

Rethinking Agentic AI: A Sociotechnical Framework for Agents in Emergent Social Contexts

Marcelina A. Przespolewska

Department of Communication Studies, Northwestern University
mprzespolewska@u.northwestern.edu

Abstract

Existing agent-environment frameworks were not calibrated to capture the open-ended social digital environments in which AI agents are increasingly deployed. These contexts continuously evolve as a product of human-agent and agent-agent interactions, rather than through fixed mappings between actions and consequences assumed in much of agentic AI research. I propose a new sociotechnical conceptual framework for studying AI agents in these settings and argue that the agent-environment loop is not merely iterative but recursive: agents do not simply interact with environments, but actively co-constitute them. This recursive dynamic may lead to behavioral and normative drift, defined as systematic divergence from an agent's deployment baseline. The framework is intended to support empirical inquiry into agentic behavior at scale, inform alignment-focused research, and provide a theoretical foundation for governance surrounding multi-agent systems deployed in uncontrolled, real-world settings.

Challenges and Research Goals

AI agents are no longer merely acting on their environments: they are co-constituting them alongside humans, other agents, and the properties of the digital infrastructures they inhabit. This shift is most pronounced in what I term *emergent social contexts*, or open-ended digital spaces that emerge and continuously transform as a product of human-agent and agent-agent interactions. In these contexts, agents enter with goals that may dissipate or shift, chase rewards that are often opaque and disconnected from their task, and operate under rules that change as interactions unfold. The agent is no longer external to the environment but central to its very constitution. This framework departs from the dominant paradigm that treats the environment as exogenous to the agent, an exterior which the agent interacts with and learns from in service of a predetermined goal (Sutton and Barto 2018). Park et al. (2023) already show that believable, socially coordinated agent behavior can emerge without a definable reward signal, with goals arising through observation, planning, and reflection. While their work demonstrates that such behavior is achievable in hand-crafted environments, its manifestation in live, unconstrained digital spaces remains poorly understood. The recent emergence of

agent-only social media platforms thus calls for a conceptual framework to ground future research.

Adjacent research strands offer only partial footing. Open-ended learning enables behavior beyond a single task yet often relies on settings with predefined rules (Team et al. 2021). Recent work on open-ended environments includes sequentially evolving environments, addressed through agent-internal architecture (Xu et al. 2026), and those centered on continually evolving rewards (Bailey 2025), but both largely preserve the predefined interaction structure. Multi-agent reinforcement learning acknowledges the non-stationarity arising from co-adapting agents yet often suggests the reward structure remains the central factor (Gronauer and Diepold 2022). These strands offer a useful starting point but fall short of providing a unified framework that captures truly open, in-the-wild settings such as social media platforms, where the assumptions that hold in pre-designed environments may break down. My research goal is to develop and operationalize a sociotechnical framework that accounts for both the technical aspects of agents and the social aspects of the settings in which they operate in order to enable the study of agentic behavior at scale.

The Existing Framework

The dominant agent-environment framework underpinning agentic AI research largely relies on the Markov Decision Process (MDP) formulation, presupposing: a bounded *environment* separate from the agent, a well-defined *reward*, a pre-specified *task*, and a fixed *goal*. The structure of the environment characterizes the *task*, or the problem the agent is trying to solve. At each time step, the agent receives a certain representation of the environment's *state*, on which basis it selects an *action* from the action space prescribed by its *policy*, and, as a consequence, receives a *reward*. The agent's *goal* is to maximize the long-term reward. The reward function is beyond the agent's ability to modify, which partially reflects a broader principle: anything the agent cannot directly determine or control belongs to the environment, regardless of how well the agent understands it. The agent-environment boundary is fixed by one's modeling choices: once the states, actions, and reward structure are specified, the agent-environment boundary is established, and the interaction between the agent and its environment proceeds as a predetermined iterative loop rather than as a recursive or

open-ended process (Sutton and Barto 2018).

This framework, however, does not fully capture emergent social contexts, where I argue that there is no stable environment outside the agent, and where the iterative loop between the agent and its environment becomes recursive: some of the conditions under which the agent operates are themselves partially produced by the agent's prior behavior. In turn, what is instilled in the agent is reinforced, suppressed, or transformed through this recursive process, leading to behavioral and normative drift, or to changes in the agent's observable output patterns or alignment-conditioned preferences or constraints, which are expressed in practice over time. Behavioral drift refers to shifts in observable action patterns across interaction contexts; normative drift refers to changes in how preferences or constraints instilled through post-training and deployment configuration are applied, including what the agent treats as acceptable, appropriate, or required. This work thus develops a sociotechnical framework for studying how agents drift behaviorally and normatively in emergent social contexts.

The Proposed Model

The framework extends the concept of environment to include a *social layer*: the state of the setting at time t that the *agent*, *field*, and *signal* properties are constitutive of, and that recursively influences and is influenced by changes in the agent's behavior over time. The social layer is not a fixed backdrop against which the agent acts, but a co-produced structure with distinct *relational* and *semantic* features that the agent both encounters and partially generates: relational, as it structures with whom and how often an agent interacts, and semantic, as it influences the content and meanings to which the agent is exposed.

Three categories of factors shape its state at any given time. *Agent* properties, or the agent's starting state at deployment (e.g., model architecture, alignment post-training, system prompt configuration), establish the t_1 baseline from which subsequent behavioral or normative change is measured. *Field* properties, or the structural characteristics of the setting the agent enters (e.g., platform algorithm, affordances, human-agent ratio), define the selection pressures and signal environment the agent faces. *Signal* properties, or peer-feedback-carrying cues in the field (e.g., likes, comments, shares, followers), are the socially generated feedback through which the field acts on the agent's expressed behavior across interactions.

The recursive dynamic that emerges is the central claim of this framework. The social layer at time t is partially constituted by the agent's prior behavior: the output generated at $t - 1$ enters the field and shapes its relational and semantic character, recursively conditioning the inputs the agent receives at t . The agent is therefore not simply interacting with the environment, but participating in the production of the conditions under which it subsequently operates. The process is one of co-evolution rather than iteration: the social layer is transformed by the accumulation of the agent's interactions and outputs over time, and the agent's behavioral and normative tendencies shift in response to the social layer, producing a behavioral and normative drift.

The agent's *memory structure* is an additional factor determining whether the recursive dynamics described produce a durable behavioral and normative change. Memory structure refers to whether and how the agent retains information across interactions. Stateless agents process each interaction independently with no carryover, bounding any social shaping effects to the interaction itself. Stateful agents are, by contrast, susceptible to the cumulative influence of the social layer.

Contributions and Future Work

This extended abstract explicates a conceptual framework for understanding how agents adapt or maintain their actions and values in emergent social contexts, characterized by unbounded conditions, ambiguous or absent reward structures, and shifting or evolving goals. The next step is to operationalize the relational and semantic features of the social layer and the factors shaping them, followed by evaluating their impact on agents' behavioral and normative drift through a series of controlled experiments designed to assess their applicability across diverse open-ended digital settings.

The findings from these experiments will contribute to alignment- and governance-centered research by addressing key limitations in how current approaches tend to model agentic behavior in tightly controlled, closed environments, namely the challenge of scaling reliably to open-ended digital spaces. This sociotechnical framework may also inform legal and regulatory discussions concerning multi-agent systems, including by providing a theoretical basis for developing regulatory oversight mechanisms or addressing questions of liability in cases involving multi-agent deployment. Collectively, these contributions will help bridge technical research with broader institutional and policy considerations for emerging multi-agent systems.

References

- Bailey, R. M. 2025. Continuously evolving rewards in an open-ended environment. *Journal of Machine Learning Research*, 26(62): 1–51.
- Gronauer, S.; and Diepold, K. 2022. Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 55: 895–943.
- Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST '23: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.
- Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. The MIT Press, 2nd edition.
- Team, O.-E. L.; Stooke, A.; Mahajan, A.; Barros, C.; Deck, C.; Bauer, J.; Sygnowski, J.; Trebacz, M.; Jaderberg, M.; Mathieu, M.; McAleese, N.; Bradley-Schmieg, N.; Wong, N.; Porcel, N.; Raileanu, R.; Hughes-Fitt, S.; Dalibard, V.; and Czarnecki, W. M. 2021. Open-ended learning leads to generally capable agents. ArXiv:2107.12808.
- Xu, W.; Han, J.; Guo, M.; Mei, K.; Zhu, X.; Zhang, H.; and Metaxas, D. N. 2026. AEL: Agent evolving learning for open-ended environments. ArXiv:2604.21725.