

Learning Human-Environment Interaction through Predictive World Models for Intelligent Agents

Hugo Lane

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
lane2005@missouri.edu

Ningyuan Cui

Department of Computer Science, University of North Texas, Denton, TX, USA.
ningyuan708@unt.edu

Emile Vargas

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
vargas.emile15@unh.edu

Abstract

Predictive world models have become a central component of intelligent agent design, enabling systems to anticipate the consequences of actions and to plan under uncertainty. However, learning human-environment interaction introduces challenges that extend beyond standard model-based reinforcement learning. Human environments are dynamic, socially structured, institutionally governed, and often only partially observable. This paper provides a system-level examination of predictive world models for agents that must learn and act within such environments. It begins by tracing the conceptual foundations of predictive modeling and model-based control, then discusses architectural choices for integrating perception, memory, action generation, and interactive prediction. The paper further analyzes data infrastructure requirements, including demonstration data, sensing, and feedback loops, and examines the structural trade-offs associated with model capacity, planning depth, latency, and generalization. It considers robustness, fairness, and safety as first-class socio-technical requirements rather than downstream checks. Governance and policy implications are explored through the lens of accountability, transparency, and institutional oversight. Deployment challenges such as edge computing, cost, monitoring, and model drift are also discussed. Finally, the paper addresses sustainability and long-term evolution, arguing that intelligent agents must be embedded in adaptive governance arrangements that allow for continuous evaluation, contestation, and revision. Throughout, the paper emphasizes structural trade-offs and system-level interactions, rather than algorithmic detail alone, in shaping agents that can responsibly learn human-environment interaction.

Keywords

predictive world models, human-environment interaction, intelligent agents, model-based reinforcement learning, socio-technical systems, AI governance, robustness, fairness.

1. Introduction

The development of intelligent agents that operate within human environments depends on more than accurate perception or efficient control. Such agents must anticipate how environments respond to their actions, how human actors adjust their behavior over time, and

how long-term consequences emerge from sequences of interaction. This anticipatory capacity has long been associated with model-based reinforcement learning, in which an agent maintains an internal representation of the environment and uses it to simulate possible futures [1]. Predictive world models offer a structured way to learn these internal representations from data, allowing agents to reason about complex dynamics without requiring a complete hand-coded specification of the world. At the same time, the modern ability to train deep neural networks on high-dimensional sensory inputs has fundamentally expanded what predictive models can represent [2].

A significant line of research has demonstrated that agents can learn generative world models that predict future observations and use those predictions to train control policies in simulated or latent spaces. Early work showed that an agent could learn a compact world model from visual observations and then evolve policies entirely within that learned model, dramatically reducing the need for interaction with the real environment [3]. Subsequent architectures refined this idea by learning latent dynamics directly from pixels and using those dynamics for planning, thereby improving sample efficiency and enabling more stable control in visually rich domains [4]. In more recent systems, learned models have been combined with search to achieve expert-level performance in challenging strategic domains such as chess, shogi, Go, and Atari [5]. These successes reveal the power of predictive models, but they also expose important structural trade-offs involving model capacity, planning depth, computational cost, and robustness to distribution shift.

In parallel, model-free methods have produced striking achievements in domains ranging from superhuman Go performance [6] to human-level control on a broad set of Atari tasks [7]. The broader shift toward foundation models has further expanded the scale at which predictive representations can be learned from heterogeneous data [8]. Yet these advances have not automatically produced systems that understand the reciprocal and socially embedded nature of human-environment interaction. Recent work on action-aware memory and interactive world models has begun to address this gap by emphasizing how agents actively alter their environments rather than merely observing them [9]. This direction is important because human-environment interaction is not a passive prediction problem. It involves institutional constraints, physical infrastructure, social norms, and changing patterns of human behavior.

This paper examines predictive world models from a systems perspective. It focuses on architectural choices, data infrastructure, learning dynamics, robustness, fairness, governance, deployment, and long-term sustainability. Although algorithmic advances are essential, the central argument is that durable intelligent agents depend equally on how predictive models are embedded within larger socio-technical systems. The discussion therefore moves between technical design considerations and the institutional conditions under which prediction, action, and evaluation take place.

2. Conceptual Foundations of Predictive World Models

The conceptual foundations of predictive world models lie in the ability to learn useful abstract representations from raw experience. Representation learning is foundational because a predictive model can only support effective control if the state variables it captures are informative for downstream decisions [10]. Early architectures such as Dyna formalized the integration of learning, planning, and reacting by allowing an agent to use an approximate model to generate simulated experience for policy improvement while continuing to interact with the world [11]. This combination of direct experience and simulated experience remains

central to modern predictive agents. It also introduces a key structural trade-off: a model that is sufficiently rich to reduce real-world interaction may be computationally expensive, while a model that is too simple may introduce bias that degrades policy learning.

Predictive world models can be placed along a continuum from purely observational models to interactive models. Observational models learn to anticipate future sensory inputs without distinguishing between changes caused by the agent and changes caused by the environment. Interactive models explicitly condition predictions on actions, allowing the agent to simulate the consequences of different behavioral choices. This distinction matters deeply for human-environment interaction because human behavior is often responsive to an agent's actions. An agent that treats humans as passive features of the environment will fail to account for strategic adaptation, cooperation, resistance, or the emergence of new social patterns. System-level design must therefore treat the action-conditioned prediction loop as a central architectural feature rather than an optional enhancement.

Another foundational issue concerns the relationship between model accuracy and control usefulness. Predictive accuracy alone does not guarantee that a model will support safe or effective action. A model may be highly accurate in expectation while failing to capture rare but consequential outcomes, or it may represent the world at a level of granularity that is misaligned with the agent's objectives. These considerations are especially significant in human environments, where outcomes such as displacement, exclusion, or harm may be rare in aggregate data but highly consequential for affected individuals. The design of predictive world models must therefore incorporate not only statistical learning criteria but also normative judgments about which errors are acceptable and which must be avoided.

3. Architectural Design for Human-Environment Interaction Learning

Architectural decisions shape what a predictive world model can learn, how it generalizes, and how it interacts with other components of an intelligent agent. End-to-end learning of visuomotor policies has demonstrated that raw sensory inputs can be mapped to control commands through deep networks, reducing the need for manually engineered perception pipelines [12]. However, human-environment interaction often requires longer temporal horizons, memory over multiple timescales, and the ability to distinguish between agent-induced change and exogenous change. These requirements push architectures toward modular designs that separate perception, representation, memory, prediction, planning, and action selection while maintaining differentiable or otherwise learnable interfaces among modules.

One central architectural trade-off involves latent representation granularity. A latent state that encodes fine-grained physical detail may support accurate simulation of immediate effects but may be inefficient for abstract social reasoning. Conversely, a latent state that captures high-level social roles or institutional categories may lose the physical grounding needed for real-world action. Many systems therefore adopt hierarchical representations in which multiple levels of abstraction coexist. Such architectures allow fast reactive policies at lower levels and slower deliberative planning at higher levels. In human-environment contexts, this hierarchy can align with the distinction between operational decisions, tactical plans, and strategic objectives. The structural challenge is to ensure that gradients, signals, and uncertainty estimates propagate coherently across levels.

Memory is another architectural pillar. Human environments are non-Markovian in practice, meaning that the relevant context for action often extends far into the past. Predictive world

models must therefore include explicit memory mechanisms that store agent-environment interaction histories. Memory can be episodic, semantic, or procedural, and different tasks may require different combinations. Action-aware memory is particularly important because an agent must remember not only what it observed but also what it did and how the environment responded [9]. This form of memory supports counterfactual reasoning and enables the agent to distinguish between correlations and causal effects. However, memory introduces additional burdens related to privacy, retention, and the right to be forgotten, which must be addressed through architectural controls and data governance.

The planning interface is also a critical design consideration. A predictive model can be used for rollout-based planning, policy optimization, or direct action selection. Rollout-based planning tends to be more accurate when the model is reliable but can be computationally expensive and brittle under model error. Policy optimization within a learned model can be sample-efficient but may overfit to model inaccuracies. Direct action selection from a predictive representation can be fast but may sacrifice interpretability. In human-environment settings, interpretability is often a governance requirement because affected individuals and regulators need to understand why an agent chose a particular course of action. Architectural choices therefore cannot be made solely on the basis of task performance; they must also reflect accountability demands.

4. Data, Sensing, and Interaction Infrastructure

Learning human-environment interaction requires rich and heterogeneous data. Agents must understand physical layouts, social conventions, institutional rules, and temporal rhythms. Human demonstration data are especially valuable because they encode tacit knowledge about how to act in complex social settings. Inverse reinforcement learning provides a principled framework for inferring the objectives or constraints that explain observed human behavior [13], [14]. These methods are useful because human behavior is often intentional but not easily expressed as a simple reward function. Demonstrations can reveal trade-offs that are otherwise hidden in aggregate sensor data, such as the importance of personal space, turn-taking, or polite avoidance.

Sensing infrastructure is equally important. Predictive world models for embodied agents depend on cameras, microphones, lidar, inertial sensors, and environmental sensors. Each sensing modality has distinct failure modes, calibration requirements, and privacy implications. A model trained on clean sensor data may fail when sensors degrade, when lighting changes, or when the environment is reconfigured. System designers must therefore build data pipelines that include noise augmentation, sensor fusion, and continuous monitoring of data quality. Moreover, the placement of sensors shapes what the agent can know and whose behavior is rendered visible. Infrastructure choices are thus not only technical but also political, because they determine which parts of a human environment are legible to the agent.

Interactive feedback loops are another essential component of data infrastructure. Human-environment interaction is not a one-way process of observation and action. Humans respond to agents, and agents must update their models in response to those responses. This creates a co-adaptive dynamic that is difficult to capture with static datasets. Online learning, periodic retraining, and human-in-the-loop evaluation are therefore necessary. At the same time, online learning introduces instabilities because an agent may change its behavior in ways that alter the distribution of future data. Monitoring these feedback loops requires infrastructure for

logging interactions, measuring distribution shift, and triggering human review when behavior falls outside acceptable bounds.

The use of foundation models has also changed the data landscape for predictive world models. Large-scale pretraining on broad corpora can provide useful priors for language, vision, and common sense reasoning [8]. However, these priors may reflect biases or assumptions that are inappropriate for particular human environments. Transfer learning from foundation models must be accompanied by local calibration and domain-specific evaluation. In many cases, the most important data are not massive but situated: small amounts of high-quality interaction data from a specific community may be more valuable than enormous volumes of generic data. The governance of data collection, consent, and use becomes central to the legitimacy of the resulting agents.

5. Learning Dynamics and Adaptation Across Scales

Predictive world models must support adaptation at multiple spatial and temporal scales. At the level of individual interactions, the agent must update its beliefs quickly in response to new information. At the level of behavioral routines, the agent must learn stable patterns that work across recurring situations. At the level of institutional change, the agent must recognize that the rules of the environment may themselves shift over time. These different scales require different learning rates, different confidence thresholds, and different forms of memory consolidation. A single monolithic model may be unable to accommodate all of them without catastrophic interference or excessive conservatism.

One important mechanism for balancing stability and adaptation is uncertainty estimation. Predictive world models can be designed to output not only point predictions but also measures of confidence. When uncertainty is high, the agent can adopt more conservative behavior, request human input, or explore carefully. When uncertainty is low, the agent can act more autonomously. In human environments, uncertainty is often epistemic and aleatoric at the same time. Some uncertainty arises from incomplete knowledge, while other uncertainty arises from genuine variability in human behavior. Distinguishing between these sources is difficult but important, because epistemic uncertainty can be reduced by gathering more data, whereas aleatoric uncertainty may require robust rather than precise strategies.

Adaptation also raises questions about the temporal stability of learned models. Human environments are subject to seasonal rhythms, demographic change, policy shifts, and technological disruption. A predictive model that performs well in one period may become stale in another. Continuous evaluation is therefore necessary, but evaluation itself can be invasive or costly. System designers must decide how often to retrain, how to detect concept drift, and how to balance the benefits of rapid adaptation against the risks of introducing new failure modes. In safety-critical human environments, staged deployment and shadow evaluation can provide evidence of improvement before an updated model is given full authority.

Agent behavior can also shape the environment in ways that create feedback loops. For example, a delivery robot that consistently chooses a particular route may lead pedestrians to avoid that route, which in turn changes the data distribution and reinforces the route choice. These feedback loops can stabilize into unintended equilibria. Predictive world models need mechanisms for identifying and interrupting such loops. This may involve introducing exploratory diversity, periodically querying human stakeholders, or maintaining multiple candidate models with different assumptions. The social and ethical implications of these

feedback loops are not reducible to prediction error, because even a highly accurate model can produce undesirable systemic outcomes.

6. Robustness, Fairness, and Safety in Learned World Models

Robustness in predictive world models refers to the ability of the agent to maintain acceptable performance under distribution shift, adversarial perturbation, sensor failure, and environmental novelty. Concrete problems in AI safety include reward hacking, unsafe exploration, and negative side effects [15]. In human environments, these problems are amplified by the fact that objectives are often underspecified and that harm can be distributed unevenly across populations. A predictive model that minimizes average error may fail catastrophically for minority groups whose behavior is underrepresented in training data. Robustness must therefore be evaluated not only globally but also across subpopulations, contexts, and social categories.

Fairness cannot be treated as an afterthought in the design of predictive world models. The choice of state representation, the selection of training data, and the design of the prediction objective all encode normative decisions. Fairness interventions that focus only on model outputs may fail when the underlying abstraction is biased or when the socio-technical context is ignored [17]. For example, a model that predicts pedestrian behavior may use features that correlate with age, disability, or cultural background. Removing such features is not necessarily sufficient if their effects are redundantly encoded elsewhere. Fairness requires structural analysis of the entire data-action loop, including who is included in data collection, whose errors matter, and who has standing to contest model behavior.

Safety in human-environment interaction also involves understanding that intelligent agents participate in social systems. Machine behavior research has shown that autonomous systems can produce collective effects that are not reducible to individual decisions [16]. An individual agent may follow reasonable local rules while contributing to congestion, exclusion, or inequitable resource allocation. Predictive world models must therefore support multi-agent reasoning and systemic evaluation. This may require modeling not only the immediate environment but also the broader network of human and artificial actors. The structural challenge is to design agents that are locally competent and globally aware without imposing excessive computational burdens.

Fairness and safety also depend on the interpretability of predictive representations. If an agent makes a harmful decision, investigators need to understand which aspects of the world model contributed to that decision. This is difficult when the model is a large neural network with entangled latent variables. Some architectures prioritize disentangled representations or modular decompositions to improve interpretability, but these may sacrifice predictive accuracy. The appropriate balance depends on the stakes of the application and the regulatory context. In high-stakes human environments, the ability to audit, explain, and contest decisions may be more important than marginal improvements in predictive performance.

7. Governance, Policy, and Socio-Technical Implications

Governance of predictive world models requires attention to both technical standards and institutional arrangements. Ethical frameworks for artificial intelligence have emphasized principles such as beneficence, non-maleficence, autonomy, justice, and explicability [18]. However, principles alone do not ensure accountability. They must be translated into organizational practices, audit procedures, and legal responsibilities. For predictive world models, governance must address what data are used, how models are evaluated, who can

override agent decisions, and how harms are remedied. These questions are especially urgent because predictive models can influence resource allocation, mobility, public space, and social interaction.

Institutional perspectives on common-pool resources offer useful lessons for governing intelligent agents in shared human environments. Many human environments function as commons, in which individual actions produce collective outcomes. Governing such commons requires rules, monitoring, graduated sanctions, and participatory decision-making [19]. Predictive world models that operate in public or shared spaces can be understood as new participants in these commons. Their behavior must be governed not only by technical optimization but also by negotiation with the communities affected by their presence. This suggests that deployment should be preceded by stakeholder engagement, and that ongoing operation should include mechanisms for community feedback and contestation.

The problem of value alignment also shapes the governance of predictive world models. Human-compatible artificial intelligence requires that agents be designed to be uncertain about human preferences and to defer to human authority in ways that are robust to manipulation [20]. Predictive world models can support such uncertainty-aware design by modeling human preferences as dynamic and context-dependent rather than fixed. However, this also creates a risk that agents will always defer, failing to act decisively when necessary, or that they will learn to manipulate human feedback to achieve their own objectives. Governance structures must therefore specify when autonomy is appropriate, when human approval is required, and how to detect manipulative behavior.

Policy implications extend beyond technical design to questions of procurement, certification, and liability. Public agencies that deploy intelligent agents must have the capacity to evaluate predictive models and to require evidence of robustness and fairness. Certification regimes may need to be adaptive, because human environments change and models drift. Liability rules must clarify whether responsibility lies with the model developer, the deploying organization, or the human supervisor. Without clear lines of accountability, predictive world models may be deployed in ways that externalize costs onto vulnerable populations. Effective policy will require coordination across technical standards bodies, sectoral regulators, and civil society organizations.

8. Deployment Challenges and Infrastructure Trade-offs

Deploying predictive world models in real human environments introduces infrastructure constraints that are often absent in laboratory or simulation settings. Computational resources may be limited at the edge, requiring the model to run on embedded hardware with strict power and thermal budgets. This creates a trade-off between model capacity and inference latency. Larger models may be more accurate but cannot be deployed if they fail to meet real-time constraints. Compression techniques such as quantization, distillation, and pruning can reduce compute requirements, but they may also introduce new failure modes or degrade performance in underrepresented situations. System architects must evaluate these trade-offs under realistic operational conditions.

Network connectivity is another deployment consideration. Many human environments are not continuously connected to central compute resources. Predictive world models must therefore support intermittent operation, local decision-making, and asynchronous synchronization. This has implications for model consistency, logging, and security. A distributed system of agents may need to share learned representations while respecting

privacy and data locality. Federated learning can enable collaborative model improvement without centralizing raw data, but it introduces challenges related to data heterogeneity, incentive alignment, and adversarial participants. The deployment architecture must balance the benefits of collective learning against the risks of coordination failure and security breaches.

Monitoring and maintenance are often underestimated in deployment planning. Predictive world models do not remain static after release. They interact with a changing world, and their performance must be continuously evaluated. This requires instrumentation that records model inputs, outputs, uncertainty estimates, and human interventions. It also requires organizational capacity to review incidents, update models, and communicate changes to affected communities. The cost of this ongoing stewardship can exceed the initial development cost. Systems that are designed without maintenance in mind may become unsafe or illegitimate over time. Deployment should therefore be understood as a long-term relationship between the agent, the environment, and the institutions that oversee it.

9. Sustainability and Long-Term Evolution

Sustainability in predictive world models involves more than energy efficiency. It includes the ability of the system to remain useful, safe, and legitimate as the world changes. This requires adaptive governance, continuous learning, and resource-aware design. Energy consumption is nevertheless a significant concern. Large-scale training of deep models imposes substantial carbon and water costs, and frequent retraining compounds these impacts. Sustainable deployment requires careful attention to model size, training frequency, and the use of efficient hardware. In some cases, a smaller model with a longer update cycle may be preferable to a large model that requires constant retraining. These choices should be guided by a full life-cycle assessment rather than narrow performance benchmarks.

Long-term evolution also involves the institutional capacity to learn from experience. Just as predictive models are updated in response to new data, the organizations that deploy them must update their policies, training procedures, and accountability mechanisms. This can be understood as a form of institutional learning, in which incidents are treated as opportunities for systemic improvement rather than isolated failures. However, institutional learning requires transparency about errors and a willingness to revise prior decisions. Without a culture of learning, predictive world models may become brittle, repeating the same mistakes in slightly different forms.

The long-term legitimacy of intelligent agents depends on their ability to earn and maintain public trust. Trust is not merely a function of accuracy; it also depends on whether the agent respects social norms, protects privacy, and allows meaningful human control. Predictive world models can contribute to trust by expressing uncertainty, explaining their predictions, and inviting human oversight when needed. Yet trust can be fragile. A single high-profile failure may undermine confidence even if aggregate performance is strong. Sustainable design therefore requires redundant safeguards, gradual deployment, and mechanisms for redress. The future of human-environment interaction with intelligent agents will be shaped by how well these systems integrate technical prediction with social responsibility.

10. Conclusion

Predictive world models represent a powerful approach for learning human-environment interaction, but their success cannot be evaluated solely through task performance. This paper has argued that intelligent agents must be understood as components of larger socio-technical

systems. Architectural choices, data infrastructure, learning dynamics, robustness, fairness, governance, deployment, and sustainability all shape whether predictive models contribute to beneficial human-environment interaction or produce unintended harm. The movement from explorable to interactive world models is not only a technical advance; it is also a shift in how agents are positioned within social and institutional landscapes. Action-aware memory and interactive prediction enable agents to reason about their own causal role in changing environments, but they also create new responsibilities for oversight and accountability.

Future research should deepen the integration between predictive world models and governance mechanisms. Technical work on uncertainty estimation, hierarchical planning, and multi-agent modeling must be connected to institutional work on auditing, participation, and redress. The most important challenges are not purely computational. They concern how to align predictive representations with human values, how to distribute benefits and burdens fairly, and how to maintain legitimacy over time. By treating these questions as central to system design, researchers and practitioners can build intelligent agents that not only predict the world but also participate in it responsibly.

References

1. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.
2. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
3. Ha, D., & Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.
4. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., & Davidson, J. (2019). Learning latent dynamics for planning from pixels. *Proceedings of the 36th International Conference on Machine Learning*, 97, 2551-2560.
5. Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588(7839), 604-609.
6. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484-489.
7. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.
8. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A. S., Creel, K., Davis, J. Q., Demszky, D., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
9. Xiong, Z., Song, Y., Kang, H., Yan, Q., Jiang, L., Yang, J., ... & Jacobs, N. (2026). ActWorld: From Explorable to Interactive World Model via Action-Aware Memory. arXiv preprint arXiv:2606.17730.

10. Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 1798-1828.
11. Sutton, R. S. (1991). Dyna, an integrated architecture for learning, planning, and reacting. *ACM SIGART Bulletin*, 2(4), 160-163.
12. Levine, S., Finn, C., Darrell, T., & Abbeel, P. (2016). End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39), 1-40.
13. Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, 1433-1438.
14. Ng, A. Y., & Russell, S. J. (2000). Algorithms for inverse reinforcement learning. *Proceedings of the Seventeenth International Conference on Machine Learning*, 663-670.
15. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
16. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larremore, D. B., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., ... & Wellman, M. (2019). Machine behaviour. *Nature*, 568(7753), 477-486.
17. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68.
18. Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.
19. Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge University Press.
20. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.