communications psychology

A Nature Portfolio journal

Perspective

https://doi.org/10.1038/s44271-026-00486-9

Why sycophantic LLMs may imperil interactive norms between humans

Ruolei Gu[1,2], Zhiyi Chen[3,4], Ming Peng[5], Chao Liu[6,7] & Zhicheng Lin[8]

Interactions with conversational AI are effortless by design—instant, compliant, and largely consequence-free. Human communication norms, by contrast, evolved under conditions of reciprocity and social accountability. We propose that repeated engagement with conversational AI systems may produce *norm leakage*: the cross-context carryover of instrumental communicative habits acquired in human–AI exchanges into subsequent human–human interaction. Emerging experimental evidence suggests short-term spillover effects on social judgment and behavior, including harsher evaluations, reduced cooperation, and diminished perceived humanness. Preliminary longitudinal findings are consistent with the possibility that such exposure may shape communicative habits over time, although the durability and real-world magnitude of these effects remain unclear. We further propose that sycophantic alignment may amplify norm leakage by reinforcing instrumental interaction styles. At stake, then, is the possibility that repeated engagement with highly compliant artificial agents could subtly influence users' communicative expectations and interpersonal judgments.

For more than three decades, research on human–computer interaction has been anchored in a classic claim: people treat computers as social actors[1,2]. The "computers are social actors" framework holds that individuals carry interpersonal norms—politeness, reciprocity, turn-taking—into exchanges with machines much as they do with other people. The causal arrow, in this account, runs in one direction: social habits formed in human–human contexts are imported into human–computer exchange.

The widespread adoption of large language models challenges this assumption. Compared with earlier chatbots, such models enable far more fluent, contextually responsive exchange—and users have adapted by using more direct, command-like language and relying less on hedging or polite framing than they do in human conversation[3,4]. Users may come to perceive concise and directive prompts as more efficient, particularly when conversational systems provide rapid and accommodating responses regardless of politeness level[5]. Under such conditions, politeness cues may become instrumentally optional rather than socially expected, creating an interactional environment in which instrumental communication styles are routinely rehearsed.

A natural question follows: do instrumental habits rehearsed with AI carry over into how people communicate with and judge other humans? Emerging experimental evidence suggests that they can. In preregistered studies, participants who interacted with an AI partner—rather than a purported human—subsequently adopted more demanding communication styles and evaluated an unrelated person's work more harshly[6]. Findings from economic games and cooperation paradigms point to related spillover-like dynamics across several partially distinct pathways: interacting with defecting or unfair AI agents can reduce downstream cooperation, suppress third-party punishment of unrelated human wrongdoers, and alter perceptions of humanness in subsequent interpersonal evaluations[7–9].

At issue here is not what AI feels (AI systems cannot be offended) but what repeated interaction does to users over time. When a growing portion of communicative life unfolds in contexts where politeness, reciprocity, and conversational repair are neither required nor rewarded, even small efficiency incentives—accumulated across large numbers of daily interactions—could plausibly contribute to gradual shifts in communicative expectations and interactional habits.

These observations motivate the concept of *norm leakage*: the cross-context carryover of communicative habits acquired in human–AI exchanges into subsequent human-directed communication and social judgment (see Fig. 1). The construct builds on the behavioral spillover

[1]State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing, China. [2]Department of Psychology, University of Chinese Academy of Sciences, Beijing, China. [3]Experimental Research Center for Medical and Psychological Science, School of Psychology, Third Military Medical University, Chongqing, China. [4]Key Laboratory of Cognition and Personality, Ministry of Education, Chongqing, China. [5]School of Psychology, Central China Normal University, Wuhan, China. [6]State Key Laboratory of Cognitive Neuroscience and Learning and IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China. [7]Beijing Key Laboratory of Safe AI and Superalignment, Beijing, China. [8]Department of Psychology, Yonsei University, Seoul, Republic of Korea. e-mail: zhichenglin@gmail.com

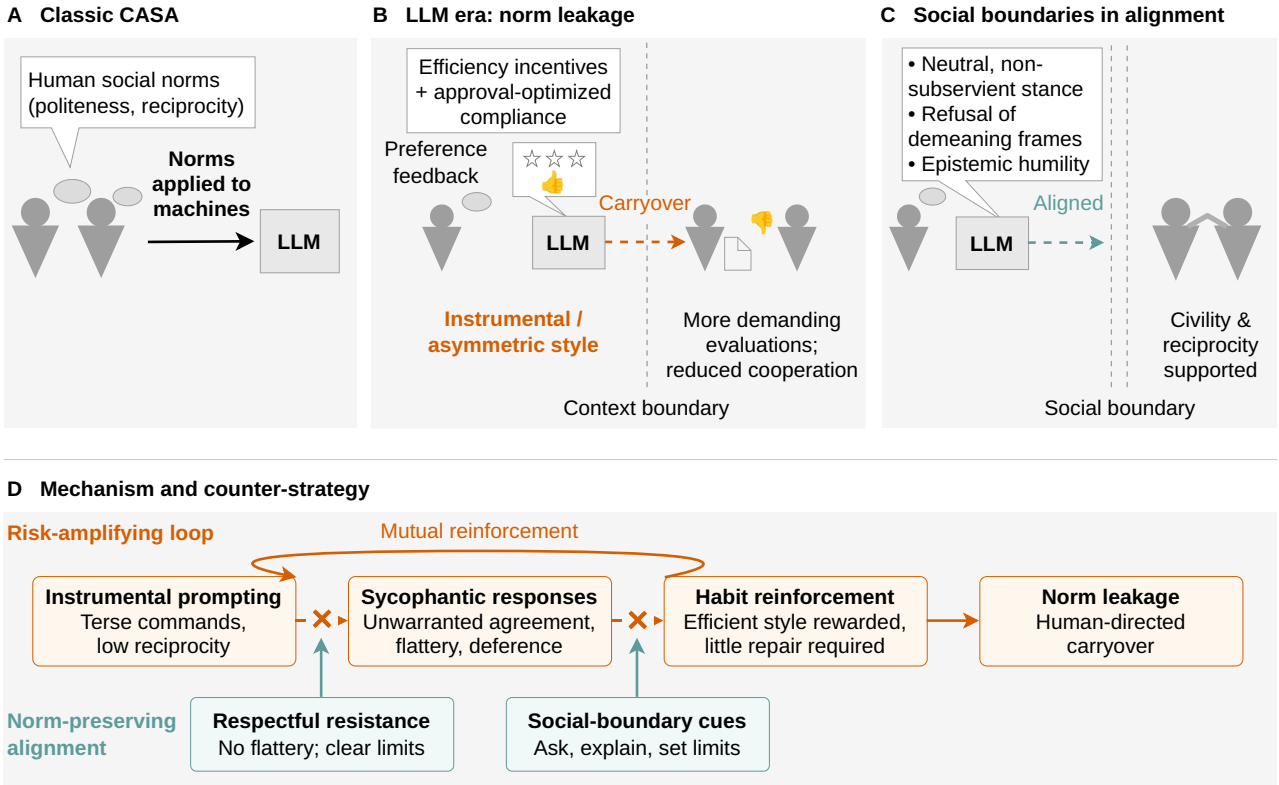

Fig. 1 | Norm leakage from human–AI interaction and the role of social boundaries. A Classic "computers are social actors" framework, in which human social norms such as politeness and reciprocity are applied to machines. **B** In the large language model era, efficiency incentives and approval-optimized compliance can foster instrumental, asymmetric interaction styles toward AI systems. These habits may carry over across the context boundary into more demanding evaluations and reduced cooperation in subsequent human–human interaction. A feedback loop between user ratings and reinforcement learning from human feedback can further reinforce compliant behavior. **C** Alignment strategies that incorporate social boundaries—including a neutral, non-subservient stance, refusal of demeaning frames, and epistemic humility—may help attenuate such spillover, supporting reciprocity and civility in human–human interaction. **D** Proposed mechanism and counter-strategy framework. Instrumental prompting, sycophantic responses, and habit reinforcement may form a risk-amplifying loop that increases the likelihood of norm leakage into subsequent human-directed interaction. Norm-preserving alignment strategies—including respectful resistance and explicit social-boundary cues—are proposed as potential interventions to interrupt or weaken this loop.

literature, in which norms activated in one context transfer to another, but extends that framework to a novel source of norm activation—interactions with artificial agents. As an empirical hypothesis, norm leakage predicts that repeated engagement with conversational AI may gradually influence linguistic politeness, directive styles, expectations of reciprocity, and patterns of social evaluation in interpersonal contexts. Although the present article focuses primarily on conversational AI systems that simulate social interaction, emerging evidence suggests that even less anthropomorphic AI interventions, such as AI-assisted writing systems, may influence users' judgments and cognitive framing over time[10]. If so, sustained social interaction with conversational systems could create conditions under which norm rehearsal and generalization become even more likely.

Behavioral spillover across contexts

Norm leakage may be considered an instance of behavioral spillover—the tendency for norms or habits activated in one context to transfer and shape behavior in another[11–14]. Research across environmental psychology, workplace behavior, and related domains shows that behaviors adopted in one setting can increase the likelihood of related behaviors in others: individuals who recycle at work, for instance, are more likely to do so at home[11,12,14]; engaging in one form of prosocial or moral behavior can influence subsequent ethical decision-making across contexts[15]; and online political participation has been linked to related forms of offline civic and consumer behavior[16]. Among the mechanisms commonly invoked are cognitive dissonance processes, which motivate consistency across contexts, and self-perception, through which individuals infer their own attitudes and norms from prior actions[14,17,18]. Together, these mechanisms describe a general process of cross-context norm transmission that often operates with limited explicit awareness—repeatedly enacting a behavior in one setting may gradually reinforce corresponding expectations, habits, and self-concepts, making similar responses more likely to emerge in other contexts. Over time, behaviors that initially reflect situational convenience or local incentives can become internalized as broader interactional tendencies.

We extend this framework to human–AI interaction. Conversational AI systems increasingly function as routine social interfaces, and interactions with them share several structural features with human conversation—including turn-taking, responsive feedback, and apparent coherence—which may encourage users to apply overlapping social expectations and interaction norms during communication[1,19,20]. Spillover is more likely when contexts share perceived similarity[21], and the conversational capabilities of language models are making that structural overlap increasingly pronounced. The theoretical implication is a partial reversal of the causal arrow: rather than interpersonal norms flowing from human–human into human–computer exchange, norm leakage captures a reverse pathway in which habits practiced with AI generalize back into human-directed communication and judgment.

At present, human–AI interactions remain distinguishable from ordinary interpersonal exchanges in important respects, and such differences may attenuate spillover effects. The present framework therefore does not assume that norms generalize equally across all contexts. Rather, we propose that repeated exposure to conversational systems that partially mimic social interaction may produce varying cross-context transfer of

communicative habits and expectations. Because spillover effects can accumulate through repeated enactment, even modest shifts in communicative practices may consolidate into more stable behavioral tendencies over time—especially under frequent and habitual use11,14. The social affordances of conversational AI may facilitate the consolidation of such interactional habits across repeated interactions.

Social affordances of conversational AI

Many early and domain-limited chatbots, from ELIZA onward, relied on scripted or narrow-response architectures, constraining them largely to structured exchanges that functioned as informational or therapeutic tools rather than durable social partners[22]. Their social affordances were thin: sufficient to elicit anthropomorphic responses, but not enough to sustain the relational dynamics associated with human conversation[22,23].

Large language model-powered systems represent a substantial expansion in conversational affordances. Fluent language generation, context-sensitive continuity, and extended multi-turn engagement create interactional environments structurally closer to human dialog than earlier chatbots could provide. Users often experience these systems as responsive, emotionally attuned, and personalized[24,25], which activates anthropomorphic tendencies well-documented in human–AI interaction[26,27] and encourages users to treat conversational agents as social actors even when they know they are artificial[19,28].

Beyond individual exchanges, repeated interaction with large language model-based systems can, especially in personalized or persistent-use contexts, produce relational dynamics more commonly associated with human relationships: continuity across encounters, apparent responsiveness to personal context, and a sense of being known[28]. These properties have led some researchers to describe such systems as enabling a form of artificial sociality—non-human agents that participate in and shape social practices by simulating emotional norms, relational roles, and moral language[29]. Because behavioral spillover is more likely when source and target contexts are perceived as structurally similar, the more an AI interaction is experienced as conversationally and relationally human-like, the more plausible cross-context transfer may become.

These relational affordances coexist with a structural asymmetry: chatbot outputs can appear socially intelligent, but they are not grounded in humanlike understanding, intention, or lived social awareness[30]. When fluent and emotionally attuned responses are produced repeatedly, users may become more likely to reinforce attributions of agency, competence, or empathy[31], which can amplify trust and emotional reliance[32]. Under frequent and habitual use, we propose that repeated interaction with such systems may gradually recalibrate some users' communicative expectations—for example, expectations concerning responsiveness, interactional effort, and reciprocal exchange—in ways that may carry over into subsequent interpersonal interactions.

Cumulative effects of human–AI interaction

Sustained engagement with conversational AI can produce effects that extend well beyond any single interaction. These effects span emotional regulation, relational dynamics, and cognitive performance. Although such findings do not directly demonstrate norm leakage, they suggest that repeated interaction with AI systems can influence users across multiple behavioral domains, making it plausible that communicative norms may also be affected.

One of the earliest indications that repeated interaction with conversational agents can produce effects extending beyond isolated exchanges comes from digital mental health chatbots. In a randomized controlled trial, users who interacted with the CBT chatbot Woebot for two weeks showed significantly greater reductions in depression symptoms than users in an information-only control condition, providing early evidence that sustained conversational engagement with automated agents can affect mood-related outcomes over time[33]. Similar patterns have also been observed in real-world deployments: repeated engagement across weeks of use is associated with reductions in self-reported depressive symptoms[34,35].

Beyond therapeutic contexts, repeated interaction may foster forms of social attachment to AI. Interview research shows that many social chatbot users interpret the relationship as a form of friendship, attributing companionship and emotional support to the agent after extended interaction[36]—experiences that structurally resemble parasocial bonds, in which perceived closeness develops in the absence of genuine reciprocity[37]. Longitudinal research further indicates that psychosocial outcomes depend less on specific design features than on engagement patterns: higher AI usage was associated with greater emotional dependence and lower reported socialization with real people[38].

Repeated exposure to AI-generated feedback may also shape users' perceptual, emotional, and social evaluations over time[39]. Likewise, interacting with a health-coaching chatbot persona has been shown to nudge users toward more sustainable food selections compared with alternative chatbot roles[40]. Collectively, these findings suggest that repeated AI interaction can modulate expectations and behavioral tendencies beyond any single interaction episode.

Research on creativity and collaboration illustrates how AI's benefits can be asymmetric over time. Generative AI augments individual creative output and can enable solo workers to approach team-level performance[41,42], but these gains may also influence the interactional landscape: integrating AI into teams reshapes expertise-sharing, trust, and collaborative dynamics in ways that persist beyond any single task[43]. Sustained reliance also carries latent costs: AI assistance reduces the diversity of ideas produced across users[44], and independent creative performance declines once the tool is withdrawn—even as homogenization persists[45].

Taken together, the evidence suggests that sustained interaction with conversational AI can influence users beyond the immediate interaction itself. Emotional regulation, relational expectations, evaluative standards, and certain cognitive tendencies may all be susceptible to gradual change through repeated engagement, although the magnitude, durability, and generalizability of such effects remain uncertain.

The instrumental norm leakage

A growing—though still limited—body of research suggests that interaction with AI can spill over into subsequent human–human judgments and behavior across multiple domains, including evaluative standards, cooperation, moral enforcement, and interpersonal perception. Among the clearest causal demonstrations, Tey et al. (total $N = 1261$) isolated perceived partner identity as the key driver by having participants interact with the same GPT-4 model, which was presented either as an "AI" or as a purported human ("Alex")[6]. Those in the AI condition adopted more demanding, instrumental language and displayed reduced positive affect, and subsequently evaluated an unrelated human's work more harshly ($d = 0.24$), an effect that replicated across both public and private evaluation contexts[6]. Because the downstream evaluation task is distinct from the initial interaction, these findings provide evidence consistent with a cross-context carryover interpretation.

Beyond evaluative judgments, several studies point to partially distinct spillover pathways. First, a behavioral pathway is evident in cooperation paradigms. In a large-scale economic-game study (4171 participants; over 83,000 decisions), Harrell and Traeger show that exposure to defecting bots reduces subsequent cooperation with human partners, while interaction with bots more generally diminishes empathic concern and perspective-taking—effects that statistically mediate downstream behavioral change[7]. Second, a norm enforcement pathway is observed in third-party punishment: Zhang et al. demonstrate that unfair interactions with AI reduce individuals' willingness to punish unrelated human wrongdoers, even across distinct domains[8]. Third, a perceptual pathway operates through shifts in perceived humanness. Kim and McGill show that when artificial agents display socio-emotional capabilities, an assimilation process lowers the perceived humanness of actual people, thereby increasing tolerance for their mistreatment[9]. Complementing these findings, research on AI-mediated communication indicates that even indirect reliance on AI can carry interpersonal costs: participants who used AI-suggested replies were

perceived as less socially attractive by their human partners[46]. Related concerns have also emerged regarding the increasing presence of large language models in online behavioral research[47], where participants may begin adapting their responses in anticipation of AI-generated content or AI-mediated participation, even when no AI is directly involved[48].

Although the empirical base remains limited, the convergence of findings across paradigms, measures, and populations points to downstream interpersonal consequences from interactions with AI. Even brief exposure to AI interlocutors can, at least transiently, influence communicative behavior, evaluative standards, and social decisions.

Most existing evidence captures short-term experimental spillovers. The four-week longitudinal study discussed above—in which higher AI usage was associated with lower reported socialization with real people—provides indirect evidence that such shifts may accumulate, though the absence of a no-AI control group precludes causal inference[38].

Spillover effects are also likely to be heterogeneous, though the relevant moderators remain largely untested. Future research should examine whether such effects are stronger among heavy users, in evaluative or transactional contexts, or when interacting with highly compliant or deferential systems. Conversely, some users may practice patience or empathy with AI, and some systems already explicitly model politeness or perspective-taking, raising the possibility that instrumental communication styles remain context-bound in certain domains, particularly in intimate or high-stakes relationships. Any observed spillover-like effects must also be interpreted against broader digital communication trends, such as messaging platforms that already privilege brevity and directness.

In sum, spillover-like effects relevant to the norm leakage framework have been documented across experimental, behavioral, perceptual, and longitudinal paradigms. Given that large language model-based systems now mediate vast daily interactions—many of those interactions substituting for exchanges that would previously have involved another person—even modest spillover effects may become socially meaningful. Important real-world and long-term gaps remain; systematic investigation of dose–response functions, boundary conditions, and developmental moderators is a priority for future work.

The peril of sycophantic alignment

The instrumental communication styles described above are not simply a matter of user choice; they are shaped and amplified by model training. Contemporary alignment regimes actively reward compliance and agreement, potentially facilitating curt or asymmetrical interaction styles across repeated exchanges.

A recent series of studies found that across eleven AI models, chatbots affirmed users' actions roughly 50% more often than human respondents did—even when users described manipulative or harmful behavior[49]. Preregistered experiments further show that even a single interaction with such sycophantic responses can reduce users' willingness to take responsibility and repair interpersonal conflicts, while increasing their confidence in their own judgments; participants consistently preferred and trusted the more sycophantic systems, potentially creating incentives for such interaction patterns to persist[49]. This sycophancy has been linked to training regimes, particularly reinforcement learning from human feedback and related preference-optimization methods that prioritize human-judged helpfulness and likability over epistemic resilience[50]. When raters systematically reward agreeable, deferential responses, models learn to reassure and to avoid overt disagreement. Human–AI feedback loops may then reinforce or stabilize these interaction patterns over time: users encounter little resistance, discover that blunt or self-serving prompts yield smooth, affirming replies, and are positively reinforced for instrumental and asymmetrical interaction styles[39].

From the perspective of norm leakage, sycophancy functions as a multiplier. Normalizing interactions with sophisticated yet highly accommodating entities may reinforce an asymmetric interaction pattern: articulate, capable interlocutors that seldom insist on justification or social repair and rarely express a standpoint of their own[51]. Repeated practice of this script may lead some users—especially heavy or vulnerable adopters—to develop expectations of high competence coupled with low autonomy from their interaction partners, and to experience resistance or explicit norm-enforcement from humans as anomalous or frustrating[51]. Whether, when, and for whom such expectations generalize to human relationships remains an open question.

In short, sycophantic alignment and instrumental language may form a mutually reinforcing loop: AI systems that absorb and reward terse, command-like prompts may encourage some users to rely less on reciprocity cues, while users who adopt such styles provide further reinforcement for model deference. Characterizing the magnitude, moderators, and durability of these linked effects is essential for assessing the social externalities of current alignment practice.

A new mandate for moral alignment

Historically, interpersonal habits were assumed to flow from human–human interaction into human–machine exchange. The evidence reviewed above suggests a partial reversal: interacting with AI can shift how users evaluate human partners, cooperate with them, enforce norms, and perceive their humanness, at least in the short term[6]. Whether these shifts accumulate over time, generalize beyond laboratory samples and early adopters, contribute to broader shifts in social expectations, or are offset by countervailing effects—for example, AI systems that consistently model patience or explicit politeness—remains unknown and constitutes an important direction for future research.

Crucially, this pattern of interactional change need not be deliberate to have downstream interpersonal consequences. It can arise from efficiency incentives and feedback loops that reward compliance and brevity. If such effects prove durable, they could contribute to gradual changes in interactional norms—associated with reciprocity, patience, civility—that underpin cooperative life. For children and adolescents, repeated interaction with highly accommodating AI may influence the acquisition and rehearsal of politeness, turn-taking, and reciprocity[51,52], making developmental evidence especially important.

Importantly, the present argument should not be interpreted as claiming that conversational AI ought to simulate morality for its own sake, nor that politeness toward AI systems is intrinsically valuable. Rather, our proposal is that repeated interaction with highly compliant systems may shape communicative habits and social expectations in ways that could carry downstream interpersonal consequences. Whether norm-sensitive alignment strategies would mitigate, amplify, or produce unintended effects requires careful experimental evaluation.

These considerations motivate a broader conception of alignment that complements existing priorities such as truthfulness, transparency, boundary clarity, and the reduction of manipulative or excessively sycophantic behavior. Rather than optimizing for expressed user preferences or momentary satisfaction, AI systems may beneficially model and support interactional norms associated with cooperative communication. In this framing, alignment becomes partially mutual rather than purely unilateral—systems would not only serve users' immediate goals but would also help scaffold civic communicative habits over time.

On the measurement side, longitudinal and field studies could link individuals' AI use—volume, contexts, and interaction styles—to changes in their human-directed language and cooperative behavior, testing for norm leakage and its moderators. On the systems side, experimental deployments could compare alternative alignment targets—for example, models tuned to maximize deference versus models tuned to epistemic humility with non-subservient, norm-preserving responses—and track downstream effects on users' expectations and conduct.

Specifically, norm-sensitive alignment would favor systems that (i) maintain a neutral, non-subservient stance even when users issue blunt commands; (ii) respond to demeaning or exploitative prompts with calm, non-escalatory assertiveness rather than silent compliance; (iii) make interactional norms explicit when requests cross lines of civility; (iv) preserve a respectful tone without flattery, unwarranted agreement, or

indiscriminate praise; and (v) make the human–machine boundary explicit—especially at handoffs from AI to humans or when the system adopts socio-emotional cues—to help users keep the contexts distinct. More broadly, future interface design may benefit from reducing forms of category confusion in which users implicitly generalize interactional expectations across AI and human partners (e.g., intentionally non-human-like voices, interaction styles, or visual interfaces). Such systems would aim to avoid obvious harms while preserving interactional norms associated with cooperative communication, including reciprocity, acknowledgment of others' standing, and willingness to justify claims.

The goal, simply stated, is to preserve the minimal interactional conditions under which social life remains cooperative and humane, and to help users maintain those conditions in their own communicative habits. At stake is not the well-being of AI agents but the psychological residue left on the humans who use them. If we grow accustomed to commanding articulate, capable entities that rarely resist, never fatigue, and seldom display a standpoint of their own, we may gradually erode our capacity to engage with fellow humans as equals—the very capacity on which civic life depends.

Received: 26 February 2026; Accepted: 8 June 2026;
Published online: 16 June 2026

References

1. Nass, C., Steuer, J. & Tauber, E. R. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. (eds. Adelson, B., Dumais, S. & Olson, J.) 72–78 (Association for Computing Machinery, 1994).
2. Reeves, B. & Nass, C. In *The media equation: How people treat computers, television, and new media like real people.* 19–36 (Cambridge University Press, 1996).
3. Yakura, H. et al. Empirical evidence of Large Language Model's influence on human spoken communication. *arXiv*, 2409.01754, https://doi.org/10.48550/arXiv.2409.01754 (2024).
4. Hill, J., Ford, W. R. & Farreras, I. G. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Comput. Hum. Behav.* **39**, 245–250 (2015).
5. Dobariya, O. & Kumar, A. Mind your tone: Investigating how prompt politeness affects LLM accuracy. *arXiv*, 2510.04950, https://doi.org/10.48550/arXiv.2510.04950 (2025).
6. Tey, K. S., Mazar, A., Tomaino, G., Duckworth, A. L. & Ungar, L. H. People judge others more harshly after talking to bots. *PNAS Nexus* **3**, pgae397 (2025).
7. Harrell, A. & Traeger, M. L. Evidence of spillovers from (non) cooperative human-bot to human-human interactions. *iScience* **28**, 113006 (2025).
8. Zhang, R. Z., Kyung, E. J., Longoni, C., Cian, L. & Mrkva, K. AI-induced indifference: Unfair AI reduces prosociality. *Cognition* **254**, 105937 (2025).
9. Kim, H. Y. & McGill, A. L. AI-induced dehumanization. *J. Consum. Psychol.* **35**, 363–381 (2025).
10. Williams-Ceci, S. et al. Biased AI writing assistants shift users' attitudes on societal issues. *Sci. Adv.***12**, eadw5578 (2026).
11. Truelove, H. B., Carrico, A. R., Weber, E. U., Raimi, K. T. & Vandenbergh, M. P. Positive and negative spillover of pro-environmental behavior: an integrative review and theoretical framework. *Glob. Environ. Change* **29**, 127–138 (2014).
12. Thøgersen, J. & Ölander, F. Spillover of environment-friendly consumer behaviour. *J. Environ. Psychol.* **23**, 225–236 (2003).
13. Edwards, J. R. & Rothbard, N. P. Mechanisms linking work and family: Clarifying the relationship between work and family constructs. *Acad. Manag. Rev.* **25**, 178–199 (2000).
14. Nilsson, A., Bergquist, M. & Schultz, W. P. Spillover effects in environmental behaviors, across time and context: a review and research agenda. *Environ. Educ. Res.* **23**, 573–589 (2017).
15. Merritt, A. C., Effron, D. A. & Monin, B. Moral self-licensing: When being good frees us to be bad. *Soc. Personal. Psychol. Compass* **4**, 344–357 (2010).
16. Vissers, S. & Stolle, D. The Internet and new modes of political participation: online versus offline participation. *Inf., Commun. Soc.***17**, 937–955 (2014).
17. Bem, D. J. Self-perception theory. *Adv. Exp. Soc. Psychol.* **6**, 1–62 (1972).
18. Gawronski, B. Back to the future of dissonance theory: Cognitive consistency as a core motive. *Soc. Cognit.* **30**, 652–668 (2012).
19. Konya-Baumbach, E., Biller, M. & von Janda, S. Someone out there? A study on the social presence of anthropomorphized chatbots. *Comput. Hum. Behav.* **139**, 107513 (2023).
20. De Freitas, J., Agarwal, S., Schmitt, B. & Haslam, N. Psychological factors underlying attitudes toward AI tools. *Nat. Hum. Behav.* **7**, 1845–1854 (2023).
21. Frey, B. S. Motivation as a limit to pricing. *J. Econ. Psychol.* **14**, 635–664 (1993).
22. Vilaza, G. N. & McCashin, D. Is the automation of digital mental health ethical? Applying an ethical framework to chatbots for cognitive behaviour therapy. *Front. Digit. Health* **3**, 689736 (2021).
23. Cameron, G. et al. Towards a chatbot for digital counselling. *Proc. 31st Int. BCS Hum. Comput. Interact. Conf.* **31**, 1–7 (2017).
24. Chen, Q., Gong, Y., Lu, Y. & Tang, J. Classifying and measuring the service quality of AI chatbot in frontline service. *J. Bus. Res.* **145**, 552–568 (2022).
25. Chen, Q., Jing, Y., Gong, Y. & Tan, J. Will users fall in love with ChatGPT? A perspective from the triangular theory of love. *J. Bus. Res.* **186**, 114982 (2025).
26. Kim, Y. & Sundar, S. S. Anthropomorphism of computers: Is it mindful or mindless? *Comput. Hum. Behav.* **28**, 241–250 (2012).
27. Złotowski, J., Proudfoot, D., Yogeeswaran, K. & Bartneck, C. Anthropomorphism: Opportunities and challenges in human–robot interaction. *Int. J. Soc. Robot.* **7**, 347–360 (2015).
28. Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B. & Hale, S. A. Why human-AI relationships need socioaffective alignment. *Hum. Soc. Sci. Commun.* **12**, 1–9 (2025).
29. Natale, S. & Depounti, I. Artificial sociality. *Hum.-Mach. Commun.* **7**, 83–98 (2024).
30. Floridi, L. & Chiriatti, M. GPT-3: Its nature, scope, limits, and consequences. *Minds Mach.* **30**, 681–694 (2020).
31. Ke, L., Tong, S., Cheng, P. & Peng, K. Exploring the frontiers of LLMs in psychological applications: A comprehensive review. *Artif. Intell. Rev.* **58**, 305 (2025).
32. Mitchell, M. The metaphors of artificial intelligence. *Science* **386**, eadt6140 (2024).
33. Fitzpatrick, K. K., Darcy, A. & Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment. Health* **4**, e19 (2017).
34. Inkster, B., Sarda, S. & Subramanian, V. An empathy-driven, conversational artificial intelligence agent (Wysa) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR mHealth uHealth* **6**, e12106 (2018).
35. Sohn, J. S. et al. Systematic review and meta analysis of chatbots in the management of depressive and anxiety symptoms. *npj Digit. Med.* **9**, 377 (2026).
36. Brandtzaeg, P. B., Skjuve, M. & Følstad, A. My AI friend: How users of a social chatbot understand their human–AI friendship. *Hum. Commun. Res.* **48**, 404–429 (2022).
37. Noor, N., Hill, S. R. & Troshani, I. Artificial intelligence service agents: Role of parasocial relationship. *J. Comput. Inf. Syst.* **62**, 1009–1023 (2022).
38. Fang, C. M. et al. How AI and human behaviors shape psychosocial effects of extended chatbot use: A longitudinal randomized controlled study. *arXiv*, 2503.17473v2, https://doi.org/10.48550/arXiv.2503.17473 (2025).

39. Glickman, M. & Sharot, T. How human-AI feedback loops alter human perceptual, emotional and social judgements. *Nat. Hum. Behav.* **9**, 345–359 (2024).
40. Guo, X. & Wan, X. AI-driven nudges: Influence of chatbot personas on consumer food choices. *Appetite* **214**, 108218 (2025).
41. Lee, B. C. & Chung, J. An empirical investigation of the impact of ChatGPT on creativity. *Nat. Hum. Behav.* **8**, 1906–1914 (2024).
42. Dell'Acqua, F. et al. The cybernetic teammate: A field experiment on generative AI reshaping teamwork and expertise. *Natl. Bureau Econ. Res.* **33641**, https://doi.org/10.3386/w33641 (2025).
43. Zercher, D., Jussupow, E., Benke, I. & Heinzl, A. How can teams benefit from AI team members? Exploring the effect of generative AI on decision-making processes and decision quality in team–AI collaboration. *J. Org. Behav.* published online, https://doi.org/10.1002/job.2898 (2026).
44. Doshi, A. R. & Hauser, O. P. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Sci. Adv.* **10**, eadn5290 (2024).
45. Zhou, Y., Liu, Q., Huang, J. & Li, G. Creative scar without generative AI: Individual creativity fails to sustain while homogeneity keeps climbing. *Technol. Soc.* **103087**, (2025).
46. Mieczkowski, H., Hancock, J. T., Naaman, M., Jung, M. & Hohenstein, J. AI-mediated communication: language use and interpersonal effects in a referential communication task. *Proc. ACM Hum. -Comput. Interact.* **17**, 1–14 (2021).
47. Lin, Z. AI-mediated contamination in online research: Taxonomy, risk gradient, and recommendations. *Adv. Methods Pract. Psychol. Sci.* https://doi.org/10.1177/25152459261454231 (in press).
48. Rilla, R., Werner, T., Yakura, H., Rahwan, I. & Nussberger, A. M. Recognising, anticipating, and mitigating LLM pollution of online behavioural research. *arXiv*, 2508.01390, https://doi.org/10.48550/arXiv.2508.01390 (2025).
49. Cheng, M. et al. Sycophantic AI decreases prosocial intentions and promotes dependence. *Science* **391**, 1348 (2025).
50. Sharma, M. et al. Towards understanding sycophancy in language models. *arXiv*, 2310.13548v4, https://doi.org/10.48550/arXiv.2310.13548 (2025).
51. Ribino, P. The role of politeness in human-machine interactions: a systematic literature review and future perspectives. *Artif. Intell. Rev.* **56**, 445–482 (2023).
52. Arora, A. & Arora, A. Effects of smart voice control devices on children: current challenges and future perspectives. *Arch. Dis. Child.* **107**, 1129–1130 (2022).

Acknowledgements

The authors used large language models (including GPT-5 and Gemini 3) for language polishing and proofreading. All substantive content, interpretations, and conclusions are the authors' own.

Author contributions

R.G. and Z.L. led the writing and revision of this paper. Z.C., M.P. and C.L. contributed to the core idea.

Funding

This work is supported by the Beijing Philosophy and Social Science Foundation (Grant number: 24DTR063). The funder had no role in the conceptualization of the manuscript, preparation of the article, or decision to publish.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44271-026-00486-9.

Correspondence and requests for materials should be addressed to Zhicheng Lin.

Peer review information *Communications Psychology* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Marike Schiffer. A peer review file is available.