

Memory-Augmented Vision-Language Models for Context-Aware Autonomous Driving in Dynamic Environments

Vidar Winberg

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.

vidar406@uab.edu

Abstract

Autonomous driving in dynamic environments remains constrained by the gap between momentary perception and durable contextual understanding. Vision-language models have expanded the capacity of vehicles to interpret scenes through natural-language concepts, yet most deployed pipelines still treat perception, prediction, and planning as loosely coupled modules operating on short temporal windows. This paper examines memory-augmented vision-language models as a system-level architecture for context-aware autonomous driving. It argues that memory is not merely an additional neural component but a socio-technical infrastructure that mediates continuity, personalization, infrastructure coordination, and regulatory accountability. The discussion develops an architectural framework in which multimodal encoders, episodic and semantic memory stores, retrieval mechanisms, world-model simulation, and safety monitors interact under latency, energy, privacy, and governance constraints. It analyzes robustness challenges in dynamic environments, including occlusion, adversarial conditions, long-tail events, and distribution shift, and explains how stale or overconfident memory can create new failure modes. The paper then considers deployment realities such as edge-cloud partitioning, fleet learning, communication dependency, and lifecycle sustainability. It further addresses fairness, interpretability, privacy, liability, and policy implications when vehicle memory becomes a durable record of places, people, and behaviors. The conclusion is that memory-augmented vision-language driving systems can improve context awareness, but only if they are designed as accountable infrastructures with bounded autonomy, transparent update practices, and explicit mechanisms for forgetting, auditing, and redress.

Keywords

memory-augmented models; vision-language models; autonomous driving; dynamic environments; context awareness; socio-technical systems; safety; governance.

1. Introduction

Autonomous driving has advanced through increasingly capable perception stacks, high-definition mapping, simulation-based training, and end-to-end learned policies [1], [2], [3], [4]. Nevertheless, dynamic environments expose a persistent structural weakness: a vehicle may recognize an object in one frame but fail to understand the evolving social and spatial context that gives that object meaning. A pedestrian waiting at a curb, a delivery vehicle double-parked near a school, or a construction worker directing traffic cannot be interpreted adequately through instantaneous object detection alone. Context is temporal, relational, and often institutionally encoded. Vision-language models offer one route toward richer

interpretation because they connect visual evidence to linguistic categories, instructions, and commonsense relations [6], [7], [8], [9], [10]. Yet language grounding without persistent memory remains fragile. The system can describe a scene but may not remember how that scene changed, why a prior maneuver was chosen, or which local norms governed earlier interactions.

Memory-augmented vision-language models address this gap by allowing driving systems to store, retrieve, and update representations across time and trips. Memory can include episodic traces of recent events, semantic summaries of recurring locations, social expectations associated with particular roads, and procedural knowledge about vehicle behavior. Such memory may support continuity in planning, personalized assistance, cooperative behavior with other road users, and more efficient learning from fleet experience [11], [12], [13], [14]. It may also enable world-model reasoning in which the vehicle simulates possible futures conditioned on past actions and observed outcomes [15], [16]. The promise is substantial: a vehicle could remember that a particular intersection has poor sightlines at dusk, that a school zone becomes congested at specific hours, or that a certain lane is routinely blocked by delivery vehicles. These memories could improve anticipation and reduce reactive braking.

The central argument of this paper is that memory augmentation should be treated as an infrastructural and governance problem, not only as a model architecture problem. A memory-augmented driving system creates durable representations of places, persons, and events. It depends on connectivity, data pipelines, update mechanisms, storage limits, and organizational incentives. It also distributes responsibility across manufacturers, fleet operators, map providers, regulators, municipalities, and users. If memory is opaque or unaccountable, it can entrench bias, violate privacy expectations, or produce cascading errors when stale information is treated as current truth. If memory is designed as a governed infrastructure, it can support robustness, fairness, sustainability, and public trust.

This paper therefore develops a system-level analysis of memory-augmented vision-language models for context-aware autonomous driving. It first reviews foundational work on vision-language representation and memory architectures. It then proposes an architectural framework that integrates multimodal perception with episodic, semantic, and procedural memory under safety constraints. Subsequent sections examine dynamic-environment robustness, deployment infrastructure and sustainability, governance and fairness, and evaluation and operational safety. The conclusion emphasizes bounded autonomy, transparent memory lifecycles, and policy mechanisms that make context awareness compatible with public accountability.

2. Foundations of Vision-Language Models and Memory in Driving Systems

Vision-language models learn joint representations between images, video, text, and sometimes action. Early contrastive approaches demonstrated that large-scale image-text supervision can produce transferable visual representations [6], while subsequent work improved multimodal alignment through noisy text supervision and unified understanding-generation objectives [7], [8]. Multimodal transformers and few-shot visual-language systems then extended this capacity toward instruction following, captioning, visual question answering, and open-vocabulary recognition [9], [10]. For autonomous driving, these capabilities are attractive because they allow a vehicle to interpret rare signs, unusual road configurations, and informal human signals that are difficult to enumerate in fixed taxonomies. A natural-language interface can also make system reasoning more inspectable to engineers,

regulators, and users, although inspectability should not be confused with full interpretability [19].

Memory architectures provide a complementary foundation. External memory networks and differentiable memory systems allow neural models to store and retrieve information beyond the immediate input window [11], [12]. Reinforcement learning agents have used recurrent and predictive memory to integrate observations across time [13], [14]. In driving, memory can be understood across several layers. Short-term episodic memory preserves the last few seconds or minutes of sensor evidence and maneuver history. Long-term semantic memory stores abstracted knowledge about locations, traffic patterns, and recurring social situations. Procedural memory encodes learned policies and safety rules. Working memory coordinates current goals, passenger preferences, and operational constraints. These layers differ in update frequency, retention policy, privacy sensitivity, and verification requirements.

World models introduce a further layer by allowing the system to predict future states and evaluate possible actions internally. Action-aware memory world models extend this line by linking memory to intervention, exploration, and counterfactual reasoning [16]. In driving, such models could support planning under uncertainty by simulating how other agents might respond to signaling, yielding, or lane changes. However, world-model predictions are not neutral. They depend on training distributions, sensor assumptions, and the remembered regularities of past environments. When memory is used to condition simulation, errors can become self-reinforcing: the system may remember a simplified version of a place and then simulate futures that ignore rare but consequential possibilities. Foundation-model research similarly warns that scale and generality do not eliminate brittleness, bias, or opaque failure modes [17].

The integration of vision-language models with memory also raises questions about grounding and truth maintenance. Language can help label a scene, but it can also produce plausible descriptions that are not causally relevant to driving. Memory can provide continuity, but it can preserve outdated facts. A robust system must therefore distinguish between memory as evidence, memory as prior, and memory as instruction. Evidence should be traceable to sensor observations. Priors should be probabilistic and revisable. Instructions should be validated against safety rules and current conditions. Without these distinctions, a vehicle may act on a remembered narrative rather than the present environment.

3. Architectural Framework for Memory-Augmented Context Awareness

A system-level architecture for memory-augmented vision-language driving can be organized around several interacting planes. The perception plane includes cameras, radar, lidar, microphones, and vehicle telemetry. The multimodal representation plane encodes sensor data and language into shared embeddings. The memory plane maintains episodic, semantic, and procedural stores. The reasoning and planning plane retrieves relevant memories, fuses them with current observations, and produces candidate trajectories. The safety and governance plane monitors constraints, audits memory use, and enforces policies for retention, access, and forgetting. These planes are not strictly sequential. They operate concurrently under real-time constraints and must degrade gracefully when components fail.

The memory plane is the most distinctive element. Episodic memory might store recent object tracks, interaction histories, and maneuver outcomes. Semantic memory might store map-linked summaries, local norms, and recurring hazards. Procedural memory might store validated driving policies, regulatory rules, and manufacturer constraints. A retrieval

mechanism selects memories based on spatial proximity, temporal relevance, semantic similarity, and uncertainty. Retrieval must be selective because storing and retrieving everything is computationally unsustainable and creates privacy risks. The system needs salience models that decide what is worth remembering, how long it remains useful, and when it should be demoted or deleted. These decisions are normative, not merely technical, because they determine which places and people become part of the vehicle's durable worldview.

Context fusion then combines retrieved memory with current vision-language representations. A vehicle approaching a school may retrieve semantic memory about pedestrian density, episodic memory about a child who ran into the street earlier that week, and procedural memory about reduced speed limits. The planner uses these signals to adjust trajectory, but the safety monitor must prevent memory from overriding direct evidence. If current sensors show an empty road, a remembered hazard should raise caution but not compel emergency braking. If current sensors are degraded by fog, memory may provide useful priors but must be marked as low-confidence. This suggests a layered trust model in which fresh sensor evidence, map data, vehicle-to-everything messages, and memory have different reliability weights depending on conditions [1], [15].

Trade-offs are unavoidable. Richer memory can improve anticipation but increases latency, energy consumption, and storage cost. Cloud-based memory can aggregate fleet knowledge but depends on connectivity and raises surveillance concerns. On-vehicle memory reduces communication risk but limits shared learning. Frequent updates improve freshness but can destabilize validated behavior. Strict forgetting protects privacy but may discard useful safety knowledge. Centralized governance can enforce consistency but may suppress local adaptation. Federated or edge-cloud architectures can balance these tensions, yet they introduce coordination complexity and new attack surfaces. The design challenge is therefore not to maximize memory but to allocate it across time, place, and authority.

The architecture also needs interfaces for human and institutional oversight. Fleet operators need dashboards showing memory retrieval events relevant to disengagements or incidents. Regulators need auditable logs of memory updates that affect safety-critical behavior. Users need understandable explanations of personalized settings and data retention. Municipalities need channels to correct erroneous semantic memories about construction zones, school hours, or traffic rules. These interfaces transform memory from a hidden model state into a governed record. They also support accountability when a vehicle's behavior is shaped by information that is outdated, biased, or unauthorized.

4. Dynamic Environments and Robustness

Dynamic environments challenge autonomous driving through occlusion, weather, lighting variation, unpredictable agents, and long-tail events. Memory can improve robustness by maintaining object permanence, anticipating recurring patterns, and smoothing perception across temporary failures. A vehicle that briefly loses sight of a cyclist behind a bus can use episodic memory to maintain a probability distribution over the cyclist's possible location. A vehicle approaching a construction zone can use semantic memory to expect temporary lane shifts. Such continuity is valuable because safe driving often depends on what is likely to happen next, not only on what is visible now. Simulation platforms and benchmarks have made progress in testing these scenarios, but they cannot exhaust the space of real-world dynamics [3], [4].

Yet memory also creates robustness risks. Stale memory can cause the vehicle to expect a pattern that no longer holds. If a road has been reconfigured, a remembered lane geometry may be dangerously wrong. If a pedestrian routinely crosses at a certain time, the system may overpredict that behavior and underreact to unusual movement. Memory can also be contaminated by rare but salient events, leading to overcaution in harmless contexts. More seriously, adversarial actors might exploit memory by repeatedly presenting misleading scenes that train the system to expect a benign pattern before staging a hazardous event. Robustness therefore requires mechanisms for memory validation, decay, conflict detection, and uncertainty propagation [18].

Vision-language grounding introduces additional failure modes. Language categories can be ambiguous, culturally specific, or context-dependent. A gesture meaning yield in one setting may mean something else elsewhere. A temporary sign may be misread as permanent. A vehicle may generate a plausible caption that omits the most safety-relevant detail. Memory can amplify these errors by storing mistaken interpretations and retrieving them later. To mitigate this, the system should separate perceptual evidence from linguistic interpretation and keep multiple hypotheses when ambiguity is high. Interpretability research offers methods for exposing model reasoning, but operational driving requires actionable diagnostics rather than post hoc explanations alone [19].

Fairness and robustness intersect when memory encodes social bias. If a vehicle's memory associates certain neighborhoods with higher risk, it may behave more cautiously or suspiciously in ways that burden residents. If pedestrian detection performs differently across skin tones, clothing, ages, or mobility aids, memory can compound the disparity by storing biased encounter histories. A fair system must evaluate memory not only for average accuracy but for distributional outcomes across locations and populations [20]. It must also allow correction when biased memories are identified. Robustness without fairness can produce systems that are reliable for some communities and hostile or inefficient for others.

5. Deployment Infrastructure and Sustainability

Deploying memory-augmented vision-language driving systems requires substantial infrastructure. Vehicles need onboard compute for real-time inference, storage for recent episodes, and communication links for updates and fleet coordination. Cloud or edge data centers support large-scale training, simulation, and memory aggregation, but they consume energy and create dependencies on network availability. Map providers, road operators, and vehicle manufacturers must exchange data through standards that preserve semantic consistency. Without such standards, memory may be fragmented across proprietary formats and difficult to audit. Infrastructure is thus not a backdrop but a constitutive part of the system's cognitive architecture [21].

Sustainability concerns arise across the lifecycle. Larger models and persistent memory increase energy use during training, inference, and data transmission. Continuous logging and cloud synchronization generate storage and cooling demands. Fleet learning can reduce redundant data collection by sharing abstracted updates, but it can also incentivize mass data extraction. A sustainable design should prioritize on-device abstraction, selective retention, efficient retrieval, and update mechanisms that transmit model improvements rather than raw personal data. It should also consider the end-of-life handling of storage media and the environmental cost of maintaining always-connected vehicles. Sustainability is therefore both an engineering objective and a governance mandate.

Deployment must also address reliability under degraded conditions. Rural areas may lack high-bandwidth connectivity, making cloud-dependent memory unreliable. Older vehicles may have limited compute, creating inequitable access to safety improvements. Municipalities may not have resources to maintain digital infrastructure for cooperative memory. A system that assumes ubiquitous connectivity risks excluding the very environments where robust autonomy could be most valuable. Edge-first architectures, intermittent synchronization, and graceful degradation are therefore essential. The vehicle should remain safe when memory services are unavailable, even if context awareness is reduced.

Fleet learning raises questions about ownership and control. If memory is shared across vehicles, who owns the aggregated knowledge? Who decides which memories are retained or deleted? How are corrections propagated? A manufacturer may treat memory as proprietary capital, while users may regard it as personal data and municipalities may see it as public infrastructure. These competing claims require contractual, regulatory, and technical arrangements. Without them, fleet memory can become a private surveillance asset rather than a shared safety resource [22]. Sustainable deployment depends on aligning incentives so that memory improves public safety without concentrating unaccountable power.

6. Governance, Fairness, and Policy

Governance of memory-augmented driving systems must address privacy, consent, purpose limitation, retention, and redress. Vehicles with persistent memory can record movement patterns, facial features, voice commands, and social interactions. Even if data is anonymized, location traces and behavioral histories can be reidentified. Privacy in context is not simply about secrecy but about appropriate flow of information according to social norms and expectations [23]. A vehicle memory system should therefore implement contextual integrity: it should retain what is necessary for safety, limit access according to role, and avoid repurposing safety data for marketing, insurance, or policing without legitimate authorization.

Liability is complicated when behavior depends on memory. If an incident occurs because a vehicle retrieved an outdated hazard memory, responsibility may be distributed among the manufacturer, fleet operator, data provider, map maintainer, and user. Regulators need clear rules for memory provenance, versioning, and audit trails. They also need authority to require recalls or updates when memory creates systematic risk. Safety cases should document how memory is validated, how conflicts are resolved, and how failures are detected. Formal models of safe driving can provide a foundation, but they must be extended to include learned memory and language-conditioned reasoning [15], [18].

Fairness requires attention to both allocation and representation. Allocation concerns who benefits from memory-enhanced autonomy and who bears its costs. Wealthier regions may receive better mapping, connectivity, and fleet updates, while poorer regions may receive older systems or none at all. Representation concerns whose behaviors, languages, and environments are included in training and memory. If vision-language models are trained predominantly on certain cultural contexts, they may misinterpret gestures, signs, or social norms elsewhere. Policy should require disaggregated performance reporting, community consultation, and mechanisms for correcting biased memory. Fairness is not an add-on but a condition for legitimate deployment [20].

Technology and policy co-produce social order. Artifacts can embody political arrangements by privileging certain users, authorities, or forms of knowledge [24]. A memory-augmented vehicle that treats official map data as authoritative while ignoring local community

corrections may reinforce existing power imbalances. Conversely, a system designed for participatory memory governance could allow residents, drivers, and municipalities to annotate hazards and correct errors. Such participation must be structured to prevent misuse, but it can improve both accuracy and trust. The governance challenge is to make memory a matter of public accountability without turning every vehicle into a site of surveillance capitalism [22]. This requires transparency, contestability, and limits on data extraction.

7. Evaluation, Validation, and Operational Safety

Evaluating memory-augmented vision-language driving systems requires metrics beyond perception accuracy and collision rate. Researchers and regulators need measures of memory precision, retrieval relevance, staleness, update latency, conflict resolution, and forgetting compliance. They need scenario tests that examine whether memory improves or harms behavior under occlusion, weather, construction, and unusual social interaction. Simulation is essential for repeatability and rare-event generation [3], but it must be complemented by closed-course and controlled road testing. Real-world deployment should begin with shadow mode, geofenced operation, and human oversight before expanding autonomy.

Validation must also examine failure interactions. A memory system may perform well when components are healthy but fail dangerously when perception is degraded, communication is lost, or memory is corrupted. Safety monitors should detect inconsistency between current evidence and retrieved memory, degrade to conservative behavior when uncertainty is high, and log the reasons for memory-influenced decisions. Formal safety models can define minimal safe behavior, but they need to be integrated with learned components that are probabilistic and context-dependent [15]. Interpretability tools can help engineers diagnose why a memory was retrieved, but they must be validated for operational usefulness [19].

Operational safety depends on update governance. Over-the-air updates can improve memory but also introduce regressions. A disciplined release process should include staged rollouts, canary fleets, rollback capability, and independent audit. Memory changes that affect safety-critical behavior should be traceable to specific triggers and test evidence. Fleet-wide learning should not automatically propagate unverified memories. Instead, abstractions should be validated against simulation, closed-course tests, and statistical monitoring before deployment. This is especially important when memories encode social norms or location-specific behavior.

Finally, evaluation should include human factors. Drivers and passengers need to understand what the vehicle remembers and why. If memory creates personalized driving styles, users should be able to inspect and adjust them. If memory is used for cooperative communication with pedestrians or other drivers, the system must avoid confusing or manipulative signals. Trust is built through predictability, transparency, and responsiveness, not through claims of intelligence. A memory-augmented vehicle should be evaluated as a participant in a social and institutional environment, not merely as a machine that maps inputs to trajectories.

8. Conclusion

Memory-augmented vision-language models offer a promising path toward context-aware autonomous driving in dynamic environments. By connecting multimodal perception with episodic, semantic, and procedural memory, vehicles can maintain continuity, anticipate recurring hazards, learn from fleet experience, and reason about possible futures. Action-aware world models and interactive memory systems further expand this potential by linking memory to exploration and counterfactual planning [16]. However, the same capacities create

risks of staleness, bias, surveillance, opaque responsibility, and unsustainable infrastructure. The decisive question is not whether memory can be added, but how it should be bounded, governed, updated, and forgotten.

A responsible architecture must treat memory as an accountable infrastructure. It should separate evidence from interpretation, enforce retention limits, support contestability, and degrade safely when memory is unavailable or uncertain. Deployment should be edge-aware, energy-conscious, and equitable across regions. Governance should protect privacy, ensure fairness, clarify liability, and enable public oversight. Evaluation should measure robustness, sustainability, and social impact alongside safety. If these conditions are met, memory-augmented vision-language driving systems can improve context awareness while remaining compatible with democratic accountability and public trust. If they are ignored, memory may become a hidden source of error and power rather than a foundation for safer mobility.

References

1. Badue, C., Guidolini, R., Carneiro, R. V., Azevedo, P., Cardoso, V. B., Forechi, A., Jesus, L., Berriel, R., Paixão, T. M., Mutz, F., de Paula Veronese, L., Oliveira-Santos, T., & De Souza, A. F. (2021). Self-driving cars: A survey. *Expert Systems with Applications*, 165, 113816. <https://doi.org/10.1016/j.eswa.2020.113816>
2. Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3354–3361). IEEE. <https://doi.org/10.1109/CVPR.2012.6248074>
3. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning* (pp. 1–16). PMLR.
4. Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation* (pp. 4693–4700). IEEE. <https://doi.org/10.1109/ICRA.2018.8460487>
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* 30 (pp. 5998–6008). Curran Associates.
6. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 8748–8763). PMLR.
7. Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q. V., Sung, Y.-H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning* (pp. 4904–4916). PMLR.
8. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning* (pp. 12888–12900). PMLR.

9. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems* 35 (pp. 23716–23736). Curran Associates.
10. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33 (pp. 1877–1901). Curran Associates.
11. Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., Colmenarejo, S. G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A. P., Hermann, K. M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., & Hassabis, D. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626), 471–476.
<https://doi.org/10.1038/nature20101>
12. Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). Meta-learning with memory-augmented neural networks. In *Proceedings of the 33rd International Conference on Machine Learning* (pp. 1842–1850). PMLR.
13. 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.
<https://doi.org/10.1038/nature14236>
14. Wayne, G., Hung, C.-C., Amos, D., Mirza, M., Ahuja, A., Grabska-Barwińska, A., Rae, J., Mirowski, P., Leibo, J. Z., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... Lillicrap, T. (2018). Unsupervised predictive memory in a goal-directed agent. *arXiv*.
<https://arxiv.org/abs/1803.10760>
15. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. *arXiv*. <https://arxiv.org/abs/1708.06374>
16. Xiong, Zhexiong, et al. "ActWorld: From Explorable to Interactive World Model via Action-Aware Memory." *arXiv preprint arXiv:2606.17730* (2026).
17. 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., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*.
<https://arxiv.org/abs/2108.07258>
18. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv*. <https://arxiv.org/abs/1606.06565>
19. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. <https://arxiv.org/abs/1702.08608>
20. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *Journal of Machine Learning Research*, 19(1), 1–58.

21. Kitchin, R., & Dodge, M. (2011). *Code/space: Software and everyday life*. MIT Press.
22. Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. PublicAffairs.
23. Nissenbaum, H. (2010). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
24. Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121–136.
25. Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society* (pp. 225–258). MIT Press.