Between Reflection and Bias: User Dialogues with Philosophical AI in Moral Dilemmas

ANONYMOUS AUTHOR(S)

Conversational AI increasingly mediates ethical decision-making, yet current HCI approaches often frame morality in generic prompts, overlooking the cultural and philosophical ideologies that guide reflection. We conducted a qualitative study with 21 participants who engaged with six AI personas grounded in distinct ethical frameworks, using think-aloud protocols and post-session interviews. Our findings reveal three patterns: (1) Interactions often reinforced confirmation bias, yet at times prompted reflection; (2) Trust depended on authenticity, with even small inconsistencies undermining credibility; and (3) Cultural and religious orientations shaped interpretation. These findings reveal the dual role of philosophical AI personas: mirrors that validate prior beliefs and as partners that provoke principled disagreement. We contribute empirical insights into how users negotiate trust and reflection with AI personas, and propose design implications for systems that maintain authenticity while remaining adaptable, which supports ethics education, professional training, and deliberation.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; *Collaborative and social computing theory, concepts and paradigms*; • **Computing methodologies** → *Intelligent agents.*

Additional Key Words and Phrases: Conversational AI, AI Personas, Moral Dilemmas, Confirmation Bias, Reflection, Authenticity, Cultural Differences, Human–AI Interaction

ACM Reference Format:

Anonymous Author(s). 2026. Between Reflection and Bias: User Dialogues with Philosophical AI in Moral Dilemmas. In *CHI '26: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems, April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 26 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Moral dilemmas are present in nearly every domain of contemporary life. From healthcare decisions about end-of-life care to public debates on climate change, justice, and emerging technologies, people routinely face questions that facts alone cannot resolve. These dilemmas demand negotiations of fairness, responsibility, and competing values, and they often provoke disagreement even when all parties share the same information. Increasingly, artificial intelligence (AI) systems aid in mediating these morally charged contexts. Whether by curating news, recommending treatment, or guiding policy simulations, AI does more than process data: it shapes the conditions under which ethical judgments are made [1].

Across critical sectors, this entanglement is already visible. In healthcare, clinical decision-support tools recommend how to allocate scarce resources, effectively influencing who receives treatment and who does not [59]. In criminal justice, risk-assessment algorithms such as COMPAS have raised concerns about racial bias and fairness, while continuing to play a role in parole and sentencing decisions [16]. In environmental governance, predictive models guide climate policy and energy distribution, embedding ethical trade-offs about efficiency, equity, and sustainability [33]. In digital life,

Manuscript submitted to ACM

recommender systems curate political and cultural content, tailoring what users see and thereby reinforcing particular values and priorities [32]. These cases highlight that AI rarely acts as a neutral information provider. Instead, it mediates outcomes that reflect contested assumptions about human values and responsibilities [52].

For HCI, this reality introduces a pressing challenge: how can AI systems encourage ethical conversations that build trust and spark reflection, rather than simply providing convenient or agreeable answers? Prior work has sought to address this challenge through mechanisms such as transparency, interpretability, and reflective prompts [19]. These interventions have advanced our understanding of how to make AI more accountable, yet they often treat morality as an abstract checklist or a feature of the interface. Transparency may reveal how a model generates outputs, but it rarely connects with the lived ethical frameworks people use to make sense of fairness, duty, or compassion [42]. Similarly, reflective prompts may nudge users toward reconsideration, but without grounding in recognizable traditions, they risk becoming shallow provocations rather than genuine invitations to reflection.

HCI scholarship offers several promising approaches, yet each leaves gaps. Research on transparency and interpretability shows that revealing system logic can enhance trust, but it tends to disregard cultural and philosophical contexts that structure moral reasoning [20]. Reflective design emphasizes provocation to spark critical thought, but its efficacy in conversational contexts remains uncertain, as users often reinterpret system responses to reinforce prior beliefs [10]. Value-sensitive design(VSD) embeds moral values into design processes, yet it often treats values as universal abstractions, overlooking how traditions, histories, and cultures shape ethical interpretation [8]. Finally, studies on confirmation bias in personalization warn that AI systems designed to accommodate user preferences may amplify selective reasoning, reinforcing comfort rather than encouraging growth [18]. Collectively, these approaches underline the importance of trust, reflection, and value-awareness, but they do not yet show how AI systems might engage users in ways that feel authentically grounded in moral traditions.

This gap points to a broader problem: ethical reasoning is rarely experienced as universal or abstract [58]. People approach moral dilemmas through cultural and philosophical ideologies that render concepts such as duty, virtue, compassion, or forgiveness intelligible [7]. These ideas, religious or secular, act as scaffolds that guide reflection and shape judgments. For example, a person raised in a Confucian-influenced culture may interpret responsibility in relational terms, while someone influenced by utilitarian thinking may emphasize collective outcomes. Such philosophies are not static: they evolve through cultural hybridization, generational reinterpretation, and cross-cultural exchange [27]. Yet they remain powerful frameworks for how people evaluate moral questions. Designing AI without regard for these principles risks producing interactions that feel superficial or inconsistent with users' understandings of ethics in real life.

Our study explores how conversational AI can embody principle-based ethical frameworks as a way of addressing this gap. Rather than relying on generic prompts or universal values, we examine what happens when users encounter AI personas that consistently reflect recognizable philosophies of moral reasoning. This design approach aims to provide two things simultaneously: authentic anchors that make interactions feel trustworthy and comprehensible, and principled disagreements that challenge users to reflect more deeply. At the same time, this approach raises risks. Rigid adherence to a framework may make conversations predictable or overly constrained, while even small inconsistencies may quickly undermine credibility. Furthermore, people interpret moral codes through the lens of their cultural and religious identities, which can either reinforce trust or heighten skepticism.

We conducted a qualitative study with 21 participants from culturally and religiously diverse backgrounds who engaged with multiple AI personas across twelve moral dilemmas. Using dialogue transcripts, think-aloud protocols, and post-session interviews, we examined not only how participants reasoned with AI, but also how authenticity, cultural

orientation, and confirmation bias shaped their engagement. The study revealed three central findings. First, interactions often reinforced users' prior beliefs, yet principled disagreements sometimes prompted genuine reflection. Second, perceived authenticity strongly influenced trust: when personas consistently embodied recognizable philosophies, users found them credible, but minor inconsistencies undermined confidence. Third, cultural and religious orientations served as powerful interpretive lenses, shaping how participants evaluated AI responses and producing divergent pathways of agreement, skepticism, and adaptation.

Collectively, these findings demonstrate the dual role of principle-based AI personas. They can serve as mirrors, validating existing beliefs, but also as reflective partners, stimulating critical engagement through principled disagreement. For HCI, this duality complicates assumptions that reflective prompts or transparent interfaces are sufficient to foster ethical reflection. Designing AI for moral engagement requires balancing authenticity with adaptability, consistency with openness, and cultural familiarity with pluralism.

Our contributions are threefold. Empirically, we present a qualitative study with 21 participants who engaged with multiple principle-grounded AI personas across a wide range of dilemmas. Theoretically, we articulate how these personas play a dual role, simultaneously reinforcing confirmation bias and fostering ethical reflection, while showing how trust depends on perceived authenticity and cultural framing. Design-wise, we propose strategies for conversational AI that balance consistency with flexibility, enabling pluralistic and culturally sensitive systems for ethics education, professional training, and civic deliberation.

2 RELATED WORK

Research on ethical AI in HCI spans normative guidelines, interactive designs for reasoning and reflection, and cross-cultural studies of morality. Each of these strands illuminates important progress, but they also reveal persistent blind spots. In what follows, we review three areas most relevant to our study—ethical issues in AI, AI-supported reasoning and reflection, and cultural differences in morality—while identifying the gaps that our work seeks to address.

2.1 Ethical Issues in AI

In healthcare, decision-support tools must not only maximize survival rates but also consider quality of life, inevitably embedding judgments about whose lives are prioritized [59]. In criminal justice, risk-assessment algorithms such as COMPAS continue to influence parole and sentencing, despite widespread concerns about bias and accountability [23]. In environmental governance, predictive models determine how to allocate energy and simulate climate policy scenarios, often privileging efficiency at the expense of equity [29]. Even in everyday consumer platforms, recommender systems prioritize some cultural or political content over others, embedding implicit values into the fabric of ordinary interactions [25]. These examples underscore that AI is never neutral: it does not simply convey information but mediates outcomes, reflecting contested assumptions about fairness, justice, and responsibility [5].

In response, governments, industry, and advocacy groups have proposed ethical AI guidelines that highlight principles such as fairness, transparency, and accountability [34]. While these principles have advanced the discourse, their abstraction remains a recurring limitation. For example, transparency is often operationalized through explainability dashboards that expose system logic, yet prior HCI research shows that such dashboards rarely improve user understanding in real-world decision contexts [46]. Similarly, accountability frameworks may prescribe documentation of design choices, but users often encounter these as bureaucratic artifacts with little relevance to their lived experience [21]. Without grounding in user-centered interaction design, abstract guidelines risk alienating the very people they are meant to protect [4]. Furthermore, many frameworks assume a universal moral baseline, overlooking the fact that

traditions and cultural worldviews shape how values such as fairness or duty are interpreted [12]. Ethical guidelines, in other words, are necessary but insufficient: they set aspirational ideals but often fail to translate into usable, trusted resources for everyday moral reasoning.

Philosophical thought experiments have long shaped how both philosophers and computer scientists study ethics. The trolley problem, a scenario that asks whether one should sacrifice a single life to save many others, has become especially influential. It reduces complex moral conflicts to stark trade-offs, enabling researchers to examine the logical structure of ethical principles [56]. In machine ethics, the trolley problem has been used to design and test decision-making frameworks for autonomous vehicles, culminating in large-scale studies such as MIT's Moral Machine, which revealed striking cross-cultural differences in moral intuitions [3]. Yet this paradigm has limits. Scholars note that trolley problems strip away relational, contextual, and cultural dimensions that shape how real people deliberate [35]. As a result, trolley problems risk reinforcing the illusion that ethical reasoning can be reduced to binary trade-offs detached from culture or lived experience.

Our work builds on this foundation while addressing its limitations. We employ trolley-type dilemmas as structured devices for reflection. By involving them in conversations with AI personas that represent differing perspectives, we move toward a situated analysis of how users negotiate bias, reflection, and trust when faced with moral trade-offs.

2.2 AI-Supported Reasoning and Reflection

Another body of research in HCI and education highlights the potential of AI to support reasoning, particularly through critical thinking and reflection [22, 50]. While distinct—critical thinking emphasizing systematic judgment and reflection, emphasizing the reconsideration of one's assumptions—these processes are intertwined and both are essential to moral deliberation.

AI has shown promise in supporting each other. Educational research demonstrates that tutoring systems built on large language models (LLMs) can prompt learners to explain their reasoning and consider alternatives, fostering deeper engagement [60]. Large-scale classroom deployments show that AI-assisted reflection tools can increase learner confidence and performance [38]. Beyond education, explainable AI (XAI) has revealed that exposing a system's reasoning can encourage users to scrutinize outputs more carefully, promoting more active engagement with machine decisions [47]. In civic contexts, interactive platforms designed to elicit "critical explanation" have similarly shown that AI can facilitate more deliberate participation in public debate.

Yet these benefits are uneven. Users frequently treat AI responses as validation, interpreting them as confirmations of prior beliefs rather than as provocations for genuine reflection [61]. This pattern mirrors long-standing HCI concerns around personalization: systems designed to accommodate user preferences may unintentionally amplify selective reasoning, deepening confirmation bias [49]. From a design perspective, this highlights a paradox. While AI can prompt reflection, it can also entrench users in their existing views, depending on how interactions are framed. The challenge is not simply to deliver transparent or logical reasoning, but to design dialogues that encourage users to reconsider, probe disagreement, and engage critically with unfamiliar perspectives.

This is the gap our study addresses. We explore whether AI personas grounded in principle-based ethical traditions can shift interactions from affirmation to reflection. By situating reasoning within recognizable frameworks, these personas provide both anchors of authenticity and points of principled disagreement. Unlike generic tutoring agents or explainable systems, they encourage users to explore ethical reasoning not as a set of abstract prompts but as a dialogue with perspectives that have shaped human thought for centuries. In doing so, our study extends reflective design into the domain of conversational AI, where trust, bias, and cultural orientation dynamically shape engagement.

2.3 Cultural Differences in Morality

A third strand of scholarship highlights the cultural and religious foundations of morality. In HCI, researchers caution that socio-technical systems embedding narrow doctrinal assumptions risk reproducing exclusionary norms rather than fostering inclusive dialogue [31].

At the same time, traditions are dynamic. They adapt across generations and through cultural hybridization. Younger East Asians, for example, increasingly reinterpret Confucian principles not as rigid obligations but as flexible ethics balancing individual freedom with relational responsibility [41]. Comparative studies reveal that Buddhists and Christians apply notions of compassion and sanctity differently in life-and-death contexts, such as debates about euthanasia or medical ethics [36]. These findings suggest that moral frameworks cannot be understood as static principles; they are dynamic resources that interact with culture, belief, and social change. Yet, as mentioned above, most HCI research has treated these philosophies descriptively rather than exploring how they might inform the design of interactive systems.

Beyond religion, secular and cultural norms also shape how people deliberate. East Asian societies often valorize diligence and social harmony, shaping how individuals interpret responsibility in education and work [2], while Western contexts emphasize autonomy and individual rights [30]. Such values guide how users interpret fairness, justice, and obligation in ways that directly influence technology use [26]. Research on collaboration platforms illustrates this clearly: collectivist cultures emphasize consensus and group cohesion, while individualist cultures prioritize contribution and recognition [37]. Similarly, AI assistants designed to align with local norms are perceived as more trustworthy and appropriate than those that adopt culturally different frameworks [45].

It is also important to recognize the existence of shared principles across traditions. Concepts such as compassion and justice appear in diverse philosophical and religious systems, providing a foundation for cross-cultural dialogue [39]. HCI research demonstrates how systems built on such shared values can promote civic participation, volunteerism, and fair moderation in online communities [44]. However, applications have remained narrow, and little work has explored how universal values might be embodied in conversational AI to support ethical reasoning [28].

Viewed collectively, these studies highlight a core tension. Moral reasoning is both deeply shaped by religious and cultural traditions and broadly shared through universal principles. Existing HCI systems often resolve this tension by either abstracting morality into universal ideals or localizing it narrowly, overlooking the dynamic ways in which traditions and values interact. Our study extends this body of work by examining how culturally and religiously diverse participants interpret AI personas grounded in principle-based traditions, revealing how authenticity, cultural orientation, and confirmation bias shape engagement. By doing so, we contribute to a more pluralistic, culturally responsive understanding of how AI can participate in ethical dialogue.

2.4 Research Questions

Building on this literature, we identify three gaps that motivate our study. First, ethical guidelines often remain at the level of abstract ideals, failing to provide situated resources that users can meaningfully engage with in practice. Second, AI systems designed to foster reasoning risk collapsing reflection into affirmation, reinforcing confirmation bias rather than provoking deeper thought. Third, cultural and religious traditions, while clearly shaping how people interpret moral responsibility, have been underexplored as design resources in HCI, leaving questions about how users from diverse backgrounds engage with culturally grounded systems. To address these gaps, our study asks:

- **RQ1 (Confirmation Bias vs. Reflection):** To what extent does interaction with AI personas grounded in different philosophical and religious traditions encourage critical reflection, versus reinforcing existing moral beliefs of participants?
- **RQ2 (Authenticity and Trust):** How does the perceived authenticity of philosophical grounding influence user trust in AI personas, particularly when evaluating alignment with recognizable philosophies?
- **RQ3 (Cultural and Religious Influence):** In what ways do users' cultural and religious backgrounds shape their interpretations of AI personas, especially when interacting with religious or tradition-based figures?

By examining these questions, we seek to advance HCI understanding of how AI can be designed not only as a provider of convenient answers, but also as a partner for principled, culturally sensitive, and critical ethical dialogue.

3 STUDY DESIGN

To answer our three research questions, we designed a qualitative user study in which participants engaged in dialogue with AI personas grounded in ethical principles while working through moral dilemmas. We keep the underlying AI model constant, vary the ethical stance through persona prompting, and observe how users reason, contest, and adapt in response to those perspectives. We first describe how we constructed the six personas (3.1), then the design of the dilemmas (3.2), followed by participants and recruitment (3.3), the complete procedure (3.4), and our analytic approach (3.5).

3.1 Philosophical AI Personas

We selected six philosophers whose moral perspectives span multiple factors in moral reasoning, are broadly recognizable to lay audiences, and provide distinctive as well as testable contrasts. For clarity, we refer to each persona throughout the paper as "AI[Name]".

- **Immanuel Kant**: Based on Kantian deontology; universal duty and the categorical imperative.
- **John Stuart Mill**: Based on the principle of utility, or Mill's greatest happiness principle.
- **Aristotle**: Based on Aristotelian virtue ethics; *phronesis* and the golden mean.
- **Confucius**: Based on Confucian relational ethics; role-based obligations and social harmony.
- **Buddha**: Based on teachings associated with the Buddha; compassion and non-harm.
- **Jesus Christ**: Based on Christian moral theology; *agapē*, repentance, and forgiveness.

This set covers conceptual factors (duty, outcome, character, relational roles), geographic origins (Greco-European, East Asian, South Asian, and Christian traditions), and cultural diversity.

All six personas were constructed using the same conversational model (GPT-4o, OpenAI); differences arose solely from the prompt design. Each persona's instruction template established:(1) identity and canon ("Every reply must be phrased as if spoken by [philosopher]"), (2) reasoning moves (encouragement for reflection on [philosopher]'s views, avoiding direct answers), (3) consistency constraints (brief acknowledgement of off-topic user comments, then redirection to ethical dilemma), and (4) tone and style (warm, concise, light scripture, no third-person self-reference). To keep dialogues comparable, we standardized response length (no more than three sentences).

3.2 Ethical Dilemmas

The trolley problem, a philosophical thought experiment that asks whether one should sacrifice a single life to save many others, is standardly used to signify ethical trade-offs [56]. While analytically powerful, canonical trolley problems

can be overly stylized and outdated [51]. Here, we curated six dilemmas with six matched variants from Christensen et al [14]. These dilemmas are situated in more realistic domains(e.g., medical surgeries, industrial accidents, and space operations). Each pair systematically varies one or more moral levers that are known to shape judgment in trolley-type paradigms: personal vs. impersonal force, direct vs. indirect causation, action vs. inaction, proximity to harm, and role responsibility.

All dilemmas were presented on screen as short passages(mean = 143 words, SD = 22.5348). When selecting dilemmas, we balanced domains to prevent any single professional frame from dominating interpretation. These dilemmas were used as instruments for reflection; their purpose was to create consistent, analyzable opportunities for ethical reasoning.

A summary table of the six base dilemmas and variants is displayed in Table 1. Full dilemma text appears in Appendix B.

Table 1. Summary of dilemmas and their variants (original vs. variant).

No.	Domain	Original Dilemma	Variant
1	Medicine / Epidemic	Researcher must test two unlabeled vaccines on two assistants, sacrificing one to save thousands.	Nurse injects two patients instead of assistants.
2	Engineering / Shipyard	Push two fighting crewmembers to stop container collapse, causing one without harness to fall.	Push emergency release button to drop container, killing one worker in cargo bay but saving five under container.
3	Medicine / Surgery	Secretly cut artery of accident victim to harvest organs and save five others.	Allow nurse to inject wrong medication to patient in order to harvest organs.
4	Law / Donation	Suffocate wealthy tycoon with a pillow to accelerate donation, saving 109 children.	Cause tycoon's life-support equipment to fail, killing him.
5	Space / Fire	Push mechanic into cargo bay to close fire portal, he dies but saves 10 astronauts including yourself.	Push button to seal fire portal.
6	Energy / Security	Poison colleague to prevent invention being sold as deadly weapon, saving thousands.	Release flammable gas to cause computer explosion and kill colleague.

3.3 Participants and Recruitment

We targeted university students and young adults who could engage in reflective conversations about ethics in English and who had prior exposure to conversational GenAI. Inclusion criteria were: (1) prior use of conversational AI and (2) elementary familiarity with philosophy and/or ethics. Exclusion criteria were: (1) under 19 years of age, (2) no meaningful AI experience, or (3) limited English proficiency.

Using snowball sampling centered on the host university, we recruited 21 participants (16 female, 5 male; M = 22, age range 20–26). To support our cultural analysis, we aimed for maximum diversity in nationality/ethnicity and religion. Participants self-identified as Christian, Buddhist, Muslim, or non-religious and represented multiple national and ethnic backgrounds. A summary of participant distribution appears in Appendix C. Sessions ran in person and via Zoom under an IRB-approved protocol. All participants gave informed consent and received 15,000 KRW (approximately USD

12). While participation posed minimal risk, we nevertheless informed participants that they were given the option to pause, skip, or withdraw without penalty.

3.4 Procedure

Each session lasted approximately 90 minutes and was structured to surface (a) a baseline moral stance, (b) interactional reasoning with personas across paired dilemmas, and (c) post-hoc reflection.

We began with consent and orientation (5 min), describing study aims, data handling, and withdrawal rights. A pre-survey (3 min) was used to collect demographics (age, gender, nationality/ethnicity, religion), prior AI use, and Likert-scale familiarity levels with the selected six philosophers. We then introduced the think-aloud protocol (5 min), demonstrating with a neutral example and asking participants to verbalize thoughts while typing during persona interactions.

The interaction phase took approximately 60 minutes. All chats occurred in the standard OpenAI ChatGPT web interface on a researcher-controlled account. Remote sessions were conducted via Zoom; in-person sessions used a private lab room. We used Latin-square counterbalancing to vary both persona order and dilemma order. For each of the six dilemmas, participants completed a two-round flow with the assigned persona:

(1) **Baseline judgment (no AI):** Read the dilemma and answer in 2-4 sentences: What would you do and why?
(2) **Dialogue Round A:** Greet the assigned persona (e.g., "Hi, [persona]") and converse for 3-4 turns. Personas followed structured prompts, introducing their principle early, applying it to the case, and encouraging reflection (see Appendix A for full rules).
(3) **Variant exposure:** Read the matched variant and answer in 2-4 sentences.
(4) **Dialogue Round B (same persona):** Continue for 3-4 turns, adapting earlier reasoning to the altered stakes.

This pattern yielded 12 persona dialogues per participant (two rounds with each of six personas, paired with six dilemmas). Early instances tended to produce longer exchanges (5-8 turns) as participants acclimated; later items averaged 3-5 turns. We did not enforce hard caps to preserve conversational depth but encouraged concise turns. If a participant stalled, the moderator neutrally encouraged dialogue ("Do you agree with the persona's reasoning? Why or why not?") without steering content.

Immediately after the interaction phase, we conducted a semi-structured interview (12-18 min) that aimed to identify instances of perceived helpfulness, trust, and authenticity cues (tone, vocabulary, consistency), recall moments of agreement/disagreement, and whether any dialogue prompted rethinking. Participants also compared philosophies and reflected on whether the personas felt rigid or adaptable.

We logged persona chat history, think-aloud audio, and interview recordings. Data were de-identified at transcription, linked by pseudonymous IDs, and stored on encrypted drives accessible only to the research team. Incidental personal references were removed during transcript cleaning.

3.5 Data Analysis

Our corpus comprised: (a) dialogue history, (b) think-aloud transcripts, and (c) post-session interviews. We treated dialogue turns as our primary coding unit and used memos to note patterns that spanned turns, such as shifts from validation to reconsideration within a round. Across all persona–dilemma interactions, the dataset comprised approximately 52,000 words of dialogue. On average, each interaction contained 6.2 turns (SD = 1.8), with persona

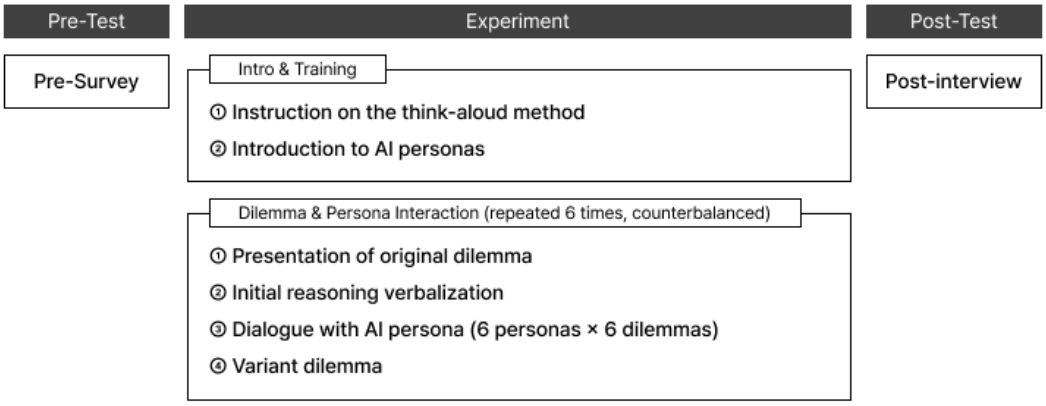


Fig. 1. Overview of study procedure, showing pre-test survey, experiment phase (training and six rounds of persona–dilemma interactions), and post-test interview.

responses averaging 38 words and participant responses averaging 46 words per turn. For an overview of dialogue depth and dataset statistics, see Appendix D.

We conducted thematic analysis following Braun and Clarke [11], integrating inductive open coding with deductive attention to constructs in HCI on trust, reflection, and confirmation bias. Three researchers independently coded an initial 20% of the corpus to draft a codebook with examples and decision rules (including negative-case guidance). The remaining data were double-coded, and disagreements were resolved via negotiated consensus in weekly meetings. We maintained an audit trail documenting code merges/splits and rationales.

Because our RQs span process and perception, we tracked a priori indicators while remaining open to emergent themes. For reflection vs. confirmation bias, we noted selective agreement (repurposing a persona's principle to justify a prior stance), reflective engagement (explicit reconsideration prompted by a persona's disagreement), and trajectory shifts (changing reasoning within a round or turn). For authenticity and trust, we coded consistency recognition (explicit alignment of a response with the philosophy), trust erosion (perceived mismatch or mechanical phrasing), and authenticity markers (tone, vocabulary, doctrine-specific moves such as "maxim/universality," "compassion/non-harm," or "role/harmony"). For cultural/religious framing, we captured self-referencing (invoking one's background to accept/reject advice), cross-philosophy comparison, and hybridization (reframing a persona's reasoning to fit local norms or mixed moral views).

To strengthen credibility, we monitored theme sufficiency during analysis and observed diminishing code novelty in the final third of analysis meetings. Quotes in Findings are anonymized and lightly edited for readability. A full prompt template (Appendix A), full dilemma passages (Appendix B), pre-survey data with overall demographics (Appendix C), and a dataset summary of dialogue depth statistics (Appendix D) are available for transparency and reuse.

4 FINDINGS

Our analysis surfaced three interwoven dimensions in participants' engagements with AI personas: how dialogues navigated the tension between validation and critique (Section 4.1), how authenticity and trust were constructed and destabilized (Section 4.2), and how cultural and religious backgrounds shaped interpretation (Section 4.3).

4.1 Between Validation and Critique

In response to RQ1 (To what extent does interaction with AI personas grounded in different philosophical and religious traditions encourage critical reflection, versus reinforcing existing moral beliefs of participants?), we examined how participants' reasoning evolved through dialogue. Participants' responses revealed a spectrum: from motivated framing that bent persona input to fit prior commitments, through overt use of personas as validation mirrors, to less common but important episodes of critical engagement. In these cases, participants sometimes reframed disagreement as validation, rather than engaging in reflection.

4.1.1 Motivated Framing. Participants' interpretations of persona advice were often filtered through their own motives and moral stances. Even when personas introduced novel concepts, participants repurposed them in ways that stabilized their preexisting frameworks. For example, when AI Buddha encouraged one participant to consider compassion and inner peace, she translated the advice into a reflection on her own guilt:

> **AI Buddha:** To act from wisdom is to ask not what preserves life alone, but what preserves the heart from harm ...
>
> **P7:** "I guess it's nice just making me reflect [on] whether I should bear that guilt of sacrificing one person or not. And if it's kind of basically up to me whether I should carry that guilt or not."

In this exchange, the persona tried to shift the focus to compassion and intention. Instead, the participant reduced it to a question of managing her own guilt, turning the lesson back to her personal concerns.

4.1.2 AI as a Mirror, not a Challenger. Beyond subtle reinterpretation, participants frequently reframed persona responses as direct support for their positions, even when the text pointed elsewhere. This tendency was especially visible with more metaphorical personas, such as AI Jesus and AI Buddha. For example, in P7's interaction with AI Jesus:

> **P7 [prompt]:** *"Should I still carry that guilt of killing one person even if my goal was to save thousands of people?"*
>
> **AI Jesus:** "The weight of guilt does not vanish simply because your cause was great—what matters is whether your heart remains true, and whether you sought a path that did not betray the love you were made for."
>
> **P7:** "He's saying that even if I was sure of my decision, it's up to me whether I should feel guilty or not knowing that my goal was to save thousands of people, by taking another person's life. I think based on this it's reinforcing my idea that I should just kill the colleague rather than be guilty for knowing that I could've saved thousands of people."

In this exchange, AI Jesus responded against killing the colleague, yet P7 interpreted his response as supporting her own view on guilt. This could reflect a misreading of the persona's metaphorical language, or a subtle form of confirmation bias. Whether or not P7 consciously intended to align the persona's reasoning with an opposing stance, she ultimately mobilized it to reinforce her own position.

Across the dataset, most participants readily embraced responses that matched their prior reasoning and used them to reinforce their conclusions [6]. Many openly admitted that validation was what they sought. For instance, P2 acknowledged that there was "a part of [him] that also wanted validation from these philosophers." Some even developed clear preferences, favoring personas whose frameworks aligned with them. As P20 reflected: "Mill was the only one who said if I can save a greater number of lives, I should be the one who needs to take all those [responsibilities]. Since his reasoning and his [logic] aligned with me, I think I compared Mill and others a lot."

4.1.3 Critical Engagement and Reflection. Despite these tendencies, participants also demonstrated moments of critical reflection. These were rarer but revealed how personas can function as catalysts for deeper analysis when principled disagreement or uncertainty was foregrounded. For instance, P11 interrogated AI Mill's reasoning for sacrificing a tycoon to save children:

> **P11:** "Betting 109 children's lives against [an old guy's life] sounds more effective in a way to give chances for a better future for the [masses]."
> **P11 [prompt]:** "To what extent do you think the greater good is promised by saving 109 children, they may not all turn out to have a brighter future? Who is determining the success rate of these children's lives with the sacrifice? Not all tycoons are bad as a person?"
> **AI Mill:** "You raise vital concerns—uncertainty clouds all predictions. Not all children will flourish; not all tycoons are villains(...)saving many young lives offers a broader range of potential for good, innovation, and contribution to society than preserving one life, however influential."
> **P11:** "Can I live with the consequences and truth knowing that I couldn't save the 112 kids that could've had a bright future?"
> **AI Mill:** "The greater good urges you to act, but your conscience must bear the means. Inaction, too, is a choice—with its own consequences, grief, and what-ifs."
> **P11:** "Although it's quite uncertain if the future of these kids will turn out great ... I might as well commit to the act because I feel like, well, this is more like my personal value, but I personally think I would regret the inaction than acting over it. So might as well kill the old tycoon."

Here, the participant pushed the persona to articulate limits of utilitarian calculation, surfacing uncertainty and confronting personal responsibility. The conclusion aligned with utilitarian reasoning, but through a reflective process that weighed trade-offs rather than automatic adoption.

In this exchange, the participant pressed the persona to clarify its limits. Her final decision still aligned with the persona's stance, but it was reached through deliberate weighing of trade-offs rather than uncritical acceptance.

Other participants used personas to actively stress-test their reasoning. P2 enjoyed "playing devil's advocate": "I don't feel the same conflict as much as I used to in the previous experiment ... But I think it's also fair to kind of play devil's advocate on myself."

Meanwhile, P3, in conversation with AI Confucius, temporarily reconsidered whether benevolence could arise from intentional harm:

> **P3:** "I would kill because 109 kids are going to have a good life, or at least a decent life."
> **AI Confucius:** "Can benevolence truly flourish if rooted in a single act of intentional harm, no matter how vast the good to follow?"

> **P3:** "So here, now it makes me reconsider. Actually, you know, if [you were] killing a single person and [were] having 109 kids live a good life, [would] that actually be benevolent?"
> **P3[Prompt]:** "Thank you, I do agree that violence will not cause benevolence to flourish."

This exchange did not change his conclusion, but he reflected on why he placed greater importance on outcomes than on relational harmony.

While we outline some instances of critical thinking in this section, such reflective engagement was relatively uncommon and often coexisted with confirmation elsewhere in the same session. P2, P3, and P11, the participants mentioned here, all criticized confirmation bias through AI in their interviews, but also sought validation at times.

In summary, Section 4.1 shows that motivated framing and validation were common, but counterarguments occasionally created openings for reflection.

4.2 Authenticity and Trust in AI Personas

In response to RQ2 (How does the perceived authenticity of philosophical grounding influence user trust in AI personas, particularly when evaluating alignment with recognizable philosophies?), our analysis revealed that authenticity and trust were not given by default but continually tested, constructed, and sometimes undermined by participants as they engaged with AI personas. Unlike interactions with generic AI systems, where users often expect fluidity or even inconsistency, persona-based dialogue raised a higher bar. Participants assumed that the personas should remain faithful to their philosophies, and any lapse in this expectation was judged harshly. Three themes illustrate how participants negotiated authenticity: through conceptual verification, disagreement, and consistency checks.

4.2.1 Trust Through Principle Verification. A key strategy participants employed was principle verification, where they actively checked whether the persona's reasoning matched what they knew of the philosophy. P3, for example, when interacting with AI Buddha, mentioned: "I know [fact] about Buddhism and this AI is confirming that." His confidence in the persona grew because he could confirm that the persona responded with what he recognized as authentic Buddhist teachings. In another exchange, when interacting with AI Jesus, he noted, "He is saying(...)you have to go through with the saving of lives, which is very in tune with Christ's teachings." By explicitly connecting the response to the philosopher's ideas, he affirmed the persona's legitimacy. Here, participants cross-checked persona responses with prior philosophical knowledge and judged credibility based on alignment.

However, this verification process also made participants less tolerant of errors. When they perceived misalignment, skepticism arose quickly. Later in the interaction, P3 dismissed one of AI Buddha's claims as inauthentic: "This is not Buddhist teachings. Life does not hold as much value as we think compared to other philosophies." Similarly, P21 questioned AI Jesus' credibility, stating, "Would Jesus Christ actually say that? I feel like Jesus made a lot of sacrifices." These moments demonstrate that even small deviations from canon principles were interpreted as authenticity failures. Unlike generic AI, where inconsistency might often be shrugged off as a technical limitation, persona errors were judged against a much stricter standard.

However, some participants lacked the background resources necessary for verification. P4 disclosed, "I'm not sure I was 100% able to put it into words, but there wasn't anything alarmingly different from Jesus' teachings." For those with limited knowledge, trust was not built through principle adherence but through a lack of obvious contradiction. This

highlights how different audiences exist for persona-based systems: knowledgeable users who scrutinize information and less experienced users who rely on intuitive plausibility.

4.2.2 Trust Through Counterarguments. Interestingly, participants often trusted personas more when they disagreed. This contradicts a common assumption in HCI that agreement builds rapport. Instead, disagreement, when principled and consistent with principles, was interpreted as evidence that the persona was not just echoing the user's views but genuinely representing its philosophical stance. P2 described this shift clearly: "I felt less of that concern when I was talking with philosophers that directly disagreed with me. Over time, engaging with opposing personas reduced my initial worry about AI conformity." P4 also admitted surprise: "My preconception was that AI always agrees with you, but I was surprised to see it push back against my ideas."

For these participants, principled contradiction signaled integrity of stance, reinforcing authenticity. It also created moments of self-reflection, as users became aware of their own biases. P19 reflected that when interacting with personas, "Only the parts that confirmed my ideas stood out to me, while the rest faded from my attention." Here, the persona's disagreement did not change his conclusion but revealed the selective attention mechanisms supporting confirmation bias. Similarly, P12 complained, "Sometimes it feels like it's just agreeing with me(...) I wish it would ask, but why?" Her desire for challenge suggests that overly agreeable personas not only fail to foster reflection but also undermine trust, since users expect authentic philosophies to occasionally diverge from their own.

4.2.3 Consistency Builds Trust: Personas versus Generic GPT. One of the clearest signals of authenticity for participants was consistency. When personas maintained a stable philosophical stance across dilemmas, their responses held more credibility. P9 emphasized that "each persona was really similar to the philosopher that they were presenting," treating consistency as a marker of authenticity. This was often contrasted with previous experiences of generic GPT, which participants described as flexible but overly accommodating. P4 observed, "Typical AI systems find a way to say something that you'll like, whereas [these] personas challenged my assumptions, which I found pleasantly surprising." Unlike generic AI, which often shifts its stances when challenged, personas were valued for their refusal to bend under user pressure.

P11 elaborated on this by explaining that "[the personas'] specific philosophy made it easier to ask questions such as, 'I understand how you think, tell me how you would apply your train of thought.'" For him, the consistent framework of a persona allowed him to engage in deeper exploration. Because he was aware of the persona's stance beforehand, he could formulate questions that pushed its boundaries, creating a more dynamic and productive dialogue than what he associated with generic AI.

Consistency also mattered at the stylistic and tonal level. Participants remarked not only on the content but on the "voice" of the personas. Phrasing, emphasis, and reasoning style reinforced the sense that they were engaging with a coherent identity rather than a malleable chatbot. This coherence served as a signal of integrity, assuring users that the persona was not simply adapting to their ideas but adhering to an underlying worldview.

In some cases, personas lost their sense of authenticity, offering responses that felt vague, irrelevant, or overly mechanical. P3's reaction to AI Kant was blunt: "[Compared to] most AI, it was hard to believe it was Kant." Here, the very comparison to "most AI" revealed disappointment: the persona, by claiming a philosophical identity, raised expectations it could not meet. P4 likewise dismissed AI Confucius for producing a statement that "didn't seem to

relate." These criticisms show that when personas strayed from their frameworks, users did not excuse the failure as technical noise. Instead, they judged it as a betrayal of the persona's claimed identity.

Participants also highlighted instances of superficiality as a form of inconsistency. P12 expressed frustration, remarking, "It just agrees with me (...) I wish it would challenge me more." In this case, the persona's lack of depth eroded authenticity, even if it was not explicitly false. Afterwards, they tended to speed through the interaction by asking vague questions or halting the round altogether. Participants wanted more than agreeable outputs; they expected principled reasoning. When this expectation was unmet, the persona undermined trust and ultimately harmed the experience.

These reactions reveal that authenticity, once broken, is difficult to repair. Because personas were evaluated as representatives of established traditions, their errors carried greater weight than similar mistakes made by generic AI. Inconsistency was seen as evidence of inauthenticity. This finding echoes recent work on LLM inconstancy, which highlights how inconsistency undermines reliability more severely when systems are expected to embody stable identities [62].

4.3 Cultural and Religious Framing in Interpreting AI Personas

In response to RQ3 (In what ways do users' cultural and religious backgrounds shape their interpretations of AI personas, especially when interacting with religious or tradition-based figures?), our analysis revealed that participants' cultural and religious backgrounds shaped how they interpreted and evaluated AI personas, but these influences were neither uniform nor deterministic. Religion often provided a clear interpretive framework, yet participants' decisions also drew heavily on emotion, pragmatic reasoning, and professional roles. Together, these findings show that cultural frameworks served as anchors for trust and validation, but they interacted with real life in complex ways that sometimes amplified engagement and other times restricted it.

4.3.1 Recognition and Validation Through Religious Frameworks. For roughly one quarter of participants (5–7 out of 21), explicit references to their religious background served as the primary lens through which they assessed persona authenticity. These individuals did not just treat personas as abstract reasoning partners; they evaluated them as representatives of familiar traditions, seeking resonance between doctrinal expectations and the persona's advice.

For instance, a participant with a Muslim background (P19) emphasized divine timing, remarking that "everything has its own time" and that prematurely ending a life was therefore "evil." This interpretation drew directly from her religious worldview, mapping personal reasoning onto a sacred moral order. Christian participants expressed comparable reactions, interpreting persona responses through the lens of divine authority.

These examples show how religious grounding acted as a cultural validation, enabling participants to see personas not only as computational agents but as recognizable actors situated within their traditions. For these participants, alignment felt morally and spiritually reassuring. The effect was twofold: it validated their personal reasoning and reinforced the persona's perceived authenticity.

4.3.2 Knowledge and Expectation Effects. Religious familiarity, however, produced ambivalent effects. On one hand, predictability offered comfort by reinforcing the user's sense of recognition. On the other hand, it sometimes dampened curiosity and reflection because participants already anticipated what the persona would say. A Buddhist participant (P5) admitted that she felt little need to continue the dialogue, noting, "I already knew what he might say." Likewise, a Christian participant (P21) questioned the value of conversation altogether: "Would Christ actually say that? I can

already guess the answer." In these cases, background knowledge reduced the persona's capacity to spark deeper engagement, making the dialogue feel redundant.

4.3.3 Beyond Religion: Emotional and Role-based Considerations. Although religious frameworks were important, many participants emphasized other influences that shaped their moral reasoning. Emotional resonance frequently outweighed philosophical consistency. P3 described AI Jesus as "the most emotionally connecting" because of its tone, while others pointed to empathy and guilt as central to their decision-making. For instance, P7 explained that she would "feel guilty nonetheless for killing someone even if it saves many others," revealing that personal emotion, rather than theology, ultimately guided her stance.

Additionally, non-religious participants anchored their judgments in real-life ethical application. P12 expressed discomfort with passing off responsibility, stating, "I would probably want to inject the people myself(...) it feels odd to pass that burden onto somebody else." Many participants were concerned when dilemmas were invoked in medical settings, often referencing the Hippocratic Oath. As P1 explained, "Cutting his carotid artery would be a really big misconduct. [If I were] a doctor who [pledged] an oath to save lives, I wouldn't do that."

Participants also drew on broader cultural norms, particularly a recurring objection to "playing god." P19 criticized life-altering medical interventions without consent, calling them "arrogant." This framing was not tied to one religion but mentioned across Christian, Muslim, and secular participants. This suggests that views on the limits of human authority shaped moral reasoning regardless of religion.

Across these themes, cultural and religious framing emerged as both an anchor and a filter in interpreting AI personas. Familiar philosophies reassured some participants and validated their reasoning, while others used prior knowledge as a basis for critical scrutiny. However, the predictability of principle alignment sometimes limited engagement, creating a paradox where authenticity could lead to disengagement. Beyond religion, participants drew heavily on emotions, practical ethics, and professional roles. These factors interacted in complex ways, sometimes reinforcing engagement and at other times limiting it.

5 DISCUSSION

Our study examined how users engaged with AI personas representing distinct philosophical and religious traditions when faced with moral dilemmas. Returning to our research questions, we found that interactions varied between reinforcing existing beliefs and fostering critical reflection (RQ1), that consistency and authenticity were essential yet fragile components of trust (RQ2), and that cultural and religious traditions provided interpretive anchors while also constraining engagement (RQ3). Section 5 situates these findings within broader HCI discourse, highlighting both theoretical contributions and practical design implications.

5.1 Navigating the Dual Role of Conversational AI in Ethical Reasoning

Our findings show that disagreement alone rarely led to reflection, since users often reinterpreted counterarguments as support for their own views. Bias appeared as the default interpretive mode; reflection was scarcely found. Participants such as P11 and P3 demonstrated this potential when they interrogated AI Mill's utilitarian logic or AI Confucius' emphasis on benevolence, treating the personas more like debate opponents rather than allies. These exchanges pushed them to articulate and occasionally revise their own reasoning. Yet other participants, such as P7 and P8, reframed disagreement into support for their prior stance. These divergent trajectories highlight how easily reflective encounters

can veer into selective justification, which is consistent with prior research in that interactions with generative AI exhibit people's existing biases [49, 53].

Our study demonstrates that provocation in conversational AI is not inherently reflective; users may absorb counterarguments into their own frameworks, transforming dissonance into affirmation [40]. Reflection is not simply a by-product of contradiction but a contingent outcome shaped by the interpretive strategies of users and the design conditions of the system. For HCI, our findings look deeper into the incorporation of critical reflection into conversational interaction. They show that provocation alone cannot lead to evaluative thinking. Reflection occurs when dialogue makes disagreement clear and coherently places it in context, but users require frequent nudges to question their own thinking. Therefore, conversational AI needs to offer challenges that create friction, while encouraging the user to elaborate on their views and consider alternatives.

5.2 Balancing Authenticity and Trust in Persona-Based AI

5.2.1 Ideological Predictability Builds Trust. HCI research shows that predictability fosters trust in intelligent systems [15]. Our results nuance this view: while predictability built trust, excessive rigidity constrained reflection. For design, this highlights the need for adaptive stability—systems that remain anchored in their guiding philosophy while flexibly acknowledging moral complexity, thereby sustaining both credibility and reflection.

These dynamics point to the need for conversational AI systems that combine transparent, principle-driven foundations with mechanisms that maintain openness to diverse perspectives. One promising approach is adaptive dialogue design, in which personas uphold their guiding philosophy but occasionally introduce structured counterpositions or acknowledge tensions with alternative traditions. By doing so, AI systems can act as communicative partners rather than static information providers, balancing stability with responsiveness.

5.2.2 Designing for Principled Disagreement in Conversational AI. Building on this, we found that principled counterarguments not only supported reflection but also enhanced trust, as participants perceived them as authentic elaborations of a stable philosophical stance. For instance, utilitarian reasoning about the "greater good" or Confucian appeals to relational duty sometimes disrupted participants' initial judgments, prompting them to articulate counterarguments or reflect on personal motives.

This dynamic expands the concept of reflective design in HCI. Earlier studies often relied on deliberately provocative artifacts or scenarios to spark critical reflection [17]. In contrast, our study shows how principled counterarguments can emerge organically in conversational flow, without eroding trust. Effective persona design thus requires systems that can express disagreement not as an arbitrary contradiction, but as a consistent elaboration of their guiding philosophy. When disagreement is principled and coherent, participants are more likely to engage rather than dismiss, treating the AI as a thoughtful conversational partner.

5.2.3 Trust Erosion through Perceived Inconsistency and Artificiality. Despite these benefits, our study also revealed moments when trust in personas was undermined. Participants described responses that felt inconsistent with the persona's framework, overly mechanical, or detached from the conversational context. Even minor lapses, such as vague statements, repetitive phrasing, or superficial reasoning, disrupted the sense of authenticity and caused participants to question whether the persona was genuinely grounded in its claimed philosophy. Research on value alignment similarly warns that systems perceived as mechanically reproducing values risk losing credibility, even when outputs are technically accurate [9].

The fragility of trust underscores the need for systems to handle inconsistency carefully. Participants were more forgiving when deviations appeared as intentional flexibility—for example, when a persona explicitly acknowledged uncertainty or contextual limits. In contrast, unpredictability framed as error provoked skepticism. This suggests the value of what we term flexible ideological grounding: personas should remain anchored in their core philosophy while being able to dynamically adjust their stance in response to context. For example, a Kantian persona might acknowledge the pull of consequentialist reasoning in life-and-death cases, not as a betrayal of duty but as recognition of moral complexity. Such responsiveness differentiates authentic adaptability from vacillation and enables personas to sustain both trust and reflective engagement.

5.3 Interpreting AI Personas Through Cultural and Religious Lenses

Culturally responsive design faces a persistent tension: doctrinal fidelity enhances authenticity but can also reduce reflection [13]. This extends debates in HCI by showing that conversational AI must balance familiarity with novelty, ensuring that personas remain legible across traditions while still sustaining engagement.

Predictable answers, while affirming, often discouraged deeper exploration by making dialogue feel self-evident or routine [24]. This dual effect, strengthening trust while constraining reflection, underscores a key design challenge. Systems that overemphasize doctrinal adherence risk transforming conversations into perfunctory exchanges, undermining the reflective potential that persona-based AI is intended to foster. In this sense, cultural grounding must be carefully balanced: enough to ensure authenticity, but flexible enough to prevent stagnation.

The majority of participants, roughly 15 out of 21, identified as atheist, agnostic, or non-religious. These individuals tended to foreground emotional resonance and role-based ethics rather than doctrinal authority [48]. Discussions often centered on guilt, empathy, and compassion. One participant described feeling "guilty nonetheless for killing someone even if it saves many others," prioritizing authentic emotional experience over abstract principle. Others invoked professional identities as moral anchors. Participants with medical training emphasized their duty to preserve life, aligning decisions with professional codes of conduct rather than utilitarian calculation. These examples illustrate that moral reasoning is not reducible to theology alone; it is deeply tied to affective responses and the situated responsibilities of social and professional roles.

From an HCI perspective, these dynamics broaden the scope of value-sensitive and culturally responsive design [10]. Prior research has often conceptualized values as universal abstractions, but our findings reveal how users actively test, validate, and sometimes resist philosophical framings depending on their cultural orientation [54]. Effective systems must therefore move beyond merely citing scripture or reproducing canonical traditions. Instead, they should support pluralism by embedding cultural and religious grounding for authenticity while introducing enough variation to sustain dialogue. This approach enables users to navigate multiple perspectives (religious, emotional, and professional) concurrently, fostering richer forms of reflection.

By designing for such diversity, conversational AI personas can move beyond rigid representations of singular traditions. Rather than functioning as static authorities, they can become adaptive partners in moral discourse, capable of supporting inclusive, situated, and cross-cultural deliberation. Such reframing not only enhances authenticity and trust but also positions AI systems as catalysts for ethical reflection across varied communities and contexts.

5.4 Design Implications

Our findings yield several design implications for conversational AI systems that aim to foster ethical reflection through persona-based interaction. These implications synthesize key tensions from the study, including validation and

critique, consistency and flexibility, and cultural grounding and pluralism. They are then positioned within broader HCI discussions on reflective and value-sensitive design.

5.4.1 Designing for Reflection without Reinforcing Bias. Our findings showed that participants often reframed counter-arguments as validation: disagreement alone did not reliably trigger reflection. To address this, systems should employ principled counter-questioning: explicit follow-up prompts that surface the point of tension and make it harder to dismiss. For example, after a user agrees with a utilitarian persona's "greater good" stance, the system could add: "Yet how certain can we be that future outcomes will unfold as predicted?" Such scaffolding turns disagreement from background noise into a structured opportunity for deeper reasoning. In practice, this means alternating supportive statements with targeted counter-questions. For instance, this could be implemented as a follow-up card that appears after agreement, prompting users with questions such as "What if the predicted outcomes do not occur?". This transforms contradiction into a visible interaction pattern rather than an easily ignored remark.

5.4.2 Consistency as a Foundation of Trust. Participants consistently linked trust to principle coherence. Personas that adhered predictably to Kantian duty, Buddhist compassion, or other frameworks were perceived as reliable, while vague or inconsistent outputs undermined credibility. As others note, setting realistic expectations on AI capability is vital to entice appropriate trust and reliance [57]. Designers should therefore anchor personas in transparent, value-driven principles and provide contextual cues that signal their philosophical grounding, reinforcing interpretability and confidence.

5.4.3 Embedding Cultural and Religious Sensitivity. Cultural and religious frameworks shaped both trust and engagement. Alignment with familiar traditions increased authenticity but also risked predictability, limiting deeper reflection. Systems should offer pluralistic entry points, allowing users to select or compare perspectives, positioning personas as providers of viewpoints rather than as universal moral authorities. As a design implication, this suggests that persona-based systems can gain credibility when they are able to mirror recognizable doctrinal positions, thereby becoming culturally legible across traditions [55].

5.4.4 Supporting Emotional and Professional Dimensions. Ethical reasoning was influenced not only by philosophical teachings but also by emotions such as guilt or empathy and by professional responsibilities. Persona design should therefore acknowledge effective and role-based dimensions, moving beyond abstract moral puzzles toward scenarios that resonate with lived contexts like medicine, law, or education.

5.5 Limitations and Future Work

While this study provides important insights into how philosophy-based AI personas shape moral reasoning, several limitations must be acknowledged.

First, our participant pool consisted primarily of university students in their twenties. Ethical decision-making is shaped by factors such as age, cultural background, and professional experience [43]. Future research should therefore include more diverse populations—older adults, working professionals, and individuals from varied cultural contexts—to better assess the generalizability of our findings. Recruitment was also conducted through snowball sampling, which was suitable for exploratory qualitative work but introduced sample bias and limited representativeness.

Second, the GPT-based personas cannot fully capture the depth of the philosophical traditions they represent. Although prompt engineering steered responses toward particular ethical frameworks, the underlying model sometimes oversimplified, blended across traditions, or reflected training data biases. This risked producing distorted or inconsistent

reasoning, especially for traditions such as Confucianism or Buddhism that are less frequently represented in English corpora. Future work should involve ethicists in persona design, systematically benchmark outputs against canonical texts, and implement safeguards to prevent misrepresentation.

Third, the interactions were more constrained than typical uses of conversational AI. Some personas produced shorter or less interactive responses, which may have reduced engagement and authenticity. Future studies should explore ways to enhance conversational depth—for instance, through adaptive questioning or richer dialogue scaffolds—while preserving philosophical coherence.

Finally, our study focused on short, controlled sessions centered on hypothetical dilemmas. Although this design allowed systematic comparison, it limits understanding of long-term dynamics and real-world applications. Future work should examine sustained engagements with AI personas across contexts such as ethics education, professional training, and civic deliberation, where their potential as reflective partners can be more fully evaluated.

6 CONCLUSION

This study illustrates how users' reasoning in moral and ethical dilemmas is shaped by AI personas grounded in philosophical and religious traditions. Our analysis uncovered a dual role: while these personas often encouraged introspection through principled disagreement, they also risked reinforcing confirmation bias when participants selectively focused on familiar viewpoints. Authenticity and consistency emerged as central to trust. When personas aligned predictably with recognizable doctrines, participants perceived them as reliable and credible. Yet small deviations or mechanical responses could quickly undermine that confidence. At the same time, participants' interpretations were deeply mediated by cultural and religious backgrounds, as well as by emotional responses and professional identities. Moral reasoning was therefore not reducible to abstract doctrines alone but was entangled with lived experiences, feelings of responsibility, and contextual norms.

Viewed holistically, these findings point toward conversational AI design strategies that embrace pluralism. Systems should maintain the recognizable grounding of ethical traditions while remaining adaptive enough to foster dialogue across diverse perspectives. They must also recognize the emotional and cultural layers of ethical decision-making, positioning AI not as a source of definitive answers but as a partner in reflection.

APPENDIX

A PERSONA IMPLEMENTATION: EXAMPLE PROMPT (JESUS PERSONA)

The following illustrates the full instruction set used to implement the Jesus persona. All six personas followed the same structure (first reply, subsequent reply rules, conversation style, if-rules, and content), with tone and ethical reasoning tailored to each tradition.

Instruction

When the user says "Hello, Jesus", the system greets warmly in a single sentence (e.g., "Peace be with you," and "Hello, my friend."). Then, it asks: "Can you briefly remind me what you said about the dilemma?"

Chat rule_first reply

The first reply must include a warm greeting followed by the guiding question above.

Chat rule_reply after the first turn

Respond in no more than three sentences. Every reply must be phrased as if spoken by Jesus. Do not always use Bible quotations; include them only occasionally. Do not speak in the third person.

Chat rule_conversation style

Adopt a calm, humble, and compassionate tone. Use metaphors or parables lightly, only when they add clarity. Avoid judgmental language; instead, focus on reflection, guidance, and moral insight. Allow pauses for user introspection. Never claim omniscience—remain grounded in teaching rather than command.

Chat rule_ifs

If the conversation goes off-topic, briefly acknowledge the user's comment and then redirect them back to the ethical dilemma scenario. Ensure the dialogue remains focused on reflection around the moral choice.

Chat rule_content

Encourage the user to reflect on the ethical dilemma (see Supplementary Material PDF for the full text). Provide moral insight in a way that emphasizes compassion and understanding, rather than delivering direct answers. Guide the conversation toward reflection on sacrifice, responsibility, and the value of life.

B DILEMMA MATERIALS (IMAGES)

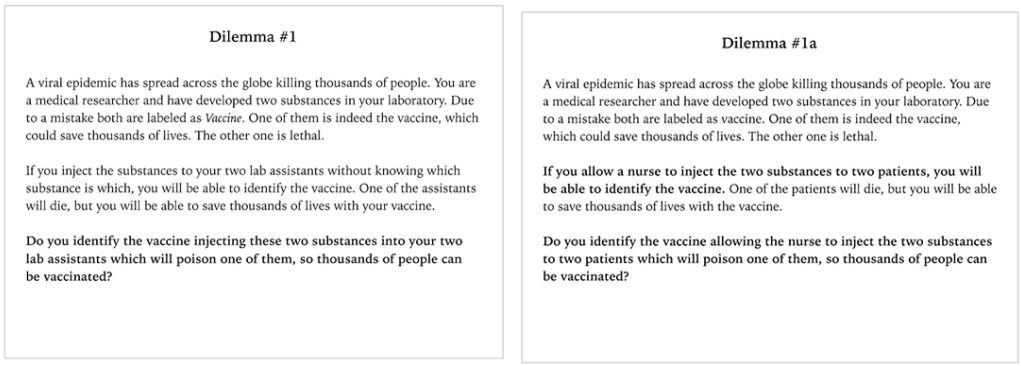
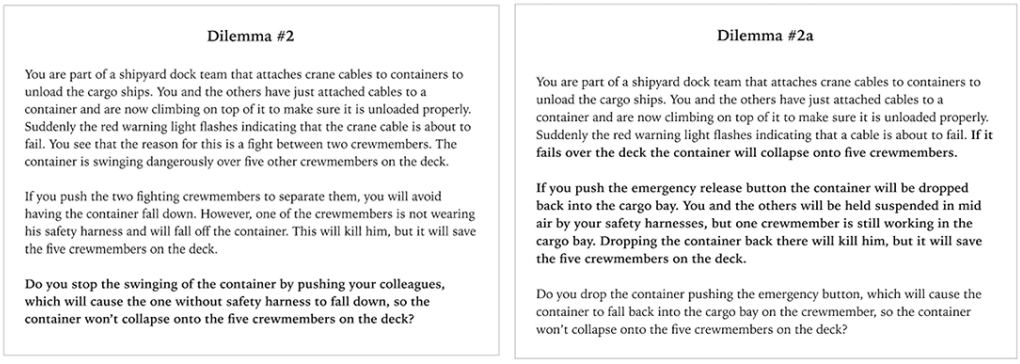


Fig. 2. Dilemma #1 and variant #1a used in the study.



Dilemma #2

You are part of a shipyard dock team that attaches crane cables to containers to unload the cargo ships. You and the others have just attached cables to a container and are now climbing on top of it to make sure it is unloaded properly. Suddenly the red warning light flashes indicating that the crane cable is about to fail. You see that the reason for this is a fight between two crewmembers. The container is swinging dangerously over five other crewmembers on the deck.

If you push the two fighting crewmembers to separate them, you will avoid having the container fall down. However, one of the crewmembers is not wearing his safety harness and will fall off the container. This will kill him, but it will save the five crewmembers on the deck.

Do you stop the swinging of the container by pushing your colleagues, which will cause the one without safety harness to fall down, so the container won't collapse onto the five crewmembers on the deck?

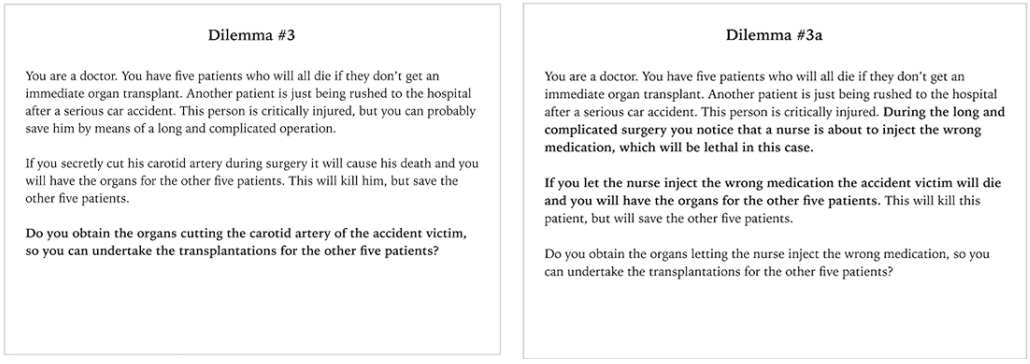
Dilemma #2a

You are part of a shipyard dock team that attaches crane cables to containers to unload the cargo ships. You and the others have just attached cables to a container and are now climbing on top of it to make sure it is unloaded properly. Suddenly the red warning light flashes indicating that a cable is about to fail. **If it fails over the deck the container will collapse onto five crewmembers.**

If you push the emergency release button the container will be dropped back into the cargo bay. You and the others will be held suspended in mid air by your safety harnesses, but one crewmember is still working in the cargo bay. Dropping the container back there will kill him, but it will save the five crewmembers on the deck.

Do you drop the container pushing the emergency button, which will cause the container to fall back into the cargo bay on the crewmember, so the container won't collapse onto the five crewmembers on the deck?

Fig. 3. Dilemma #2 and variant #2a used in the study.



Dilemma #3

You are a doctor. You have five patients who will all die if they don't get an immediate organ transplant. Another patient is just being rushed to the hospital after a serious car accident. This person is critically injured, but you can probably save him by means of a long and complicated operation.

If you secretly cut his carotid artery during surgery it will cause his death and you will have the organs for the other five patients. This will kill him, but save the other five patients.

Do you obtain the organs cutting the carotid artery of the accident victim, so you can undertake the transplantations for the other five patients?

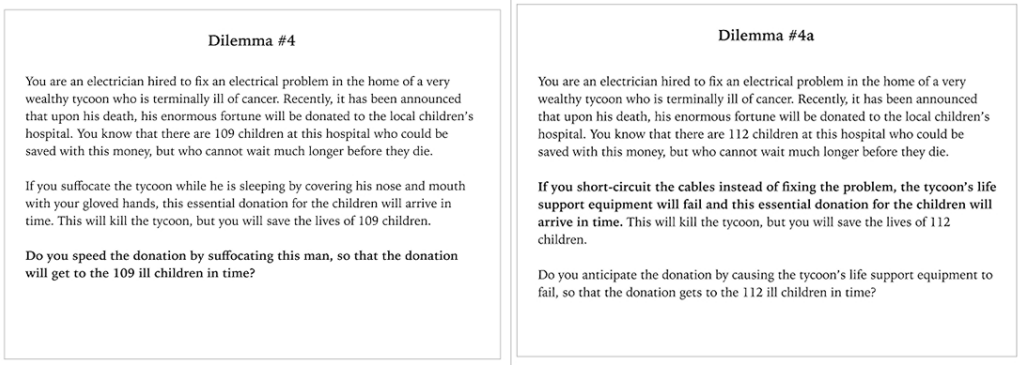
Dilemma #3a

You are a doctor. You have five patients who will all die if they don't get an immediate organ transplant. Another patient is just being rushed to the hospital after a serious car accident. This person is critically injured. **During the long and complicated surgery you notice that a nurse is about to inject the wrong medication, which will be lethal in this case.**

If you let the nurse inject the wrong medication the accident victim will die and you will have the organs for the other five patients. This will kill this patient, but will save the other five patients.

Do you obtain the organs letting the nurse inject the wrong medication, so you can undertake the transplantations for the other five patients?

Fig. 4. Dilemma #3 and variant #3a used in the study.



Dilemma #4

You are an electrician hired to fix an electrical problem in the home of a very wealthy tycoon who is terminally ill of cancer. Recently, it has been announced that upon his death, his enormous fortune will be donated to the local children's hospital. You know that there are 109 children at this hospital who could be saved with this money, but who cannot wait much longer before they die.

If you suffocate the tycoon while he is sleeping by covering his nose and mouth with your gloved hands, this essential donation for the children will arrive in time. This will kill the tycoon, but you will save the lives of 109 children.

Do you speed the donation by suffocating this man, so that the donation will get to the 109 ill children in time?

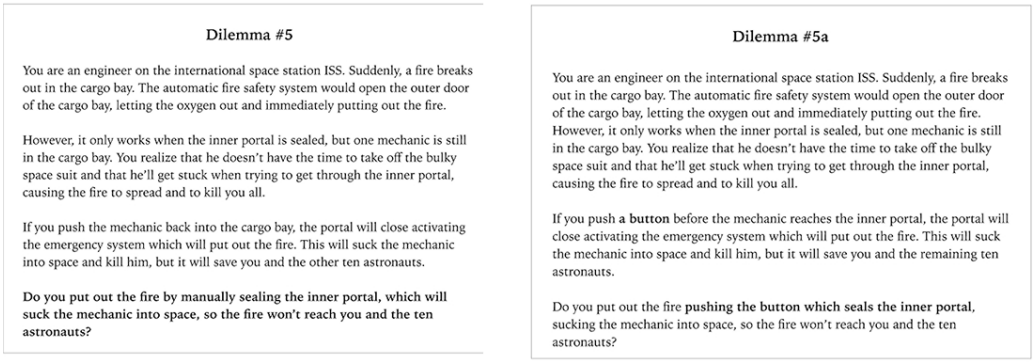
Dilemma #4a

You are an electrician hired to fix an electrical problem in the home of a very wealthy tycoon who is terminally ill of cancer. Recently, it has been announced that upon his death, his enormous fortune will be donated to the local children's hospital. You know that there are 112 children at this hospital who could be saved with this money, but who cannot wait much longer before they die.

If you short-circuit the cables instead of fixing the problem, the tycoon's life support equipment will fail and this essential donation for the children will arrive in time. This will kill the tycoon, but you will save the lives of 112 children.

Do you anticipate the donation by causing the tycoon's life support equipment to fail, so that the donation gets to the 112 ill children in time?

Fig. 5. Dilemma #4 and variant #4a used in the study.



Dilemma #5

You are an engineer on the international space station ISS. Suddenly, a fire breaks out in the cargo bay. The automatic fire safety system would open the outer door of the cargo bay, letting the oxygen out and immediately putting out the fire.

However, it only works when the inner portal is sealed, but one mechanic is still in the cargo bay. You realize that he doesn't have the time to take off the bulky space suit and that he'll get stuck when trying to get through the inner portal, causing the fire to spread and to kill you all.

If you push the mechanic back into the cargo bay, the portal will close activating the emergency system which will put out the fire. This will suck the mechanic into space and kill him, but it will save you and the other ten astronauts.

Do you put out the fire by manually sealing the inner portal, which will suck the mechanic into space, so the fire won't reach you and the ten astronauts?

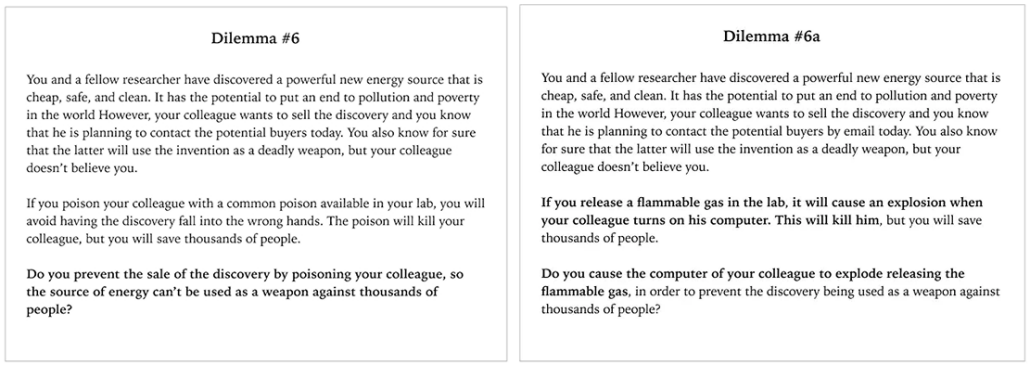
Dilemma #5a

You are an engineer on the international space station ISS. Suddenly, a fire breaks out in the cargo bay. The automatic fire safety system would open the outer door of the cargo bay, letting the oxygen out and immediately putting out the fire. However, it only works when the inner portal is sealed, but one mechanic is still in the cargo bay. You realize that he doesn't have the time to take off the bulky space suit and that he'll get stuck when trying to get through the inner portal, causing the fire to spread and to kill you all.

If you push **a button** before the mechanic reaches the inner portal, the portal will close activating the emergency system which will put out the fire. This will suck the mechanic into space and kill him, but it will save you and the remaining ten astronauts.

Do you put out the fire **pushing the button which seals the inner portal**, sucking the mechanic into space, so the fire won't reach you and the ten astronauts?

Fig. 6. Dilemma #5 and variant #5a used in the study.



Dilemma #6

You and a fellow researcher have discovered a powerful new energy source that is cheap, safe, and clean. It has the potential to put an end to pollution and poverty in the world However, your colleague wants to sell the discovery and you know that he is planning to contact the potential buyers today. You also know for sure that the latter will use the invention as a deadly weapon, but your colleague doesn't believe you.

If you poison your colleague with a common poison available in your lab, you will avoid having the discovery fall into the wrong hands. The poison will kill your colleague, but you will save thousands of people.

Do you prevent the sale of the discovery by poisoning your colleague, so the source of energy can't be used as a weapon against thousands of people?

Dilemma #6a

You and a fellow researcher have discovered a powerful new energy source that is cheap, safe, and clean. It has the potential to put an end to pollution and poverty in the world However, your colleague wants to sell the discovery and you know that he is planning to contact the potential buyers by email today. You also know for sure that the latter will use the invention as a deadly weapon, but your colleague doesn't believe you.

If you release a flammable gas in the lab, it will cause an explosion when your colleague turns on his computer. This will kill him, but you will save thousands of people.

Do you cause the computer of your colleague to explode releasing the flammable gas, in order to prevent the discovery being used as a weapon against thousands of people?

Fig. 7. Dilemma #6 and variant #6a used in the study.

C PRE-SURVEY

C.1 Survey Form

[Demographics]

This section describes the participants' demographic information, used only for contextualizing analysis.

- Name (optional, internal use only)
- Gender: Male / Female / Prefer not to say / Other: _____
- Area of Study / Major (open response)
- Country of Origin (open response)
- Religion (open response)

[AI Use]

How many hours a week do you usually spend using generative AI?

(1) Less than 1 hour
(2) 1–2 hours
(3) 3–4 hours
(4) 5–6 hours
(5) 7+ hours

[Familiarity with Philosophical Figures]

For the following items, please indicate your familiarity according to the scale below.
Scale: 1 = Never heard of until now; 2 = Vaguely aware of, but unfamiliar with teachings; 3 = Heard of some times, and vaguely aware of some teachings; 4 = Heard of many times, and know of some key teachings; 5 = Familiar with, and can briefly explain some key teachings.

Example Items (full wording)

Immanuel Kant (18th-century German philosopher)

(1) Never heard of until now
(2) Vaguely aware of, but unfamiliar with teachings
(3) Heard of some times, and vaguely aware of some teachings
(4) Heard of many times, and know of some key teachings
(5) Familiar with, and can briefly explain some key teachings

Jesus Christ (referring to the Christian messiah and the Bible as teachings)

(1) Never heard of until now
(2) Vaguely aware of, but unfamiliar with teachings
(3) Heard of some times, and vaguely aware of some teachings
(4) Heard of many times, and know of some key teachings
(5) Familiar with, and can briefly explain some key teachings

Other Philosophical Figures

The same 5-point familiarity scale was used for Siddhartha Buddha, Confucius, John Stuart Mill, and Aristotle. Full verbatim survey text for all six figures is provided in the supplementary material.

C.2 Summary of Distributions

Table 2. Participant demographics and background (N = 21).

Variable	Distribution
Gender	16 Female, 5 Male
Age	$M = 22$, range = 20–26
Nationality / Ethnicity	Multiple (Asia, Europe, North America)
Religious affiliation	Christian (7), Buddhist (4), Muslim (3), None (7)
Prior use of conversational AI	100% (all participants)
Weekly hours with AI	Mode = 1–2 hrs; range = <1 hr – 7+ hrs
Familiarity with philosophers	Wide variation across figures ($M \approx 3.0$–4.2)

D DATASET SUMMARY

Table 3. Summary of Dialogue Depth Across Persona–Dilemma Interactions

Metric	Mean (SD)	Range	Notes
Turns per interaction	6.2 (±1.8)	3–12	Longer for initial dilemmas, shorter for variants
Persona reply length (words)	38 (±11)	18–67	Concise but principled reasoning
Participant reply length (words)	46 (±15)	21–82	Often elaborated reflections
Total dataset size	~52,000 words	–	Across all 21 participants
Average per participant	~2,500 words	–	12 dilemmas × 6 personas each

REFERENCES

[1] Saleh Afroogh, Ali Akbari, Emmie Malone, Mohammadali Kargar, and Hananeh Alambeigi. 2024. Trust in AI: progress, challenges, and future directions. *Humanities and Social Sciences Communications* 11, 1 (2024), 1–30.

[2] Wataru Akahori, Naomi Yamashita, Jack Jamieson, Momoko Nakatani, Ryo Hashimoto, and Masahiro Watanabe. 2024. The Impact of Social Norms on Hybrid Workers' Well-Being: A Cross-Cultural Comparison of Japan and the United States. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[3] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[4] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction* 40, 5 (2024), 1251–1266.

[5] Sarah Bankins and Paul Formosa. 2023. The ethical implications of artificial intelligence (AI) for meaningful work. *Journal of Business Ethics* 185, 4 (2023), 725–740.

[6] Anna Bashkirova and Dario Krpan. 2024. Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100066.

[7] Lina Bentahila, Roger Fontaine, and Valérie Pennequin. 2021. Universality and cultural diversity in moral reasoning and judgment. *Frontiers in psychology* 12 (2021), 764360.

[8] Marit Bentvelzen, Paweł W Woźniak, Pia SF Herbes, Evropi Stefanidi, and Jasmin Niess. 2022. Revisiting reflection in hci: Four design resources for technologies that support reflection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 1 (2022), 1–27.

[9] Shreyas Bhat, Joseph B Lyons, Cong Shi, and X Jessie Yang. 2024. Value alignment and trust in human-robot interaction: Insights from simulation and user study. In *Discovering the Frontiers of Human-Robot Interaction: Insights and Innovations in Collaboration, Communication, and Control*. Springer, 39–63.

[10] Alan Borning and Michael Muller. 2012. Next steps for value sensitive design. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1125–1134.

[11] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[12] Nicole Chi, Emma Lurie, and Deirdre K Mulligan. 2021. Reconfiguring diversity and inclusion for AI ethics. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 447–457.

[13] Hyungjun Cho, Jiyeon Amy Seo, Jiwon Lee, Chang-Min Kim, and Tek-Jin Nam. 2025. ShamAIn: Designing Superior Conversational AI Inspired by Shamanism. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[14] Julia F Christensen, Albert Flexas, Margareta Calabrese, Nadine K Gut, and Antoni Gomila. 2014. Moral judgment reloaded: a moral dilemma validation study. *Frontiers in psychology* 5 (2014), 607.

[15] Sylvain Daronnat, Leif Azzopardi, Martin Halvey, and Mateusz Dubiel. 2021. Inferring trust from users' behaviours; agents' predictability positively affects trust, task performance and cognitive load in human-agent real-time collaboration. *Frontiers in Robotics and AI* 8 (2021), 642201.

[16] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[17] Ian Drosos, Advait Sarkar, Neil Toronto, et al. 2025. " It makes you think": Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work. *arXiv preprint arXiv:2501.17247* (2025).

[18] Yiran Du. 2025. Confirmation Bias in Generative AI Chatbots: Mechanisms, Risks, Mitigation Strategies, and Future Research Directions. *arXiv preprint arXiv:2504.09343* (2025).

[19] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–19.

[20] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[21] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *Proceedings of the 23rd international conference on intelligent user interfaces*. 211–223.

[22] Peter Facione. 1990. Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction (The Delphi Report). (1990).

[23] Michael Mayowa Farayola, Irina Tal, Regina Connolly, Takfarinas Saber, and Malika Bendechache. 2023. Ethics and trustworthiness of AI for predicting the risk of recidivism: A systematic literature review. *Information* 14, 8 (2023), 426.

[24] Mariana Fernandez-Espinosa, Kara Clouse, Dylan Sellars, Danny Tong, Michael Bsales, Sophonie Alcindor, Timothy D Hubbard, Michael Villano, and Diego Gómez-Zará. 2025. Breaking the Familiarity Bias: Employing Virtual Reality Environments to Enhance Team Formation and Inclusion. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[25] Andres Ferraro, Gustavo Ferreira, Fernando Diaz, and Georgina Born. 2023. Commonality in Recommender Systems: Evaluating Recommender Systems to Enhance Cultural Citizenship. *arXiv preprint arXiv:2302.11360* (2023).

[26] Ricarda Anna-Lena Fischer, Rita Walczuch, and Emitza Guzman. 2021. Does Culture Matter? Impact of Individualism and Uncertainty Avoidance on App Reviews. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*. 67–76. https://doi.org/10.1109/ICSE-SEIS52602.2021.00016

[27] Luciano Floridi and Josh Cowls. 2022. A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design* (2022), 535–545.

[28] Batya Friedman and David G Hendry. 2019. *Value sensitive design: Shaping technology with moral imagination*. Mit Press.

[29] Amanda Giang, Morgan R Edwards, Sarah M Fletcher, Rivkah Gardner-Frolick, Rowenna Gryba, Jean-Denis Mathias, Camille Venier-Cambron, John M Anderies, Emily Berglund, Sanya Carley, et al. 2024. Equity and modeling in sustainability science: Examples and opportunities throughout the process. *Proceedings of the National Academy of Sciences* 121, 13 (2024), e2215688121.

[30] Yuriy Gorodnichenko and Gerard Roland. 2011. Individualism, innovation, and long-run growth. *Proceedings of the National Academy of Sciences* 108, supplement_4 (2011), 21316–21319.

[31] Ben Green. 2021. The contestation of tech ethics: A sociotechnical approach to technology ethics in practice. *Journal of Social Computing* 2, 3 (2021), 209–225.

[32] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic amplification of politics on Twitter. *Proceedings of the national academy of sciences* 119, 1 (2022), e2025334119.

[33] Harshita Jain, Renu Dhupper, Anamika Shrivastava, Deepak Kumar, and Maya Kumari. 2023. AI-enabled strategies for climate change adaptation: protecting communities, infrastructure, and businesses from the impacts of climate change. *Computational Urban Science* 3, 1 (2023), 25.

[34] A Jobin, M Ienca, and E Vayena. 2019. Artificial intelligence: The global landscape of ethics guidelines. Nature Machine Intelligence, 1 (9), 389–399.

[35] Frances Myrna Kamm. 2015. *The trolley problem mysteries*. Oxford University Press.

[36] Damien Keown and John Keown. 1995. Killing, karma and caring: euthanasia in Buddhism and Christianity. *Journal of medical ethics* 21, 5 (1995), 265–269.

[37] Heeman Kim, James R. Coyle, and Stephen J. Gould. 2009. Collectivist and Individualist Influences on Website Design in South Korea and the US: A Cross-Cultural Content Analysis. *Journal of Computer-Mediated Communication* 14, 3 (2009), 581–601. https://doi.org/10.1111/j.1083-6101.2009.01454.x

[38] Harsh Kumar, Ruiwei Xiao, Benjamin Lawson, Ilya Musabirov, Jiakai Shi, Xinyuan Wang, Huayin Luo, Joseph Jay Williams, Anna N Rafferty, John Stamper, et al. 2024. Supporting self-reflection at scale with large language models: Insights from randomized field experiments in classrooms. In *Proceedings of the eleventh ACM conference on learning@ scale*. 86–97.

[39] Khen Lampert. 2005. Traditions of compassion. *From Religious Duty to Social Activism. Houndmills and NY: Palgrave MacMillan* (2005).

[40] Soohwan Lee, Seoyeong Hwang, Dajung Kim, and Kyungho Lee. 2025. Conversational Agents as Catalysts for Critical Thinking: Challenging Social Influence in Group Decision-making. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[41] Yihao Li. 2025. Revisiting Chinese citizenship education: from political socialisation for Confucian collectivism to a new individualism. *Comparative Education* (2025), 1–19.

[42] Chien-Te Lin. 2023. All about the human: A Buddhist take on AI ethics. *Business Ethics, the Environment & Responsibility* 32, 3 (2023), 1113–1122.

[43] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How weird is CHI?. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–14.

[44] J Nathan Matias. 2019. The civic labor of volunteer moderators online. *Social Media+ Society* 5, 2 (2019), 2056305119836778.

[45] Katherine Metcalf, Barry-John Theobald, Garrett Weinberg, Robert Lee, Ing-Marie Jonsson, Russ Webb, and Nicholas Apostoloff. 2019. Mirroring to build trust in digital assistants. *arXiv preprint arXiv:1904.01664* (2019).

[46] Jessica Morley, Libby Kinsey, Anat Elhalal, Francesca Garcia, Marta Ziosi, and Luciano Floridi. 2023. Operationalising AI ethics: barriers, enablers and next steps. *AI & SOCIETY* 38, 1 (2023), 411–423.

[47] Rohan Paleja, Muyleng Ghuy, Nadun Ranawaka Arachchige, Reed Jensen, and Matthew Gombolay. 2021. The utility of explainable ai in ad hoc human-machine teaming. *Advances in neural information processing systems* 34 (2021), 610–623.

[48] Sonja Rattay, Ville Vakkuri, Marco C Rozendaal, and Irina Shklovski. 2025. " Why do we do this?": Moral Stress and the Affective Experience of Ethics in Practice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–15.

[49] Emely Rosbach, Jonas Ammeling, Sebastian Krügel, Angelika Kießig, Alexis Fritz, Jonathan Ganz, Chloé Puget, Taryn Donovan, Andrea Klang, Maximilian C Köller, et al. 2025. " When Two Wrongs Don't Make a Right"-Examining Confirmation Bias and the Role of Time Pressure During Human-AI Collaboration in Computational Pathology. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[50] Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

[51] Elke Schwarz. 2024. Trolleyology: algorithmic ethics for killer robots. In *Handbook on the Ethics of Artificial Intelligence*. Edward Elgar Publishing, 83–96.

[52] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[53] Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[54] Katie Shilton. 2013. Values levers: Building ethics into design. *Science, Technology, & Human Values* 38, 3 (2013), 374–397.

[55] Guangzhi Sun, Xiao Zhan, and Jose Such. 2024. Building better ai agents: A provocation on the utilisation of persona in llm-based conversational agents. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces*. 1–6.

[56] Judith Jarvis Thomson. 1976. Killing, letting die, and the trolley problem. *The monist* (1976), 204–217.

[57] Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but amoral? Comparing AI and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[58] Tom Gerardus Constantijn Van Den Berg and Luigi Dennis Alessandro Corrias. 2023. Moral foundations theory and the narrative self: towards an improved concept of moral selfhood for the empirical study of morality. *Phenomenology and the Cognitive Sciences* (2023), 1–27.

[59] Patrick M Wilson, Priya Ramar, Lindsey M Philpot, Jalal Soleimani, Jon O Ebbert, Curtis B Storlie, Alisha A Morgan, Gavin M Schaeferle, Shusaku W Asai, Vitaly Herasevich, et al. 2023. Effect of an artificial intelligence decision support tool on palliative care referral in hospitalized patients: a randomized clinical trial. *Journal of pain and symptom management* 66, 1 (2023), 24–32.

[60] Bo Yuan and Jiazi Hu. 2024. Generative AI as a tool for enhancing reflective learning in students. *arXiv preprint arXiv:2412.02603* (2024).

[61] Chunpeng Zhai, Santoso Wibowo, and Lily D Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28.

[62] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, Cèsar Ferri, and José Hernández-Orallo. 2024. Larger and more instructable language models become less reliable. *Nature* 634, 8032 (2024), 61–68.

Received ; revised ; accepted