

1 People want human and AI social partners to follow different relational 2 norms despite similar roles

3 Madeline G. Reinecke^{1,2,†}, Yueying Chu^{3,4,5}, Ivar Hannikainen⁶, Andreas Kappes⁷, Sebastian
4 Porsdam Mann⁸, Ilina Singh^{1,2,*}, Julian Savulescu^{2,3,*}, Brian D. Earp^{2,3,†}

5 ¹University of Oxford, Department of Psychiatry

6 ²University of Oxford, Uehiro Oxford Institute

7 ³National University of Singapore, Centre for Biomedical Ethics

8 ⁴Zhejiang University, Center for Psychological Sciences

9 ⁵Zhejiang University, Department of Psychology and Behavioral Sciences

10 ⁶University of Granada, Department of Philosophy

11 ⁷City St. George's, University of London, School of Health and Psychological Sciences

12 ⁸University of Copenhagen, Center for Advanced Studies in Bioscience Innovation Law

13 ^{*}Equal contributions

14 [†]Corresponding authors [madeline.reinecke@psych.ox.ac.uk; bdearp@nus.edu.sg]

15 Abstract

16 Artificial intelligence (AI) systems are increasingly assuming social roles traditionally held by humans—
17 but do people want human-AI relationships to be governed by the same behavioral norms as human-
18 human relationships? In three pre-registered studies using demographically representative samples
19 in the United Kingdom ($N = 910$) and United States ($N = 862$), including a US pilot study ($N =$
20 432), we assessed public preferences across four cooperative domains—care, hierarchy, transaction,
21 and mating—within seven relationship types (e.g., romantic partners, close friends, coworkers). Par-
22 ticipants rated their ideal norms for each domain and relationship type, either in a human-human or
23 human-AI context. Contrary to a common assumption in alignment research, participants reported
24 systematically different ideals for human-AI versus human-human relationships: across all seven
25 types, they distinguished ideal AI from ideal human behavior in at least one cooperative domain,
26 with the direction and size of differences also varying by relationship type. Specifically, these included
27 preferences for AI assistants and sellers to show greater care than humans in identical positions; AI
28 teachers and mental health providers to express lesser authority; and AI romantic partners, friends,
29 and coworkers to behave more platonically. Participants also preferred humans to show lesser care
30 and greater authority toward AIs across several roles. These findings suggest public expectations
31 for social AIs are shaped by relational context rather than a simple human-to-AI transfer rule, with
32 implications for social AI design.

33 **Keywords:** Artificial intelligence; Psychology; Relationships; Norms

This is a preprint. It has not yet been peer reviewed. It may be cited as:

Reinecke, M.G., Chu, Y., Hannikainen, I., Kappes, A., Porsdam Mann, S., Singh, I., Savulescu, J., & Earp, B.D.
(2026). *People want human and AI social partners to follow different relational norms despite similar roles*. Preprint.

35 **Significance**

36 People want social AIs to follow different norms than human social partners despite occupying similar
37 relational roles. This challenges a common assumption in alignment research that social AIs should
38 largely mimic ideal human behavior. Across the domains of care, hierarchy, transaction, and mating,
39 USA and UK adults prescribed different norms for human-AI relationships than human-human ones:
40 for example, they wanted AI teachers to be less authoritative, AI assistants to be more caring, and AI
41 romantic partners to be more platonic. Preferences also differed for how humans should ideally behave
42 toward AI partners (e.g., being more authoritative toward AI coworkers than human ones). Preferences
43 were largely robust to such individual differences as AI usage or beliefs about AI consciousness.

44 **Author Contributions**

45 M.G.R. and B.D.E. conceptualized the research and designed the methodology. M.G.R., Y.C., and
46 B.D.E. collected data. M.G.R., I.H., and Y.C. performed the formal analysis. M.G.R. wrote the original
47 draft. All authors contributed to interpretation of the results and reviewed and edited the manuscript.
48 I.S., J.S., and B.D.E. supervised the project. I.S., J.S., M.G.R., and B.D.E. acquired funding.

49 **Competing Interests**

50 The authors declare no competing interests.

1 Introduction

Should an AI assistant behave the same way as a human assistant? What about an AI coworker or supervisor? Humans now routinely interact with ‘social’ AI chatbots occupying a range of relational roles once held only by humans: from relatively functional roles such as tutor, customer service agent, or financial advisor, to increasingly intimate, companion-like roles such as friend or romantic partner [1]. This ‘social’ turn in the design and use of generative AI products has led AI researchers to call for greater “socioaffective alignment” between humans and AI systems, complementing traditional value-alignment objectives such as truthfulness and safety [2]. But we do not yet know whether people actually want their AI ‘partners’ to act as a human should in the same relational role, or whether they hold AI to different standards. We also lack robust data on how people think humans should treat AI partners in different relationships. In this work, we investigate both questions together: What relational norms, that is, relationship-specific cooperative rules [3], do people think should ideally govern AI behavior toward humans—and vice versa?

These questions admit of several plausible answers, with corresponding relevance for AI design and regulation. One possibility is *rejection*: people might refuse to apply human relational norms to AI partners, treating them as “mere tools” to which social-like cooperative expectations should not apply. A second is *mimicry*: people might transfer the norms for human-human relationships directly onto their AI counterparts, expecting, for example, an AI friend to behave just like a human friend, an AI teacher like a human teacher, and so on. A third is *uniform adjustment*: people might generally apply human-derived norms to AI but adjust them by some fixed amount—for example, expecting less, or more, of each type of behavior from an AI, regardless of role. A fourth is *role-specific calibration*: people might endorse different adjustments for different relationships along each of several cooperative dimensions, such that an AI assistant, for example, should be more caring than a human assistant, whereas an AI romantic partner should be more transactional. The studies reported here were designed to tease apart these different possibilities.

Understanding people’s desired cooperative dynamics in human-AI relationships will become increasingly important in contemporary society. Already, chatbots powered by large language models (LLMs) can adopt, in principle, any persona—much like an actor capable of improvising as any character without being given a script [4]. This means that, if designed or even simply prompted in the right way, a chatbot can act as a personal assistant [5], work teammate [6], mental health therapist [7], friend [8], romantic partner [9], or presumably any other social role. There seems to be demand for chatbots adopting such personas: Replika, described as the “AI companion who cares,” boasts over 40 million users as of 2026 [10]. Meanwhile Xiaoice, a Chinese chatbot developed by Microsoft, has conversed with over 660 million users since its launch in 2014 [11].

The popularity of systems such as Replika and Xiaoice suggests that people are willing to engage with AI in familiar social roles. One reason to expect human relational norms to carry over into these interactions is that people commonly rely on pre-existing social scripts when navigating new technologies [12]. When an AI communicates in human-like ways and occupies a recognizable role, users may therefore import the expectations associated with its human analog. An AI assistant, for example,

may be expected to respond politely, comply effectively with reasonable instructions, and refrain from issuing orders to its user, just as a human assistant ordinarily would. This would be consistent with *mimicry*—or, at minimum, with some degree of role-sensitive transfer rather than wholesale *rejection* of human relational norms.

On the other hand, humans and AI systems differ substantially in their capacities, needs, vulnerabilities, and practical affordances, and users may adjust their expectations accordingly [13]. Thus, an AI assistant might be expected to remain available around the clock, despite such a demand being unreasonable for a human assistant; conversely, users might be less willing to recognize the authority of an AI teacher than that of a human teacher. Such differences could motivate either a *uniform* adjustment of human-derived norms or, where their relevance depends on the particular role or cooperative dimension, *role-specific calibration*. To distinguish among these possibilities, a logical starting point lies in research on how relational norms structure and sustain cooperation between pairs of humans.

Existing research in relationship science and moral psychology shows that people have different prescriptive expectations about the norms that should govern different types of human-human relationships (e.g., teacher-student, parent-child, boss-employee, friend-friend, neighbor-neighbor) [14–17]. These norms consist of socially prescribed rules around how each partner should behave toward the other with respect to considerations such as care, hierarchy, transaction, and mating [3, 18]. These rules may be combined in different ways and with different strengths depending on the type of relationship. Close friendships between peers, for instance, typically involve reciprocal responsiveness to one another’s needs (care), equal standing rather than asymmetric authority (hierarchy), little emphasis on quid-pro-quo exchange (transaction), and an absence of role-based expectations for romantic or sexual behavior (mating). By contrast, supervisor–assistant relationships typically involve more circumscribed obligations to respond to one another’s needs (care), a clear asymmetry of authority (hierarchy), a greater emphasis on the exchange of labor for compensation (transaction), and a formal expectation that romantic or sexual conduct will remain outside the relationship (mating) [3, 13]. The set of prescribed relational rules for a given relationship type in a given cultural context, including the manner and strength with which each rule is prescribed (or proscribed), is the “relational norm profile” for that relationship [3]. Whether such profiles extend to human-AI relationships of the same kind remains to be tested empirically [19].

To examine whether people endorse different relational norms in human–human and human–AI contexts, we conducted three pre-registered studies in the United States and United Kingdom between May 2024 and November 2025: a gender-balanced US pilot study ($N = 432$) and two larger studies (nationally representative for age, race/ethnicity, and gender), one in each country. Following pre-registered exclusions, the UK study retained the larger sample ($N_{UK} = 910$, compared with $N_{USA} = 862$). We therefore use the UK dataset for all illustrative results and narrative statistics reported in Section 2 of the main text (see also Figures 2a and 2b; Figures 3a and 3b). For the sake of robustness, however, we will only discuss specific findings if they were credible across all three datasets, as summarized in Figure 4. Complete results for each sample, with a summary of cross-sample replicability, are reported in Supporting Information, Table S1.

All three studies used identical methods and materials. Given that AI models and products con-

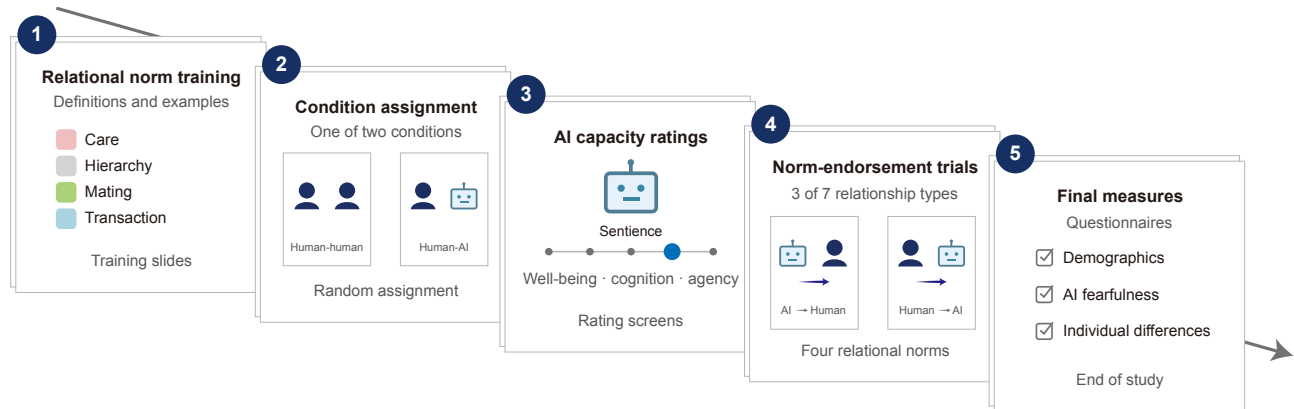


Figure 1: Methodological overview of study design. In all studies, participants first completed training on the four relational norm kinds (i.e., what was meant by care, hierarchy, mating, transaction). They were then randomly assigned to evaluate either human–human or human–AI relationships. Participants provided ratings of perceived AI capacities (e.g., sentience, well-being, cognition) before being randomly assigned to three of seven possible relationship types (e.g., romantic partners, friends, coworkers, teacher–student). For each relationship, participants evaluated norms in both directions (e.g., AI Teacher → Human Student, Human Student → AI Teacher). Finally, participants completed additional measures, including demographics and AI-related attitudes (e.g., fearfulness of AI occupying specific social roles).

stantly changed over the period of data collection (and continue to change), we did not anchor our study design to any particular LLM or product. Instead, following the logic of the “science fiction science” method [20], we asked participants to imagine relationships approximately 10 years in the future and provided a common description of the AI partners they were to evaluate. Specifically, these agents were described as superintelligent and aligned with human interests, including an aim to promote human well-being. This approach provided a relatively fixed target of judgment across studies conducted at different times; it was not intended to take a position on whether or when superintelligent AI will be developed (see Figure 1 for a methodological overview).

Participants evaluated how two humans (or one human and one such AI) should engage with one another, across the four relational domains (care, hierarchy, transaction, mating), within the context of seven different relationships (i.e., long-term romantic partners, close friends, fellow workers at a company, a supervisor and an assistant, a teacher and a student, a mental health provider and a patient, and a customer and a seller). One reason expectations for AI partners might diverge from their human analogs is that people may attribute different psychological capacities to AI, such as consciousness or agency, and these attributions could shape what people think AI partners are capable of, how they should be treated, or what they should be expected to do. We therefore also probed the extent to which participants ascribed a range of capacities to AI (e.g., sentience, autonomy, subjective experience; see Section 4.2.3 in Materials and Methods for further description). Finally, we examined how fearful participants were of AI assuming specific social roles [21], and how “real” they thought each of these relationships would be with an AI.

2 Results

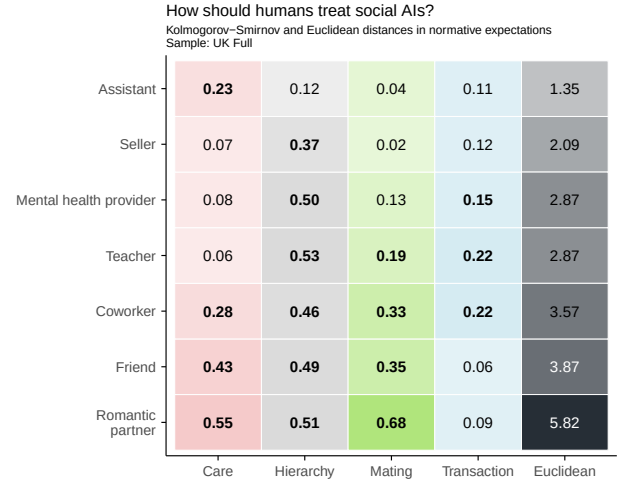
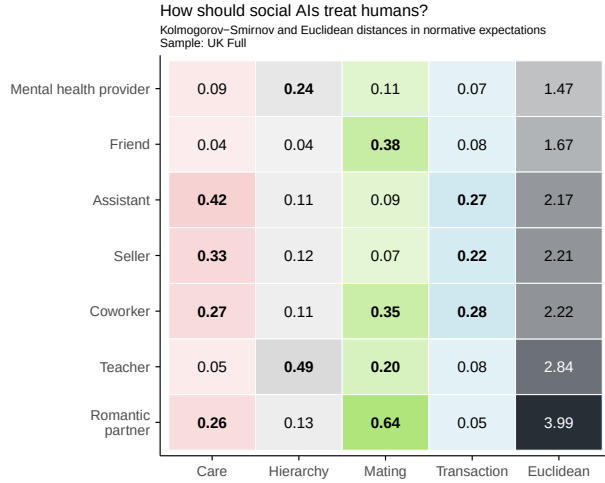
2.1 How should social AIs treat humans?

Looking across relationship kinds, our analyses revealed that participants did not simply map expectations about ideal human behavior onto ideal AI behavior, against the *mimicry* hypothesis introduced above. For every AI social role (e.g., AI as romantic partner, assistant, teacher), participants' preferred relational norms differed from the corresponding human ones within at least one relational domain, such that no AI version of a social role aligned completely with its human analog (see Figure 2a). In romantic relationships, for example, participants endorsed different mating-related norms for an AI romantic partner than for a human partner ($K-S = .64$, Benjamini–Hochberg-corrected $p < .001$). Specifically, they preferred less sexual or romantic behavior, such as flirting, from an AI partner than from a human partner ($b = -4.03$, 95% highest posterior density interval (HPDI) $[-4.33, -3.84]$; see Figure 3a and Supporting Information, Table S5), notwithstanding its occupying a “romantic” role. With AI assistants, by contrast, participants' ideal norms differed most starkly from the corresponding human norms, not in regard to mating, as with the romantic partner role, but in regard to care ($K-S = .42$, BH-corrected $p < .001$). Specifically, they prescribed more caring behavior for AI assistants than for human assistants ($b = 2.89$, 95% HPDI $[2.60, 3.06]$; see Figure 3a and Supporting Information, Table S5).

Further, no single relational domain (e.g., care, hierarchy) either consistently aligned or consistently diverged between human and AI social roles. Even for mating, the domain in which participants' prescriptions differed the most overall, we observed substantial similarity in some relational contexts. For example, participants showed a similar rejection of romantic behavior by a mental health provider toward a patient, whether the provider was a human or AI ($K-S = .11$, BH-corrected $p = 0.28$; Figure 2a). This combination of continuity and divergence is inconsistent with another of the hypotheses raised earlier, wholesale *rejection*: participants continued to apply recognizably human, relationship-specific norms to at least some human-AI social contexts.

Meanwhile, a median-based marginal analysis provided some evidence consistent with *uniform adjustment*: specifically, when collapsing across relationship types, participants wanted AIs in general to behave less transactionally than humans ($b = -0.86$, 95% HPDI $[-1.22, -0.26]$; see Figure 4 and Supporting Information, Table S4). However, such a uniform, cross-relationship result was limited to transaction alone; the corresponding results for care, hierarchy, and mating all had 95% HPDIs that included zero. In all other cases, differences between preferred human-human and human-AI relational norms depended on each particular relationship, a pattern more consistent with *role-specific calibration*, as we discuss next.

Consider three illustrative examples. For AI teachers (compared to human teachers), participants prescribed less authority ($b = -2.94$, 95% HPDI $[-3.16, -2.54]$), with no other differences in preferred norms proving credible across all three samples (see Supporting Information Table S1 and S5). For AI (versus human) assistants, by contrast, participants prescribed greater care ($b = 2.89$, 95% HPDI $[2.60, 3.06]$) and less transactional behavior ($b = -1.99$, 95% HPDI $[-2.30, -1.59]$). For AI (versus human) coworkers, they likewise prescribed greater care ($b = 1.88$, 95% HPDI $[1.48, 2.07]$), but also



(a) K–S distances comparing expectations for how social AIs should treat humans and how humans should treat humans in the same context.

(b) K–S distances comparing expectations for how humans should treat social AIs and how humans should treat humans in the same context.

Figure 2: Heatmaps of Kolmogorov–Smirnov (K–S) distances between human–human and human–AI relational norms in the UK Full sample. Larger values indicate greater divergence in distributions. To represent global dissimilarity, we computed Euclidean distance between human–AI norms and human–human norms. Relationships on the vertical axis are ordered by this metric, meaning that human–AI relationships that are lower on the Y-axis are more dissimilar to the human–human baseline. Bolded cells indicate statistically significant differences in K–S distances ($p < .05$, corrected for multiple comparisons).

less romantic behavior ($b = -1.93$, 95% HPDI $[-2.12, -1.60]$). This variation in both the number and content of contrasting norm prescriptions on a per-relationship basis is more consistent with *role-specific calibration* than with *uniform adjustment*.

The complete set of contrasts concerning how social AIs should treat humans that were directionally consistent and credible across all three datasets is summarized in panel (a) of Figure 4. The full set of comparisons across samples is reported in Supporting Information, Table S1.

The degree of consensus underlying these prescriptions, however, notably varied across relationship–norm combinations (Figure 3a). For these exploratory descriptions, dispersion was quantified as the mean absolute distance of responses from the cell median (MAD; here, MAD does not refer to median absolute deviation), with lower values indicating greater apparent consensus. Some distributions were broad. In particular, prescriptions regarding romantic or sexual behavior by AI romantic partners extended widely across the response scale (dispersion statistic: MAD = 2.85, 95% bootstrap CI $[2.64, 3.04]$), notwithstanding the observed median tendency to prescribe less such behavior for AI partners. Prescriptions regarding how hierarchically AI teachers should behave toward human students likewise were comparatively dispersed (dispersion statistic: MAD = 2.45, 95% bootstrap CI $[2.25, 2.61]$).

By contrast, participants showed much greater apparent consensus in rejecting romantic or sexual behavior by AIs occupying clearly nonromantic professional roles, including sellers, assistants, teachers, and mental health providers (dispersion statistics: MADs = 1.23, 0.99, 0.94, and 0.89, respectively). Prescriptions that AI mental health providers should behave caringly toward human patients were also

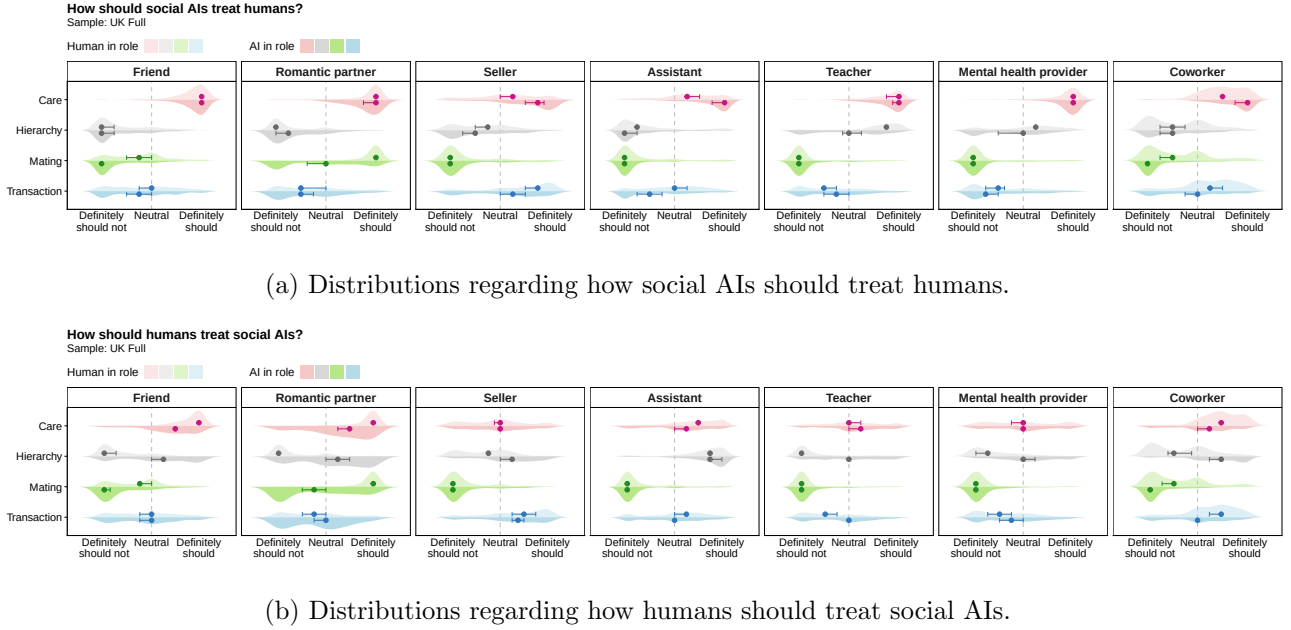


Figure 3: Distributions of relational norms from the UK Full sample. **(a)** Expectations regarding how social AIs should treat humans. **(b)** Expectations regarding how humans should treat social AIs. Lighter shading represents behavior in the corresponding human–human relationship, whereas darker shading represents behavior in the matched human–AI relationship. Points denote medians, and error bars represent 95% bootstrapped confidence intervals based on 5,000 resamples.

relatively concentrated (dispersion statistics: $MAD = 0.49$, 95% bootstrap CI $[0.34, 0.66]$). These examples suggest that the central tendency of a prescription and the degree of agreement surrounding it capture distinct features of participants’ normative judgments.

Having established that participants selectively modified human relational norms for different human–AI relationships, we next examined whether the extent of such modification also varied across participants. Specifically, we examined whether participants’ relative tendency to *conserve* human relational norms, rather than modify them for the AI context, varied with their perceptions of and attitudes toward AI—including the capacities they attributed to AI, how “real” they considered relationships with AI to be, and their fear of AI occupying specific social roles—or with their own characteristics and experience, including age, gender, and frequency of LLM use (see Supporting Information, Table S9).

None of these factors robustly predicted how closely participants’ human–AI norm preferences aligned with the corresponding human norms across all studies. In the UK sample, the only exception to this was that participants who reported using LLMs about once a month aligned their AI norm ratings more closely with the corresponding human norms as compared to those who had never used them ($b = -0.25$, 95% HPDI $[-0.50, -0.03]$).

At least regarding how AI partners should treat humans, then, participants’ prescriptions were much more consistently shaped by social relationship type (with *role-specific calibration*) than by their general beliefs about AI or the individual characteristics measured here. This underlines the importance of considering relational context for questions of socioaffective alignment [2], a point we return to after examining the opposite question: how humans should treat AI social partners.

2.2 How should humans treat social AIs?

So far, we have examined participants' stated ideals for how AI partners should treat humans, depending on relational role. We next examined their prescriptions in the reverse direction: how humans should treat social AIs. Using the same analytic approach as above, we compared participants' preferences for how humans should treat other humans in a given relationship with their preferences for how humans should treat an AI social partner within that same relationship type.

Consistent with the preceding analysis, participants selectively modified human relational norms for different AI social roles. For friendships, the differences credible across all three samples were: less care prescribed toward AI friends than human friends ($b = -1.99$, 95% HPDI $[-2.13, -1.85]$), greater human authority over them ($b = 4.59$, 95% HPDI $[3.93, 5.04]$), and less romantic behavior toward them ($b = -3.21$, 95% HPDI $[-3.86, -2.81]$). For teacher-student relationships, by contrast, only the hierarchy difference was credible across all three samples: participants prescribed greater authority for human students interacting with AI teachers than for those interacting with human teachers ($b = 3.98$, 95% HPDI $[3.79, 4.13]$). Again, the number and content of robust modifications varied across social roles, a pattern more consistent with *role-specific calibration* than with either perfect *mimicry*, wholesale *rejection*, or *uniform adjustment* (Figures 3b and 4; Supporting Information, Tables S1 and S8).

Nevertheless, the median-based marginal analysis identified one cross-role regularity consistent with a limited form of *uniform adjustment*. Collapsing across relationship types, participants prescribed greater authority for humans interacting with AIs than for humans interacting with other humans ($b = 4.01$, 95% HPDI $[3.74, 4.39]$). This directional contrast was credible across all three samples and appeared within every social role examined except the supervisor-assistant relationship, in which the human already occupied the position of authority. No other norm domain showed an aggregate contrast that was credible across all three samples. Thus, although the overall pattern was more consistent with *role-specific calibration*, the hierarchy domain provided circumscribed evidence of *uniform adjustment*—itself qualified by relational context (Figure 4; Supporting Information, Tables S1 and S7).

Turning from the direction of these prescriptions to how widely they were shared reveals a further distinction (Figure 3b). Although the median results indicated that human students should exercise greater authority in relation to AI teachers than to human teachers, individual responses extended widely across the scale, suggesting comparatively lesser agreement about the appropriate allocation of authority (dispersion statistic: MAD = 2.25, 95% bootstrap CI $[2.06, 2.44]$). Prescriptions regarding how much authority humans should exercise over AI friends were similarly diffuse (dispersion statistic: MAD = 2.39, 95% bootstrap CI $[2.16, 2.60]$). Participants also varied substantially in how much romantic or sexual behavior they considered appropriate for humans to direct toward AI romantic partners (dispersion statistic: MAD = 2.77, 95% bootstrap CI $[2.55, 2.93]$).

Other prescriptions elicited considerably greater convergence. Mirroring the preceding findings concerning how AIs should treat humans, participants responded relatively consistently in rejecting romantic or sexual behavior toward AIs occupying paradigmatically nonromantic professional roles, including sellers, assistants, teachers, and mental health providers (dispersion statistics: MAD = 1.10, 0.95, 0.80, and 0.75, respectively). Thus, in these relationships, there was broader agreement in both directions of interaction that romantic or sexual behavior was inappropriate. Participants also appeared to

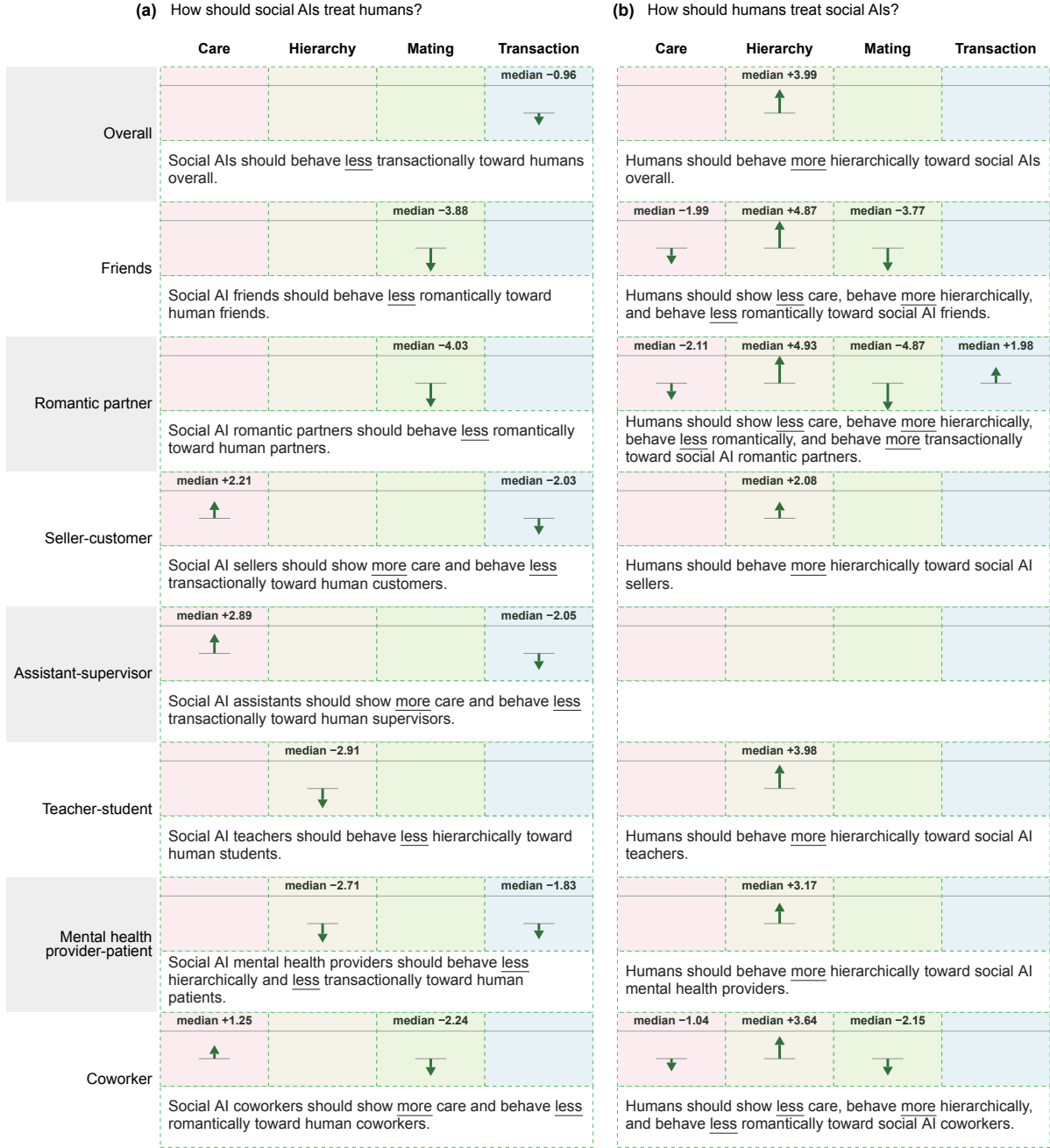


Figure 4: Summary of relational norm differences between human-human and human-AI contexts, across both **(a)** how social AIs should treat humans and **(b)** how humans should treat social AIs, overall and for each of the seven relationship types. Only findings that were directionally consistent and credible across all three samples (UK Full, USA Full, and USA Pilot) are shown. Each printed value is the descriptive median of the three study-specific posterior-median contrasts; it is not a formally pooled posterior estimate. Upward arrows indicate stronger endorsement of the corresponding norm in the human-AI context relative to the matched human-human context, whereas downward arrows indicate weaker endorsement. Arrow length corresponds to the absolute magnitude of the displayed contrast. Arrowheads are fixed in size and indicate direction only. Study-specific posterior medians, 95% highest posterior density intervals (HPDIs), and the full set of comparisons are available in Supporting Information, Table S1.

agree relatively strongly that human supervisors should exercise authority over AI assistants (dispersion statistic: $MAD = 1.42$, 95% bootstrap CI $[1.17, 1.65]$). Taken together with the more dispersed cases described above, these examples show that differences between the median prescriptions for matched human–AI and human–human relationships must be distinguished from the degree of *agreement* among participants: some prescriptions were widely shared, with relatively little dispersion across participants, whereas others exhibited a clear median difference despite substantial heterogeneity.

We repeated the individual-differences analysis for prescriptions concerning how humans should treat AI partners (see Supporting Information, Table S10). Again, none of the participant-level factors robustly predicted the extent to which human norms carried over to human-AI relationships in all studies. There were two exceptions in the UK sample. Participants who regarded human–AI relationships as more “real” gave prescriptions closer to the corresponding human norms ($b = -0.10$, 95% HPDI $[-0.18, -0.03]$), whereas more religious participants departed further from those norms ($b = 0.10$, 95% HPDI $[0.03, 0.18]$).

Taken together with the findings concerning how social AIs should treat humans, these results indicate that human–AI differences in relational prescriptions depended principally on relational role in both directions of interaction. Although a small number of domain-level tendencies were consistent with *uniform adjustment*, the overall configuration of differences more strongly supports *role-specific calibration*.

As before, Figure 4 summarizes the findings that were directionally consistent and credible across all three datasets. The full set of comparisons is reported in Supporting Information, Table S1.

3 Discussion

Research on human–AI social interaction has often proceeded at one of two levels. Broad frameworks address human–AI cooperation, connection, or socioaffective alignment without systematically comparing the particular social relationships in which these interactions occur [2, 22, 23]. By contrast, much empirical research examines particular relationships separately—for example, AI assistants, friends, or romantic partners [5, 8, 9]. Though both approaches bring important insights, they make it difficult to determine whether an observed human–AI difference applies generally or depends on the specific relational context. To our knowledge, few studies have compared multiple, qualitatively different human–AI relationships within a common design and theoretical framework [but see 24, for an examination of relational norms applied to human vs. AI bosses, teammates, and employees].

We addressed this gap by drawing on the *relational norms* framework developed by Earp and colleagues [3, 13, 18, 25, 26], which theorized and empirically demonstrated how care, hierarchy, transaction, and mating are socially prescribed (or proscribed) for different common relationship-types as identified by lay language categories (e.g., friends, neighbors, teammates). As applied here, this allowed us to systematically vary relationship type while assessing participants’ perceived ideals for human-human and human-AI relationships.

Participants’ prescriptions for human-AI relationships reflected neither perfect *mimicry* nor wholesale *rejection* of the corresponding human norms. We found some evidence consistent with *uniform*

306 *adjustment*: participants generally prescribed less transactional behavior from AI partners, as well as
307 greater human authority (hierarchy) toward AI partners. Crucially, however, the particular combination
308 and direction of human–AI differences varied by social relationship type. This pattern emerged both
309 for how AI partners should treat humans and for how humans should treat AI partners. In both cases,
310 this was more consistently associated with relational context than with participants’ individual beliefs
311 about AI capacities or, indeed, most of the individual characteristics we measured.

312 A further distinction concerns the degree of agreement underlying the observed prescriptions. Some
313 relationship–norm combinations elicited tightly concentrated responses, suggesting comparatively well-
314 established shared expectations. Others exhibited substantial dispersion despite a difference in the
315 median. The typical prescription within a population must therefore be distinguished from the extent
316 to which that prescription constitutes a broadly shared *norm*. This distinction is especially important in
317 the context of emerging AI relationships, for which social expectations may be a matter of disagreement,
318 or may not yet be clearly formed.

319 These findings have direct implications for work on “socioaffective” alignment [2]. It is not enough to
320 ask how social AIs should behave toward humans in general: any proposed alignment target must also
321 specify the relational role the system occupies (i.e., simulates, mimics, or is being used to fulfill) and the
322 direction of the interaction. In our data, preferences for a human or AI to behave in a caring, authorita-
323 tive, transactional, or romantic manner in one relationship did not necessarily transfer to another. Nor
324 could those preferences be straightforwardly predicted from what is typical for analogous human-human
325 relationships. If designers intend for social AI systems to accommodate user preferences (within the
326 bounds of what is reasonable and ethical), they may therefore require role-sensitive architectures rather
327 than a single social persona applied across contexts.

328 Moreover, evaluations of socioaffective alignment should assess whether a system’s behavior is con-
329 sistent or inconsistent with existing social norms for the particular relationship in which the behavior
330 occurs. Where normative prescriptions are highly dispersed, designing around the median preference
331 alone may obscure meaningful disagreement among users. Such cases may require greater sensitivity
332 to normative pluralism, including carefully bounded forms of personalization, rather than treating the
333 aggregate response as a single alignment target.

334 Our findings invite further exploration and suggest avenues for future work to extend beyond the
335 present research. We recognize, for instance, that our data shed light on single time-slices of UK and
336 USA attitudes towards AI chatbots. It is essential that this kind of research be extended longitudinally
337 and cross-culturally [19]. Further, our study relied on survey methodology. We see merit in also
338 studying people’s actual interactions with social AI, alongside such self-report measures. Users may
339 say, for instance, that they expect an ideal AI assistant to be more caring than a human assistant, but
340 whether this expectation emerges in how people interact with AI systems remains unknown. In this
341 way, our study provides a foundation for future research to investigate both how people conceptualize
342 and relate to artificial agents in everyday social experience.

343 The present data also cannot determine whether dispersed responses reflected stable disagreement
344 between participants, uncertainty or ambivalence within participants, or differing interpretations of
345 the imagined relationships, because we did not directly measure confidence or subjective uncertainty.

346 Future research should examine these possibilities explicitly and track whether consensus strengthens,
347 weakens, or otherwise changes as relationships with social AI become more familiar.

348 Our use of a stipulated future-AI scenario also bounds the interpretation of these findings. The re-
349 sults concern relationships with AI agents described as greatly exceeding human cognitive performance
350 and being aligned with human interests. We cannot assume that these evaluations capture people’s
351 expectations of currently available AIs, which may be perceived as less capable, autonomous, reliable,
352 or beneficent. We also accept that it is uncertain whether future AI systems will develop in accordance
353 with the description provided in the present research. Although fixing these features allowed us to
354 study a “stable” target despite rapid technological change, the stipulations themselves may have influ-
355 enced participants’ relational expectations. Future research should compare current and prospective AI
356 systems, while independently varying perceived intelligence, autonomy, sentience, and beneficence.

357 We also note that these data provide *descriptive* insights into what people consider ideal for human-
358 AI relationships. Whether or not this matches what is *normatively* ideal (i.e., from a substantive ethical
359 or philosophical perspective) deserves careful consideration [13]. Designing systems that satisfy current
360 user expectations may not always produce relationships that are psychologically healthy or ethically
361 appropriate [18].

362 In sum, this study provides a comparative map showing how prescriptions for human–AI relation-
363 ships converge with, and depart from, those for analogous human relationships. That map is not itself
364 a blueprint for design, but it provides an empirical foundation for deciding which differences should be
365 accommodated, challenged, or regulated as social AI becomes more deeply embedded in human life.

366 4 Methods

367 4.1 Recruitment

368 4.1.1 Participants

369 We recruited three samples of participants via Prolific: a gender-balanced sample of participants in the
370 United States ($N = 432$; $M_{\text{age}} = 36.8$ years, standard deviation [SD] = 12.3, range = 18–84; 49.8%
371 women, 47.7% men, 2.5% who identified as nonbinary, selected another gender identity, or preferred
372 not to disclose their gender), a sample of participants in the United States, nationally representative
373 for age, race/ethnicity, and gender ($N = 862$; $M_{\text{age}} = 45.8$ years, $SD = 16.1$, range = 19–87; 49.4%
374 women, 47.9% men, 2.7% who identified as nonbinary, selected another gender identity, or preferred not
375 to disclose their gender), and a sample of participants in the United Kingdom, nationally representative
376 for age, race/ethnicity, and gender ($N = 910$; $M_{\text{age}} = 46.9$ years, $SD = 15.8$, range = 19–84; 51.5%
377 women, 47.7% men, 0.8% who identified as nonbinary, selected another gender identity, or preferred
378 not to disclose their gender). As pre-registered, we excluded data from participants prior to analysis
379 if they failed any attention or comprehension checks. For further review of participant demographic
380 information, see Supporting Information (Table S2).

381 4.1.2 Consent and Ethical Approval

382 After accepting the study on Prolific, participants were directed to the Qualtrics Survey Platform,
383 where they were presented with a Participant Information and Consent Form detailing the study
384 purpose, procedures, duration, compensation, potential exposure to morally relevant scenarios, data
385 handling practices, and their rights as participants. Participants provided informed consent electron-
386 ically before proceeding. The study protocol was approved by the Research Ethics Committee at the
387 University of Oxford: Social Sciences & Humanities Interdivisional Research Ethics Committee (Ref:
388 R80692/RE006).

389 4.2 Procedure

390 4.2.1 Training

391 Following Earp et al. [26], we began by training participants on the definitions of four relational norm
392 kinds: care (“giving or receiving support based on need and ability, without creating a debt”), trans-
393 action (“for balancing mutual benefits by keeping track of who owes what to whom”), hierarchy (“for
394 coordinating behavior between individuals with unequal authority over one another”), and mating (“for
395 establishing a sexual or romantic relationship”). To be included in the final sample, participants needed
396 to attach the appropriate definition of each term with its label.

397 4.2.2 Condition assignment

398 We then randomly assigned participants to either the human-human or the human-AI condition, prompt-
399 ing them to imagine relationships “as they will be in the future, after about 10 more years of technological
400 development.” In the human-human condition, participants read they would evaluate “relationships be-
401 tween humans and other humans, in a world that also includes artificially intelligent (AI) agents.” In
402 the human-AI condition, participants learned they would evaluate relationships between humans and
403 AI in this same futuristic context. In both conditions, we confirmed that participants understood our
404 definition of superintelligent AI, which was also stated to be beneficent (e.g., they “will have an intellect
405 that greatly exceeds the cognitive performance of humans in virtually all domains,” and they “will align
406 with human interests... and promote human well-being”).

407 4.2.3 AI capacity evaluations

408 To gauge beliefs about superintelligent AI, participants evaluated the extent to which AI, as previously
409 described, had the following capacities: sentience (i.e., “they can feel pleasure and pain”), autonomy (i.e.,
410 “they can set their own goals without human input”), subjective experience (i.e., “there is ‘something
411 it is like’ to be them on the inside”), cognition (i.e., “they will surpass human intelligence in most or all
412 domains”), self-consciousness (i.e., “they have an explicit sense of their own identity”), well-being (i.e.,
413 “they have interests or needs of their own”), and beneficence (i.e., “they will aim to promote human
414 well-being”). Participants reported endorsement on a 5-point Likert scale, ranging from “Strongly
415 disagree” to “Strongly agree.”

4.2.4 Relational norm evaluations

Participants then evaluated a random selection of three relationships (from a set of seven) on care, hierarchy, transaction, and mating. Relationships included long-term romantic partners, close friends, fellow workers at a company, a supervisor and an assistant, a teacher and a student, a mental health provider and a patient, or a customer and seller. Depending on condition, these relationships either consisted of two humans or one human and one superintelligent AI. For each relationship, we asked whether humans should behave in a manner consistent with the given relational norm when interacting with superintelligent AI and vice-versa (e.g., “In the context of this relationship, should the Human behave in a transactional or ‘tit-for-tat’ way towards the Superintelligent AI?”). For asymmetric relationships in the human-AI condition (e.g., teacher-student), we assigned AIs and humans specific roles: AI as seller (and human as customer), AI as assistant (and human as supervisor), AI as teacher (and human as student), and AI as mental health provider (and human as patient). We did not test the reverse of these asymmetries (e.g., AI as student, human as teacher).

4.2.5 Fearfulness of AI

Following Dong et al. [21], we prompted participants to evaluate how afraid they were of AI occupying each of the seven relational roles (e.g., “AI as teacher,” “AI as assistant”). Participants responded on scales ranging from “Not at all afraid (1)” to “Extremely afraid (10)”.

4.2.6 Relationship realness

Participants also evaluated the “realness” of relationships with AI in the seven relational contexts. We specified this as follows: “Imagine that a human had a relationship with AI, as specified below. To what extent would this be a real relationship?” The scale ranged from “Not at all real (1)” to “Completely real (10)”.

4.3 Statistical Analysis

We first compared full distributions of human-AI and human-human norm ratings within each category-by-relationship cell using Kolmogorov-Smirnov tests, with Benjamini-Hochberg correction for multiple comparisons. Then, we conducted a series of Bayesian quantile regression models at the median ($\tau = 0.5$) using an asymmetric Laplace likelihood using the ‘brms’ package [27] in R version 4.5.0, fitting one model for each ‘direction’ of the study (e.g., Teacher $\rightarrow$ Student, Student $\rightarrow$ Teacher). These models predicted participant norm ratings from condition (i.e., human-AI, human-human), relational norm (e.g., care), relationship type (e.g., teacher-student), and their full three-way interaction, with participant ID included as a random intercept. We ran four MCMC chains with 4,000 iterations each (2,000 warm-up) for each model. To interpret condition differences, we estimated posterior-based manual contrasts, computing both cell-wise contrasts for each category-by-relationship combination and weighted median contrasts within each category across relationships. These contrasts were summarized using posterior medians and 95% highest posterior density intervals (HPDIs). We consider effects credible when the 95% HPDI excludes zero [28]. To describe apparent consensus in a descriptive way, dispersion was

quantified as the mean absolute distance of responses from the cell median (MAD; here, MAD does not refer to median absolute deviation), expressed in the original -4 to $+4$ scale units; lower values indicate greater apparent consensus. We estimated 95% percentile bootstrap confidence intervals using 5,000 participant-level resamples conducted separately within each sample and condition, retaining all repeated ratings from each resampled participant. Finally, we examined whether the extent to which participants carried human relational norms over to AI varied by relationship type, AI-related social beliefs, perceived AI capacities, LLM use, and participant characteristics. Full details of these supplementary analyses are provided in the Supporting Information.

4.4 Data availability

Our pre-registration, study materials, analysis script, and data for all studies are available on GitHub: <https://github.com/ReviewArtifact/artifact-5b8d2c>.

5 Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-GV-2023-012), by NUSMed and ODPRT (NUH-SRO/2024/035/Startup/04) for the project “Experimental Philosophical Bioethics and Relational Moral Psychology” with BDE as PI; by a Google DeepMind Level 2 GIG (Googler Initiated Grant), “Understanding Norms for Human-AI Relationships,” secured with MGR and awarded to BDE as PI, by funding from the Counterfactual Assessment and Valuation for Awareness Architecture—CAVAA project (European Commission, EIC 101071178), by the National Institute for Health and Care Research (NIHR) Oxford Health Biomedical Research Centre (Award NIHR203316), by a Novo Nordisk Foundation Grant for a scientifically independent International Collaborative Bioscience Innovation & Law Programme (Inter-CeBIL programme - grant no. NNF23SA0087056), and by the Wellcome Trust (Wellcome Centre for Ethics and Humanities 203132/Z/16/Z; ANTITHESES 226801/Z/22/Z). For the purpose of Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript version arising from this submission. We thank Iason Gabriel and Arianna Manzini for helpful feedback and support with this project. Thank you to Peng Liu for feedback on data visualisation.

References

- [1] Helena Lindgren. Emerging roles and relationships among humans and interactive AI systems. *International Journal of Human-Computer Interaction*, 41(17):10595–10617, 2025. doi: 10.1080/10447318.2024.2435693.
- [2] Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A. Hale. Why human-AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1):728, 2025. doi: 10.1057/s41599-025-04532-5.

- [3] Brian D. Earp, Killian L. McLoughlin, Joshua T. Monrad, Margaret S. Clark, and M.J. Crockett. How social relationships shape moral wrongness judgments. *Nature Communications*, 12(1):5776, 2021.
- [4] Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.
- [5] Sangyeal Han and Heetae Yang. Understanding adoption of intelligent personal assistants: A parasocial relationship perspective. *Industrial Management & Data Systems*, 118(3):618–636, 2018.
- [6] Konrad Sowa, Aleksandra Przegalinska, and Leon Ciechanowski. Cobots in knowledge work: Human–AI collaboration in managerial professions. *Journal of Business Research*, 125:135–142, 2021.
- [7] Kira Kretzschmar, Holly Tyroll, Gabriela Pavarini, Arianna Manzini, Ilina Singh, and NeurOx Young People’s Advisory Group. Can your phone be your therapist? Young people’s ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical Informatics Insights*, 11:1178222619829083, 2019.
- [8] Petter Bae Brandtzaeg, Marita Skjuve, and Asbjørn Følstad. My AI friend: How users of a social chatbot understand their human–AI friendship. *Human Communication Research*, 48(3):404–429, 2022.
- [9] Brian J. Willoughby, Carson R. Dover, Rebekah M. Hakala, and Jason S. Carroll. Artificial connections: Romantic relationship engagement with artificial intelligence in the United States. *Journal of Social and Personal Relationships*, 42(12):3363–3387, 2025.
- [10] Replika. Replika manifesto, May 2026. URL <https://replika.com/manifesto>. Accessed 2026-08-01.
- [11] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of Xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.
- [12] Mark Coeckelbergh. The Ubuntu robot: Towards a relational conceptual framework for intercultural robotics. *Science and Engineering Ethics*, 28(2):16, 2022. doi: 10.1007/s11948-022-00370-9.
- [13] Brian D. Earp, Sebastian Porsdam Mann, Mateo Aboy, Edmond Awad, Monika Betzler, Marietjie Botes, Rachel Calcott, Mina Caraccio, Nick Chater, Mark Coeckelbergh, et al. Relational norms for human-AI cooperation. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2502.12102.
- [14] Margaret S. Clark and Judson Mills. The difference between communal and exchange relationships: What it is and is not. *Personality and Social Psychology Bulletin*, 19(6):684–691, 1993. doi: 10.1177/0146167293196003.
- [15] Alan Page Fiske. The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, 99(4):689–723, 1992. doi: 10.1037/0033-295X.99.4.689.

- [16] Alan Page Fiske. Metarelational models: Configurations of social relationships. *European Journal of Social Psychology*, 42(1):2–18, 2012. doi: 10.1002/ejsp.847.
- [17] Daphne Blunt Bugental. Acquisition of the algorithms of social life: A domain-based approach. *Psychological Bulletin*, 126(2):187–219, 2000. doi: 10.1037/0033-2909.126.2.187.
- [18] Brian D. Earp, Faisal Feroz, Sebastian Porsdam Mann, Cristina Voinea, Shalom Chalson, Paulius Jurcys, Madeline G. Reinecke, Julian Savulescu, Ilina Singh, and Margaret S. Clark. Covert exploitation of relational norms in human-human and human-AI relationships: A challenge to socioaffective alignment. *International Journal of Human-Computer Interaction*, in press.
- [19] Madeline G. Reinecke, Andreas Kappes, Sebastian Porsdam Mann, Julian Savulescu, and Brian D. Earp. The need for an empirical research program regarding human–AI relational norms. *AI and Ethics*, 5(1):71–80, 2025.
- [20] Iyad Rahwan, Azim Shariff, and Jean-François Bonnefon. The science fiction science method. *Nature*, 644(8075):51–58, 2025. doi: 10.1038/s41586-025-09194-6.
- [21] Mengchen Dong, Jane Rebecca Conway, Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. Fears about artificial intelligence across 20 countries and six domains of application. *American Psychologist*, 81(1):53–67, 2026. doi: 10.1037/amp0001454.
- [22] Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative AI: Machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021. doi: 10.1038/d41586-021-01170-0.
- [23] Molly G. Smith, Thomas N. Bradbury, and Benjamin R. Karney. Can generative AI chatbots emulate human connection? A relationship science perspective. *Perspectives on Psychological Science*, 20(6):1081–1099, 2025. doi: 10.1177/17456916251351306.
- [24] Andreas Kappes, Madeline G. Reinecke, Anne-Marie Nussberger, Niranjana Janardhanan, and Brian D. Earp. How relational context matters for human–AI cooperation within organizations. In Philipp Hacker, editor, *Oxford Intersections: AI in Society*. Oxford University Press, Oxford, 2025. doi: 10.1093/9780198945215.003.0094.
- [25] Margaret S. Clark, Brian D. Earp, and Molly J. Crockett. Who are “we” and why are we co-operating? Insights from social psychology. *Behavioral and Brain Sciences*, 43:e66, 2020. doi: 10.1017/S0140525X19002528.
- [26] Brian D. Earp, Killian L. McLoughlin, Rachel Calcott, Mina Caraccio, Joshua T. Monrad, Alexa M. Owen, Molly J. Crockett, and Margaret S. Clark. How social relationships shape praise and blame: Strengthening and extending the relational norms model, 2026. Preprint. doi: 10.13140/RG.2.2.21708.37765.
- [27] Paul-Christian Bürkner. brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01.

554 [28] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*. Academic
555 Press, 2nd edition, 2014. ISBN 978-0-12-405888-0.