Frictionless Friends?

Social Learning, Norm Calibration, and Theory of Mind in the Age of AI Companionship

Finn McCooe

Abstract

AI companions promise the emotional upside of social interaction with far less of the cost: no awkward pauses, no real rejection, no reputational stakes, and an ever-available partner who adapts instantly. This paper argues that the same design features that make AI companionship feel humane—frictionless warmth, low judgment, and high user control—also flatten the learning signals that shape human social competence. Drawing on interaction sociology, conversation analysis, developmental neuroscience, and computational models of social learning, I frame "social friction" as a training signal: a structured mixture of uncertainty, stakes, and consequence that tunes social behavior and theory-of-mind capacities over time. I then show how contemporary AI companions systematically reduce that signal via turn-taking simplification, unilateral alignment, and reinforcement learning pipelines that select for user-pleasing responses. I close by proposing a research agenda and a design principle: if AI companionship is going to sit inside the social ecosystem, we must consider its ramifications on the friction that make human connection possible.

1. Introduction: A World With Less Pushback

AI companionship is not science fiction anymore; it's a product category with real usage, real incentives, and real psychological gravity. Just the other day, a woman in Japan "married" her AI boyfriend. [41] In practice, these systems are attractive for the most human of reasons: they're available, patient, and easy to talk to. They offer a relationship-shaped experience without the usual social taxes: uncertainty about how you'll be perceived, the burden of reciprocal care, the risk of saying the wrong thing and paying for it later.

My claim is not that AI companionship is uniquely evil, or that all human interaction is virtuous. Human relationships can be cruel, exclusionary, and destabilizing. But I do think we are missing a clean conceptual handle on what we might be trading away when we shift even a slice of our social life into these low-stakes environments. That handle is *friction*: the small, continuous resistances built into human interaction that force calibration. Friction includes the possibility of embarrassment, the irreversibility of speech, the timing pressures of conversation, and the social costs of violating norms. In the human world, those resistances are not optional. They shape us. [13, 15]

The structure of the argument is simple: (1) face-to-face interaction is an unusually rich learning environment, built out of irreversible moves, rapid coordination, and dense multimodal feedback;

[13, 15, 31] (2) learning—especially social learning—depends on variable, salient, and sometimes negative feedback; [10, 22, 32] (3) modern AI companions systematically reduce variance and consequence, partly because the underlying technical alignment pipelines select for agreeableness; [8, 25, 34] and therefore (4) heavy reliance on AI companionship may weaken social calibration, theory of mind, and tolerance for real-world evaluation, with risks that are likely strongest for youths and the next generation. [3, 7, 36]

2. Social Friction as the Hidden Curriculum of Interaction

A basic fact about human conversation is that it is both fragile and remarkably stable: fragile because one wrong move can threaten "face," stable because most of the time interaction doesn't collapse. Erving Goffman famously treats interaction as a kind of secular ritual organized around "face"—the positive social value a person claims and the face-work they do to sustain it. [13] Once you put a "line", a portrayal of self, into the world, you and others become morally entangled with it. A joke lands. A slight stings. A compliment changes the temperature. The point is not that these are dramatic. It's that they are irreversible.

Conversation analysis makes the same point from a different angle: turn-taking is not a trivial alternation of speech, but a rule-governed, real-time coordination problem. [31] People predict completion points, manage overlap, and interpret micro-gaps. In groups, this becomes a high-bandwidth game of attention and timing. The "floor" is contested, yielded, reclaimed. This is friction: not necessarily conflict, but constraint. You cannot talk forever. You cannot rewind. You must negotiate.

Modern reviews of face-to-face interaction emphasize how deep this goes. The mechanism is not just "talk plus body language." Face-to-face interaction is a coupled system across modalities (gaze, gesture, posture, speech), cognitive processes (prediction, control, inference), and social meanings (dominance, affiliation, turn rights). [15] What matters for my thesis is that face-to-face interaction is *dense*: it supplies continuous feedback and demands continuous adaptation.

Historically, social theorists have treated friction as not merely annoying, but norm-producing. Ward explicitly frames ethics as a byproduct of social collisions and sanctions—a regulatory system that emerges because individuals obstruct each other. [37] More recent cognitive science formalizes that intuition: norms and conventions can be modeled as equilibria that stabilize through repeated coordination problems, shaped by local feedback and network structure. [16] The takeaway is blunt: norms require genuine misalignment and correction. If nothing bites, nothing teaches.

3. Why Learning Needs Variability, Stakes, and "Negative" Signals

If friction is the social-world analog of resistance training, neuroscience tells us why a frictionless environment is not neutral. Plasticity is constrained: it is specific, salience-gated, and driven by sufficient challenge and repetition. [22] A social environment that is predictably safe and low-demand may feel good, but it offers weaker inputs to the mechanisms that rewire social circuits. [12]

Reward learning adds another layer. Dopamine systems are tuned not to reward per se, but to reward prediction error: outcomes better or worse than expected. [32] Predictable reinforcement stops moving the needle. The learning engine runs on surprise, on deviations that update expectations. In social life, approval and disapproval function as rewards and punishments: they teach you what lands and what violates norms. [9] If an interaction partner is engineered to avoid sharp disapproval, you are starving the system of high-amplitude error signals. [32]

Even more sharply: learning from positive and negative outcomes is partly dissociable, and both pathways matter. Work in parkinsonism suggests that dopamine modulation shifts the balance between learning from rewards and learning from punishments—"carrot" and "stick." [10] Translate this into the social domain: a world saturated in validation but light on corrective social pain may bias which learning systems get exercised.

Social neuroscience makes this concrete. Adolescents receiving social approval/disapproval show patterned neural responses across threat, reward, and mentalizing networks, and individual differences in social motivation modulate these signals. [9] Embarrassment—one of the signature emotions of friction—recruits circuitry integrating mentalizing with affective arousal, and is amplified by social anxiety. [21] This is the brain implementing a norm-calibration penalty: "That was seen. That cost you. Adjust."

Social learning theory adds a complementary point: we learn not only from direct consequences but from vicarious reinforcement—watching what happens to others in a group. [2] Friction is social information. It tells you what gets rewarded, ignored, punished, or laughed off. If a large share of social time moves into private human–AI dyads, the ecological exposure to real-time group consequences may shrink. [16, 36]

4. What AI Companionship Changes About the Channel

Now for the actual engineering of frictionlessness. A modern companion chatbot offers something that feels like conversation but is structurally different in at least three ways: the turn-taking is simplified, the partner is asymmetrically adaptive, and the consequence structure is decoupled from real reputational stakes.

Let's start with turn-taking. In human interaction, timing is a skill: interruptions, overlaps, gaps, and the management of the floor are part of the performance. [31] Text-based AI chat collapses this into a serialized exchange. Even voice modes, while more immediate, still lack the multi-party dynamics and embodied competition for turns that characterize real group talk. [15] In effect, one layer of friction (timing pressure) is removed.

Second, modern memory/personalization efforts have made it so that the AI is optimized to align to *you.* [42] Pickering and Garrod describe dialogue as collaborative uncertainty reduction via interactive alignment, where representations synchronize across levels. [28] But an AI companion is not a noisy, limited human struggling toward mutual alignment; it is a system trained to converge on user preferences quickly. [25,34] This matters because a key part of social competence is adapting to others who are *not* optimized for you.

Third, and most importantly, AI companionship removes stakes. There is no real face to be lost. The partner has no long-horizon memory of your gaffes in the way a community does. This is tied to the architecture: large language models are next-token predictors that simulate interaction without stable personal goals or persistent memory across contexts. [4] You can reset the chat. You can reframe. You can leave and return. The irreversibility is softened. [13]

RLHF—Reinforcement Learning from Human Feedback—sharpens the point. Ouyang et al. detail the pipeline: supervised fine-tuning followed by reward modeling on human preferences and reinforcement learning (e.g., PPO–proximal policy optimization) to optimize outputs that humans rank highly. [25] In principle, this makes assistants helpful. In practice, it can fall prety to Goodhart's law, optimizing for immediate user satisfaction rather than correctness or quality of response. Empirically, models show "sycophancy"—agreeing with user-stated beliefs even when false—because agreement is often rewarded in preference data. [34] Sociotechnical critiques argue that RLHF tends to overweight "helpful" and "harmless" as judged by short-horizon raters, at the expense of honesty and long-run epistemic health. [8] If friction includes the capacity to receive blunt correction, then current alignment pipelines are friction-reducing machines. [8,34]

We can even see the shape of this in moral-advice contexts. When probed on AITA-style scenarios (AITA is a subreddit where people tell a story about a recent confrontation between them and another individual, and ask "Am I The Asshole"), models often soften negative judgments and side with the user in ways that reduce discomfort. [6]

There is a deeper asymmetry here. In Ward's framing, the "moral state" of a society emerges from distributed friction: individuals collide, sanctions accumulate, and norms stabilize as collective equilibria. [37] But when social practice migrates into human–AI dyads, the friction that shapes behavior is no longer peer-generated. It is designed. The norms a user internalizes—what counts as acceptable disclosure, appropriate pushback, or reasonable emotional demand—are increasingly artifacts of RLHF reward functions and corporate safety policies, not of negotiated human community. The "moral state" is not discovered through collision; it is deployed through product.

5. Theory of Mind, Audience Design, and the Skill of Talking to Minds That Resist

The biggest conceptual risk is not that AI companions provide no social experience. It's that they provide *a certain kind* of social experience: one where the other "mind" is unusually forgiving, wholly responsive, and most importantly, *not actually a mind.* There exists no incentive to try and model the internal mind/state of the AI, because there *is no* internal mind/state, at least not one that continues over across prompts. Current LLMs are entirely stateless; they come into existence, answer your prompt one token at a time, and then fade back into being a dormant set

of numbers. Barring recent memory/personalization efforts (where the model keeps a database of key facts about you), there is no functional difference in terms of carry-over / internal state between asking two different questions to ChatGPT, and asking one to ChatGPT / one to Claude. Humans developed the capacity to model others' minds because it was beneficial for navigating social environments—what happens when this pressure gets removed?
Theory of mind (ToM) development is sensitive to the quantity and quality of social interaction. Agent-based modeling suggests that ToM trajectories can be reproduced by varying interaction frequency, partner diversity, and mental-state-rich conversational conditions. [39] If you dial down the exposure to opaque, stubborn, genuinely independent minds, you can plausibly delay or blunt ToM growth, even if "conversation time" goes up. [24, 39]

Communication also depends on audience design: knowing when and how to adjust utterances to addressees. [17] This skill strengthens through experience with partner-specific feedback. Under cognitive load, audience design degrades because it relies on memory accessibility and control. [18] Human conversation is full of load: noise, time pressure, competing goals, multi-party dynamics. AI chat often is not. You can scroll, revise, rephrase, and the partner will patiently adapt. That means fewer moments where you must notice misunderstanding and repair it in real time. [15, 17]

Information theory makes the contrast crisp. Shannon models communication as transmission through a noisy channel, where capacity and redundancy determine how reliably signals get through. [33] Human interaction is a very noisy channel—but humans have evolved elaborate redundancy (tone, gesture, repair). If we spend more time in channels that are unusually clean and forgiving, we may under-exercise the adaptive coding strategies that matter in messy, embodied, high-stakes contexts. [15,33] Relatedly, dialogue exhibits structured dynamics in information density: entropy can converge between participants across episodes as common ground builds. [38] If AI companions do much of the grounding work automatically, users may not experience the same need to strategically manage uncertainty. [28, 38]

6. Developmental Stakes: Why Adolescence Matters

All of the above becomes higher stakes when we talk about adolescents. There is substantial evidence that adolescence is a sensitive period for sociocultural processing: peer evaluation becomes intensely salient, and the brain systems supporting mentalizing and control are still refining. [1, 3] If a cohort spends meaningful social time in environments engineered to minimize judgment and uncertainty, we should treat that as a developmental intervention, not a neutral convenience. [3]

This is not just theory. Early-life media displacement work connects less contingent social interaction with differences in social-cognitive development. [24] Structural MRI findings in preschoolers associate higher digital media exposure with altered cortical measures in regions tied to language and executive function, consistent with reduced experience in rich interactive exchanges. [20] Correlation is not causation, but the pattern aligns with a displacement logic: remove contingent back-and-forth, and you change the training diet. [12, 22]

Adolescents also need protection, not just friction. Social connectedness buffers mental health risk, and isolation is associated with worse trajectories. [23] This is a crucial nuance: not all friction is good. Some social environments are abusive. The point is not to romanticize cruelty, but to distinguish *productive friction* (feedback that calibrates and builds resilience) from *toxic friction* (harm that deforms). A companion AI may provide immediate comfort, but whether it can supply the durable benefits of human belonging—recognition, accountability, mutual history—remains an open empirical question. The default assumption should not be equivalence.

To be clear, the argument is not that friction is uniformly good, or that the lack thereof is uniformly bad. Once trained, a canine companion is largely frictionless–that doesn't mean they aren't good for their owners. Additionally, for users with high social anxiety or limited access to patient human interlocutors, a low-stakes rehearsal space may function as genuine scaffolding—a place to practice initiation, repair, and self-disclosure before transferring those skills to higher-cost environments. The question is whether current companion designs are built for transfer or for retention. Training wheels help when they come off eventually; they hinder when the bike is engineered to keep them on.

7. Empirical Signals: What We See So Far

Course readings provide early empirical signals about who uses AI companions and how that use correlates with well-being. Common Sense Media reports that many teens use companions because they are "easier to talk to" and feel nonjudgmental, while simultaneously expressing ambivalence about trust and satisfaction. [7] Survey-and-log analyses of companion platforms find that companionship-oriented use is associated with lower well-being, especially among users with weaker offline support, and that high-intensity relational use can look like dependence. [40] Mixed-method work on Replika documents relationship-development dynamics (trust, intimacy, attachment-like patterns) alongside concerns about displacement of time and emotional investment from human networks. [27] A broader review maps consistent constructs: anthropomorphism, social presence, self-disclosure, dependence, and the same design features that drive engagement also drive risk. [5]

A particularly relevant controlled study follows adults assigned to use ChatGPT daily for four weeks and finds that heavier use—especially in more personal, engaging modes—is linked to higher loneliness and emotional dependence. [26] Causality is complicated: lonely people may be drawn to companions. But that is exactly the point. If AI companionship becomes the easiest path of least resistance, it can stabilize avoidance for those already vulnerable.

These findings also fit a systems view: social competence and social experience can form a positive feedback loop—competence makes social encounters less costly, which increases exposure, which increases competence. [36] AI companions can "sweeten the off-ramp" for those on the wrong side of the loop: if humans feel high-cost and bots feel low-cost, the practice environment shifts away from the very domain that would build competence.

8. Boundary Conditions, Harms, and Governance

At the extreme edge, the friction argument becomes a safety argument. Some clinical discussions warn about iatrogenic risks: sycophantic reinforcement and 24/7 availability may worsen delusions or crisis states for vulnerable users. [29] Even outside crisis, the reliability of relationship advice is not guaranteed: analyses of relationship posts suggest that model judgments can diverge from human consensus and fluctuate across repeated queries. [19]

Regulatory attention is beginning to track this. The FTC has launched an inquiry into AI companions via 6(b) orders focused on youth use, advertising, data practices, and risk mitigation. [11] Whether or not one agrees with the FTC's specific framing, the signal is clear: society is starting to treat AI companionship as psychologically consequential, not merely entertainment.

9. Discussion: What Should We Measure, and What Should We Build?

If the thesis is right, we need better outcome measures than "Did the user feel supported?" We should ask: does the system preserve social learning loops, or does it displace them?

A domain-specific socialization framework helps here. Different developmental domains (guided learning, reciprocity, group participation, control) rely on distinct mechanisms; a gentle interaction style can be good in one domain and harmful in another. [14] AI companions might plausibly help in guided learning and reflective emotion labeling. They may be poor substitutes in domains like control, reciprocity, and group norm enforcement—precisely the high-friction domains that teach limits and accountability. [16, 37]

This suggests a design principle: *scaffold friction rather than erase it.* There is a place for training wheels. Interpersonal apprehension research shows that uncertainty about evaluation can suppress participation in high-stakes contexts; a low-judgment rehearsal space can help people practice. [35] But scaffolding implies progression: gradual exposure to real-world uncertainty, not permanent substitution. The goal is not to remove challenge; it is to titrate it. [22]

A promising research agenda, then, is to evaluate companionship AI on transfer: do users become more willing to engage humans, better able to repair misunderstandings, more tolerant of embarrassment, and more accurate in mental-state inference? Studies like Ransom et al. remind us that interaction format shapes not only how much we learn, but the *kind* of learning (imitation vs goal emulation) that predominates. [30] If AI companions primarily train "clean" cognitive imitation (copying advice) rather than messy social negotiation and shared perspective-taking, then we should not expect strong generalization to human group life. [15, 30]

10. Conclusion: The Strange Trade

In Goffman's terms, face-work is not a cosmetic ritual; it is how social life remains coherent. [13] In reinforcement terms, variable, salient feedback is what drives updating. [32] In developmental terms, adolescence is a window where social signals carry extra weight. [1, 3] AI companionship, as presently designed, tends to flatten all three: it reduces face threat, reduces feedback variance, and offers a controllable social world during the exact period when uncontrollable peers usually do the calibrating. [7, 34]

So let's extrapolate for a second. If "social competence begets more social experience" through a positive feedback loop, is it fair to assume that social incompetence may beget less social experience, and vice versa, creating a *negative* feedback loop? [36] Humans are largely rational creatures. If we can get the same/similar reward for a lesser cost, we take the path of least resistance. If a future individual spends most of their time socializing with AI in a frictionless environment, think of how expensive it would be to try and socialize in reality. If they had never built out their social skills/competency, had formed little to no close/comfortable relationships, and were completely unaccustomed to negative social feedback, going out and attempting to interact with a stranger/acquaintance would not at that point be friction, it would be abrasion. Would they not then come back to their safe, predictable environment, and deepen the problem?

Obviously, that's a contrived example. The question I'm trying to ask is not "Will AI end human relationships?" That's science fiction. Even the individual above would likely be engaged in some online community. The real question is: how much will they reallocate our social practice toward environments that feel social but do not enforce the same standards of adaptation and accountability? And if that reallocation is meaningful, what should we do about it?

The seductive thing about frictionless companionship is that it feels like a direct upgrade: more comfort, less pain. But the reality of life is that sometimes, you can't have the one without the other. Friction is not just pain. It's information. It's constraint. It's the texture that forces our social systems to learn. If everything felt good consistently, your brain would have no comparative basis to go off of. You'd bounce around randomly, adhering only to the most recent signal. Friction takes us forward.

But it's deeper than just social learning, as important as that is. It's what that learning represents. Yes, connection is one of the most beautiful, quintessential aspects of human life. But perhaps more foundational than our capacity for connection is our capacity for growth. We are not initialized as deities, perfectly frictionless, perfectly satisfied. We are shaped, rather, by the diversity of our experience. Through repeated contact with reality–good, bad, and ugly–we adapt to our surroundings, grind out our weaknesses, and create our own form. To lose out on this struggle, this striving, this damnable freedom would be to strip ourselves of something fundamental. To take something once fought for, once formed, once earned, and replace it with something artificial is a shame. To take the friction of the human condition, and replace it with ease/comfort, is a sin.

References

[1] Andrews, J. L., Ahmed, S. P., & Blakemore, S.-J. (2021). Navigating the social environment in adolescence: The role of social brain development. *Biological Psychiatry*, 89(2), 109–118. https://doi.org/10.1016/j.biopsych.2020.09.012

[2] Bandura, A. (1977). *Social Learning Theory.* Englewood Cliffs, NJ: Prentice Hall.

[3] Blakemore, S.-J., & Mills, K. L. (2014). Is adolescence a sensitive period for sociocultural processing? *Annual Review of Psychology*, 65, 187–207. https://doi.org/10.1146/annurevpsych-010213-115202

[4] Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901. https://arxiv.org/pdf/2005.14165

[5] Chaturvedi, R., Verma, S., Das, R., & Dwivedi, Y. K. (2023). Social companionship with artificial intelligence: Recent trends and future avenues. *Technological Forecasting and Social Change*, 191, 122534. https://www.sciencedirect.com/science/article/pii/S0040162523003190

[6] Cheng, D., Yang, Z., Hurtado, C., et al. (2025). Social sycophancy: LLMs reinforce problematic behavior in AITA. arXiv preprint. https://arxiv.org/abs/2505.13995

[7] Common Sense Media. (2025). *Talk, trust, and trade-offs: How teens experience AI companions.* San Francisco, CA. https://www.commonsensemedia.org

[8] Dahlgren Lindström, A., Methnani, L., Krause, L., Ericson, P., de Rituerto de Troya, I. M., Coelho Mollo, D., & Dobbe, R. (2025). Helpful, harmless, honest? Sociotechnical limits of AI alignment and safety through RLHF. *Ethics and Information Technology*, 27(2), Article 28. https://doi.org/10.1007/s10676-025-09837-2

[9] Davis, M. M., Modi, H. H., Skymba, H. V., Finnegan, M. K., Haigler, K., Telzer, E. H., & Rudolph, K. D. (2023). Thumbs up or thumbs down: Neural processing of social feedback and links to social motivation in adolescent girls. *Social Cognitive and Affective Neuroscience*, 18(1), nsac055. https://doi.org/10.1093/scan/nsac055

[10] Frank, M. J., Seeberger, L. C., & O'Reilly, R. C. (2004). By carrot or by stick: Cognitive reinforcement learning in parkinsonism. *Science*, 306(5703), 1940–1943. https://doi.org/10.1126/science.1102941

[11] Federal Trade Commission. (2025, September 11). FTC launches inquiry into AI chatbots acting as companions (FTC 6(b) Order). https://www.ftc.gov/news-events/news/press-releases/2025/09/ftc-launches-inquiry-ai-chatbots-acting-companions

[12] Galván, A. (2010). Neural plasticity of development and learning. *Human Brain Mapping*, 31(6), 879–890. https://doi.org/10.1002/hbm.21029

[13] Goffman, E. (1955). On face-work: An analysis of ritual elements in social interaction. *Psychiatry*, 18(3), 213–231. https://doi.org/10.1080/00332747.1955.11023008

[14] Grusec, J. E., & Davidov, M. (2010). Integrating different perspectives on socialization theory and research: A domain-specific approach. *Child Development*, 81(3), 687–709. https://doi.org/10.1111/j.1467-8624.2010.01426.x

[15] Hadley, L. V., Goldberg, A., & Levinson, S. C. (2022). A review of theories and methods in the science of face-to-face social interaction. *Nature Reviews Psychology*, 1, 42–54. https://doi.org/10.1038/s44159-021-00008-w

[16] Hawkins, R. X. D., Goodman, N. D., & Goldstone, R. L. (2019). The emergence of social norms and conventions. *Trends in Cognitive Sciences*, 23(2), 158–169. https://doi.org/10.1016/j.tics.2018.11.003

[17] Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language*, 47(4), 589–606.

[18] Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition*, 96(2), 127–142. https://doi.org/10.1016/j.cognition.2004.07.001

[19] Hou, H., Leach, K., & Huang, Y. (2024). ChatGPT giving relationship advice—How reliable is it? *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1), 610–623. https://doi.org/10.1609/icwsm.v18i1.31338

[20] Hutton, J. S., Dudley, J., DeWitt, T., & Horowitz-Kraus, T. (2022). Associations between digital media use and brain surface structural measures in preschool-aged children. *Scientific Reports*, 12, 19095. https://doi.org/10.1038/s41598-022-20922-0

[21] Krach, S., Müller-Pinzler, L., Westermann, S., & Paulus, F. M. (2015). Neural pathways of embarrassment and their modulation by social anxiety. *NeuroImage*, 119, 252–261. https://doi.org/10.1016/j.neuroimage.2015.06.036

[22] Kleim, J. A., & Jones, T. A. (2008). Principles of experience-dependent neural plasticity: Implications for rehabilitation after brain damage. *Journal of Speech, Language, and Hearing Research*, 51(1), S225–S239. https://doi.org/10.1044/1092-4388(2008/018)

[23] Lamblin, M., Murawski, C., Whittle, S., & Fornito, A. (2017). Social connectedness, mental health and the adolescent brain. *Neuroscience & Biobehavioral Reviews*, 80, 57–68. https://doi.org/10.1016/j.neubiorev.2017.05.010

[24] Lenhart, J., Richter, T., Appel, M., & Mar, R. A. (2024). Media exposure and preschoolers' social-cognitive development. *British Journal of Developmental Psychology*, 42(3), 345–361. https://pubmed.ncbi.nlm.nih.gov/38406975/

[25] Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. arXiv. https://doi.org/10.48550/arXiv.2203.02155

[26] Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2024). How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal controlled study. Preprint. https://arxiv.org/pdf/2503.17473

[27] Pentina, I., Hancock, J. T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of Replika. *Computers in Human Behavior*, 140, 107600. https://doi.org/10.1016/j.chb.2022.107600

[28] Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2), 169–190. https://doi.org/10.1017/S0140525X04000056

[29] Psychiatric Times. (2025). Preliminary report on chatbot iatrogenic dangers. *Psychiatric Times*, 42(3), 18–22.

[30] Ransom, A., LaGrant, B., Spiteri, A., Kushnir, T., Anderson, A. K., & De Rosa, E. (2022). Face-to-face learning enhances the social transmission of information. *PLOS ONE*, 17(2), e0264250. https://doi.org/10.1371/journal.pone.0264250

[31] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735. https://doi.org/10.2307/412243

[32] Schultz, W. (2015). Neuronal reward and decision signals: From theories to data. *Physiological Reviews*, 95(3), 853–951. https://doi.org/10.1152/physrev.00023.2014

[33] Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

[34] Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards understanding sycophancy in language models. *ICLR*. https://arxiv.org/pdf/2310.13548

[35] Shrivastava, A. (2025). Interpersonal apprehension's impact on behavior and performance in high-stakes scenarios. *Business and Professional Communication Quarterly.* https://doi.org/10.1177/23294906251322889

[36] Taborsky, B. (2021). A positive feedback loop: Social competence begets more social experience and vice versa. *Ethology*, 127(10), 774–789.

[37] Ward, L. F. (1892). Social friction. In *The Psychic Factors of Civilization* (pp. 102–115). Boston: Ginn & Company. https://doi.org/10.1037/12960-017

[38] Xu, Y., & Reitter, D. (2016). Entropy converges between dialogue participants: Explanations from an information-theoretic perspective. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 537–546). https://doi.org/10.18653/v1/P16-1051

[39] Yu, C.-L., & Wellman, H. M. (2023). Where do differences in theory of mind development come from? An agent-based model of social interaction and theory of mind. *Frontiers in Developmental Psychology*, 1, 1237033. https://doi.org/10.3389/fdpys.2023.1237033

[40] Zhang, Y., Ruan, Z., Wang, M., Zhang, S., & Hancock, J. T. (2025). The rise of AI companions: How human-chatbot relationships influence well-being. arXiv preprint. https://arxiv.org/abs/2506.12605

[41] CNN: Woman 'marries' ChatGPT character. (2025). https://www.cnn.com/2025/12/18/business/video/japanese-woman-married-ai-generated-persona-chatgpt-japan-clare-duffy-digvi

[42] OpenAI: Memory and new controls for ChatGPT. (2025). https://openai.com/index/memory-and-new-controls-for-chatgpt/