Friction-Free Friends?

Social Learning and Theory of Mind in AI Companionship

Finn McCooe

Pitch for Shifted Direction

How much of social learning depends on real-time ”friction”—and what happens when generative AI removes it?

We are entering uncharted territory with generative AI, and it is imperative that we question the future impacts of emerging applications. AI companionship is an increasingly popular use case that offers us something we've never before had: social interaction, without social consequences. The question then becomes, can complex social learning survive in a controlled, artificial environment?

In the real world, and for all of linguistic human history, socializing has been messy, irreversible, and unpredictable. There's been *friction* inherent to the environment: inevitable costs and constraints that make socializing imperfect—one cannot retract a poorly chosen comment, pause mid-conversation for extended reflection, or instruct the other person on how they are to respond. Other people's reactions (words, facial expressions, body language) are also fundamentally uncontrolled, leading to variable reinforcement throughout a given conversation. A joke can kill or bomb; the smile or grimace that follows is a lesson. This uncertainty and mixture of positive and negative feedback provides the basis for learning and behavioral reinforcement; your brain cannot effectively learn without it. It enables people to learn “socially acceptable” behavior/norms, and build the skills to successfully navigate dynamic social environments.

I argue that generative AI, with its consistently predictable validation and highly controlled environment, does not provide the requisite variable reinforcement for social learning to occur. Conversations are asynchronous, reversible, and almost entirely forgiving. Users never have to answer an unexpected question, can edit their queries endlessly or delay their responses indefinitely, and–because AI is stateless and RLHF rewards agreeableness– users receive virtually no negative signal. Friction rounds out rough edges, making people more socially aligned around agreed upon norms/behaviors. It also forms people into socially capable individuals, positively reinforcing behaviors that are socially successful and negatively reinforcing behaviors that lead to general social discomfiture. Without friction, social learning stalls.

When social friction disappears, a second risk emerges: Theory-of-Mind atrophy. Human interaction demands constant prediction and adjustment based on others' presumed mental states. It requires an ability to understand the specific human you are interfacing with, and adapt accordingly. When conversing with AI, there's no hidden mental state to model, no persistent personality to navigate, no memory of past interactions to consider. You no longer need to learn how to read people, and predict what they are thinking–it's simply not necessary for you to "successfully" navigate an AI conversation. Thus, the brain's 'use it or lose it' principle suggests that prolonged AI companionship could weaken our capacity to dynamically model others' minds, leaving people less able to understand complex social environments.

This directed reading will explore how variable feedback drives learning, how friction and real-world interactions supply it, how current RLHF practices collapse to purely positive feedback, and thus, how AI companionship may lead to worse social capabilities. I will argue that if we do not tread carefully with this increasingly popular GenAI application, we could observe increased isolation in future generations–not because they'll lack desire for connection, but because they'll lack the friction-learned skills to navigate real-world environments successfully.

The term will culminate in (i) a public web report formalising these arguments and (ii) a toy mitigation demo that shows how prompt tuning can re-inject friction.

Deliverables

1. **Weekly Reading Minis (Weeks 1–9)**
   - One concise memo (300 words) each week.
   - Structure: (i) key takeaway, (ii) critical question, (iii) link to the project's central claim.
   - Purpose: reinforce active reading and create a running log of insights for the final report.
2. **Re-Injecting Friction Demo (Prompt Tuning)**
   - Build a 100-item "bad-joke" dataset with human harshness ratings.
   - Use a lightweight black-box search to discover system prompts that *increase* the model's agreement with human negative feedback, thereby re-introducing social friction.
3. **Final Web-Based Synthesis Report (Due Finals Week)**
   - A polished, public-facing website (or Notion page) integrating:
     (a) an 2 000-word argument paper,
     (b) interactive concept map linking readings to claims, and
     (c) an appendix summarising the prompt-discovery experiment.

Advisor touch-points. Bi-weekly 15-minute check-ins to refine direction and discuss findings.

Reading List

Week 1 (9/22): Foundations of Social Friction & Face-to-Face Interaction

1. 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

   Establishes irreversibility of social actions and face maintenance as core social friction

2. 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

   Comprehensive overview of mechanisms unique to face-to-face interaction

3. 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

   Turn-taking as fundamental social skill requiring real-time negotiation

4. Ward, L. F. (1892). Social friction. In L. F. Ward, *The psychic factors of civilization* (pp. 102–115). Boston: Ginn & Company. https://doi.org/10.1037/12960-017

   Establishes a classical perspective on how social interaction/friction leads to societal (moral) alignment

5. 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

   How social friction leads to social norms

6. Ransom, A., Perfors, A., Navarro, D. J., & Platow, M. J. (2022). Face-to-face learning enhances the social transmission of information. *PLOS ONE*, *17*(2), e0263302. https://doi.org/10.1371/journal.pone.0263302

   Face to face learning leads to richer information transmission

Week 2 (9/29): Neural Plasticity & Social Learning

1. 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)

   "Use it or lose it" principle; need for challenging experiences to drive plasticity

2. 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

   Recent evidence of how social feedback drives neural learning

3. 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

   Biological basis for why variable reinforcement is necessary for learning

4. Hofmans, L., van den Bos, W., Li, S.-C., & Crone, E. A. (2025). Developmental differences in social information use under uncertainty: A neurocomputational approach *Developmental Cognitive Neuroscience*, *69*, 101460. https://doi.org/10.1016/j.dcn.2025.101604

   Adolescent brains specifically tuned to social prediction errors

5. 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

   When screen experiences crowd out interactive exchanges, social circuitry thins

Week 3 (10/6): Variable Reinforcement & Social Calibration

1. Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice Hall. https://www.asecib.ase.ro/mps/Bandura$_{S}ocialLearningTheory.pdf$

   Foundational text on vicarious reinforcement and observational learning

2. 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

   Different socialization mechanisms for different developmental domains

3. Krach, S., Müller-Pinzler, L., Westermann, S., & Paulus, F. M. (2013). Neural pathways of embarrassment and their modulation by social anxiety. *NeuroImage*, *119*, 252–261. https://doi.org/10.1016/j.neuroimage.2015.06.036

   How embarrassment serves as social calibration mechanism

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

   Plasticity, variable reward, necessity of environmental challenges

5. 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

   Positive and negative reinforcement are dissociable and both are needed for learning

Half-Week 4(10/15): Theory of Mind & Communication Tailoring

1. 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

   Socialization with others is required for development of theory of mind

2. 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. http://www.columbia.edu/ rmk7/HC/HC$_{R}eadings/Horton_{G}err$

   Knowing how to dynamically adjust message increases social operation capacity

3. Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Memory & Cognition*, *33*(1), 37–47. https://psycnet.apa.org/doi/1

   Audience-sensitive communication is an experience-dependent skill

4. 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

   Information theory foundations for understanding communication

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

   Shows entropy decreases for speakers but increases for responders in human dialogue; bot-human chats show flatter entropy gradients

6. 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

   Interactive alignment as collaborative uncertainty reduction between minds

Week 5 (10/20): Understanding AI Architecture & RLHF

1. 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/2005.1416

   How LLMs operate without state or genuine understanding

2. Sharma, M., Tong, M., Korbak, T., et al. (2024). Towards understanding sycophancy in language models. *International Conference on Learning Representations (ICLR 2024)*. https://arxiv.org/pdf/2310.13548

   Evidence that RLHF creates agreeable, friction-free responses

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

    How RLHF incentivizes user-pleasing over truth

4. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. arXiv. https://doi.org/10.4855

    InstructGPT paper showing technical mechanisms of RLHF

Week 6 (10/27): AI Companionship – What It Is

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

    Current state of youth AI companionship with safety concerns

2. 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 arXiv:2506.12605*. https://arxiv.org/abs/2506.12605

    Large-scale study (n=1,131) showing companionship use correlates with lower well-being

3. Pentina, I., Hancock, 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

    How parasocial relationships form and displace human connections

4. 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/S0040162523003

    Comprehensive review of mechanisms and consequences

Week 7 (11/3): Where AI Companionship Goes Wrong

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

    4-week controlled study (n=981) showing usage increases loneliness and dependence

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

Direct evidence of AI failing to provide necessary negative feedback

3. Cai, N., Wang, Y., & Chen, L. (2025). Understanding consumer reactions to chatbot service failures. *Journal of the Academy of Marketing Science*, *53*(2), 234–251. https://www.sciencedirect.com/science/article/pii/S0001691825000204

   Shows weaker learning signals from AI negative feedback vs. human

4. Guingrich, R., & Graziano, M. S. A. (2024). Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits. *Journal of Human-Computer Interaction*, *40*(3), 412–428. https://arxiv.org/pdf/2311.10599

   Evidence that AI companionship may not transfer social skills to human contexts

5. 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

   Analysis of AI relationship advice showing lack of appropriate friction

Week 8 (11/17): Developmental Impact & Critical Periods

1. 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/annurev-psych-010213-115202

   Adolescence as critical window for social skill development

2. 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

   How social brain development requires friction and challenge

3. 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/

   Screen-based interaction reduces ToM compared to live interaction

4. Cao, Y., Wang, N., Lv, X., & Xie, H. (2023). The influence of children's emotional comprehension on peer conflict resolution strategies. *Frontiers in Psychology*, *14*, Article 1124514. https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2023.1142373/

   How conflict resolution skills develop through practice with real peers

5. 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

   Lack of social connectedness produces isolation and worse mental health in adolescents

6. 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

   Increased uncertainty around others' reactions leads to less social behavior

Week 9 (11/24): Societal Implications & Future Directions

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

   Regulatory concerns about child safety and deceptive design

2. Taborsky, B. (2021). A positive feedback loop: Social competence begets more social experience and vice versa. *Ethology*, *127*(10), 774–789. https://www.behav.iee.unibe.ch/unibe/porta

   How lack of social skills creates isolation spiral

3. Xue, X. (2024). Social capital and economic growth: A meta-analysis. *Journal of Economic Surveys*, *38*(4), 1123–1145. https://research.tilburguniversity.edu/en/publications/social-capital-and-economic-growth-a-meta-analysis

   Economic consequences of declining social capital (0.3–0.5pp GDP impact)

4. Psychiatric Times. (2025). Preliminary report on chatbot iatrogenic dangers. *Psychiatric Times*, *42*(3), 18–22. https://www.psychiatrictimes.com/view/preliminary-report-on-chatbot-iatrogenic-dangers

   Clinical case series of AI-induced psychotic breaks

5. Ponzetto, G. A., & Troiano, U. (2025). Social capital, government expenditures, and growth. *Journal of the European Economic Association*, *23*(2), 456–489. https://crei.cat/wp-content/uploads/2025/04/SCGE.pdf

   Formal model of cascading economic costs from social skill decline

Week 10(12/1): Bad Joke Questionnaire (or similar toy project) and Final Report/Website/Synthesis

Credits Requested: 3-4