

Multimodal Large Language Agents for Real-World Interactive Applications

Shea Heraghty

Department of Computer Science, George Mason University, Fairfax, VA, USA.
sheah@gmu.edu

Alannah Lappin

Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
contactalannah@colostate.edu

Eshana Arya

School of Computing, Clemson University, Clemson, SC, USA.
eshana1985@clemson.edu

Arnaude Lamy

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

Abstract

Multimodal large language agents represent a convergence of foundation models, interactive perception, tool use, and autonomous control in ways that increasingly shape real-world applications. Unlike conventional language systems that operate primarily over text, these agents integrate visual, auditory, spatial, and procedural signals into continuous decision loops, allowing them to act within physical and digital environments. This paper examines the system-level architecture, deployment infrastructure, governance requirements, and socio-technical implications of such agents. It argues that the central challenge is not merely model scale or modality fusion, but the design of reliable, sustainable, and accountable interactive systems that can operate under uncertainty. Architectural trade-offs include latency versus deliberation, centralization versus edge distribution, closed versus open tool ecosystems, and general-purpose reasoning versus domain-specific constraints. The paper further analyzes robustness, fairness, accessibility, and sustainability as interdependent properties rather than isolated metrics. It discusses how multimodal agents alter institutional workflows, regulatory expectations, and the economics of compute and maintenance. Through cross-domain comparisons spanning robotics, healthcare, education, public services, and industrial automation, the analysis shows that real-world deployment requires layered safeguards, human oversight, and lifecycle governance. The conclusion emphasizes that multimodal large language agents should be understood as infrastructural systems whose legitimacy depends on transparency, contestability, inclusive design, and long-term operational resilience. The paper contributes a systems-oriented agenda for researchers, engineers, and policymakers seeking to move beyond prototype demonstrations toward durable interactive applications.

Keywords

multimodal agents; large language models; interactive systems; system architecture; governance; robustness; fairness; sustainability; deployment infrastructure; socio-technical systems.

1. Introduction

The rapid maturation of large language models has shifted artificial intelligence from isolated prediction tasks toward interactive systems capable of planning, tool use, and situated action [1], [2]. Early language models demonstrated strong few-shot generalization, but their confinement to textual input and output limited their ability to operate in real-world settings. Multimodal systems extended these capabilities by aligning vision, language, audio, and increasingly action representations within shared computational substrates [3], [4]. The resulting agents can interpret scenes, follow natural-language instructions, call external services, and adapt behavior across tasks. However, this expansion also introduces system-level complexity that cannot be resolved by model scaling alone.

Real-world interactive applications differ from benchmark environments in several decisive ways. They involve noisy sensors, incomplete information, heterogeneous software stacks, human collaborators, legal constraints, and long operational horizons. A multimodal large language agent deployed in a hospital, factory, vehicle, classroom, or public-service portal must not only answer correctly but also act safely, explain decisions when required, degrade gracefully, and remain maintainable over time. These requirements transform the design problem from one of isolated model performance into one of infrastructure, governance, and institutional fit.

This paper examines multimodal large language agents as socio-technical systems. It focuses on structural trade-offs in architecture, deployment, robustness, fairness, sustainability, and policy. The paper does not propose a single optimized design. Instead, it maps the design space and identifies the tensions that arise when interactive agents move from demonstration to routine operation. In doing so, it treats agents as assemblies of models, tools, interfaces, data pipelines, human roles, and regulatory arrangements. The central claim is that the viability of multimodal agents depends less on achieving human-like intelligence than on constructing accountable systems that remain useful, safe, and legitimate under real-world constraints.

2. Architectural Foundations and Multimodal Integration

The architecture of a multimodal large language agent typically combines a language model backbone with modality-specific encoders, fusion layers, memory components, tool interfaces, and an action controller [5], [6]. Vision-language models such as those trained through contrastive and generative objectives have shown that aligned representations can support captioning, visual question answering, and instruction following [4]. More recent embodied and multimodal systems integrate sensor streams with language to support planning and control [7]. A recent full-scenario agentic scaling effort further illustrates the ambition to generalize across tasks and environments [8]. Yet architectural integration is not simply additive; each added modality introduces synchronization, calibration, and failure-mode challenges.

The first major trade-off concerns modality fusion. Early fusion, in which modalities are combined at the input or representation level, can support rich cross-modal reasoning but increases training complexity and can make individual modalities harder to debug. Late fusion, in which modality-specific modules produce outputs that are combined at decision time, offers modularity and easier maintenance but may miss subtle cross-modal dependencies. Hierarchical fusion attempts to balance these extremes by allowing early interaction in some layers and later arbitration in others. In real-world systems, the

appropriate choice depends on latency budgets, safety requirements, data availability, and the cost of retraining. A medical triage agent may prioritize late fusion for auditability, while a robotic manipulation agent may require early fusion for tight perception-action coupling.

The second trade-off concerns memory and state. Interactive agents must maintain context across turns, sessions, and possibly physical locations. Short-term context windows support conversational continuity, while long-term memory enables personalization and task persistence. However, persistent memory raises privacy, consent, and security concerns. It also creates technical debt because memory stores must be versioned, migrated, and reconciled with changing policies [19]. A system that remembers too much may become intrusive and vulnerable; a system that remembers too little may frustrate users and repeat errors. The architecture must therefore treat memory as a governed resource rather than an unlimited buffer.

The third trade-off concerns modularity versus end-to-end optimization. End-to-end multimodal models can learn representations that are difficult to hand-engineer, but they are often opaque and expensive to update. Modular architectures composed of perception modules, language planners, tool adapters, and safety filters are easier to inspect and replace, but they can suffer from error propagation and interface mismatch. In practice, hybrid architectures are likely: learned components handle perception and reasoning, while explicit interfaces and controllers enforce safety and operational constraints. The key design question is where to place deterministic boundaries around probabilistic components. This question is as much institutional as technical, because it determines who can audit, override, or certify the system.

3. Agentic Control, Tool Use, and Embodied Interaction

Multimodal agents become truly interactive when they can act. Tool use allows language models to query databases, invoke application programming interfaces, operate software, and coordinate with other agents [9], [10]. Multi-agent frameworks extend this capability by assigning roles, enabling negotiation, and supporting complex workflows [11]. In embodied settings, agents connect language to perception and action through affordances, policies, and code generation [12], [13], [14]. These developments point toward systems that can translate high-level goals into sequences of operations across digital and physical domains.

The control architecture of such agents must balance reactivity and deliberation. Reactive policies respond quickly to sensor input and are essential for safety-critical loops such as collision avoidance or emergency shutdown. Deliberative planners reason over longer horizons and can optimize for complex objectives, but they introduce latency and may rely on uncertain world models. Real-world agents often adopt layered control, where fast low-level reflexes are supervised by slower high-level planners. The challenge is to define clear escalation rules: when does the agent act autonomously, when does it ask for clarification, and when does it hand control to a human? These rules cannot be derived from model accuracy alone; they depend on domain risk, reversibility of actions, and regulatory context.

Tool use also raises questions of trust and provenance. An agent that calls external services may depend on data it cannot verify, interfaces that change without notice, and third-party policies beyond its control. Tool ecosystems therefore require versioning, authentication, rate limiting, and fallback behavior. In enterprise settings, an agent might need to operate across customer relationship management systems, scheduling platforms, and internal knowledge bases. Each integration creates a potential point of failure and a governance obligation. The

more capable the agent, the more important it becomes to log tool invocations, record assumptions, and support post hoc review.

Embodied interaction intensifies these concerns. Robots and embodied agents must respect physical constraints, human proximity, and environmental unpredictability. Language can provide flexible task specification, but it is not a substitute for robust perception and control. Systems that ground language in affordances, executable code, or learned policies can improve generalization, yet they also create new failure modes when language misinterprets the scene or when generated code produces unsafe actions [13], [14]. The design implication is that embodiment demands conservative action selection, extensive simulation, and staged deployment. A multimodal agent that works well in a virtual environment may fail catastrophically when sensor noise, actuator delay, and human unpredictability are introduced.

4. Deployment Infrastructures and Scalability Trade-offs

Deploying multimodal large language agents at scale requires computational infrastructure that differs substantially from single-model serving. Inference may involve multiple encoders, retrieval systems, memory stores, policy modules, and external tools. Latency accumulates across these components, and availability depends on many services. Centralized cloud deployment offers elastic scaling, easier updating, and consolidated monitoring, but it can create bandwidth bottlenecks, privacy risks, and dependency on network connectivity. Edge deployment reduces latency and can improve privacy by keeping data local, but it constrains model size and complicates fleet management. Hybrid architectures attempt to distribute workloads according to sensitivity, urgency, and computational demand.

Scalability is not only a matter of hardware. It also involves data pipelines, evaluation harnesses, deployment tooling, and human support. Multimodal data are expensive to store and process, especially video, audio, and sensor streams. Continuous learning from deployment data can improve agents, but it also risks feedback loops, distributional drift, and privacy violations. Organizations must decide whether to update models frequently, how to validate changes, and how to roll back failures. These decisions resemble the hidden technical debt of machine learning systems, where entanglement, undeclared consumers, and configuration complexity accumulate over time [19]. For multimodal agents, technical debt is compounded by modality-specific dependencies and the difficulty of reproducing failures that involve real-world context.

Interoperability is another deployment challenge. Real-world applications often span legacy systems, cloud services, mobile devices, and operational technology. Agents must communicate through standards that may be incomplete or inconsistently implemented. In healthcare, for example, an agent may need to interact with electronic health records, imaging systems, scheduling tools, and clinical communication platforms. Each interface has its own security model and data semantics. Without careful abstraction, the agent becomes a fragile integration layer that is difficult to certify. Interoperability therefore requires not only technical protocols but also governance agreements about data use, liability, and change management.

The economics of deployment also shape architecture. Large models are expensive to train and serve, and multimodal inputs increase bandwidth and storage costs. Organizations may choose smaller specialized models for routine tasks and reserve large general models for complex reasoning. This tiered approach can improve cost efficiency but adds routing complexity and potential inconsistency. It also raises equity concerns if only well-resourced

institutions can afford high-quality multimodal agents. Sustainable deployment thus depends on designing systems that are not only capable but also economically accessible and operationally maintainable over years.

5. Robustness, Safety, and Governance

Robustness in multimodal agents encompasses perception, reasoning, action, and interaction. Perception may fail under occlusion, poor lighting, domain shift, or adversarial perturbation. Reasoning may fail when the agent overgeneralizes from limited data or confuses correlation with causation. Action may fail when the environment changes or when tool outputs are misinterpreted. Interaction may fail when the agent misreads human intent, fails to ask for clarification, or provides explanations that are fluent but misleading. Safety therefore cannot be treated as a final filter; it must be embedded throughout the architecture.

Governance frameworks for language models have identified risks such as misinformation, bias, privacy violation, and malicious use [16]. Multimodal agents add new dimensions because they can observe and act in the world. A system that can see a room, hear a conversation, and control a device has a broader attack surface than a text-only chatbot. Security must address model weights, training data, inference endpoints, tool credentials, and physical actuators. Safety must address both unintended harm and adversarial manipulation. Governance must define who is accountable when an agent causes harm, how affected parties can seek redress, and what evidence is retained for investigation.

Human oversight is often proposed as a safeguard, but it is not automatically effective. Oversight requires meaningful information, sufficient time, and authority to intervene. If an agent operates at machine speed or presents complex multimodal evidence, human reviewers may experience automation bias or information overload. Effective oversight therefore requires interface design that surfaces uncertainty, highlights conflicting evidence, and supports rapid escalation. It also requires organizational roles such as safety officers, auditors, and operational supervisors. Governance is thus not a document but a set of practices, incentives, and technical affordances.

Regulatory approaches are beginning to address these issues, but they remain uneven across jurisdictions. Some frameworks emphasize transparency, risk assessment, and human oversight, while others focus on sector-specific rules. For multimodal agents, regulation must grapple with cross-border data flows, liability allocation among model providers and deployers, and the difficulty of auditing learned systems. Certification may be feasible for narrow applications with stable environments, but general-purpose agents challenge traditional approval models. A more practical path may involve continuous monitoring, incident reporting, and adaptive standards that evolve with deployment experience. The goal is not to eliminate risk but to make it visible, bounded, and contestable.

6. Fairness, Accessibility, and Socio-Technical Inclusion

Fairness in multimodal agents extends beyond statistical parity in predictions. It includes access to services, quality of interaction, representation in training data, and distribution of benefits and harms. Bias can enter through datasets, modality-specific sensors, language variation, cultural context, and deployment environments [17], [18]. A vision system trained primarily on certain skin tones or geographic settings may perform unevenly across populations. A speech interface may fail for accents, dialects, or disabled users. A language model may reproduce stereotypes when describing images or generating recommendations.

These failures are not merely technical; they can reinforce exclusion in healthcare, education, employment, and public services.

Accessibility is both a design requirement and a fairness concern. Multimodal agents can potentially improve accessibility by offering speech, text, vision, and gesture interfaces. However, poorly designed systems can create new barriers. For users with visual impairments, reliance on visual confirmation may be problematic. For users with motor impairments, gesture-based control may be inaccessible. For users with cognitive disabilities, complex multi-turn interactions may be overwhelming. Inclusive design requires multimodal redundancy, customizable interaction modes, and evaluation with diverse participants. It also requires attention to language diversity and cultural norms, because what counts as a polite or clear interaction varies across communities.

The socio-technical dimension of fairness concerns institutional power. Organizations that deploy agents may use them to automate decisions, monitor workers, or allocate resources. If affected communities lack representation in design and governance, agents may entrench existing inequalities. Participatory methods, community review boards, and impact assessments can help, but they must be backed by enforceable accountability. Transparency about data sources, model limitations, and performance across groups is necessary but not sufficient. Fairness also requires mechanisms for appeal, correction, and redress. In this sense, multimodal agents should be understood as governance instruments as much as technical artifacts.

Cross-domain comparisons reveal different fairness priorities. In healthcare, equity may focus on diagnostic accuracy across demographics and access to care. In education, it may focus on language inclusion and avoiding surveillance. In public services, it may focus on procedural fairness and non-discrimination. In industrial settings, it may focus on worker safety and job displacement. A single fairness metric cannot capture these varied concerns. Instead, deployment requires domain-specific deliberation about values, harms, and trade-offs. The architecture must support this deliberation by enabling disaggregated evaluation, audit trails, and configurable constraints.

7. Sustainability and Lifecycle Economics

Sustainability in multimodal agents has environmental, economic, and organizational dimensions. Training and serving large multimodal models consume significant energy and compute. Inference costs scale with modality richness, context length, and tool usage. While efficiency techniques can reduce resource consumption, they may trade off against accuracy or robustness. The environmental impact of widespread agent deployment depends on data center energy sources, hardware lifecycles, and utilization rates. Organizations should therefore consider carbon accounting, workload scheduling, and model right-sizing as part of system design.

Economic sustainability concerns the total cost of ownership. Multimodal agents require ongoing data curation, integration maintenance, safety monitoring, and user support. These costs often exceed initial development budgets. A system that performs well in a pilot may become unsustainable when scaled to thousands of users or when external tools change. Lifecycle governance should include versioning, deprecation policies, and migration plans. It should also consider the cost of human oversight and the risk of automation-induced skill loss. If organizations become dependent on agents they cannot maintain or explain, they accumulate operational fragility.

Organizational sustainability involves trust and legitimacy. Users are more likely to adopt agents that respect their privacy, explain limitations, and provide recourse. Institutions need processes for handling incidents, updating policies, and incorporating feedback. Sustainability is therefore not only about environmental footprint but about the social license to operate. A multimodal agent that saves costs but erodes trust may be economically rational in the short term and disastrous in the long term. Designers should treat maintainability, transparency, and contestability as core features rather than overhead.

8. Policy, Regulation, and Institutional Adaptation

Policy frameworks for artificial intelligence increasingly emphasize human-centered values, accountability, and precaution [20]. Multimodal large language agents challenge these frameworks because they cross sectoral boundaries and operate through emergent behavior. A single agent might combine medical advice, scheduling, and insurance interaction, thereby falling under multiple regulatory regimes. Policymakers must decide whether to regulate models, applications, deployers, or outcomes. Each choice has different implications for innovation, liability, and enforcement. Model-level regulation can set baseline safety requirements but may struggle with context-specific risks. Application-level regulation can be more precise but may lag behind technical change.

Institutional adaptation is equally important. Hospitals, schools, factories, and government agencies must develop procurement standards, incident response plans, and staff training. They must also negotiate with vendors about data rights, model updates, and audit access. Contracts should specify performance expectations, security requirements, and remedies for failure. Professional bodies may need to update codes of conduct and certification requirements. These adaptations take time and resources, which means that early deployments should be designed to generate evidence and build capacity rather than assume immediate transformation.

International coordination is complicated by divergent values and legal systems. Data protection rules, content moderation standards, and liability doctrines vary widely. Multimodal agents that operate across borders may need to comply with multiple regimes simultaneously. Technical mechanisms such as data localization, privacy-preserving computation, and configurable policy layers can help, but they cannot resolve fundamental conflicts. Diplomacy and standard-setting will be necessary to avoid fragmentation. In the absence of coordination, organizations may default to the most restrictive jurisdiction or the least protective one, depending on market incentives.

9. Case Illustrations and Cross-Domain Comparisons

In healthcare, multimodal agents could assist with triage, documentation, and coordination. They might interpret images, summarize patient histories, and suggest next steps while clinicians retain decision authority. The benefits include reduced administrative burden and faster information retrieval. The risks include privacy breaches, diagnostic error, and automation bias. Successful deployment would require clinical validation, audit trails, and clear escalation pathways. It would also require integration with existing workflows rather than adding another screen or alert.

In manufacturing and logistics, agents could monitor equipment, coordinate robots, and optimize schedules. Multimodal perception can detect defects, while language interfaces allow operators to query system status and issue commands. The risks include physical injury, production disruption, and cybersecurity vulnerabilities. Safety certification and deterministic

shutdown mechanisms become essential. The economics favor edge deployment for latency and reliability, but fleet management and updates remain challenging.

In education, multimodal agents could provide tutoring, translation, and accessibility support. They might observe student work, respond to speech, and adapt materials. The risks include surveillance, bias, and over-reliance on automated judgment. Governance should prioritize student privacy, teacher oversight, and equitable access. In public services, agents could help citizens navigate benefits, permits, and health information. The risks include exclusion, misinformation, and opaque decision-making. Transparency, multilingual support, and human appeals are critical.

Across these domains, a common pattern emerges. The most successful deployments are likely to be bounded, monitored, and co-designed with affected stakeholders. General-purpose agents may be powerful, but they are difficult to certify and govern. Domain-specific agents may be less flexible, but they can be integrated into existing accountability structures. The optimal architecture may combine a general reasoning core with domain-specific constraints, tools, and oversight. This hybrid approach reflects the broader lesson that real-world intelligence is distributed across people, institutions, and infrastructures, not contained solely in a model.

10. Conclusion

Multimodal large language agents promise to reshape interactive applications by combining perception, language, reasoning, and action. Yet their real-world value depends on how they are architected, deployed, governed, and sustained. This paper has argued that the decisive challenges are system-level: managing architectural trade-offs, ensuring robustness, distributing benefits fairly, reducing environmental and economic burdens, and adapting policy to fast-moving technical change. These challenges cannot be solved by scaling alone. They require deliberate choices about modularity, human oversight, data governance, and institutional responsibility.

The analysis suggests that multimodal agents should be treated as infrastructural systems rather than standalone tools. Infrastructure implies long lifespans, public dependencies, and governance obligations. It also implies that design must account for maintenance, interoperability, and failure. A system that cannot be audited, repaired, or contested is not ready for high-stakes deployment, regardless of benchmark performance. Conversely, a more modest agent embedded in strong accountability structures may deliver greater value than a highly capable but opaque system.

Future research should focus on evaluation methods that capture real-world robustness, fairness, and sustainability. It should also examine participatory design, regulatory sandboxes, and lifecycle governance. Engineers should build systems that expose uncertainty, support human intervention, and log decisions in ways that enable learning. Policymakers should create adaptive frameworks that protect rights without freezing innovation. Ultimately, the legitimacy of multimodal large language agents will depend on whether they enhance human agency, distribute benefits equitably, and remain accountable to the communities they serve.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.

2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
3. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
4. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748–8763.
5. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35, 23716–23736.
6. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning*, 202, 19730–19742.
7. Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., ... & Florence, P. (2023). PaLM-E: An embodied multimodal language model. *Proceedings of the 40th International Conference on Machine Learning*, 202, 8469–8488.
8. Jiang, H., Xin, G., Huang, Y., Zhu, G., Zhang, J., Lin, X., ... & Cao, Z. (2026). AgentOmnia: Scaling Agentic Models for Full-Scenario Applications. *arXiv preprint arXiv:2607.23124*.
9. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*.
10. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., ... & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.
11. Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., ... & Wang, C. (2023). AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*.
12. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.
13. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., ... & Zeng, A. (2022). Do as I can, not as I say: Grounding language in robotic affordances. *Proceedings of the 6th Conference on Robot Learning*, 205, 287–318.
14. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., ... & Zeng, A. (2023). Code as policies: Language model programs for embodied control. *Proceedings of the 40th International Conference on Machine Learning*, 202, 19730–19742.
15. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

16. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., ... & Gabriel, I. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214–229.
17. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1–35.
19. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
20. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>