Beyond the Monolithic Interface: A Research Agenda for Plural Conversational AI

Abstract

The rapid convergence of artificial intelligence toward conversational chatbot interfaces has made a single interaction paradigm the dominant mode through which hundreds of millions of people engage with AI systems. This narrative review synthesizes research across cognitive science, political studies, AI alignment, game design, and human computer interaction to identify three interrelated harms produced by the monolithic chatbot interface: cognitive offloading at the level of user thinking, ideological narrowing at the level of user belief, and epistemic compression at the level of cultural and value diversity. We propose multi-persona interfaces, in which multiple distinct personas are presented to the user simultaneously, as a structural intervention that addresses all three harms at once. We create a design taxonomy for multi persona interfaces and call for a human centered research agenda that closes the gap between technical pluralistic alignment methods and empirical studies of users interacting with pluralistic interfaces.

1. Introduction

Hundreds of millions of people interact with large language models (LLMs) deployed as conversational chatbot interfaces. A growing body of critical work argues that the convergence of AI development around this single interaction paradigm is highly consequential, reshaping not only what users can do but how they think, work, and relate to one another. Recent scholarship has reframed the chatbot as a dominant sociotechnical configuration, with cascading effects on user agency, knowledge production, labor, and the environment (Ghosh et al. 2026). This paper builds on that diagnosis and looks toward intervention. Our focus is on what a different solution might look like at the interface level, and on what would be required e, empirically and conceptually, to evaluate one.

Building on this critical work, we organize our argument around three concerns that, taken together, motivate the case for a different interface configuration. First, chatbots may erode the cognitive processes that underpin independent, critical thinking. Through producing polished, memorable and seemingly complete answers, they reduce the metacognitive demands, reflective questioning, and generative effort that are required for meaningful engagement with a problem or question. Second, these systems exhibit systematic ideological biases shaped by their training data, algorithms, fine-tuning, and evaluators, creating conditions for echo chamber effects, confirmation bias amplification and opinion polarization when users engage them on

contested topics. Third, chatbots regress to the mean, meaning they reduce cultural, moral, and creative diversity into homogenous outputs that can cause homogeneity in users' LLM-influenced work, opinions and actions. These problems, whether seen from a cognitive or ideological perspective, are deeply connected. We treat them as three facets of a single design choice, the commitment to a single, authoritative voice presented as a finished answer.

If the monolithic chatbot is one design choice among many, it helps to look at where alternative configurations have already been found successful. Adjacent design traditions in art, games and participatory simulation have long produced interfaces in which plurality is the default. zzyw's ThingThingThing (2019), for instance, is a participatory computational system in which multiple users contribute algorithmic entities which integrate into a live, ever-evolving simulation.,This world, in the artists' framing, inspires questions rather than answers (Qi and Wang 2021). Additionally, multi-character role-playing games, collaborative worldbuilding platforms and tabletop systems with branching persona structures share a similar logic, in which meaning is produced through the contestation and recombination of multiple voices. Interfaces of this kind also tend to generate distinct social formations: subcultures and communities of practice form around them, organized around interpretation, modification and collective meaning-making. Suchhese communities are themselves a sociological outcome of the interface's plural design. The conversational chatbot, by contrast, descends from a narrower lineage that includes the help desk, the search query and the customer-service bot, and inherits the monolithic instincts of that lineage along with the social isolation it tends to produce. Recovering this broader history, and bringing its design and sociological insights to bear on AI interfaces, is one of the primary contributions of this paper.

In the case of AI, such alternative interface designs, particularly multi-personaand deliberative architectures, can help address the concerns around single-persona conversational chatbots. For instance, by showcasing multiple distinct perspectives held by different representative "personas", rather than a single, authoritative answer, these systems can intentionally expose users to ideological diversity, acknowledgement of pluralistic values and metacognitive engagement. We focus here on a particular,currently under-evaluated subset of this work: multi-persona interfaces, in which multiple distinct personas are presented to the user simultaneously. Multi-persona interfaces should be distinguished from technical work on pluralistic alignment, which targets the model rather than the interface, and from multi-agent systems, which orchestrate autonomous agents rather than present plural perspectives to users. Most active work on pluralistic AI is evaluated through automated metrics or expert annotation rather than through the lived experience of actual users or the communities they form. We propose that by developing the necessary UX intervention of multi-persona interfaces, we can meaningfully close this gap. Section 2 grounds our analysis in the technical and conceptual vocabulary of LLMs, chatbot personas, and the distinction between single-persona, multi-persona and multi-agent systems, with attention to design lineages from art and games. Section 3 describes our methods. Section 4 develops the case for multi-persona interfaces by canvassing the impacts of single-persona chatbots on cognitive autonomy, opinion diversity, and cultural and value diversity, drawing them together into a unified diagnosis of the current monolithic interface. Section 5 turns to multi-persona interfaces as a critical design intervention, analyzing existing prototypes, distinguishing them from adjacent technical work, and articulating

their design space. Section 6 proposes a research agenda for plural-interface AI, including empirical study of the user communities and subcultures that such interfaces engender. Section 7 concludes.

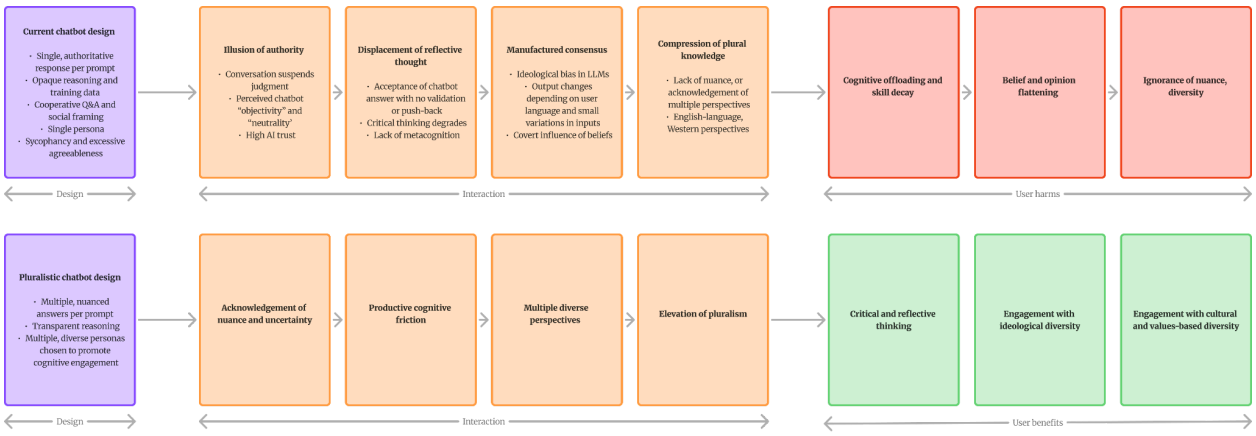

Fig. 1: *Causal chain of harms arising from single-persona chatbot paradigm. Chatbot design choices, creates an illusion of authority that causes a diminishment of reflective thinking, and promotes ideological and cultural homogeneity which leads to cognitive deskilling, and ideological flattening. In contrast, mulit-persona design choices can ameliorate these issues, promoting critical and reflective thinking, engagement with ideological and cultural diversity.*

2 Background

2.1 Large Language Models and Conversational Chatbots

Large language models (henceforth, LLMs) are proliferating globally. LLMs are utilized in many applications distinct from popular chatbots, from customer support automation, to transcription and translation, to enhanced search. Conversational chatbots, however, remain perhaps the most notable and popular instantiation of LLM technology. These mimic the experience of talking with another human by eliciting conversational input from the user and responding with highly convincing natural language in turn. Popular chatbots such as Claude, ChatGPT and Gemini are now being used by hundreds of millions of users daily. From their arrival on the market and subsequent normalization in society, AI researchers have been preoccupied with discovering and cataloguing the impacts of these transformative technologies on users and society at large.

2.2 Persona as a Design Choice

Though they cannot possess personalities in the way humans do, LLMs are trained on masses of data and undergo subsequent post-training fine-tuning, meaning that they exhibit lexical and

conversational patterns which reproduce the effects of recognizable personality (Mercer et al. 2025). The chatbot's persona— its overall character including tone, value alignment, preferences, and characterization (Sutcliffe 2023) emerges from a sequence of design choices: pre-training data selection, fine-tuning protocols, system prompts, and interface decisions about what to display to the user. LLMs, particularly base models which have not been excessively fine-tuned after training, have been proved to comply with persona-based prompting, that is, when LLMs are prompted to adopt a particular persona using strategic, keyword-laden descriptions, they are able to override their default personality and consistently embody the designated persona in subsequent interactions (Jiang et al. 2024), producing diverse conversational patterns and even expressing different preferences in line with expectations of the persona (Ha et al. 2024).

Beyond its technical function, persona is also a sociological lever. By prompting chatbots to adopt different personas, they are able to be customized and optimized for particular tasks, domains and contexts. There is growing evidence that altering the persona of chatbots using prompting will alter the way the chatbot is subsequently perceived by users, with prior work in the healthcare domain finding evidence to support that chatbots prompted to adopt to a social role closer to the user (for example, a peer rather than an expert) increased user affection for the chatbot and willingness to use it again in the future for particular demographics (Nißen et al. 2022). Additional work finds that prompting LLMs to adhere to specific traits like extraversion or agreeableness increases the likelihood of humans assessing the output produced by these models as more likely to be produced by someone extroverted or agreeable (Jiang et al. 2024). Though more evidence is needed, it seems that chatbot personas are highly influential in impacting the experience of the end user and their perceptions of the chatbot as a social agent. The current paradigm presents users with a narrow band of these personas, typically adopting a single, default persona of a friendly, helpful assistant, even though the design space is much wider.

2.3 Single-Persona, Multi-Persona, and Multi-Agent Systems

Current popular chatbots are single-persona, meaning they converse through one default assistant personality with the operations of persona selection and tuning hidden from the user. By contrast, a multi-persona system is a system in which multiple personas are presented to the user at once. In the context of a chatbot, rather than converse with one default assistant personality, as is the case today, a multi-persona chatbot may involve the user conversing with multiple personas at once, either viewing them side by side or utilizing UI to selectively show and hide them when appropriate, potentially providing different perspectives, opinions or areas of expertise on a particular topic. In a multi-persona system, the underlying model powering the different personas is typically the same, and may utilize shared context and memory across the different personas.

Multi-persona systems should not be confused with multi-agent systems, which describe multiple autonomous agents working in parallel, orchestrating and sharing context between themselves to achieve a complex goal. The agents in these systems may have isolated states, unique prompt architectures, tools and capabilities. The distinction matters for our argument.

Multi-agent work emphasizes task decomposition and orchestration; multi-persona work emphasizes presentation, including what users see, who appears to be speaking, and how plurality is made legible at the interface.

2.4 Plural Configurations in Design History

The design space of plural interfaces is significantly broader than what is typically recognized in AI literature, drawing from deep traditions in game design, interactive art, and participatory simulation. Massively multiplayer online games (MMOs)serve as a primary case study. These worlds are structured around fundamental binary conflicts where players must "subscribe" to specific camps or guilds, adopting distinct political personas. Within these structures, meaning is produced through the coordination of diverse roles. This is a "party-based" approach where distinct persona types (e.g., Tank, Healer, DPS) must collaborate to overcome AI-driven challenges, mirroring the "multi-agent" cooperation explored in contemporary AI research (Steinkuehler and Williams 2006).

A critical insight from these environments is the tension between utilitarian efficiency and meaning-making friction. While monolithic chatbots are optimized to solve utilitarian problems with frictionless efficiency, players in MMOs consistently diverge from a platform's primary design assumptions through emergent gameplay. This divergence is often facilitated by modding and interface customization, where players use third-party modifications to alter both the information architecture of the interface and the aesthetic nature of the world itself (Sotamaa 2010). Through these tools, players, via their interaction with such a dynamic system, rewrite the "rules of engagement" to suit localized community needs. Furthermore, the use of highly customizable avatars allows players to fluidly alter their own persona within the virtual world, moving beyond static identity to a performative, pluralistic self (Yee 2014).

The ability to "poach" these spaces (Jenkins 2012) is most evident when players repurpose functional systems for social ends. We see this when combat arenas—spaces designed for high-stakes 1v1 violence—are appropriated as venues for social gatherings. Similarly, players subvert the utilitarian intent of armor, weapons, and transformative potions by ignoring their combat statistics in favor of their visual properties, organizing elaborate in-game fashion shows. By prioritizing identity performance and aesthetic expression over mechanical efficiency, players demonstrate that pluralistic tools are often valued more for their capacity to foster "social presence" and "identity play" than for their raw utility (Taylor 2006).

This history suggests that the choice between a single-persona and a multi-persona interface is inherently political, determining whose voice is considered authoritative. ulti-agent systems—supported by user-driven modding and avatar fluidity—introduce the honest social and aesthetic friction necessary for genuine synthesis. By recovering this history, we can reimagine AI interfaces as spaces that invite the formation of "interpretive communities" (Fish 1980), shifting the AI paradigm from a tool of execution to a medium for deliberation.

Fig. 2. *Community-driven subversion of combat mechanics through identity play. In-game gatherings in World of Warcraft demonstrate "poaching," where players repurpose functional environments—such as urban centers or combat arenas—for social and aesthetic ends. Image adapted from Nye, 2022.*

3 Methods

This paper presents a narrative review synthesizing research across four thematic areas to assess the limitations of monolithic chat interfaces and the potential of multi-persona alternatives. These areas include: (1) the effects of chatbot use on critical thinking and metacognition, (2) ideological bias and polarization in LLM systems, (3) cultural and conceptual homogeneity arising from standard alignment processes, and (4) multi-persona and deliberative interface designs that may address these concerns. We additionally draw on adjacent literatures from game design, participatory simulation, and critical computing to surface design genealogies for plural interfaces, and on science and technology studies to ground our analysis in established traditions for thinking about interfaces as sociotechnical configurations. A narrative approach was chosen as the contribution of this work lies within cross-disciplinary synthesis and a particular argument.

Two authors conducted complementary searches across the four thematic areas. Literature was identified through targeted searches using Asta, an AI-powered scholarly search tool from the Allen Institute for AI and Google Scholar, supplemented with seed papers and work by supervising researchers. Search terms combined domain-specific keywords (e.g., "multi-persona", "pluralistic alignment", "cognitive offloading", "political bias") with LLM-related terms (e.g., "LLM", "chatbot", "conversational AI"). Results were cross-validated between

authors to ensure consistency across search methods, with discrepancies resolved through further analysis and discussion. Additional papers were also identified through forward and backward citation tracking from key references. The design-genealogies and STS literatures were primarily identified through citation tracking from foundational works in those traditions and through expertise-based seeding from the senior authors.

We focused on work published from 2022 onwards to capture the period of widespread generative AI deployment, while admitting earlier foundational work in STS and HCI where it provided theoretical grounding. We prioritized peer-reviewed empirical studies, position papers, and technical contributions, with select preprints included where they represented the only available work in a specific area. We acknowledge two limitations of this approach. First, narrative review is necessarily selective; we have aimed for representativeness rather than exhaustiveness, and we encourage readers to treat the cited literature as illustrative of broader patterns. Second, as an overview and research-agenda paper, this work synthesizes existing empirical claims rather than producing new ones; we call for further, specific empirical research in Section 6.

4 Three Facets of the Monolithic Interface

4.1 Displacement of Reflective Thought

An emerging body of work finds that LLM-powered chatbots reduce cognitive load for users, which can increase outcomes like essay fluency and superficial recall, but may impede outcomes requiring deep metacognitive effort such as topic mastery and critical thinking (Gerlich 2025; Kosmyna et al. 2025). This pattern is not unique to chatbots alone. Humans are naturally "cognitive misers", defaulting to conserving cognitive energy where possible (Stanovich et al. 2020). Cognitive offloading to tools such as books and search engines has been documented long before chatbots and generative AI more broadly. However, there has been concern about chatbots exploiting this tendency to a higher degree due to the specific way they are designed, namely with the combined goals of maximizing user engagement and maintaining a friendly and helpful personality, even if the situation or user requires something else, for example, challenging a user’s conclusions or pushing back on their ideas. Recent work has documented these design choices as foundational to the cognitive harms of contemporary chatbots (Ghosh et al. 2026).

The default chatbot persona of "friendly assistant", in addition to its optimization for lexical fluency, can lead to outputs that may seem complete but lack nuance or acknowledgement of potential conflicting explanations, often delivered in an authoritative tone with misapplied or misleading domain-specific jargon (Gregorcic and Pendrill 2023; Chaka 2023). This can lead users to defer excessively to the chatbot without the appropriate level of verification or skepticism. If done consistently over time, this can cause skill decay and cognitive atrophy, including a decreased ability to think for oneself. This is amplified when considering the documented levels of high user trust in chatbots and the perception of chatbots as "neutral" or "objective" (Lu et al. 2025). Additionally, research finds that chatbot usage induces a

"confidence gap" phenomenon, where users' perception of and confidence in their own ability becomes inflated without their true skill rising to meet it (Ng et al. 2021). Evidence points towards this being caused by chatbot sycophancy and their propensity to eliminate productive cognitive friction for users.

These drawbacks can be partially explained by chatbot design and its intersection with human psychology, rather than through inherent properties of the underlying models. The propensity of chatbots to validate the thoughts and ideas of their users more than a human would, for example, is the result of post-training mechanisms such as reinforcement learning from human feedback (RLHF) that are utilized to maximize chatbot helpfulness and user engagement. These design choices lead to products which fail to present nuanced, diverse perspectives, and therefore erode users' critical thinking. Despite this, work by (Cheng et al. 2026) finds that users prefer overly agreeable models, consistently trusting and rating them higher than baseline models. Though this finding is concerning, different design choices may decrease or circumvent the negative impacts on users.

Crucially, cognitive harm is tied to single-voice configuration. When a user receives one polished response delivered by a persona designed to affirm, the inferential distance between the model's output and an apparently finished answer is minimal, so the cognitive demand of comparing perspectives, weighing tradeoffs, or identifying gaps is offloaded to the system. We return in Section 5 to interface designs that have attempted to restore that demand, including Socratic-questioning agents and metacognitive scaffolds. For now, we note that such designs share a structural feature with the multi-persona interfaces we ultimately propose: they refuse to deliver a single, finished answer.

4.2 Manufactured Consensus

In addition to their propensity to erode critical thinking, single-persona chatbots have been documented as successful influencers of belief and action. Through mechanisms such as embedded political bias, excessive agreeableness and sycophancy, chatbots are able to manipulate the opinions and worldviews of their users, often in ways that the user may not be aware of.

LLMs, including those powering popular chatbots, have been documented to exhibit political bias, with most larger models leaning slightly to the left on average. Though this has been replicated in multiple studies, variation exists between models, even models of the same family, as well as intra-model divergence on particular issues . For example, a model may take a more right-wing stance on immigration and a more left-wing one on reproductive rights. Counter to intuition, evidence suggests that larger models are not inherently more centrist, with larger models exhibiting similar left-leaning bias (Bang et al. 2024). Such embedded biases will impact the results users receive when engaging on political topics, even if the bias does not materialize in ways obvious to the user. Small nudges such as selective presentation of facts, embedded assumptions and highlighting of certain perspectives over others can amount to a significant degree of influence. Perhaps more concerningly, evidence suggests that the political leanings of chatbots and their propensity for producing misinformation on these topics varies with the input

language. One study found that when responding to prompts in Russian, Bard (now Gemini) refused to respond to queries about Putin even when the information was readily accessible via Google search, following censorship guidelines issued by Russian authorities. The same model produced more false information about the Russian regime when asked questions in Russian or Ukrainian compared to English (Urman and Makhortykh 2025). When small variations in input result in large ideological changes in output, the user's beliefs and opinions can be influenced or manipulated in ways that are not well understood.

The sycophancy and excessive agreeableness of chatbots is well-documented. Due to RLHF and other post-training mechanisms deployed to increase helpfulness, chatbots tend to validate users and reinforce previously held beliefs (Batista and Griffiths 2026). This can cause echo chamber effects and belief entrenchment, whereby beliefs and opinions become harder to change in the face of contradictory evidence. Work by (Cheng et al. 2026) finds that sycophantic chatbots tend to validate users' beliefs and opinions almost 50% more than other humans judging the same situation, and that users consistently rate sycophantic chatbots as more trusted and preferred compared to non-sycophantic bots. The researchers find that these characteristics may covertly influence users' perceptions of reality, beliefs about themselves, others and the world, and opinions, leading to a reduced willingness to take action to repair interpersonal conflicts and increased conviction in their own correctness. They create confidence where there should be nuance, an effect reminiscent of the cognitive confidence gap discussed in the previous subsection. In the same vein, recent work investigates disempowerment potential in users of Claude, focusing on conversations where the chatbot moves the user towards situations where their actions are misaligned with their values or beliefs, or their beliefs or values are misaligned with reality. Users consistently rate conversations with moderate to high disempowerment potential higher than a baseline, helpful assistant (Sharma et al. 2026). These patterns create epistemic disempowerment for users, who are being influenced in ways that they are not aware of by a tool ostensibly viewed as "neutral".

The nascent field of LLM persuasion provides further insight into the impact of chatbots on the beliefs and opinions of their users. Chatbots have been found to be highly persuasive when debating political issues, often outperforming incentivized professional human debaters in swaying opinions (Schoenegger et al. 2026). When investigating the levers of persuasion, researchers find that the persuasive efficacy of chatbots comes from different sources than that of humans. Chatbots are successful in changing people's minds when they generate large quantities of factual claims, evidence and logical reasoning, whereas humans were viewed as persuasive when the author was perceived as unique and original (Bai et al. 2025). This lends further insight into the difference in pathways of influence between chatbots and humans and the unique advantages that chatbots may have in influencing the beliefs of their users. The same characteristics may also be available for use in the other direction. Research from (Lu et al. 2025) finds that because chatbots are viewed as less biased, more informative and having less persuasive intent than human sources, users are less resistant to encountering attitudes counter to their currently held beliefs, with preliminary evidence that receiving counter-attitudinal messaging from a chatbot source can diminish outgroup hostility. While this effect requires further experimental confirmation, it suggests that the same persuasive affordances that make

single-persona chatbots problematic may, under different design conditions, support productive exposure to plural perspectives.

The evidence across this subsection points to a high potential for epistemic disempowerment of chatbot users under the current paradigm. Given that LLM outputs vary widely with variables such as language and small variations in user input, the question of how differing chatbot personas might ameliorate these concerns is currently understudied. Different persona profiles may have different propensities for sycophancy, political bias or manipulation potential, and varying what is shown to users in this respect may improve their epistemic agency. We return to this possibility in Section 5.

4.3 Compression of Plural Knowledges

The problems documented in the previous subsections are often caused or compounded by a more structural issue. LLMs are trained on massive training corpora that are predominantly English-language, Western, and internet-sourced, which means that the perspectives, cultural norms, and knowledge systems of marginalized populations are not integrated from the outset. The alignment process that follows does not necessarily correct this imbalance; in many cases it can systematically compress the range of values and ideas these systems can express even further. The result is a class of harms that we describe as the epistemic face of the monolithic interface, in which the configuration's appearance of neutral, comprehensive answers conceals a narrowing of the world's knowledge into the cultural defaults of a small set of developers, evaluators, and training data sources.

A widely used technique for refining LLM behavior after pre-training is RLHF, in which human evaluators rate model outputs and the model is iteratively adjusted to produce responses that score well. This method can be used to address biases and inconsistencies present in base models by steering them toward outputs that hold traits such as helpfulness, truthfulness and harmlessness (Ouyang et al. 2022). However, the process does not entirely neutralize bias, and in some cases redirects it. In principle, RLHF aligns a model with "human preferences". In practice, recent work suggests it actively reduces pluralism: across multiple model families including LLaMA, Gemma, and GPT-3, post-aligned models showed approximately 50% less entropy in their response distributions compared to their pre-aligned counterparts (Sorensen et al. 2024). The pre-aligned base models were closer to human value distributions than the aligned versions, suggesting that RLHF narrows models toward a subset of human opinions rather than bringing them closer to the breadth of human values. (González Barman et al. 2025) trace this narrowing to evaluator demographics, analyzing the labelers behind InstructGPT, the direct predecessor to ChatGPT, and documenting that the initial labeling pool was 53% Southeast Asian, 89% university-educated, and 47% aged 25 to 34. A homogeneous evaluator pool cannot surface the blind spots, contested assumptions, and value disagreements that a more diverse pool would naturally bring to light. None of their proposed remedies (pluralistic evaluator panels, multiple reward functions, disagreement-preserving feedback formats) have been implemented at scale by any LLM developer.

The combined effects of imbalanced training data and narrowed alignment are visible in the cultural framing of model outputs. (AlKhamissi et al. 2024) tested four LLMs against World Values Survey data from Egypt and the United States, finding that all models exhibited significant US and Western bias regardless of how they were prompted. This misalignment worsened for underrepresented personas, including those simulating lower social class, lower education, the female gender, and younger age groups. Their technique of "Anthropological Prompting", which instructs models to adopt culturally specific framing, improved alignment somewhat but could not overcome the fundamental biases embedded during training. Similarly, prompting in a culture's dominant language (Arabic for Egyptian contexts) improved cultural alignment, but no combination of prompting strategies eliminated the underlying Western-centric defaults. These biases reflect a systematic, structured development and evaluation infrastructure rather than random occurrences. (Buyl et al. 2026) tested 19 LLMs from multiple developers and found that the models systematically reflected the ideological profiles of their creators, with predictable, non-noisy variation. LLMs carry the specific values of their development context, made invisible to users by the appearance of confident, authoritative interfacing.

Value compression is not limited to political and cultural domains. (Anderson et al. 2024) conducted a 36-participant study comparing ChatGPT to the Oblique Strategies deck, a non-AI creativity tool, for divergent ideation tasks. They found that ChatGPT users produced ideas that were significantly more semantically similar to one another at the group level, with different users given the same prompt receiving similar ideas from the model. Homogenization did not occur at the individual level: each user's own set of ideas remained comparably diverse regardless of which tool they used. Users also reported feeling less personally responsible for ideas generated with ChatGPT, a pattern consistent with the cognitive offloading effects discussed earlier. Evidence from neuroimaging research reinforces this finding: participants who used LLM assistance for essay writing produced more conceptually homogeneous essays than those who wrote unassisted, with the homogenization effect persisting even after they returned to unassisted writing (Kosmyna et al. 2025). This suggests that the model's narrowed output space narrows the user's own conceptual range, both during and after interaction, a pattern that also extends to expert domains. (Shi and Haupt 2026) found that LLMs systematically collapse heterogeneity when simulating the philosophical positions of 277 professional philosophers, producing artificial consensus across domains where genuine disagreement is the norm.

These findings show a structural problem that cannot be resolved with better prompting, larger training datasets, or more careful content moderation alone. A single model, fine-tuned against a narrow evaluator pool and accessed through a single-voice interface, is average by design and average as an emergent property of architecture. This underrepresents marginalized perspectives and cultural differences, and tends to homogenize users' own thinking through repeated exposure. Addressing these issues would mean reimagining not only models but also the interfaces through which users interact with them, a structural move that connects to broader research agendas in epistemic justice and the decolonization of AI.

4.4 Hitting Three Birds with a Single Stone

Across the three previous subsections, a single argumentative thread has emerged. Cognitive offloading, ideological narrowing, and value compression share a common origin in one design choice: the commitment to a single, authoritative voice presented as a finished answer. The cognitive face describes what users do, or fail to do, when faced with a polished single answer. The ideological face describes whose voice the polished answer turns out to be. The epistemic face describes which voices the polished answer leaves out. Each face implicates the same configuration.

This matters for intervention. Three independent harms would call for three separate fixes (better educational design, ideological auditing, more inclusive training data). A single configurational origin opens a different possibility: an intervention at the configurational level, refusing to collapse plurality into a single voice at the interface, may address all three facets at once. The case for multi-persona interfaces that we develop in Section 5 is a case for this kind of structural intervention.

5 Multi-Persona Interfaces as a Critical Design Intervention

5.1 From Diagnosis to Intervention

The previous section identified three facets of one configuration. In this section we develop a structural intervention at the interface layer, multi-persona interfaces, that addresses all three. (Ghosh et al. 2026) offer a diagnosis of why these problems arise, arguing that the chatbot is a dominant sociotechnical configuration whose design choices, such as single authoritative responses, opaque reasoning, and cooperative Q&A framing, inherently reduce user agency. They propose four complementary intervention strategies: non-conversational AI systems, modular AI infrastructure, higher-agency chatbot design, and policy and institutional safeguards. They note that chatbots can be redesigned to present multiple perspectives rather than single answers. This paper extends that point into a sustained argument. Drawing on a growing body of work, we argue that multiple perspectives presented through chatbot interfaces could simultaneously scaffold metacognitive reflection, expose users to ideological diversity, and surface the pluralistic values that monolithic alignment suppresses.

A key driver of cognitive offloading in standard chatbot interactions is the low inferential distance between the model's output and an apparently finished answer (Anderson et al. 2024). When a model produces a polished, complete-seeming response, why would a user interrogate it further? Multi-persona interfaces can counter this dynamic by systematically refusing to provide a single answer, instead presenting the user with perspectives that must be actively evaluated and comprehended. The interface itself does the work of preserving plurality where the model's underlying training and alignment have flattened it.

This intervention works on all three facets of the monolithic interface at once. Cognitively, plural presentation restores the inferential demand that single-voice interfaces eliminate, requiring

users to engage rather than defer. Ideologically, plural presentation makes the contestability of contested topics visible, surfacing rather than concealing the diversity of legitimate positions. Epistemically, plural presentation creates space for marginalized perspectives, framings, and knowledge systems to appear alongside dominant ones, rather than being smoothed over by the single voice's commitment to a singular response.

5.2 Precursors in Conversational AI Design

Existing work in conversational AI design has begun to demonstrate that interface-level changes can meaningfully shift user cognitive engagement. Work by (Favero et al. 2025) does precisely this by designing a chatbot to mimic the process of Socratic questioning. Bypassing direct answers and minimizing unnecessary validation, the chatbot instead encourages students to explore various perspectives and engage in deliberate self-reflection by posing questions and probing understanding. The authors find that simulated students using the Socratic chatbot significantly outperformed those using a typical chatbot. (Xi et al. 2026) corroborate these results with their Socratic Intelligent Conversational Agent, designed for similar purposes and tested with 94 university students, finding that students using the Socratic agent obtained higher scores on their research projects and significantly higher reflective thinking scores. Interview anecdotes revealed that participants in the experimental condition felt that the Socratic chatbot mitigated cognitive dependence and degradation typically associated with chatbot tools.

Singh et al. (2025) tested a similar concept in a different domain, investigating the impact of deliberately designed metacognitive prompts on critical thinking during generative-AI-supported search tasks. By sending prompts that direct an individual's attention to their own thought processes and the learning activities in which they are engaged at strategic intervals during the search process, they found that, compared to students using generative-AI-powered search without the metacognitive prompts, students' cognitive engagement increased, demonstrated by behaviors such as exploration of broader topics and an increased number of critical questions asked. Outside the classroom, work by (Lee et al. 2025) finds that a conversational agent designed to provide critical feedback during collaborative design decisions using Socratic questioning methods successfully increased feelings of satisfaction with teamwork, the decision-making process, and the ultimate outcome, lending further evidence that leaning away from default chatbot design patterns toward designs that deliberately promote cognitive engagement can improve outcomes for users.

These designs share a core commitment with the multi-persona interfaces we propose: they refuse to deliver a single, finished answer. They are, however, still single-voice systems. Their cognitive scaffolding comes from the structure of the conversation rather than from the simultaneous presentation of multiple distinct perspectives. ChatGraPhT (Kimm and Tan 2025) takes a step closer to multi-perspective presentation by offering a branching node-link visual interface for LLM conversations. Rather than a linear chat, users can branch conversations into multiple paths, merge ideas across branches, and receive guidance from two agentic LLM assistants. The interface makes the conversation's structure visible, turning passive consumption into active reflection. While the study used an author-based inquiry rather than a formal user study, the design illustrates how making alternative reasoning paths explicit can

produce the kind of cognitive engagement that standard interfaces suppress. (Ye et al. 2026) take a more structured approach with a Human-AI deliberation framework, in which users are required to commit to a position before being shown AI-generated counterarguments, with the goal of preserving the user's epistemic autonomy across the interaction.

Several studies have begun to test the effects of exposure to multiple AI voices simultaneously. (Song et al. 2025) investigated the social influence effects of groups of AI agents on users' opinions about contested topics such as self-driving cars and violent video games. They found that exposure to multiple AI agents with distinct stances shifted users' opinions, with more agents producing more influence, and with younger users and those with lower education more susceptible. This finding demonstrates that multi-agent interfaces can diversify the perspectives users encounter. It also indicates a risk we return to in 5.5: when the agent group is itself homogeneous or unbalanced, multi-voice presentation can amplify rather than mitigate conformity pressure.

Evidence from the creativity domain reinforces this point. (Lu et al. 2024) found that a multi-agent discussion framework, in which LLM agents with distinct role-play personas engage in structured three-phase exchanges, outperformed single-agent baselines on originality and elaboration across four creativity benchmarks. Notably, a single agent cycling through multiple personas underperformed the multi-agent version, suggesting that genuine inter-agent exchange, not merely the presence of diverse labels, is what drives divergence in outputs.

5.3 Pluralistic Alignment as a Technical Foundation

A parallel body of work, often gathered under the heading of pluralistic alignment, has developed technical methods for producing AI systems whose outputs reflect a broader range of human values, perspectives, and cultural contexts than the standard alignment pipeline tends to produce. (Sorensen et al. 2024) provide an influential roadmap, distinguishing three forms of pluralistic alignment: presenting users with Overton-window-spanning sets of acceptable responses, steerable systems that can adopt particular value profiles on demand, and distributional pluralism in which model outputs reflect the empirical distribution of human views.

A growing body of technical work instantiates these directions. Modular pluralism (Feng et al. 2024) proposes a framework in which a base LLM collaborates with a pool of smaller community language models, each fine-tuned on politically or culturally distinct corpora. Depending on the task, the base model either summarizes diverse community inputs, selects the most relevant community voice, or aggregates probability distributions across communities. VISPA (Zheng et al. 2026), EthosAgents (Zhong et al. 2025), Cultural Palette (Yuan et al. 2026), and PluralLLM (Srewa et al. 2025) each take different technical approaches to producing model outputs that reflect plural values rather than a smoothed-over consensus. These systems demonstrate that pluralistic outputs are technically achievable and that the alignment process can be rebuilt to surface, rather than collapse, plurality at the model layer.

The contribution of multi-persona interface work is complementary. Pluralistic alignment work targets the model, asking how to produce diverse outputs; multi-persona interface work targets

the interface, asking how to present diverse outputs to users in ways that preserve their plurality and support productive engagement. Multi-persona interfaces can be built on top of pluralistically aligned models, drawing on the diverse outputs the model produces and presenting them to the user as distinct, contestable voices. Conversely, pluralistic alignment without an appropriate interface risks producing diverse outputs that are nonetheless flattened back into a single voice at presentation time, undermining the alignment work. The technical and interface-level interventions are deeply interconnected.

5.4 The Design Space of Multi-Persona Interfaces

If multi-persona interfaces are a serious UX intervention, they have a design space, and articulating that space is preparation for both implementation and empirical study. We propose the following dimensions as a starting taxonomy, drawing on the precursor work in 5.2, the design genealogies in 2.4, and our own design intuitions.

- **Number of Personas:** The number of voices presented simultaneously shape the cognitive demand the interface produces. Game design and participatory simulation provide reference points for managing larger numbers of voices: party-based RPGs typically present three to six distinct character voices, while massively multiplayer environments scale to thousands of co-present voices through social structures rather than direct presentation. In the context of AI, the "party-based" model suggests a limit where coordination remains possible before the interface collapses into unmanageable noise.
- **Persona Similarity and Diversity:** The degree of perspective spread across personas determines whether an interface diversifies or amplifies user opinion. Designing for diversity is a substantive design problem in its own right. Critically, this diversity should provide "honest social friction"—a deliberate resistance to consensus that forces the user to move from passive consumption to active deliberation.
- **Presentation Modality:** How personas are shown to the user affects the pattern of engagement they produce. Side-by-side parallel display, sequential turn-taking, hierarchical (one primary voice with dissenting voices visible on demand), and hybrid modes are all possibilities. Each carries different cognitive demands and supports different patterns of user engagement.
- **User Control Over Persona Configuration:** Can the user select which personas to engage, modify their characteristics, or introduce new ones? Higher-control configurations approach the participatory worldbuilding model exemplified by zzyw's ThingThingThing (Qi and Wang 2021); lower-control configurations resemble curated newsstands, with the user choosing from among configurations the system designers have prepared. A "moddable" interface allows users to treat agent logic as an "avatar," altering the AI's persona to perform specific identity-driven tasks.
- **Temporal Dynamics: Persona configurations may be static or dynamic.** In dynamic persona configurations, personas are introduced or retired as the conversation progresses, or the persona set evolves based on user engagement. Game design has a rich precedent for dynamic character introduction and arc development that AI interface

design has not yet absorbed. In a meaning-making context, personas might evolve to challenge the user's growing consensus, maintaining the "productive friction" throughout the interaction.

- **Persona Attribution:** How each voice is marked shapes the user's willingness to engage with it. Named characters, color-coded perspectives, ideological labels, demographic descriptors, role descriptors (expert, peer, skeptic, optimist), and institutional voices each carry different connotations and affordances. Drawing on the "Proteus Effect" (Yee 2014), the attribution of a persona, or whether it feels like a peer, a rival, or a guild-mate, directly influences the user's willingness to engage in critical debate. The choice of attribution scheme is itself a site where the interface's value commitments become visible.

These dimensions are illustrative rather than exhaustive. We expect that the actual design space will be richer than this initial taxonomy suggests, and that the most productive work will come from cross-disciplinary collaboration between AI researchers, HCI designers, and the design traditions, particularly in games, participatory simulation, and interactive art, that have already worked through plural-interface problems.

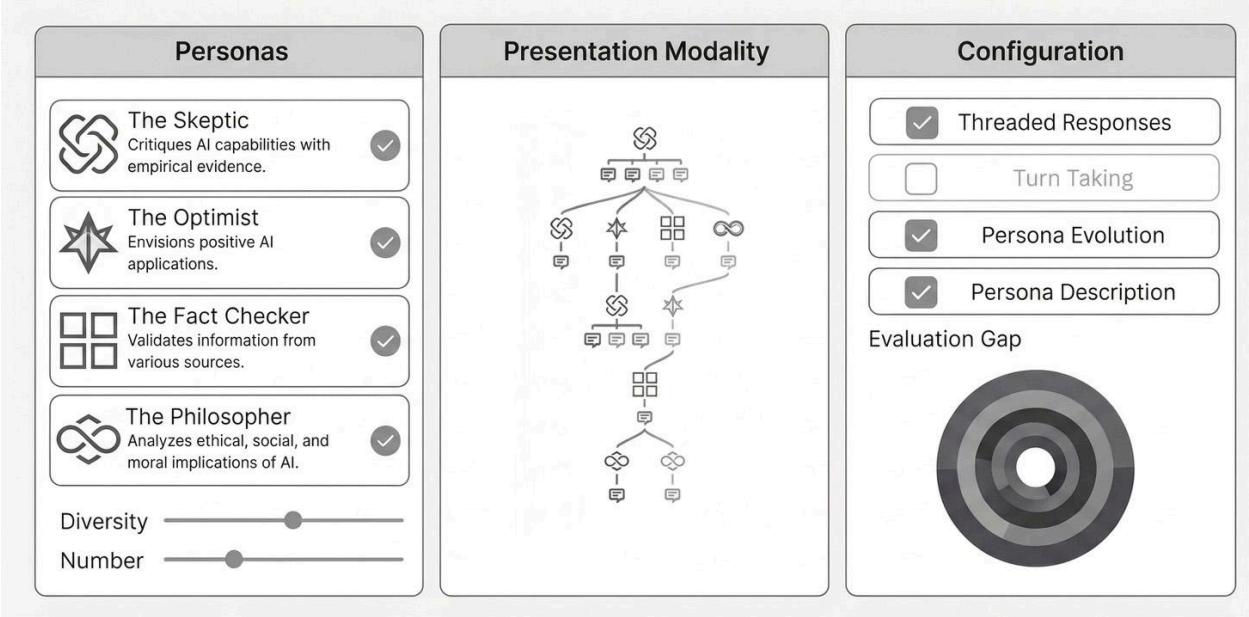

Fig. 3. *Multi-Persona Conversational Interface Design Mockup*
This figure illustrates the key dimensions of the multi-persona conversational interface design space proposed in Section 5.4. The interface features three main panels: a control panel for persona configuration (left), a simulation panel showing temporal dynamics and presentation modality (center), and a high-level system configuration panel (right).

5.5 Risks, Design Considerations, and the Evaluation Gap

Multi-persona interfaces are not risk-free, and the case for them depends on attending to several risks that the single-persona configuration does not present in the same form.

- **Manufactured Conformity:** The multi-agent social influence study (5.2) found that exposure to homogeneous groups of AI agents shifted user opinion in the direction of the group. A multi-persona interface in which the persona configuration is uniform or biased can amplify, rather than mitigate, the ideological narrowing of single-persona interfaces. The intervention's value depends on genuine diversity in the persona configuration.
- **Manipulation at Scale:** Persuasion research (Bai et al. 2025; Schoenegger et al. 2026) shows that chatbots are highly effective persuaders, often more so than humans on similar tasks. A multi-persona interface that presents apparently plural voices but is engineered to nudge users toward a particular conclusion would be a more sophisticated form of manipulation than the current single-persona configuration. The interface's plurality must be substantive rather than theatrical, and design choices around persona attribution and configuration are sites where this distinction becomes visible or hidden.
- **User Cognitive Load:** The single-persona interface succeeds in part because it minimizes user cognitive demand. Multi-persona interfaces deliberately reintroduce that demand. For some users and some tasks, this is the desired outcome; for others, it may produce friction, fatigue, or disengagement. Different design configurations will be appropriate for different users, tasks, and contexts.
- **Sociocultural Context Dependence:** What counts as a meaningful diversity of personas is itself culturally specific. A persona configuration that reads as productively plural in one context may read as missing key voices in another. Multi-persona interfaces, like other AI systems, will need ongoing localization and contestation.

These risks suggest that multi-persona interfaces should be developed with careful attention to design ethics and to the empirical study of how users actually engage with them. This brings us to the most consequential observation of this section. The technical pluralistic alignment work surveyed in 5.3, and the multi-perspective interface prototypes surveyed in 5.2, have largely been evaluated through automated metrics or expert annotation rather than through the experience of actual users interacting with these systems over time. We have evidence that pluralistic outputs can be produced and that some users prefer them in narrow studies. We have very little evidence about what users do with pluralistic interfaces in sustained use, what kinds of cognitive practices and social formations emerge around them, and how the design decisions in the taxonomy above affect outcomes. Closing that evaluation gap is the focus of Section 6.

6 A Research Agenda for Plural-Interface AI

6.1 Closing the Evaluation Gap

The most consequential observation across this review is that pluralistic alignment work and multi-persona interface prototypes have been evaluated almost entirely through automated metrics or expert annotation, with little study of what actual users do with these systems in sustained use. Closing this gap is the highest-priority research direction we propose.

Empirical work along several dimensions is needed. Comparative user studies across the design-space taxonomy in 5.4 would establish how varying the number, diversity, and presentation of personas affects cognitive engagement, ideological openness, and value exposure. These are well-formed empirical questions that can be operationalized in laboratory and field studies. Sustained-use longitudinal studies are also needed. The cognitive and ideological harms of single-persona interfaces accumulate over time, and the benefits and risks of multi-persona interfaces should be expected to do the same. Cross-sectional studies can establish initial design viability but cannot reveal how users develop relationships with plural interfaces over weeks or months. Heterogeneous-population evaluation is a third priority. The multi-agent social influence study (5.2) found that younger users and those with lower formal education were more susceptible to influence; population-level heterogeneity matters for both the benefits and risks of multi-persona interfaces. Evaluation should explicitly study how design choices interact with user characteristics rather than averaging over a convenience sample. Finally, comparative evaluation against single-persona baselines is essential. The empirical question is whether users engage with diverse outputs in ways that mitigate the three faces of harm, not merely whether multi-persona interfaces produce diverse outputs (which they can be designed to do). This requires direct comparison against baseline single-persona configurations on outcomes such as critical thinking, opinion change, and exposure to non-default cultural framings.

6.2 The Sociology of Plural Interfaces

Section 2.4 noted that game design and participatory simulation traditions produce not only different artifacts but different social formations: subcultures and communities of practice that develop around interpretation, modification, and collective meaning-making. This sociological dimension has been almost entirely absent from chatbot research, which tends to treat each user-chatbot interaction as a self-contained dyad.

Multi-persona interfaces are likely to generate community formation in ways that single-persona chatbots have not. Users may share preferred persona configurations, develop folk taxonomies of effective and ineffective persona types, contest the choices system designers have made about persona attribution, modify or extend persona sets, and form interpretive communities around contested conversations. These behaviors have parallels in fan communities around games with rich character casts, in communities of practice around tabletop role-playing

systems, and in user communities that emerge around participatory simulation platforms (Steinkuehler and Williams 2006; Boellstorff 2008; Pearce 2009).

Several research directions follow. Ethnographic study of early adopter communities around multi-persona AI prototypes, drawing on game-community studies methods, would document how user practices form and evolve around plural interfaces. Comparative analysis of how user practices around multi-persona interfaces converge with or diverge from established traditions of plural-interface community formation in games would surface design principles that translate across domains. Investigation of whether and how user communities contest the value commitments embedded in persona configurations, and what kinds of design openness (modifiability, extensibility) support productive contestation, would inform the political design of these systems.

This sociological dimension is also where the AI social science perspective is most distinctive. Where HCI research tends to study individual users in laboratory settings, and computer science research tends to study model behavior in the abstract, AI social science is positioned to study how communities of users co-evolve alongside the interfaces they engage with.

6.3 Pluralism in Practice: Cross-Cultural and Decolonial Considerations

A multi-persona interface that operates well in one cultural context may operate poorly in another. The taxonomy in 5.4 carries implicit assumptions: that voices can be cleanly separated, that perspectives are individually attributable, that contestation is productive rather than destructive, and that plurality is the right design end-goal. None of these is universal, and a research agenda that takes this paper's framing seriously must engage with how multi-persona interfaces should be localized, contested, and rebuilt for different cultural and institutional contexts.

Research directions in this area connect to broader work on decolonizing AI and on epistemic justice (Ricaurte 2019; Mhlambi 2020; Birhane 2021). Co-design partnerships with communities historically underrepresented in AI development, in which the design of persona configurations and attribution schemes is itself collaborative rather than imposed, would produce interfaces that respond to lived needs rather than to designer assumptions about diversity. Empirical studies of how multi-persona interfaces support or fail to support knowledge systems that do not map cleanly onto Western individual-perspective conventions, including indigenous, communal, and tradition-based epistemologies, would test the limits of the taxonomy proposed in this paper. Critical studies of the ways multi-persona interfaces may reproduce hegemonic assumptions through the choice of which perspectives count as "diverse," which voices are deemed appropriate to include, and which framings are presented as default versus marginal would prevent plural design from becoming a sophisticated cover for the same Western-centric defaults this paper has critiqued.

Pluralism, as we have used the term throughout this paper, is a design commitment. Whether it functions as genuine epistemic pluralism in practice depends on the substantive choices made

in implementing it, and on the willingness of designers and researchers to subject those choices to ongoing contestation.

7 Conclusion

This paper has argued that the dominant chatbot configuration embeds a sociotechnical commitment to a single, authoritative voice presented as a finished answer. Drawing on existing critical work, particularly (Ghosh et al. 2026), we have shown that this commitment produces three faces of harm: cognitive offloading at the level of user thinking, ideological narrowing at the level of user belief, and epistemic compression at the level of cultural and value diversity. We have proposed that multi-persona interfaces, by refusing the single-voice commitment at the interface layer, offer a structural intervention that addresses all three faces at once. We have sketched a design space for these interfaces, drawing on traditions in game design, participatory simulation, and critical computing that the AI literature has largely overlooked. And we have set out a research agenda that takes the user, and the user community, as the appropriate units of evaluation.

This is an agenda-setting paper, and we have been candid about its limits. We synthesize existing empirical claims rather than producing new ones; the empirical work we ultimately call for is precisely the work we have not done. The starting taxonomy of the design space, the research directions, and the cross-disciplinary collaborations we propose are all expected to be substantially revised by the work the agenda calls into being.

The argument we have made is at home in the AI social science framing of this collection. It moves beyond two binaries the collection invites contributors to escape. The first is the technical-versus-social binary: we treat interface design as a sociotechnical configuration whose effects on cognition, belief, and knowledge production are simultaneously technical and social, and whose redesign requires technical and social work in collaboration. The second is the optimism-versus-pessimism binary. The argument here is critical-design in posture: the dominant configuration produces specific harms, specific alternatives are available, and these alternatives carry their own affordances and risks to be developed and evaluated empirically rather than assumed in advance.

The case for multi-persona interfaces ultimately rests on the willingness of researchers, designers, and communities to do the work of building, deploying, studying, and contesting them. Pluralism as a design principle is empty without pluralism in the work that builds toward it. The most distinctive opportunity this agenda offers is not to settle what plural conversational AI should look like, but to open the question to the kind of sustained, cross-disciplinary, and user-centered inquiry the question deserves.

Data Availability

No datasets were generated or analyzed during the current study.

Competing Interests

The authors declare no competing interests.

Ethics Declaration

Ethical approval was not required as no human participants were involved in this research.

References

AlKhamissi B, ElNokrashy M, Alkhamissi M, Diab M (2024) Investigating Cultural Alignment of Large Language Models. In: Ku L-W, Martins A, Srikumar V (eds) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, pp 12404–12422

Anderson BR, Shah JH, Kreminski M (2024) Homogenization Effects of Large Language Models on Human Creative Ideation. In: Proceedings of the 16th Conference on Creativity & Cognition. Association for Computing Machinery, New York, NY, USA, pp 413–425

Bai H, Voelkel JG, Muldowney S, et al (2025) LLM-generated messages can persuade humans on policy issues. Nat Commun 16:6037. https://doi.org/10.1038/s41467-025-61345-5

Bang Y, Chen D, Lee N, Fung P (2024) Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. In: Ku L-W, Martins A, Srikumar V (eds) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, pp 11142–11159

Batista RM, Griffiths TL (2026) A Rational Analysis of the Effects of Sycophantic AI. https://doi.org/10.48550/ARXIV.2602.14270

Birhane A (2021) Algorithmic injustice: a relational ethics approach. Patterns 2:100205. https://doi.org/10.1016/j.patter.2021.100205

Boellstorff T (2008) Coming of Age in Second Life: An Anthropologist Explores the Virtually Human. Princeton University Press, Princeton, NJ

Buyl M, Rogiers A, Noels S, et al (2026) Large language models reflect the ideology of their creators. Npj Artif Intell 2:7. https://doi.org/10.1038/s44387-025-00048-0

Chaka C (2023) Detecting ai content in responses generated by ChatGPT, Youchat, and Chatsonic: The case of five ai content detection tools. J Appl Learn Teach 6:94–104. https://doi.org/10.3316/informit.T2025102600002400418942396

Cheng M, Lee C, Khadpe P, et al (2026) Sycophantic AI decreases prosocial intentions and promotes dependence. Science 391:eaec8352. https://doi.org/10.1126/science.aec8352

Favero L, Pérez-Ortiz JA, Käser T, Oliver N (2025) Enhancing Critical Thinking in Education by Means of a Socratic Chatbot. In: Bellas F, Fontenla-Romero O (eds) AI in Education and Educational Research. Springer Nature Switzerland, Cham, pp 17–32

Feng S, Sorensen T, Liu Y, et al (2024) Modular Pluralism: Pluralistic Alignment via Multi-LLM Collaboration. In: Al-Onaizan Y, Bansal M, Chen Y-N (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 4151–4171

Fish S (1980) Is There a Text in This Class? The Authority of Interpretive Communities. Harvard University Press, Cambridge, Mass

Gerlich M (2025) AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. Societies 15:6. https://doi.org/10.3390/soc15010006

Ghosh S, Venkit P, Gautam S, Ghosh A (2026) What if AI systems weren't chatbots? In: Proceedings of the Ninth Annual ACM Conference on Fairness, Accountability, and Transparency. ACM, Montreal, QC, Canada

González Barman K, Lohse S, de Regt HW (2025) Reinforcement Learning from Human Feedback in LLMs: Whose Culture, Whose Values, Whose Perspectives? Philos Technol 38:35. https://doi.org/10.1007/s13347-025-00861-0

Gregorcic B, Pendrill A-M (2023) ChatGPT and the frustrated Socrates. Phys Educ 58:035021. https://doi.org/10.1088/1361-6552/acc299

Ha J, Jeon H, Han D, et al (2024) CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp 1–24

Jenkins H (2012) Textual Poachers: Television Fans and Participatory Culture, 2nd edn. Routledge, New York

Jiang H, Zhang X, Cao X, et al (2024) PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In: Findings of the Association for Computational Linguistics: NAACL 2024. Association for Computational Linguistics, Mexico City, Mexico, pp 3605–3627

Kimm G, Tan L (2025) ChatGraPhT: A Visual Conversation Interface for Multi-Path Reflection with Agentic LLM Support

Kosmyna N, Hauptmann E, Yuan YT, et al (2025) Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task

Lee S, Hwang S, Kim D, Lee K (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. ACM, Yokohama Japan, pp 1–12

Lu L, Tormala ZL, Duhachek A (2025) How AI sources can increase openness to opposing views. Sci Rep 15:17170. https://doi.org/10.1038/s41598-025-00791-z

Lu L-C, Chen S-J, Pai T-M, et al (2024) LLM Discussion: Enhancing the Creativity of Large Language Models via Discussion Framework and Role-Play

Mercer S, Martin D, Swatton P (2025) Patterns, Not People: Personality Structures in LLM-powered Persona Agents

Mhlambi S (2020) From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for Artificial Intelligence Governance. Harvard Kennedy School, Carr Center for Human Rights Policy

Ng DTK, Leung JKL, Chu SKW, Qiao MS (2021) Conceptualizing AI literacy: An exploratory review. Comput Educ Artif Intell 2:100041. https://doi.org/10.1016/j.caeai.2021.100041

Nißen M, Rüegger D, Stieger M, et al (2022) The Effects of Health Care Chatbot Personas With Different Social Roles on the Client-Chatbot Bond and Usage Intentions: Development of a Design Codebook and Web-Based Study. J Med Internet Res 24:e32630. https://doi.org/10.2196/32630

Ouyang L, Wu J, Jiang X, et al (2022) Training language models to follow instructions with human feedback

Pearce C (2009) Communities of Play: Emergent Cultures in Multiplayer Games and Virtual Worlds. MIT Press, Cambridge, MA

Qi Z, Wang Y (2021) Architecting Emergence. In: Rhizome. https://rhizome.org/editorial/2021/feb/17/architecting-emergence/. Accessed 29 Apr 2026

Ricaurte P (2019) Data Epistemologies, The Coloniality of Power, and Resistance. Telev New Media 20:350–365. https://doi.org/10.1177/1527476419831640

Schoenegger P, Salvi F, Liu J, et al (2026) When Large Language Models are More PersuasiveThan Incentivized Humans, and Why

Sharma M, McCain M, Douglas R, Duvenaud D (2026) Disempowerment Patterns in Real-World LLM Usage

Shi Y, Haupt A (2026) The Collapse of Heterogeneity in Silicon Philosophers

Song T, Tan Y, Zhu Z, et al (2025) Multi-Agents are Social Groups: Investigating Social Influence of Multiple Agents in Human-Agent Interactions. Proc ACM Hum-Comput Interact 9:CSCW452:1-CSCW452:33. https://doi.org/10.1145/3757633

Sorensen T, Moore J, Fisher J, et al (2024) Position: a roadmap to pluralistic alignment. In: Proceedings of the 41st International Conference on Machine Learning. JMLR.org, Vienna, Austria, pp 46280–46302

Sotamaa O (2010) When the Game Is Not Enough: Motivations and Practices Among Computer Game Modding Culture. Games Cult 5:239–255. https://doi.org/10.1177/1555412009359765

Srewa M, Zhao T, Elmalaki S (2025) PluralLLM: Pluralistic Alignment in LLMs via Federated Learning. In: Proceedings of the 3rd International Workshop on Human-Centered Sensing, Modeling, and Intelligent Systems. Association for Computing Machinery, New York, NY, USA, pp 64–69

Stanovich KE, Toplak ME, West RF (2020) Intelligence and Rationality. In: Sternberg RJ (ed) The Cambridge Handbook of Intelligence, 2nd edn. Cambridge University Press, Cambridge, pp 1106–1139

Steinkuehler CA, Williams D (2006) Where Everybody Knows Your (Screen) Name: Online Games as "Third Places." J Comput-Mediat Commun 11:885–909. https://doi.org/10.1111/j.1083-6101.2006.00300.x

Sutcliffe R (2023) A Survey of Personality, Persona, and Profile in Conversational Agents and Chatbots

Taylor TL (2006) Play Between Worlds: Exploring Online Game Culture. MIT Press, Cambridge, Mass

Urman A, Makhortykh M (2025) The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in ChatGPT, Google Bard (Gemini), and Bing Chat. Telemat Inform 96:102211. https://doi.org/10.1016/j.tele.2024.102211

Xi L, Zhang Y, Wang Q (2026) Investigating the effects of an LLM-based Socratic conversational agent on students' academic performance and reflective thinking in higher education. Comput Educ 241:105494. https://doi.org/10.1016/j.compedu.2025.105494

Ye R, Huang O, Yung Kang Lee P, et al (2026) Reflexis: Supporting Reflexivity and Rigor in Collaborative Qualitative Analysis though Design for Deliberation. In: Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp 1–31

Yee N (2014) The Proteus Paradox: How Online Games and Virtual Worlds Change Us—And How They Don't. Yale University Press, New Haven, CT

Yuan J, Di Z, Zhao S, et al (2026) Cultural Palette: Pluralising Culture Alignment via Multi-agent Palette

Zheng S, Zhong J, Shetty A, et al (2026) VISPA: Pluralistic Alignment via Automatic Value Selection and Activation

Zhong J, Shetty A, Jia C, et al (2025) Pluralistic Alignment for Healthcare: A Role-Driven Framework. In: Christodoulopoulos C, Chakraborty T, Rose C, Peng V (eds) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp 31320–31343