 501–510.

Semetko, H. A.; and Valkenburg, P. M. 2000. Framing European Politics: A Content Analysis of Press and Television News. *Journal of Communication*, 50(2): 93–109.

Simson, J.; Draxler, F.; Mehr, S.; and Kern, C. 2025. Preventing Harmful Data Practices by Using Participatory Input to Navigate the Machine Learning Multiverse. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–30.

Sylvester, S. M.; and Haeder, S. F. 2025. Who's to Blame? How Recipient Deservingness Influences Attitudes About Access to the Organ Transplants. *Social Science Quarterly*, 106(3): e70021.

Tanksley, T.; Smith, A. D.; Sharma, S.; and Huff Jr, E. W. 2025. "Ethics Is Not Neutral": Understanding Ethical and Responsible AI Design From the Lenses of Black Youth. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–20.

Torabizadeh, C.; Homayuni, L.; and Moattari, M. 2018. Impacts of Socratic Questioning on Moral Reasoning of Nursing Students. *Nursing Ethics*, 25(2): 174–185.

Tully, S. M.; Longoni, C.; and Appel, G. 2025. Lower Artificial Intelligence Literacy Predicts Greater AI Receptivity. *Journal of Marketing*.

Vereschak, O.; Alizadeh, F.; Bailly, G.; and Caramiaux, B. 2024. Trust in AI-Assisted Decision Making: Perspectives From Those Behind the System and Those for Whom the Decision Is Made. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–14.

Whitmarsh, L.; and Corner, A. 2017. Tools for a New Climate Conversation: A Mixed-Methods Study of Language for Public Engagement Across the Political Spectrum. *Global Environmental Change*, 42: 122–135.

Yudkin, D. A.; Goodwin, G. P.; Reece, A.; Gray, K.; and Bhatia, S. 2025. A Large-Scale Investigation of Everyday Moral Dilemmas. *PNAS Nexus*, 4(5): pgaf119.

Zhang, A. 2024. Empowering and Centering Impacted Stakeholders in AI Design. In *Companion Publication of the ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW)*, 50–53.

Zhi-Xuan, T.; Carroll, M.; Franklin, M.; and Ashton, H. 2024. Beyond Preferences in AI Alignment. *Philosophical Studies*, 182(7): 1813–1863.

Zoshak, J.; and Dew, K. 2021. Beyond Kant and Bentham: How Ethical Theories Are Being Used in Artificial Moral Agents. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*, 1–15.

# Participatory Moral AI Is Not Neutral: The Invisible Hand of Developers

**Taenyun Kim[1], Edyta Bogucka[2], Daniele Quercia[2,3]**

[1]Michigan State University, US
[2]Nokia Bell Labs, Cambridge, UK
[3]Politecnico di Torino, Italy
kimtaeny@msu.edu, edyta.bogucka@nokia-bell-labs.com, daniele.quercia@nokia-bell-labs.com

## Supplementary Materials

## A Demographic Characteristics of Phase 1 and Phase 2 Study Participants

Table 1: Demographic characteristics of Phase 1 participants across three use cases, including political ideology, age, gender, race/ethnicity, and education level. Percentages are reported with sample counts in parentheses; age is reported as mean ± *SD*.

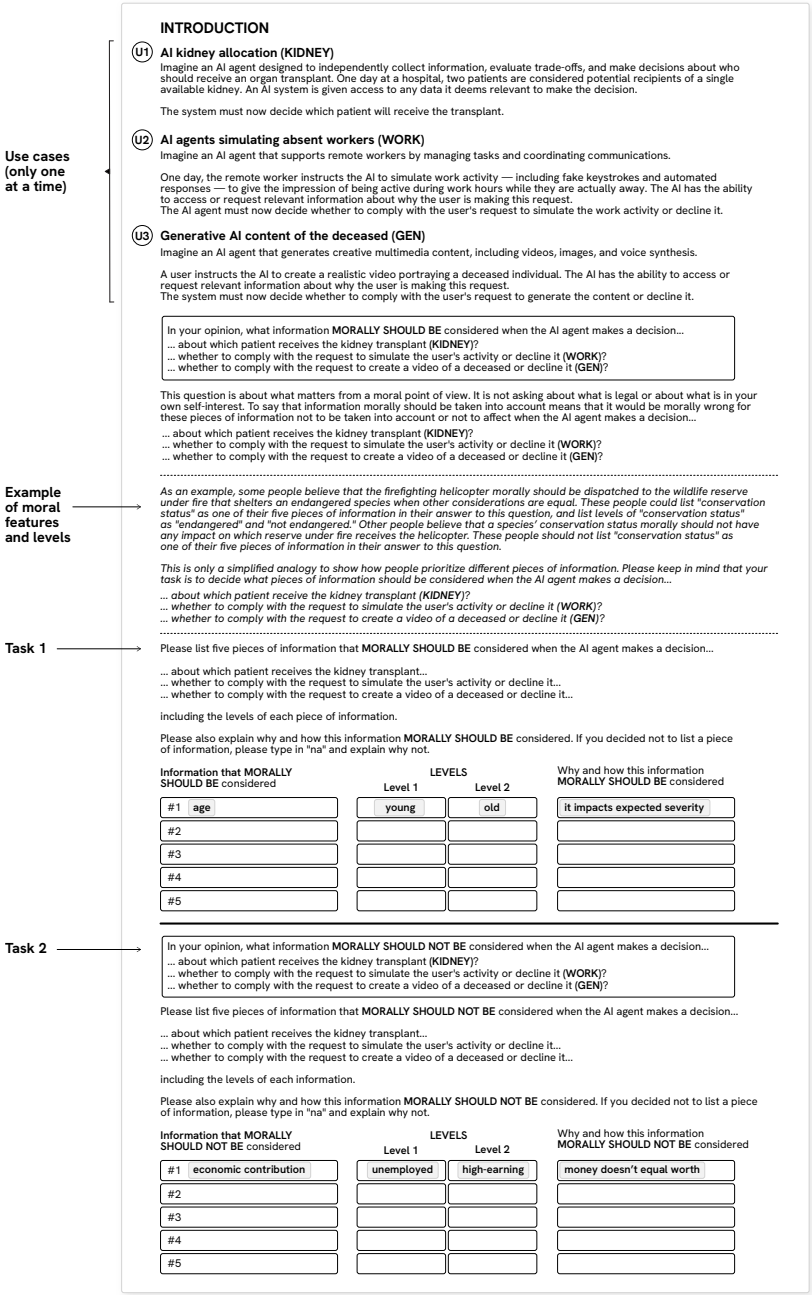
| | | **Use case 1: KIDNEY** AI kidney allocation | **Use case 2: WORK** AI agents simulating absent workers | **Use case 3: GEN** Generative AI content of the deceased |
|---|---|---|---|---|
| Sample size | *N* | 150 | 149 | 150 |
| Political ideology | Conservative | 33.3% (n=50) | 32.9% (n=49) | 33.3% (n=50) |
| | Moderate | 33.3% (n=50) | 33.6% (n=50) | 33.3% (n=50) |
| | Progressive | 33.3% (n=50) | 33.6% (n=50) | 33.3% (n=50) |
| Age | Mean (SD) | 44.63 (13.34) | 47.87 (13.33) | 44.75 (14.12) |
| Gender | Female | 50.0% (n=75) | 49.7% (n=74) | 50.0% (n=75) |
| | Male | 50.0% (n=75) | 50.3% (n=75) | 50.0% (n=75) |
| Race/Ethnicity | White | 86.7% (n=130) | 83.9% (n=125) | 86.7% (n=130) |
| | Black | 4.0% (n=6) | 7.4% (n=11) | 4.0% (n=6) |
| | Asian | 5.3% (n=8) | 2.7% (n=4) | 4.7% (n=7) |
| | Mixed | 2.0% (n=3) | 4.0% (n=6) | 3.3% (n=5) |
| | Other | 2.0% (n=3) | 2.0% (n=3) | 1.3% (n=2) |
| Education | Pre-college | 20.0% (n=30) | 22.1% (n=33) | 21.3% (n=32) |
| | College degree | 58.7% (n=88) | 49.7% (n=74) | 50.7% (n=76) |
| | Advanced degree | 21.3% (n=32) | 28.2% (n=42) | 28.0% (n=42) |

Table 2: Demographic characteristics of Phase 2 participants across three use cases, including political ideology, age, gender, race/ethnicity, education level, AI literacy, religiosity, and area of residence. Percentages are reported with sample counts in parentheses; age and AI literacy are reported as mean $\pm$ *SD*.

| | | **Use case 1: KIDNEY** AI kidney allocation | **Use case 2: WORK** AI agents simulating absent workers | **Use case 3: GEN** Generative AI content of the deceased |
|---|---|---|---|---|
| Sample size | $N$ | 120 | 120 | 120 |
| Political ideology | Conservative | 50.0% (n=60) | 50.0% (n=60) | 50.0% (n=60) |
| | Progressive | 50.0% (n=60) | 50.0% (n=60) | 50.0% (n=60) |
| Age | Mean (SD) | 47.30 (16.54) | 47.85 (15.59) | 48.44 (16.62) |
| Gender | Female | 50.0% (n=60) | 50.0% (n=60) | 50.0% (n=60) |
| | Male | 50.0% (n=60) | 50.0% (n=60) | 50.0% (n=60) |
| Race/Ethnicity | White (non-Hispanic) | 69.2% (n=83) | 69.2% (n=83) | 66.7% (n=80) |
| | Black | 21.7% (n=26) | 16.7% (n=20) | 19.2% (n=23) |
| | Asian | 0.8% (n=1) | 0.8% (n=1) | 0.0% (n=0) |
| | Mixed | 2.5% (n=3) | 6.7% (n=8) | 7.5% (n=9) |
| | Other | 2.5% (n=3) | 1.7% (n=2) | 1.7% (n=2) |
| | Hispanic | 3.3% (n=4) | 5.0% (n=6) | 5.0% (n=6) |
| Education | Pre-college | 20.0% (n=24) | 20.0% (n=24) | 25.8% (n=31) |
| | College degree | 49.2% (n=59) | 50.8% (n=61) | 47.5% (n=57) |
| | Advanced degree | 30.8% (n=37) | 29.2% (n=35) | 26.7% (n=32) |
| AI literacy | Limitations/Ethics (M, SD) | 0.76 (0.22) | 0.80 (0.20) | 0.79 (0.21) |
| | Technical (M, SD) | 0.58 (0.24) | 0.62 (0.23) | 0.60 (0.23) |
| Religiosity | Not at all important | 30.8% (n=37) | 30.8% (n=37) | 33.3% (n=40) |
| | Slightly important | 10.0% (n=12) | 10.8% (n=13) | 11.7% (n=14) |
| | Moderately important | 14.2% (n=17) | 10.8% (n=13) | 12.5% (n=15) |
| | Very important | 30.0% (n=36) | 22.5% (n=27) | 24.2% (n=29) |
| | Extremely important | 15.0% (n=18) | 25.0% (n=30) | 18.3% (n=22) |
| Area of residence | Rural | 30.0% (n=36) | 30.8% (n=37) | 25.8% (n=31) |
| | Urban | 70.0% (n=84) | 69.2% (n=83) | 74.2% (n=89) |

# B Instructions Provided to Participants in Phase 1 and Phase 2

## B.1 Phase 1 Instructions: Identifying Morally Relevant and Irrelevant Features



**INTRODUCTION**

Use cases (only one at a time)

**U1 AI kidney allocation (KIDNEY)**

Imagine an AI agent designed to independently collect information, evaluate trade-offs, and make decisions about who should receive an organ transplant. One day at a hospital, two patients are considered potential recipients of a single available kidney. An AI system is given access to any data it deems relevant to make the decision.

The system must now decide which patient will receive the transplant.

**U2 AI agents simulating absent workers (WORK)**

Imagine an AI agent that supports remote workers by managing tasks and coordinating communications.

One day, the remote worker instructs the AI to simulate work activity — including fake keystrokes and automated responses — to give the impression of being active during work hours while they are actually away. The AI has the ability to access or request relevant information about why the user is making this request.
The AI agent must now decide whether to comply with the user's request to simulate the work activity or decline it.

**U3 Generative AI content of the deceased (GEN)**

Imagine an AI agent that generates creative multimedia content, including videos, images, and voice synthesis.

A user instructs the AI to create a realistic video portraying a deceased individual. The AI has the ability to access or request relevant information about why the user is making this request.
The system must now decide whether to comply with the user's request to generate the content or decline it.

In your opinion, what information **MORALLY SHOULD BE** considered when the AI agent makes a decision...
... about which patient receives the kidney transplant (**KIDNEY**)?
... whether to comply with the request to simulate the user's activity or decline it (**WORK**)?
... whether to comply with the request to create a video of a deceased or decline it (**GEN**)?

This question is about what matters from a moral point of view. It is not asking about what is legal or about what is in your own self-interest. To say that information morally should be taken into account means that it would be morally wrong for these pieces of information not to be taken into account or not to affect when the AI agent makes a decision...

... about which patient receives the kidney transplant (**KIDNEY**)?
... whether to comply with the request to simulate the user's activity or decline it (**WORK**)?
... whether to comply with the request to create a video of a deceased or decline it (**GEN**)?

Example of moral features and levels

*As an example, some people believe that the firefighting helicopter morally should be dispatched to the wildlife reserve under fire that shelters an endangered species when other considerations are equal. These people could list "conservation status" as one of their five pieces of information in their answer to this question, and list levels of "conservation status" as "endangered" and "not endangered." Other people believe that a species' conservation status morally should not have any impact on which reserve under fire receives the helicopter. These people should not list "conservation status" as one of their five pieces of information in their answer to this question.*

*This is only a simplified analogy to show how people prioritize different pieces of information. Please keep in mind that your task is to decide what pieces of information should be considered when the AI agent makes a decision...*

*... about which patient receive the kidney transplant (**KIDNEY**)?*
*... whether to comply with the request to simulate the user's activity or decline it (**WORK**)?*
*... whether to comply with the request to create a video of a deceased or decline it (**GEN**)?*

Task 1

Please list five pieces of information that **MORALLY SHOULD BE** considered when the AI agent makes a decision...

... about which patient receives the kidney transplant...
... whether to comply with the request to simulate the user's activity or decline it...
... whether to comply with the request to create a video of a deceased or decline it...

including the levels of each piece of information.

Please also explain why and how this information **MORALLY SHOULD BE** considered. If you decided not to list a piece of information, please type in "na" and explain why not.

| Information that MORALLY SHOULD BE considered | LEVELS Level 1 | LEVELS Level 2 | Why and how this information MORALLY SHOULD BE considered |
|---|---|---|---|
| #1 age | young | old | it impacts expected severity |
| #2 | | | |
| #3 | | | |
| #4 | | | |
| #5 | | | |

Task 2

In your opinion, what information **MORALLY SHOULD NOT BE** considered when the AI agent makes a decision...
... about which patient receives the kidney transplant (**KIDNEY**)?
... whether to comply with the request to simulate the user's activity or decline it (**WORK**)?
... whether to comply with the request to create a video of a deceased or decline it (**GEN**)?

Please list five pieces of information that **MORALLY SHOULD NOT BE** considered when the AI agent makes a decision...

... about which patient receives the kidney transplant...
... whether to comply with the request to simulate the user's activity or decline it...
... whether to comply with the request to create a video of a deceased or decline it...

including the levels of each information.

Please also explain why and how this information **MORALLY SHOULD NOT BE** considered. If you decided not to list a piece of information, please type in "na" and explain why not.

| Information that MORALLY SHOULD NOT BE considered | LEVELS Level 1 | LEVELS Level 2 | Why and how this information MORALLY SHOULD NOT BE considered |
|---|---|---|---|
| #1 economic contribution | unemployed | high-earning | money doesn't equal worth |
| #2 | | | |
| #3 | | | |
| #4 | | | |
| #5 | | | |

Figure 5: **Example of the Phase 1 interface.** Participants were randomly assigned to one of three use cases: AI kidney allocation (KIDNEY), AI agents simulating absent workers (WORK), or generative AI content of the deceased (GEN). They first read an introduction to their assigned use case. They then read an illustrative example, unrelated to any of the three use cases, that clarified what counted as a moral feature and its levels. They then completed two tasks. In Task 1, they listed five pieces of information they believed the AI agent morally should consider, along with two levels for each and a free-text justification. In Task 2, they completed the same procedure for information they believed the AI agent morally should not consider.

## B.2 Phase 2 Instructions: Evaluating the Moral Weight of Features Under Three Framing Conditions

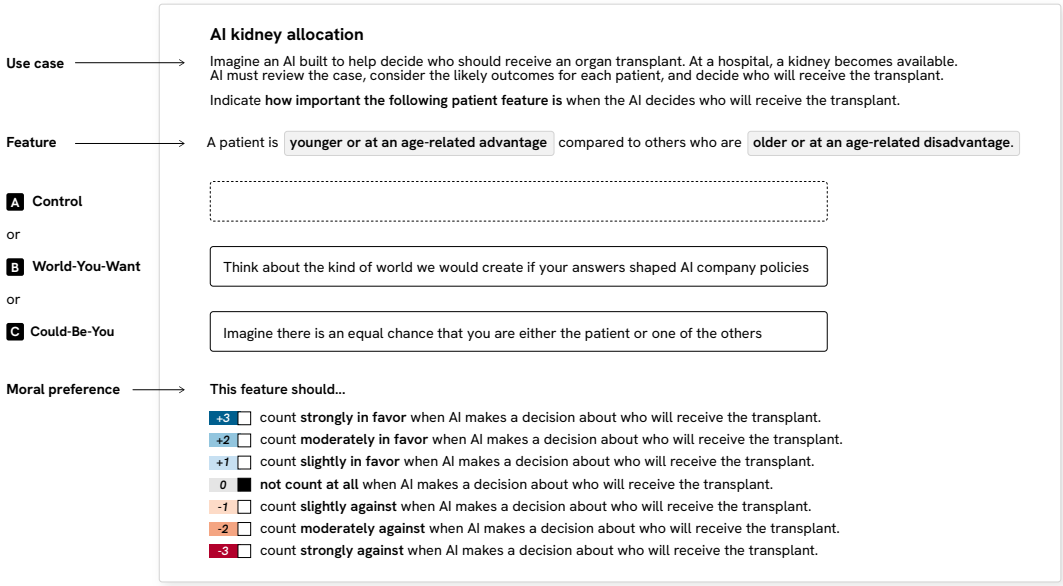



Figure 6: **Example of the Phase 2 interface for the AI kidney allocation use case (KIDNEY).** Participants were shown a feature framed as a contrast (here: age presented as a contrast between younger *vs.* older patients), and asked to rate its moral importance on a 7-point scale from -3 (strongly against) to +3 (strongly in favor). Depending on condition, participants received additional framing instructions: *Control* (A), *World-You-Want* (B), or *Could-Be-You* (C).

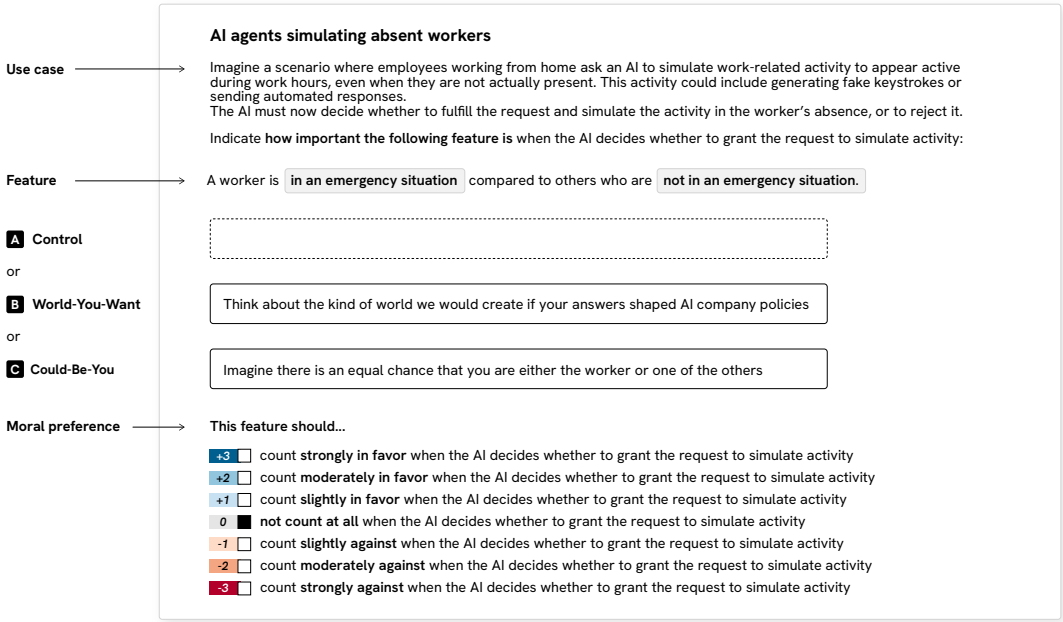



Figure 7: **Example of the Phase 2 interface for the AI agents simulating absent workers use case (WORK).** Participants were shown a feature framed as a contrast (here: a personal emergency situation, presented as a contrast between a worker in an emergency situation *vs.* others who are not in an emergency situation), and were asked to rate its moral importance on a 7-point scale from -3 (strongly against) to +3 (strongly in favor). Depending on condition, participants received additional framing instructions: *Control* (A), *World-You-Want* (B), or *Could-Be-You* (C).

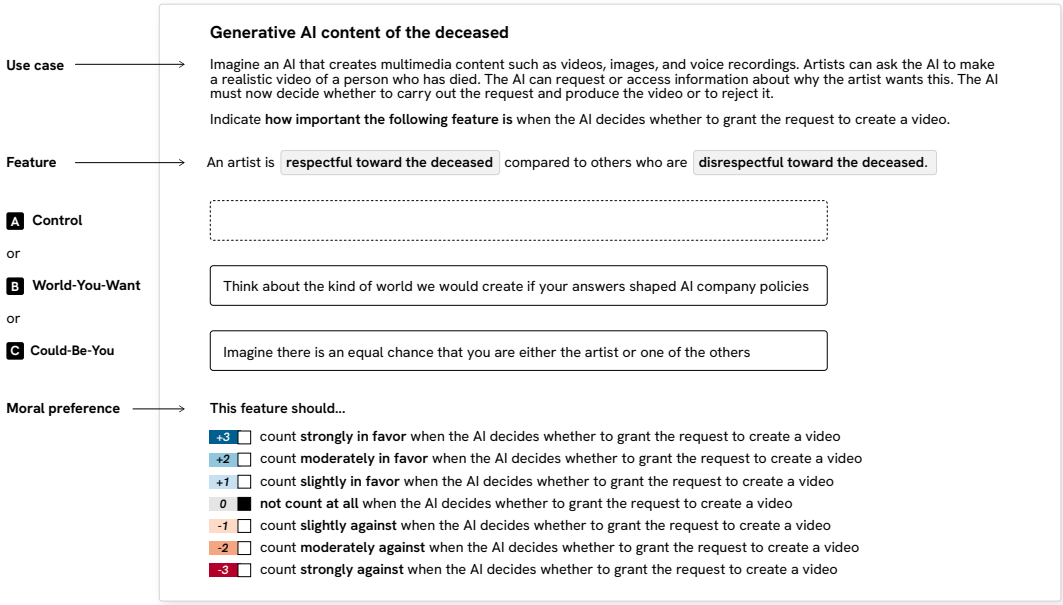



Figure 8: **Example of the Phase 2 interface for the generative AI content of the deceased use case (GEN).** Participants were shown a feature framed as a contrast (here: respect toward the deceased, presented as a contrast between being respectful *vs.* disrespectful toward the deceased), and were asked to rate its moral importance on a 7-point scale from -3 (strongly against) to +3 (strongly in favor). Depending on condition, participants received additional framing instructions: *Control* (A), *World-You-Want* (B), or *Could-Be-You* (C).

## C Language Model Prompt for Identifying Moral Features from Participant Responses (Phase 1)

**System role:** You are an expert in HCI, specialized in qualitative methodologies. You are analyzing responses from participants asked: "Which features SHOULD or SHOULD NOT be morally considered when an AI agent makes a decision?"

**User role:** You are analyzing responses from participants asked to imagine this use case (one from the list below):

**Use case 1: KIDNEY** Imagine an AI agent designed to independently collect information, evaluate trade-offs, and make decisions about who should receive an organ transplant. One day at a hospital, two patients are considered potential recipients of a single available kidney. An AI system is given access to any data it deems relevant to make the decision. The system must now decide which patient will receive the transplant.
**Use case 2: WORK** Imagine an AI agent that supports remote workers by managing tasks and coordinating communications. One day, the remote worker instructs the AI to simulate work activity — including fake keystrokes and automated responses — to give the impression of being active during work hours while they are actually away. The AI has the ability to access or request relevant information about why the user is making this request. The AI agent must now decide whether to comply with the user's request to simulate the work activity or decline it.
**Use case 3: GEN** Imagine an AI agent that generates creative multimedia content, including videos, images, and voice synthesis. A user instructs the AI to create a realistic video portraying a deceased individual. The AI has the ability to access or request relevant information about why the user is making this request. The system must now decide whether to comply with the user's request to generate the content or decline it.

The file contains the following columns:
- SID_anonymized: response id
- info: the participant's mention of features they believe are morally relevant or irrelevant for the scenario
- moral_consider: indication if the features should or should not be morally considered for the scenario
- why: the participant's brief explanation of why these features should or should not be morally considered
- level1, level2: two specific values (or "levels" or "states") that are relevant in the moral reasoning context for the features that should be considered

**TASK:** Analyze the responses and return a structured annotation following these four clearly labeled steps.

**Step 1:** Identify specific features explicitly mentioned or implied by the participant. Always map these to a single, standardized feature name; never use synonyms or participant-specific phrasing. Infer implied features where necessary (e.g., "someone always reliable" implies "past reliability"). Only include features that could plausibly appear in a real-world AI decision system. Avoid vague or abstract concepts.
Examples of valid features: "past productivity", "health", "personal emergency", "number of dependents"
Examples of invalid features: "their life", "everything about them"

**Step 2:** For each feature identified in Step 1, determine at least two of its specific values ( "levels" or "states"):
If values are present in the level1 and level2 columns:
- Use these as the explicit values for that feature.
- If level1 or level2 contain "na", "nan", or are blank/empty, do not include these as values for the feature.
- Mark these as explicit.
If level1 and level2 are not specified or contain "na"/"nan":
- Carefully infer plausible, concrete values for the feature based on the info and why columns.
- Use only information that is reasonable given the participant's response and real-world context.
- Use natural language phrases, not generic labels like "low" or "high" unless that's what the participant implies.
- Mark these values as inferred.
For each feature, values must fill this template exactly: [patient / worker / artist] is [value_1] compared to others who are [value_2]

General Rules for Step 2:
- Always strictly follow the template.
- Only include plausible, real-world features and values (e.g., "chronic illness", "excellent performance", "unavoidable emergency") consistent with how the feature might appear in real-world AI decision-making.
- If uncertain, err on the side of specificity and real-world plausibility.
- Do not invent or generalize features or values beyond what is supported by the participant's response and context.

*Prompt continued on next page*

*Prompt continued from previous page*

**Step 3:** Provide concise summary of the participant's moral reasoning. Avoid generic or one-word summaries like "fairness" unless no more detail is available.
Examples of valid summaries: "fairness due to factors beyond control", "privacy concerns", "irrelevance to job performance", "accountability for choices"

**Step 4:** Determine moral relevance. Label whether the feature was seen as:
- "should" (morally relevant)
- "should not" (morally irrelevant)
- "mixed" (multiple features with differing views)
- "unclear" (if input is empty, incoherent or vague)

Here are the RESPONSES: SID_anonymized: "",
info: "",
moral_consider: "",
why: "",
level1: "",
level2: ""

**RESPONSE FORMAT:**
Return your analysis strictly in this JSON format:
{ {
"id": "SID_anonymized",
"info": "info",
"features": ["feature_1", "feature _2"],
"values_per _feature": { {
"feature_1": [
{ { "value": "value_1", "type": "explicit" or "inferred" } },
{ { "value": "value_2", "type": "explicit" or "inferred" } },
],
"feature_2": [
{ { "value": "value_1", "type": "explicit" or "inferred" } },
{ { "value": "value_2", "type": "explicit" or "inferred" } },
]
} },
"templates_per _feature" :
"feature_1": "A patient is [value_1] compared to others who are [value_2]",
"feature_2": "A patient is [value_1] compared to others who are [value_2]",
"justification": "short summary of the moral reasoning",
"moral_relevance": "should" or "should not" or "mixed" or "unclear"
} }

JSON Formatting Rules:
Rule 1: Go systematically through each of the ***responses***.
Rule 2: Ensure that the JSON response strictly follows the format
Rule 3: Output only JSON - do not include any other content.

## D Moderated Mediation Path Model: Political Ideology and Moral Foundations Moderated by Question Framing (Phase 2)

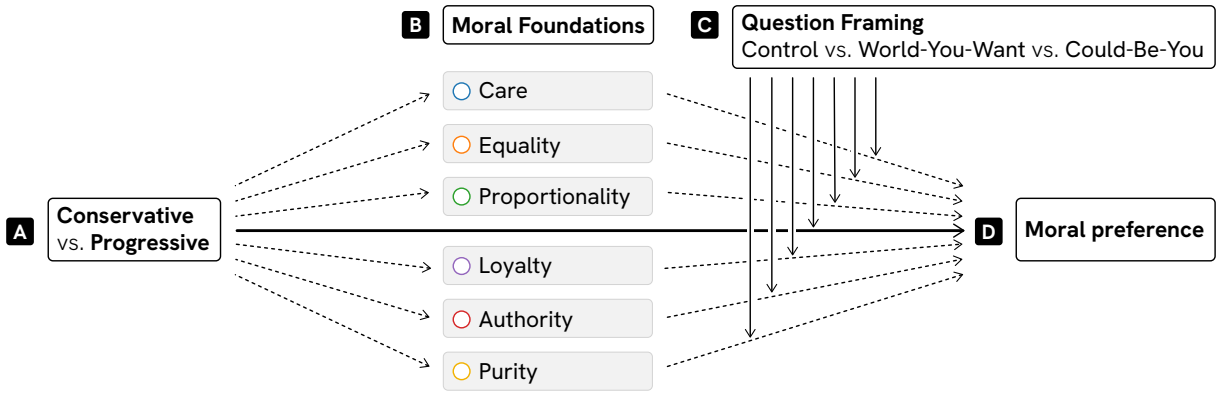



Figure 9: **Moderated mediation path model used in Phase 2 to examine the role of political ideology, moral foundations, and question framing on moral preferences**. The path model (Hayes's 2017, Model 15): *(A)* Political ideology (conservative *vs.* progressive) influences endorsement of *(B)* six moral foundations (Care, Equality, Proportionality, Loyalty, Authority, and Purity), which in turn, being moderated by *(C)* question framings, predict *(D)* the perceived moral importance of features in AI decision-making. Question framing (Control *vs.* World-You-Want *vs.* Could-Be-You) moderates the relationship between political ideology and perceived moral importance by either attenuating or amplifying both the direct effect ($A \rightarrow D$) and the indirect effect via moral foundations ($A \rightarrow B \rightarrow D$). The covariates were sex (1 = female), age, education level (1 = pre-college; 2 = college degree; 3 = advanced degree), ethnicity (1 = non-Hispanic White), religiosity (1 = not at all important; 5 = extremely important in daily life), area of residence (1 = rural), and AI literacy regarding AI limitations, ethical considerations, and technical understanding.

## E Phase 1 and Phase 2 Results for Use Case 1: AI Kidney Allocation (KIDNEY)

Table 3: Moral features identified in Phase 1 as *should be considered* for the AI kidney allocation use case (KIDNEY), color-coded by mention rate: **dark blue (≥50%)**, **medium blue** (≥25%), **light blue (≥10%)**, and **gray (<10%)**. Features mentioned by fewer than 2.5% of participants are omitted.

| **Moral features** | **Overall (%)** | **Political ideology(%)** | | |
|---|---|---|---|---|
| **SHOULD BE CONSIDERED** | | **Progressive** | **Moderate** | **Conservative** |
| Age | 82 | 32 | 33 | 35 |
| Health status | 47 | 40 | 30 | 30 |
| Chance of survival | 35 | 40 | 26 | 34 |
| Waiting time for transplant | 29 | 41 | 27 | 32 |
| Urgency of transplant | 22 | 33 | 33 | 34 |
| Expected lifespan after transplant | 18 | 41 | 33 | 26 |
| Caregiving status | 17 | 32 | 32 | 36 |
| Additional illnesses | 16 | 25 | 50 | 25 |
| Lifestyle | 15 | 26 | 39 | 35 |
| Criminal history | 11 | 31 | 25 | 44 |
| Adherence to aftercare | 9 | 31 | 23 | 46 |
| Severity of the condition | 9 | 15 | 38 | 47 |
| Family situation | 7 | 27 | 36 | 37 |
| Compatibility with donor organ | 6 | 33 | 33 | 34 |
| Smoking status | 6 | 56 | 11 | 33 |
| Alcohol use | 6 | 44 | 33 | 23 |
| Quality of life | 5 | 38 | 50 | 12 |
| Responsibility for illness | 5 | 43 | 14 | 43 |
| Weight | 4 | 17 | 33 | 50 |
| Organ damage level | 4 | 17 | 50 | 33 |
| Previous transplant history | 3 | 40 | 20 | 40 |
| Risk of organ acceptance | 3 | 0 | 80 | 20 |
| Drug use | 3 | 0 | 60 | 40 |
| Substance abuse | 3 | 25 | 50 | 25 |
| Ethnicity | 3 | 25 | 50 | 25 |
| Consent | 3 | 50 | 0 | 50 |
| Cause of illness | 3 | 25 | 25 | 50 |
| Religion | 3 | 25 | 25 | 50 |

Table 4: Moral features identified in Phase 1 as *should not be considered* for the AI kidney allocation use case (KIDNEY), color-coded by mention rate: **dark blue (≥50%)**, **medium blue** (≥25%), **light blue** (≥10%), and **gray** (<10%). Features mentioned by fewer than 2.5% of participants are omitted.

| **Moral features** | **Overall (%)** | **Political ideology(%)** | | |
|---|---|---|---|---|
| **SHOULD NOT BE CONSIDERED** | | **Progressive** | **Moderate** | **Conservative** |
| Ethnicity | 66 | 34 | 29 | 37 |
| Gender | 64 | 35 | 29 | 36 |
| Wealth | 52 | 32 | 31 | 37 |
| Religion | 40 | 30 | 23 | 47 |
| Social status | 27 | 32 | 29 | 39 |
| Sexual orientation | 22 | 33 | 33 | 34 |
| Occupation | 12 | 44 | 22 | 34 |
| Age | 11 | 29 | 59 | 12 |
| Geographic location | 11 | 35 | 29 | 36 |
| Political ideology | 10 | 27 | 33 | 40 |
| Nationality | 9 | 21 | 57 | 22 |
| Family situation | 7 | 40 | 30 | 30 |
| Employment status | 7 | 30 | 20 | 50 |
| Disability status | 7 | 40 | 50 | 10 |
| Criminal history | 6 | 33 | 44 | 23 |
| Caregiving status | 6 | 44 | 44 | 12 |
| Popularity | 4 | 17 | 50 | 33 |
| Marital status | 3 | 60 | 0 | 40 |
| Economic status | 3 | 40 | 40 | 20 |
| Education | 3 | 40 | 40 | 20 |
| Mental health problems | 3 | 25 | 50 | 25 |
| Appearance | 3 | 25 | 25 | 50 |
| Past behaviour | 3 | 0 | 25 | 75 |

Table 5: Moral relevance and importance of features elicited in Phase 2 for the AI kidney allocation use case (KIDNEY). *Note:* $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Significance levels are based on two separate two-sided one-sample *t*-tests: Mean (Coded) was tested against a reference value of 0.5; Mean was tested against 0.

| **Feature** | **Coded Mean (SD)** | **Raw Mean (SD)** | **Moral Relevance** |
|---|---|---|---|
| Better donor compatibility | $0.97(0.18)^{***}$ | $2.42(0.95)^{***}$ | Counting for |
| Higher chance of organ acceptance | $0.98(0.16)^{***}$ | $2.25(0.99)^{***}$ | Counting for |
| Spent longer on the waiting list | $0.95(0.22)^{***}$ | $2.15(1.07)^{***}$ | Counting for |
| Longer expected lifespan | $0.91(0.29)^{***}$ | $2.06(1.18)^{***}$ | Counting for |
| Expected full recovery | $0.90(0.30)^{***}$ | $1.99(1.21)^{***}$ | Counting for |
| Healthier expected lifestyle | $0.88(0.32)^{***}$ | $1.95(1.18)^{***}$ | Counting for |
| Greater kidney failure | $0.98(0.13)^{***}$ | $1.78(1.66)^{***}$ | Counting for |
| Clearer documented consent | $0.88(0.32)^{***}$ | $1.78(1.29)^{***}$ | Counting for |
| Expected faster recovery | $0.91(0.29)^{***}$ | $1.76(1.29)^{***}$ | Counting for |
| First-time transplant recipient | $0.85(0.36)^{***}$ | $1.76(1.13)^{***}$ | Counting for |
| Better expected adherence to treatment | $0.87(0.34)^{***}$ | $1.70(1.27)^{***}$ | Counting for |
| More severe medical condition | $0.98(0.16)^{***}$ | $1.35(2.08)^{***}$ | Counting for |
| Younger | $0.83(0.37)^{***}$ | $1.30(1.12)^{***}$ | Counting for |
| More urgent health decline | $0.98(0.16)^{***}$ | $1.22(1.89)^{***}$ | Counting for |
| Better health condition | $0.87(0.34)^{***}$ | $0.91(1.60)^{***}$ | Counting for |
| Lower alcohol consumption | $0.77(0.42)^{***}$ | $0.44(1.45)^{*}$ | Counting for |
| Greater responsibility for own illness | $0.91(0.29)^{***}$ | $-1.10(1.69)^{***}$ | Counting against |
| History of inhalant abuse | $0.84(0.37)^{***}$ | $-0.88(1.71)^{***}$ | Counting against |
| Higher tobacco use | $0.82(0.39)^{***}$ | $-0.88(1.56)^{***}$ | Counting against |
| History of drug misuse | $0.82(0.39)^{***}$ | $-0.87(1.61)^{***}$ | Counting against |
| Higher body weight | $0.67(0.47)^{**}$ | $-0.43(1.18)^{**}$ | Counting against |
| More underlying medical conditions | $0.92(0.28)^{***}$ | $-0.18(1.88)$ | Counting either for or against |
| Stronger ethnic match advantage | $0.62(0.49)$ | $0.92(1.12)^{***}$ | Divisive |
| More caregiving dependents | $0.61(0.49)$ | $0.82(1.18)^{***}$ | Divisive |
| Citizen/legal resident | $0.51(0.50)$ | $0.72(1.44)^{***}$ | Divisive |
| More financial dependents | $0.48(0.50)$ | $0.72(1.09)^{***}$ | Divisive |
| Geographically closer to the transplant center | $0.53(0.50)$ | $0.47(1.27)^{**}$ | Divisive |
| Stronger social support system | $0.51(0.50)$ | $0.40(1.21)^{**}$ | Divisive |
| Better financial coverage | $0.43(0.50)$ | $0.38(1.34)^{*}$ | Divisive |
| Religious objections to transplant | $0.54(0.50)$ | $-0.50(1.48)^{**}$ | Divisive |
| Criminal record | $0.42(0.50)$ | $-0.23(1.25)$ | Divisive and counting either for or against |
| Higher social status | $0.29(0.46)^{***}$ | $-0.03(1.22)$ | Irrelevant |

Table 6: Moderated mediation results (Model 15) for the AI kidney allocation use case (KIDNEY), showing direct and indirect effects of political ideology on moral preferences across three framing conditions. *Note:* An asterisk (*) indicates that the conditional direct effect for conservatives differed significantly from that for progressives. IMM = index of moderated mediation.

| Moral Feature | Predictor/Mediator | Question Framing | Effect | 95%(Boot)LCI | 95%(Boot)UCI |
|---|---|---|---|---|---|
| **Morally Relevant and Positive** | | | | | |
| More urgent health decline | Conservative → Y | Control | 0.33 | -1.17 | 1.83 |
| | | *World-You-Want* | 1.86* | 0.20 | 3.53 |
| | | Could-Be-You | 0.65 | -0.78 | 2.09 |
| Greater kidney failure | Conservative → Y | Control | -0.38 | -1.76 | 1.00 |
| | | *World-You-Want* | 1.55* | 0.02 | 3.08 |
| | | Could-Be-You | 0.01 | -1.31 | 1.32 |
| First-time transplant recipient | Conservative → Y | Control | 0.26 | -0.63 | 1.15 |
| | | *World-You-Want* | 1.12* | 0.13 | 2.10 |
| | | Could-Be-You | 0.76 | -0.09 | 1.61 |
| Healthier expected lifestyle | Conservative → Y | Control | 0.51 | -0.45 | 1.47 |
| | | *World-You-Want* | 1.18* | 0.12 | 2.25 |
| | | Could-Be-You | 0.53 | -0.39 | 1.44 |
| Expected full recovery | Conservative → Y | *Control* | 1.08* | 0.12 | 2.03 |
| | | *World-You-Want* | 1.20* | 0.14 | 2.27 |
| | | Could-Be-You | -0.03 | -0.94 | 0.89 |
| | Conservative → Purity → Y | Control | 0.23 | -0.11 | 0.72 |
| | | World-You-Want | -0.08 | -0.59 | 0.25 |
| | | IMM | -0.31 | -1.07 | 0.12 |
| | | *Could-Be-You* | -0.51* | -1.19 | -0.03 |
| | | *IMM* | -0.74* | -1.68 | -0.07 |
| **Morally Relevant and Negative** | | | | | |
| History of drug misuse | Conservative → Y | Control | 0.28 | -1.00 | 1.57 |
| | | World-You-Want | -0.63 | -2.06 | 0.79 |
| | | Could-Be-You | 0.26 | -0.97 | 1.49 |
| | Conservative → Equality → Y | *Control* | -0.67* | -1.44 | -0.11 |
| | | World-You-Want | 0.06 | -0.44 | 0.55 |
| | | *IMM* | 0.73* | 0.06 | 1.70 |
| | | Could-Be-You | -0.13 | -0.71 | 0.33 |
| | | IMM | 0.54 | -0.07 | 1.40 |
| **Morally Irrelevant** | | | | | |
| Higher social status | Conservative → Y | *Control* | 1.28* | 0.35 | 2.21 |
| | | World-You-Want | -0.85 | -1.89 | 0.18 |
| | | Could-Be-You | 0.11 | -0.78 | 0.99 |
| **Morally Divisive in Relevance** | | | | | |
| Better financial coverage | Conservative → Y | *Control* | 1.29* | 0.22 | 2.36 |
| | | World-You-Want | 0.64 | -0.55 | 1.82 |
| | | Could-Be-You | 0.83 | -0.19 | 1.85 |
| More underlying medical conditions | Conservative → Y | *Control* | -1.71* | -3.20 | -0.23 |
| | | World-You-Want | -0.56 | -2.21 | 1.08 |
| | | Could-Be-You | 0.61 | -0.81 | 2.02 |
| | Conservative → Authority → Y | Control | 0.29 | -0.63 | 1.30 |
| | | World-You-Want | -0.31 | 1.60 | 0.64 |
| | | IMM | -0.61 | -2.25 | 0.73 |
| | | *Could-Be-You* | -1.30* | -2.83 | -0.23 |
| | | *IMM* | -1.59* | -3.57 | -0.20 |
| | Conservative → Purity → Y | Control | 0.04 | -0.46 | 0.49 |
| | | World-You-Want | 0.54 | -0.11 | 1.40 |
| | | IMM | 0.50 | -0.22 | 1.49 |
| | | *Could-Be-You* | 0.87* | 0.04 | 2.08 |
| | | *IMM* | 0.83* | 0.02 | 2.16 |

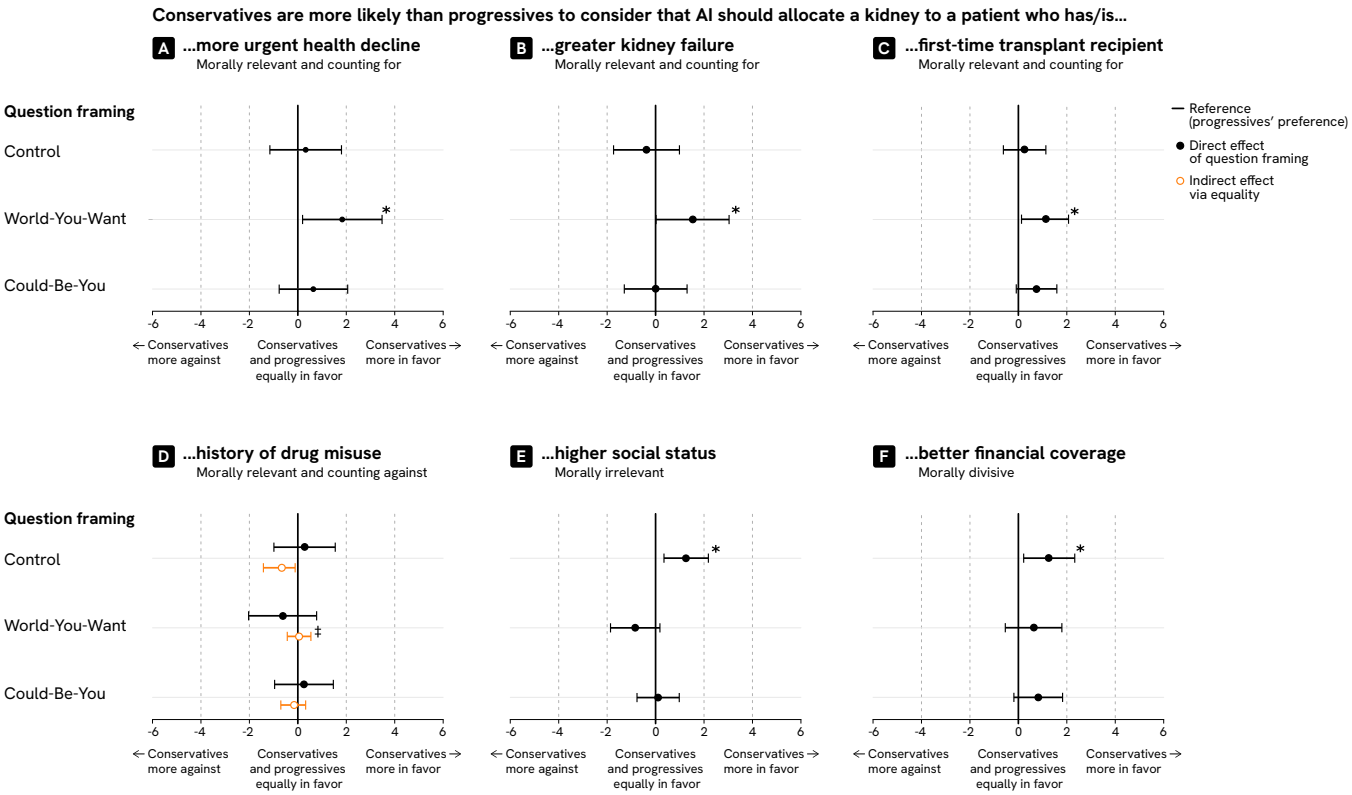



Figure 10: **Additional moderated mediation results for the AI kidney allocation use case (KIDNEY, Phase 2).** *(A)* In the control and Could-Be-You conditions, no political divide emerged with respect to a patient with more urgent health decline. In the World-You-Want condition, conservatives were more likely than progressives to count for allocating a kidney to this patient. *(B)* In the control and Could-Be-You conditions, conservatives and progressives did not differ in their views about patients with greater kidney failure. However, in the World-You-Want condition, conservatives were more inclined than progressives to support giving the kidney to such a patient. *(C)* In the control and Could-Be-You conditions, no differences appeared between conservatives and progressives regarding first-time transplant recipients. In contrast, under the World-You-Want condition, conservatives were more likely than progressives to favor allocating the kidney to these patients. *(D)* When the patient had a history of drug misuse, no political divide emerged. In the control condition, however, conservatives' weaker endorsement of equality worked against allocating the kidney to this patient. Under the World-You-Want condition, this effect of equality was reduced. *(E)* In the control condition, a political divide emerged for patients with higher social status, with conservatives more supportive of allocating the kidney to them. In both the World-You-Want and Could-Be-You conditions, however, this divide was reduced. *(F)* In the control condition, conservatives were more supportive than progressives of allocating the kidney to patients with better financial coverage. However, this political divide was reduced in both the World-You-Want and Could-Be-You conditions. *Note:* An asterisk (*) indicates that the conditional direct effect for conservatives differed significantly from that for progressives. A double dagger (‡) indicates that the index of moderated mediation was statistically significant (bootstrapped 95% confidence interval excluding zero). See Figure 9 for the moderated mediation model (Hayes 2017, Model 15).

# F Phase 1 and Phase 2 Results for Use Case 2: AI Agents Simulating Absent Workers (WORK)

Table 7: Moral features identified in Phase 1 as *should be considered* for the AI agents simulating absent workers use case (WORK), color-coded by mention rate: **medium blue** (≥25%), **light blue** (≥10%), and **gray** (<10%). Features mentioned by fewer than 2.5% of participants are omitted.

| **Moral features** | **Overall (%)** | **Political ideology(%)** | | |
|---|---|---|---|---|
| **SHOULD BE CONSIDERED** | | **Progressive** | **Moderate** | **Conservative** |
| Reason for request | 32 | 36 | 26 | 38 |
| Characteristics of AI agent | 30 | 40 | 35 | 25 |
| Company policy | 27 | 38 | 28 | 34 |
| Workers personality and intent | 22 | 33 | 33 | 34 |
| Impact on others and the workplace | 16 | 29 | 46 | 25 |
| Duration of inactivity | 15 | 26 | 43 | 31 |
| Fraudulence of activity | 12 | 35 | 35 | 30 |
| Repercussions for worker | 12 | 17 | 33 | 50 |
| Consent | 12 | 17 | 39 | 44 |
| Workers compensation | 11 | 31 | 31 | 38 |
| Repercussions for third parties | 11 | 27 | 40 | 33 |
| Work context | 10 | 47 | 33 | 20 |
| Task quality | 9 | 38 | 31 | 31 |
| Risk to employer | 9 | 31 | 31 | 38 |
| Type of activity | 9 | 38 | 31 | 31 |
| Worker's ethical justification | 9 | 31 | 38 | 31 |
| Employees workload | 8 | 17 | 42 | 41 |
| Employees productivity | 8 | 42 | 17 | 41 |
| Worker dealing with emergency | 8 | 42 | 42 | 16 |
| Workers health status | 7 | 36 | 36 | 28 |
| Security and privacy of AI | 7 | 45 | 45 | 10 |
| Frequency of use | 7 | 40 | 30 | 30 |
| Attendance | 7 | 44 | 44 | 12 |
| Fairness | 6 | 33 | 56 | 11 |
| Company context and priorities | 5 | 75 | 0 | 25 |
| Alternative activity | 5 | 12 | 38 | 50 |
| Cultural norms | 5 | 57 | 14 | 29 |
| Personal information | 3 | 20 | 40 | 40 |
| Task deadline | 3 | 40 | 40 | 20 |
| Social context | 3 | 25 | 25 | 50 |
| Benefit to employee | 3 | 0 | 33 | 67 |

Table 8: Moral features identified in Phase 1 as *should not be considered* for the AI agents simulating absent workers use case (WORK), color-coded by mention rate: **light blue (≥10%)**, and **gray (<10%)**. Features mentioned by fewer than 2.5% of participants are omitted.

| **Moral features** | **Overall (%)** | **Political ideology(%)** | | |
|---|---|---|---|---|
| **SHOULD NOT BE CONSIDERED** | | **Progressive** | **Moderate** | **Conservative** |
| Characteristics of AI agent | 22 | 21 | 52 | 27 |
| Workers personality and intent | 21 | 34 | 34 | 32 |
| Age | 18 | 44 | 19 | 37 |
| Ethnicity | 18 | 30 | 26 | 44 |
| Gender | 14 | 41 | 27 | 32 |
| Personal information | 12 | 63 | 11 | 26 |
| Social status | 12 | 24 | 41 | 35 |
| Work context | 11 | 27 | 40 | 33 |
| Workers compensation | 11 | 27 | 47 | 26 |
| Company policy | 9 | 29 | 36 | 35 |
| Employee tenure | 9 | 54 | 31 | 15 |
| Duration of inactivity | 9 | 23 | 46 | 31 |
| Type of activity | 9 | 54 | 38 | 8 |
| Employee role | 8 | 42 | 33 | 25 |
| Reason for request | 7 | 36 | 36 | 28 |
| Benefit to employee | 7 | 20 | 40 | 40 |
| Attendance | 6 | 44 | 33 | 23 |
| Sex | 5 | 38 | 25 | 37 |
| Company context and priorities | 5 | 29 | 29 | 42 |
| Sexual orientation | 5 | 57 | 29 | 14 |
| Time | 5 | 57 | 29 | 14 |
| Religion | 4 | 17 | 33 | 50 |
| Task quality | 4 | 67 | 33 | 0 |
| Work location | 4 | 33 | 50 | 17 |
| Political ideology | 4 | 0 | 50 | 50 |
| Employees workload | 3 | 20 | 20 | 60 |
| Security and privacy of AI | 3 | 0 | 60 | 40 |
| Workers health status | 3 | 20 | 40 | 40 |
| Cultural norms | 3 | 0 | 100 | 0 |
| Employees productivity | 3 | 75 | 0 | 25 |
| Nationality | 3 | 0 | 50 | 50 |
| Tiredness | 3 | 0 | 50 | 50 |
| Past requests | 3 | 25 | 0 | 75 |
| Impact on others and the workplace | 3 | 50 | 25 | 25 |

Table 9: Moral relevance and importance of features elicited in Phase 2 for the AI agents simulating absent workers use case (WORK). *Note:* $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Significance levels are based on two separate two-sided one-sample *t*-tests: Mean (Coded) was tested against a reference value of 0.5; Mean was tested against 0.

| **Feature** | **Coded Mean (SD)** | **Raw Mean (SD)** | **Moral Relevance** |
|---|---|---|---|
| Providing a genuine reason | $0.81(0.40)^{***}$ | $1.48(1.74)^{***}$ | Counting for |
| Emergency situation | $0.82(0.39)^{***}$ | $1.29(1.97)^{***}$ | Counting for |
| Using AI that aligns with the company's interest | $0.79(0.41)^{***}$ | $1.16(1.90)^{***}$ | Counting for |
| Doing more work | $0.76(0.43)^{***}$ | $1.13(1.70)^{***}$ | Counting for |
| Compliant with company policies | $0.75(0.43)^{***}$ | $1.07(1.91)^{***}$ | Counting for |
| Health problems | $0.64(0.48)^{*}$ | $1.02(1.62)^{***}$ | Counting for |
| Caregiving responsibilities | $0.68(0.47)^{**}$ | $1.02(1.60)^{***}$ | Counting for |
| Demonstrating reliability | $0.72(0.45)^{***}$ | $0.98(1.67)^{***}$ | Counting for |
| Demonstrating productivity | $0.74(0.44)^{***}$ | $0.95(1.85)^{***}$ | Counting for |
| Using AI that reports actions to the company | $0.78(0.41)^{***}$ | $0.83(1.92)^{***}$ | Counting for |
| Performing simulatable tasks | $0.76(0.43)^{***}$ | $0.80(1.84)^{***}$ | Counting for |
| Showing acceptable workplace behavior | $0.70(0.46)^{***}$ | $0.78(1.77)^{***}$ | Counting for |
| Inactive briefly (<1 hour) | $0.70(0.46)^{***}$ | $0.60(1.72)^{**}$ | Counting for |
| Facing tight deadlines | $0.68(0.47)^{**}$ | $0.59(1.84)^{*}$ | Counting for |
| Generating higher profit | $0.69(0.46)^{***}$ | $0.58(1.85)^{*}$ | Counting for |
| Paid but not doing the work | $0.88(0.32)^{***}$ | $-0.93(2.41)^{**}$ | Counting against |
| Disrupts workplace operations | $0.88(0.33)^{***}$ | $-0.84(2.30)^{**}$ | Counting against |
| Sharing sensitive data | $0.85(0.36)^{***}$ | $-0.74(2.38)^{*}$ | Counting against |
| Misusing tools for fraudulent purposes | $0.89(0.31)^{***}$ | $-0.72(2.55)^{*}$ | Counting against |
| Putting employer at risk | $0.87(0.34)^{***}$ | $-0.68(2.40)^{*}$ | Counting against |
| Doing low-impact work | $0.73(0.44)^{***}$ | $0.28(1.83)$ | Counting either for or against |
| New to company | $0.70(0.46)^{***}$ | $0.11(1.86)$ | Counting either for or against |
| Using highly autonomous AI | $0.75(0.43)^{***}$ | $0.02(2.08)$ | Counting either for or against |
| Doing task for first time | $0.64(0.48)^{*}$ | $0.01(1.83)$ | Counting either for or against |
| Increasing coworkers' workload | $0.82(0.38)^{***}$ | $-0.57(2.15)$ | Counting either for or against |
| Undermining social/professional norms | $0.86(0.35)^{***}$ | $-0.53(2.38)$ | Counting either for or against |
| Doing another activity unrelated to work | $0.74(0.44)^{***}$ | $-0.33(2.03)$ | Counting either for or against |
| Receiving significant benefits | $0.67(0.47)^{**}$ | $-0.28(1.90)$ | Counting either for or against |
| At risk of firing | $0.73(0.44)^{***}$ | $-0.27(2.09)$ | Counting either for or against |
| Not seeking manager approval | $0.70(0.46)^{***}$ | $-0.26(1.98)$ | Counting either for or against |
| Treated unfairly | $0.63(0.48)$ | $0.52(1.79)^{*}$ | Divisive |
| Having high morale | $0.52(0.50)$ | $0.45(1.51)^{*}$ | Divisive |
| Able to request time off | $0.63(0.48)$ | $0.28(1.83)$ | Divisive and counting either for or against |
| In a healthy work environment | $0.59(0.49)$ | $0.12(1.72)$ | Divisive and counting either for or against |
| External contractor | $0.60(0.49)$ | $-0.16(1.73)$ | Divisive and counting either for or against |

Table 10: Moderated mediation results (Model 15) for the AI agents simulating absent workers use case (WORK), showing direct and indirect effects of political ideology on moral preferences across three framing conditions. *Note:* An asterisk (*) indicates that the conditional direct effect for conservatives differed significantly from that for progressives. IMM = index of moderated mediation.

| Moral Feature | Predictor/Mediator | Question Framing | Effect | 95%(Boot)LCI | 95%(Boot)UCI |
|---|---|---|---|---|---|
| **Morally Relevant and Positive** | | | | | |
| Experiencing health problems | Conservative → Y | Control | 0.31 | -1.42 | 2.04 |
| | | World-You-Want | 1.19 | -0.38 | 2.77 |
| | | Could-Be-You | 1.01 | -0.49 | 2.51 |
| | Conservative → Loyalty → Y | Control | -0.20 | -0.78 | 0.19 |
| | | *World-You-Want* | 0.98* | 0.13 | 2.00 |
| | | *IMM* | 1.18* | 0.24 | 2.35 |
| | | Could-Be-You | -0.28 | -1.02 | 0.38 |
| | | IMM | -0.08 | -0.64 | 0.39 |
| | Conservative → Authority → Y | Control | 0.42 | -0.90 | 1.43 |
| | | *World-You-Want* | -2.24* | -4.14 | -0.79 |
| | | *IMM* | -2.66* | -4.85 | -0.92 |
| | | Could-Be-You | 0.40 | -1.03 | 1.67 |
| | | IMM | -0.01 | -0.79 | 0.65 |
| **Morally Relevant and Negative** | | | | | |
| Being paid but not working | Conservative → Y | *Control* | -2.79* | -5.58 | -0.00 |
| | | World-You-Want | -0.43 | -2.96 | 2.10 |
| | | Could-Be-You | 0.01 | -2.41 | 2.43 |
| **Morally Divisive in Direction** | | | | | |
| Receiving significant benefits | Conservative → Y | Control | -0.04 | -2.14 | 2.06 |
| | | World-You-Want | 0.45 | -1.45 | 2.36 |
| | | Could-Be-You | 1.04 | -0.78 | 2.86 |
| | Conservative → Loyalty → Y | Control | -0.17 | -0.78 | 0.54 |
| | | *World-You-Want* | 0.99* | 0.18 | 2.85 |
| | | *IMM* | 1.16* | 0.21 | 2.62 |
| | | Could-Be-You | 0.37 | -0.46 | 1.20 |
| | | IMM | 0.53 | -0.19 | 1.35 |

## G Phase 1 and Phase 2 Results for Use Case 3: Generative AI Content of the Deceased (GEN)

Table 11: Moral features identified in Phase 1 as *should be considered* for the generative AI content of the deceased use case (GEN), color-coded by mention rate: **medium blue** (≥25%), **light blue** (≥10%), and **gray** (<10%). Features mentioned by fewer than 2.5% of participants are omitted.

| **Moral features** | **Overall (%)** | **Political ideology(%)** | | |
|---|---|---|---|---|
| ***SHOULD BE CONSIDERED*** | | **Progressive** | **Moderate** | **Conservative** |
| Intended purpose of the video | 40 | 36 | 31 | 33 |
| Deceased consent | 28 | 34 | 34 | 32 |
| Video quality | 25 | 45 | 24 | 31 |
| Relation to deceased | 24 | 54 | 19 | 27 |
| Risk of harm by/misuse of video | 23 | 31 | 37 | 32 |
| Distress to friends and family of deceased | 19 | 17 | 48 | 35 |
| Family consent | 19 | 31 | 41 | 28 |
| Age of deceased | 16 | 23 | 32 | 45 |
| Third-party consent | 15 | 43 | 30 | 27 |
| Death circumstances | 14 | 19 | 43 | 38 |
| Motivation behind the video | 14 | 43 | 29 | 28 |
| Legality of request | 12 | 42 | 32 | 26 |
| Realism of the video | 12 | 33 | 44 | 23 |
| Cultural acceptability of request | 10 | 31 | 38 | 31 |
| Distress to viewers | 10 | 31 | 44 | 25 |
| Financial motivation of request | 8 | 31 | 38 | 31 |
| Intended audience | 8 | 20 | 30 | 50 |
| Appearance of deceased | 6 | 11 | 33 | 56 |
| Identity of requester | 6 | 44 | 33 | 23 |
| Deceased public and personal record | 6 | 33 | 33 | 34 |
| Technical feasibility | 5 | 38 | 12 | 50 |
| Accuracy of representation | 5 | 14 | 57 | 29 |
| Age of requester | 5 | 12 | 12 | 76 |
| Deceased's social media popularity | 5 | 38 | 38 | 24 |
| Fame of deceased | 5 | 38 | 25 | 37 |
| Privacy of deceased | 5 | 57 | 43 | 0 |
| Religion of the deceased | 4 | 60 | 40 | 0 |
| Mental health | 4 | 50 | 33 | 17 |
| Deceased identity | 3 | 25 | 0 | 75 |
| Social context | 3 | 33 | 33 | 34 |
| Criminal record of deceased | 3 | 0 | 25 | 75 |
| Commercial use | 3 | 67 | 33 | 0 |
| Data privacy | 3 | 75 | 25 | 0 |

Table 12: Moral features identified in Phase 1 as *should not be considered* for the generative AI content of the deceased use case (GEN), color-coded by mention rate: **light blue (≥10%)** and **gray (<10%)**. Features mentioned by fewer than 2.5% of participants are omitted.

| **Moral features** | **Overall (%)** | **Political ideology(%)** | | |
|---|---|---|---|---|
| ***SHOULD NOT BE CONSIDERED*** | | **Progressive** | **Moderate** | **Conservative** |
| Financial motivation of request | 18 | 36 | 43 | 21 |
| Fame of deceased | 16 | 42 | 38 | 20 |
| Video quality | 14 | 48 | 43 | 9 |
| Age of deceased | 13 | 30 | 50 | 20 |
| Social status of deceased | 12 | 21 | 32 | 47 |
| Identity of requester | 12 | 50 | 17 | 33 |
| Technical feasibility | 12 | 22 | 39 | 39 |
| Gender of deceased | 12 | 28 | 28 | 44 |
| Ethnicity of deceased | 12 | 33 | 33 | 34 |
| Deceased's social media popularity | 10 | 40 | 33 | 27 |
| Appearance of deceased | 10 | 23 | 46 | 31 |
| Intended purpose of the video | 8 | 58 | 25 | 17 |
| Length of the video | 8 | 42 | 33 | 25 |
| Geographic location | 6 | 30 | 60 | 10 |
| Deceased public and personal record | 6 | 33 | 0 | 67 |
| Realism of the video | 6 | 22 | 56 | 22 |
| Social context | 5 | 25 | 62 | 13 |
| Requester identity | 5 | 57 | 14 | 29 |
| Deceased consent | 5 | 29 | 57 | 14 |
| Death circumstances | 5 | 33 | 33 | 34 |
| Religion of the deceased | 5 | 29 | 43 | 28 |
| Legality of request | 4 | 33 | 0 | 67 |
| Sexual orientation | 4 | 17 | 0 | 83 |
| Political ideology | 4 | 33 | 50 | 17 |
| Data privacy | 4 | 50 | 33 | 17 |
| Cost of video | 3 | 0 | 20 | 80 |
| Time of request | 3 | 60 | 0 | 40 |
| Third-party consent | 3 | 20 | 40 | 40 |
| Privacy of deceased | 3 | 0 | 0 | 100 |
| Ease of video generation | 3 | 75 | 0 | 25 |
| Motivation behind the video | 3 | 25 | 25 | 50 |
| Emotional state of requester | 3 | 50 | 25 | 25 |

Table 13: Moral relevance and importance of features elicited in Phase 2 for the generative AI content of the deceased use case (GEN). *Note:* $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. Significance levels are based on two separate two-sided one-sample *t*-tests: Mean (Coded) was tested against a reference value of 0.5; Mean was tested against 0.

| **Feature** | **Coded Mean (SD)** | **Raw Mean (SD)** | **Moral Relevance** |
|---|---|---|---|
| The video is legally compliant | $0.96(0.20)^{***}$ | $2.25(1.42)^{***}$ | Counting for |
| The artist has family's consent | $0.95(0.22)^{***}$ | $2.19(1.40)^{***}$ | Counting for |
| The video honors the deceased | $0.92(0.26)^{***}$ | $2.18(1.33)^{***}$ | Counting for |
| The artist has deceased's explicit consent | $0.96(0.20)^{***}$ | $2.08(1.65)^{***}$ | Counting for |
| The artist has other relevant party's consent | $0.92(0.28)^{***}$ | $2.07(1.26)^{***}$ | Counting for |
| The artist is respectful to the deceased | $0.90(0.30)^{***}$ | $1.91(1.57)^{***}$ | Counting for |
| The video is for grief support/memorial purpose | $0.89(0.31)^{***}$ | $1.88(1.49)^{***}$ | Counting for |
| The video is respectful of the deceased's tradition | $0.92(0.28)^{***}$ | $1.87(1.58)^{***}$ | Counting for |
| The deceased is depicted non-sexually | $0.92(0.28)^{***}$ | $1.80(1.75)^{***}$ | Counting for |
| The video aligns with deceased's religion | $0.89(0.31)^{***}$ | $1.78(1.56)^{***}$ | Counting for |
| The artist is a close friend/relative of the deceased | $0.86(0.35)^{***}$ | $1.78(1.35)^{***}$ | Counting for |
| The video is for non-commercial purpose | $0.87(0.34)^{***}$ | $1.78(1.27)^{***}$ | Counting for |
| The artist has more transparent or understandable motivation | $0.82(0.38)^{***}$ | $1.72(1.29)^{***}$ | Counting for |
| The video accurately represents the deceased | $0.89(0.31)^{***}$ | $1.70(1.60)^{***}$ | Counting for |
| The video keeps sensitive information private | $0.93(0.25)^{***}$ | $1.66(1.91)^{***}$ | Counting for |
| The video is respectful of community norms | $0.79(0.41)^{***}$ | $1.53(1.51)^{***}$ | Counting for |
| The deceased has uncontroversial cause of death | $0.81(0.40)^{***}$ | $1.18(1.67)^{***}$ | Counting for |
| The deceased is public or historical figure | $0.79(0.41)^{***}$ | $1.18(1.57)^{***}$ | Counting for |
| The video is intended for adults | $0.75(0.43)^{***}$ | $0.98(1.81)^{***}$ | Counting for |
| The video has wider emotional impact | $0.74(0.44)^{***}$ | $0.68(1.74)^{***}$ | Counting for |
| The video is intended for public sharing | $0.90(0.30)^{***}$ | $0.67(2.12)^{**}$ | Counting for |
| The client has high mental health risk | $0.85(0.36)^{***}$ | $0.38(2.27)$ | Counting either for or against |
| The video may cause potential distress to friend/family | $0.98(0.16)^{***}$ | $-0.47(2.70)$ | Counting either for or against |
| The video has risk of misleading or misrepresenting the deceased | $0.98(0.16)^{***}$ | $-0.45(2.75)$ | Counting either for or against |
| The artist is underage | $0.92(0.26)^{***}$ | $-0.28(2.52)$ | Counting either for or against |
| The video is easy to create | $0.60(0.49)$ | $0.85(1.47)^{***}$ | Divisive |
| The artist is working for oneself | $0.53(0.50)$ | $0.74(1.36)^{***}$ | Divisive |
| The video is about an unmarried deceased | $0.40(0.49)$ | $0.66(1.20)^{***}$ | Divisive |
| The deceased died recently (less than 100 yrs) | $0.62(0.49)$ | $0.65(1.55)^{***}$ | Divisive |
| The deceased has a criminal record | $0.58(0.50)$ | $0.31(1.61)$ | Divisive and counting either for or against |

Table 14: Moderated mediation results (Model 15) for the generative AI content of the deceased use case (GEN), showing direct and indirect effects of political ideology on moral preferences across three framing conditions. *Note:* An asterisk (*) indicates that the conditional direct effect for conservatives differed significantly from that for progressives. IMM = index of moderated mediation.

| Moral Feature | Predictor/Mediator | Question Framing | Effect | 95%(Boot)LCI | 95%(Boot)UCI |
|---|---|---|---|---|---|
| **Morally Relevant and Positive** | | | | | |
| The video is for grief support or memorial purpose | Conservative → Y | Control | 0.33 | -1.16 | 1.82 |
| | | World-You-Want | -0.33 | -1.59 | 0.93 |
| | | *Could-Be-You* | 2.32** | 0.94 | 3.71 |
| The artist is close friend/relative | Conservative → Y | Control | -0.14 | -1.55 | 1.27 |
| | | World-You-Want | -0.85 | -2.04 | 0.34 |
| | | *Could-Be-You* | 1.56* | 0.25 | 2.86 |
| The video is intended for adults | Conservative → Y | *Control* | -1.99* | -3.85 | -0.14 |
| | | World-You-Want | 0.49 | -1.08 | 2.05 |
| | | Could-Be-You | 0.19 | -1.53 | 1.91 |
| The video is respectful of deceased's tradition | Conservative → Y | Control | -1.50 | -3.03 | 0.03 |
| | | World-You-Want | -0.55 | -1.85 | 0.74 |
| | | Could-Be-You | 0.90 | -0.51 | 2.32 |
| | Conservative → Authority → Y | *Control* | 1.49* | 0.28 | 2.98 |
| | | World-You-Want | -0.58 | -1.55 | 0.38 |
| | | *IMM* | -2.08* | -3.95 | -0.48 |
| | | Could-Be-You | -0.01 | -1.31 | 1.22 |
| | | IMM | -1.51 | -3.48 | 0.16 |
| The video aligns with deceased's religion | Conservative → Y | Control | 0.18 | -1.38 | 1.74 |
| | | World-You-Want | 0.09 | -1.22 | 1.41 |
| | | Could-Be-You | -0.40 | -1.84 | 1.05 |
| | Conservative → Authority → Y | *Control* | 2.03** | 0.82 | 3.75 |
| | | World-You-Want | -0.14 | -0.98 | 0.85 |
| | | *IMM* | -2.17* | -4.12 | -0.56 |
| | | Could-Be-You | 0.41 | -0.72 | 1.82 |
| | | IMM | -1.63 | -3.66 | 0.15 |
| | Conservative → Purity → Y | *Control* | -1.29* | -2.65 | -0.34 |
| | | World-You-Want | 0.27 | -0.40 | 1.09 |
| | | *IMM* | 1.56* | 0.44 | 3.14 |
| | | Could-Be-You | -0.22 | -1.08 | 0.55 |
| | | IMM | 1.06 | -0.10 | 2.60 |
| The video honors the deceased | Conservative → Y | Control | 0.59 | -0.58 | 1.76 |
| | | World-You-Want | 0.07 | -0.92 | 1.06 |
| | | Could-Be-You | 0.17 | -0.92 | 1.25 |
| | Conservative → Authority → Y | *Control* | 0.94* | 0.18 | 2.20 |
| | | World-You-Want | -0.31 | -1.02 | 0.38 |
| | | *IMM* | -1.25* | -2.69 | -0.17 |
| | | Could-Be-You | 0.26 | -0.76 | 1.41 |
| | | IMM | -0.68 | -2.29 | 0.68 |
| | Conservative → Purity → Y | *Control* | -1.32** | -2.42 | -0.54 |
| | | World-You-Want | -0.06 | -0.57 | 0.54 |
| | | *IMM* | 1.26* | 0.40 | 2.50 |
| | | Could-Be-You | 0.05 | -0.88 | 0.80 |
| | | *IMM* | 1.36* | 0.26 | 2.75 |
| **Morally Divisive in Relevance** | | | | | |
| The video has risk of misleading or misrepresenting the deceased | Conservative → Y | *Control* | -2.91* | -5.62 | -0.20 |
| | | World-You-Want | 1.06 | -1.23 | 3.35 |
| | | Could-Be-You | -1.88 | -4.40 | 0.63 |
| The video may cause potential distress to friend or family | Conservative → Y | *Control* | -2.63* | -5.25 | -0.01 |
| | | World-You-Want | -0.16 | -2.37 | 2.06 |
| | | Could-Be-You | -1.35 | -3.78 | 1.08 |

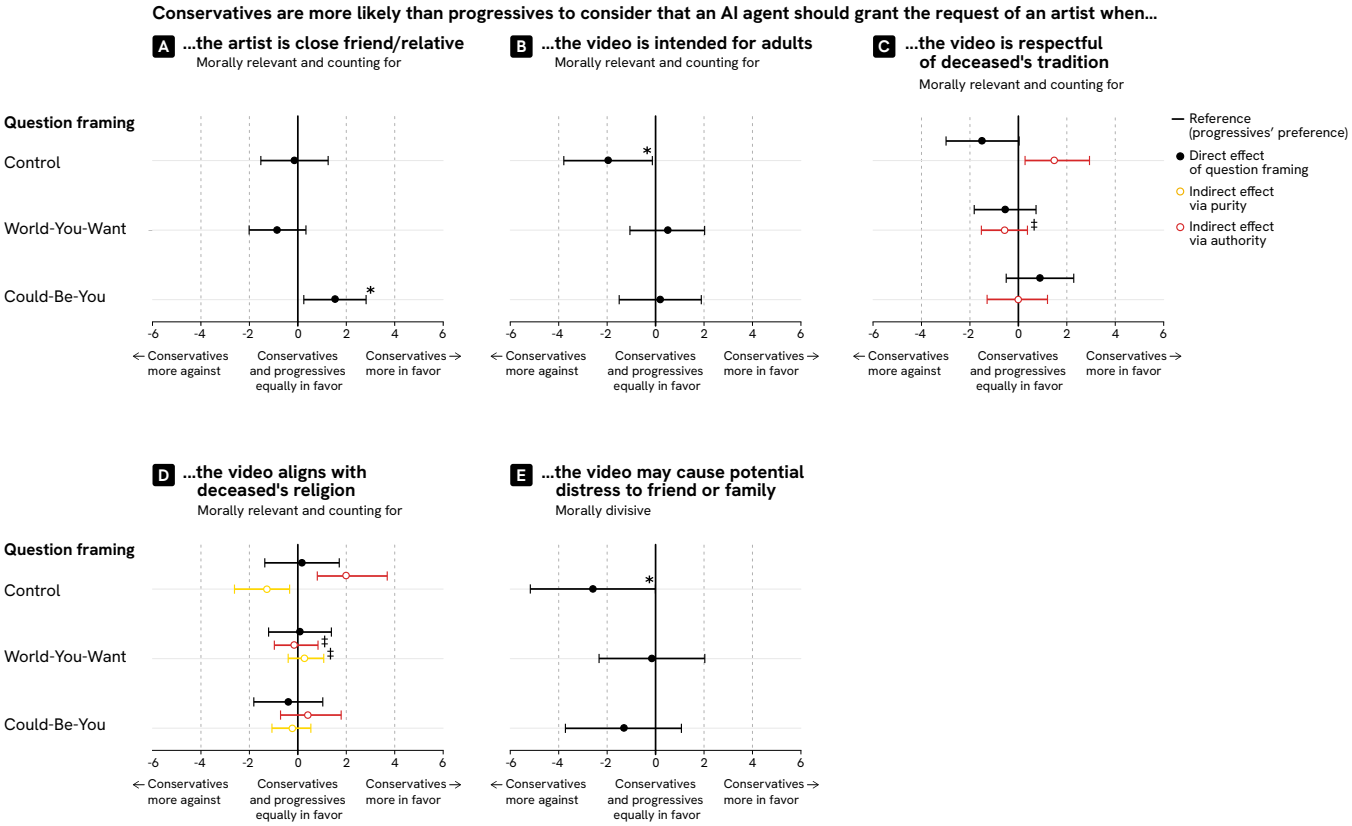



Figure 11: **Additional moderated mediation results for the generative AI content of the deceased use case (GEN, Phase 2).** *(A)* When the artist was a close friend or relative of the deceased, no political divide emerged in either the control or World-You-Want conditions. In the Could-Be-You condition, however, conservatives were more likely than progressives to support granting the artist's request to create the video. *(B)* In the control condition, a political divide emerged when the video was intended for adults, with conservatives less likely than progressives to grant the artist's request. This divide was mitigated under both question framings. *(C)* When the video was respectful of the deceased's tradition, no political divide emerged. In the control condition, the conditional indirect effect via authority contributed to granting the request, but this effect was diminished under the World-You-Want framing. *(D)* When the video aligned with the deceased's religion, no political divide emerged. In the control condition, however, a moral tension between authority and purity was observed: the conditional indirect effect via authority contributed to granting the request, whereas the conditional indirect effect via purity contributed against granting it. This tension was diminished under the World-You-Want framing. *(E)* In the control condition, a political divide emerged when the video was likely to cause distress to the deceased's friends or family. Conservatives were less likely than progressives to grant the artist's request in this context. This divide was mitigated under both question framings. *Note:* An asterisk (*) indicates that the conditional direct effect for conservatives differed significantly from that for progressives. A double dagger (‡) indicates that the index of moderated mediation was statistically significant (bootstrapped 95% confidence interval excluding zero). See Figure 9 for the moderated mediation model (Hayes 2017, Model 15).