

Continual User Intent Learning for Adaptive Human–AI Interaction in Long-Term Conversational Systems

Enzo R. Warner

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
erwarner@binghamton.edu

Miguel Burns

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
miguelburns@uc.edu

Zhouqing Gu

Department of Computer Science and Engineering, University of Nevada, Reno, NV, USA.
hellozhouqing@unr.edu

Abstract

Long-term conversational systems increasingly mediate complex activities in health care, education, finance, customer service, and personal assistance. Static intent classifiers trained on fixed datasets cannot adequately support interactions in which user goals, preferences, constraints, and domain knowledge shift over months or years. This paper presents a system-level analysis of continual user intent learning for adaptive human–AI interaction. It examines architectural strategies that combine foundation-model representations with persistent user memory, episodic retrieval, and controlled model updating. The analysis argues that continual intent learning is not only a machine learning problem but also an infrastructure design and governance problem. Three structural tensions organize the discussion: the need for adaptation against the need for stability, the value of personalization against privacy obligations, and the ambition of automation against the requirement for meaningful human oversight. The paper considers memory lifecycle, intent drift, alignment, fairness, robustness, deployment constraints, evaluation regimes, and policy implications. It further compares conversational systems across domains to illustrate how organizational capacity and contextual harm shape acceptable adaptation. The paper concludes that sustainable adaptive conversational platforms require continual intent learning to be embedded within accountable sociotechnical infrastructure, with feedback loops that connect model behavior to institutional review, user control, and long-term evaluation.

Keywords

continual learning; user intent; human–AI interaction; conversational systems; adaptive systems; memory architecture; governance; fairness; sustainability.

1. Introduction

Contemporary conversational systems are shaped by transformer-based sequence models [1], pretrained bidirectional encoders [2], and large autoregressive language models that demonstrate few-shot generalization across tasks [3]. Instruction tuning and reinforcement

learning from human feedback have further improved the ability of these systems to follow user instructions in constrained settings [4]. However, these advances primarily address generalization at training or alignment time rather than longitudinal adaptation after deployment. Continual lifelong learning research has long emphasized that systems operating in nonstationary environments must balance new information with previously acquired knowledge [5]. In conversational settings, this challenge is especially acute because user intent is expressed through incomplete, ambiguous, and evolving language. A user may begin by asking factual questions about a medical condition and later seek emotional reassurance, treatment planning support, or help communicating with a clinician. A static intent model cannot account for these shifting goals without either overgeneralizing or relying on rigid task definitions.

Long-term human–AI interaction requires a different orientation. Instead of treating intent recognition as a classification event at each turn, system designers must treat intent as a continuously reconstructed representation that draws on interaction history, current context, explicit user signals, and inferred preferences. This perspective reframes the conversational system as a learning infrastructure rather than a fixed inference pipeline. It raises system-level questions about where user memory should live, how it should be updated, who can access it, how long it should persist, and how to prevent harmful feedback loops. These questions cannot be answered by optimizing a single model architecture in isolation. They require governance mechanisms, monitoring systems, human oversight, and deployment strategies that respect privacy, fairness, and contextual sensitivity. This paper therefore develops a structural account of continual user intent learning with attention to architectural foundations, memory representation, adaptation loops, robustness, fairness, infrastructure, sustainability, and policy.

2. The Problem of Long-Term Intent Dynamics

Long-term conversational systems face a distinct problem compared with single-session assistants. In a short interaction, the primary goal is to recover a relatively stable intent from one or a few utterances. In long-term settings, the relevant construct is not a single intent but an evolving distribution of preferences, constraints, and task orientations. User goals may shift because of life events, domain changes, or changing trust in the system. For example, a financial assistant may initially support budgeting and later be asked to explain retirement planning options. An educational assistant may move from basic grammar correction to helping a learner compose research arguments. Without adaptation, the system either anchors too strongly to early behavior or treats every session as independent, losing accumulated knowledge that could improve interaction quality.

Recent work on memory-based query matching and intent modeling has demonstrated that user-oriented memory can improve search efficiency in conversational systems [6]. The core insight is that past queries, corrections, and preferred formulations are not just interaction traces but can serve as structured memory that conditions future retrieval and understanding. However, memory introduces a new set of system-level risks. If the memory representation is too rigid, the system may reinforce outdated preferences. If it is too fluid, the system may fail to develop stable user models. If it is too opaque, users may lose trust because they cannot understand why the system appears to remember or forget certain information. Therefore, continual intent learning must be understood as a controlled process of accumulation, consolidation, revision, and forgetting rather than as indefinite retention.

The long-term intent problem also includes temporal asymmetry. Early interactions may be sparse and highly ambiguous. Later interactions may involve richer context but also higher stakes. A health assistant that fails to distinguish between a request for general information and a request for urgent symptom evaluation may create risk. A customer service system that misreads long-standing account preferences may cause financial inconvenience. A tutoring system that fails to detect a learner moving from procedural exercises to conceptual understanding may provide poor pedagogy. Across these settings, intent is not a static label but a relation among user state, domain state, and system capability. Consequently, no single ground truth exists for many long-term intent modeling tasks. The system must learn to act under uncertainty while preserving opportunities for user correction and human intervention.

3. Architectural Foundations for Continual Intent Learning

A robust architecture for continual intent learning separates the representation of user intent from the decision-making components that consume it. A modular design typically includes an intent encoder, a user memory store, a retrieval mechanism, a policy layer, and an interaction interface. The intent encoder maps the current utterance and contextual signals into a representation that can be compared with memory entries. The memory store maintains user-specific summaries, episodic interaction records, and preference signals. The retrieval mechanism selects relevant memory content to condition the response. The policy layer determines how strongly the memory should influence the current output and when the system should seek clarification. The interaction interface exposes appropriate control mechanisms so that users can inspect, correct, or delete memory.

This modularity is important for system governance because it allows different components to be updated at different rates. The intent encoder may remain relatively stable to preserve behavioral consistency. The memory store may be updated frequently as new interactions occur. The policy layer may be adjusted by domain experts without retraining the entire model. Such differential update rates reduce the risk that localized adaptation destabilizes the entire conversational system. They also make it possible to audit which component contributed to a particular failure. Foundation models can serve as the shared substrate for intent encoding, but they should not be treated as complete solutions to continual adaptation. Their scale and opacity call for architectural boundaries that preserve accountability and allow targeted revision [7].

At the systems level, retrieval-augmented architectures provide a useful middle ground between full model retraining and fixed model inference. Rather than continuously fine-tuning a large model, the system can encode user-specific knowledge in a separate memory index and retrieve relevant entries at inference time. This reduces the frequency of expensive model updates and creates a more manageable unit of personalization. It also supports deletion and correction, which are essential for privacy and user agency. However, retrieval introduces its own trade-offs. Retrieval errors can surface outdated or irrelevant memory, and the combination of generative output with retrieved memory can produce confident yet incorrect responses. System designers must therefore include verification mechanisms, confidence calibration, and fallback behavior that seeks clarification when memory and current context conflict.

Foundational model deployment also influences adaptation capacity. Large models may be hosted by external providers, which creates constraints on fine-tuning, data access, and latency. Smaller task-specific models may be deployed locally but provide weaker generalization. Hybrid designs can use a large model for interpretation and a smaller

specialized model for frequent adaptation. This division of labor reduces dependence on a single provider and supports privacy-sensitive deployments. Nonetheless, hybrid systems increase complexity in monitoring, versioning, and failure attribution. A continual intent learning architecture must therefore be evaluated not only by accuracy but also by its operational manageability, ability to support user control, and compatibility with organizational governance structures [8].

4. User Intent Representation and Memory Lifecycle

User intent representation in long-term systems must balance richness with interpretability. Rich vector representations can capture subtle semantic similarity and support flexible retrieval. However, they are difficult for users and auditors to inspect directly. Structured representations, such as named preference fields or task-specific schemas, are more interpretable but may be too rigid to capture evolving goals. A practical design often combines an interpretable summary layer, such as a set of recent goals and constraints, with an embedding-based memory index for deeper retrieval. The summary layer supports user inspection and correction, while the embedding layer supports automatic adaptation. This layered representation mirrors broader debates about using language models as components in sociotechnical systems, where semantic fluency can obscure underlying limitations [9].

The memory lifecycle determines how information enters, persists, consolidates, and exits the system. Entry can occur through explicit user statements, implicit behavioral signals, or system-initiated clarification. Implicit signals are valuable because they reduce user burden, but they are also noisy and may reflect momentary preferences rather than stable intent. Consolidation involves grouping related interaction episodes into durable user-level representations. Without consolidation, the memory store becomes fragmented and increasingly expensive to search. Without forgetting, the system may accumulate contradictory preferences and become less responsive to genuine change. Effective memory management therefore requires decay mechanisms, conflict detection, and periodic reconciliation between old and new signals.

Memory lifecycle decisions also affect user trust. Users are more likely to trust systems that appear to remember important commitments while forgetting irrelevant or time-bound constraints. For example, a tutoring assistant should remember a learner's long-term goal of passing a certification exam, but it should not treat a single frustrated comment as a permanent preference. A health assistant should remember medication allergies, but it should revisit noncritical preferences if they have not been confirmed for a long period. In each case, the system must communicate its understanding and offer opportunities for revision. Memory is not simply a technical cache; it is a social and institutional artifact because it produces effects that extend beyond the immediate conversation.

The representation of intent across long time horizons also intersects with identity and self-presentation. Users may experiment with different roles, explore hypothetical scenarios, or express conflicting desires. A well-designed system should not prematurely collapse these explorations into a fixed user model. It should preserve enough uncertainty to allow future clarification. This is particularly important in health, legal, and financial domains, where inaccurate personalization can lead to concrete harm. System-level design must therefore include uncertainty thresholds, confidence indicators, and explicit mechanisms for users to reset or correct memory. These mechanisms are not secondary features but central components of responsible adaptive interaction.

5. Adaptive Decision Loops and Interaction Governance

Continual intent learning becomes operational through adaptive decision loops that connect observation, memory update, response generation, and user feedback. These loops operate at multiple timescales. Turn-level loops adjust the system's immediate interpretation when the user corrects or clarifies. Session-level loops update short-term memory about the current task. Long-term loops consolidate information into persistent preferences and goals. Each loop introduces different forms of risk. Turn-level adaptation may be too reactive to noisy input. Long-term adaptation may be too slow to detect sudden shifts. A well-governed system coordinates these loops so that rapid adjustments do not immediately become durable user model changes unless sufficient evidence accumulates.

Governance mechanisms are needed because adaptation can amplify errors. If the system updates its user model after every interaction, a single misclassification can become embedded in future behavior. If the system never updates, it cannot support longitudinal tasks. One approach is to separate candidate intent hypotheses from confirmed user model commitments. Candidates can influence the current response provisionally, while confirmed commitments influence long-term behavior. Promotion from candidate to commitment requires repeated evidence, high confidence, or explicit user confirmation. This separation reduces the risk of silent drift and creates an audit trail for model changes. It also supports explainability, because the system can distinguish between a tentative interpretation and a durable assumption.

Human oversight is especially important in high-stakes domains. Rather than requiring humans to approve every intent update, the system can route only ambiguous, consequential, or contested updates for review. For example, a health assistant might proceed with a provisional interpretation for a low-stakes informational query, but flag a change in medication preference for human review. A financial assistant might allow routine categorization changes but require human approval before altering risk tolerance profiles. These intervention thresholds must be domain-specific and should be evaluated against the consequences of both under-adaptation and over-adaptation. Human oversight is not simply a safeguard; it also shapes the data that the system uses for future learning, creating a governance loop between human judgment and model behavior.

Interaction governance also includes user-facing controls. Users need mechanisms to inspect what the system believes about them, challenge incorrect inferences, and request deletion. Without such controls, continual intent learning can become coercive rather than helpful. Long-term systems may accumulate sensitive information about mental health, financial circumstances, legal exposure, or personal relationships. Users may not recall what they disclosed months earlier. The conversational interface should therefore provide periodic memory summaries, natural language explanations of remembered preferences, and accessible commands for correction. These controls must be integrated into the interaction model rather than buried in external privacy settings, because the relevant harms arise during ordinary use.

6. Robustness, Fairness, and Sociotechnical Risk

Continual intent learning systems face robustness problems that differ from static models. They must remain effective under distributional shift, noisy user feedback, adversarial misuse, and concept drift. Memory retrieval can become a vulnerability if users intentionally or unintentionally inject misleading information. For example, a user may state a preference that

later changes, or a malicious actor may manipulate a shared assistant by inserting harmful instructions into memory. Robustness therefore requires ongoing validation of memory content, conflict resolution strategies, and mechanisms to bound the influence of any single interaction. The system must also be robust to missing data. Users may stop interacting for long periods, during which their preferences and circumstances may change. Re-engagement routines should treat historical memory as informative but potentially stale, prompting clarification rather than immediately acting on outdated assumptions.

Fairness concerns arise because intent models may systematically misclassify language varieties, cultural expressions, or disability-related communication styles. If the system treats nonstandard phrasing as lower confidence, it may offer poorer adaptation to already marginalized users. In long-term settings, this can produce compounding disadvantage, because early misclassifications influence later memory. Fairness cannot be achieved solely by balanced training data; it requires ongoing monitoring of interaction outcomes across user groups and contexts. The conceptual distinction between individual fairness and group fairness remains relevant, but it is complicated by the temporal nature of user modeling [10]. A user may be treated fairly at one moment and unfairly later because of accumulated assumptions. Fairness metrics must therefore include longitudinal measures of equitable adaptation, not only aggregate predictive equality.

Bias in user intent learning is not only a property of the model but also a property of the feedback environment. Conversational systems may be trained on interaction data that reflects existing disparities in service quality, user patience, or institutional filtering. When deployed, the system may continue to learn from interactions shaped by its own behavior, creating feedback loops that reinforce inequality. For example, if users from a particular group receive less useful responses, they may disengage or provide less corrective feedback, which the system may interpret as lower need for adaptation. This dynamic resembles broader sociotechnical phenomena in which algorithmic systems produce self-fulfilling outcomes [11]. Addressing such feedback requires more than model retraining; it requires interventions in the interaction pipeline, the memory lifecycle, and the organizational incentives that shape deployment.

Transparency mechanisms such as model cards and system documentation play an important role in managing these risks [12]. For continual intent learning, documentation must go beyond static model performance and describe how user data are stored, updated, retrieved, and deleted. It should identify the domains in which adaptation is enabled, the thresholds for human review, the expected failure modes, and the rights users have over their inferred profiles. Fairness and accountability are not properties of a fixed artifact but of an ongoing sociotechnical relationship. Drawing on political philosophy, fairness in machine learning often involves choices among conflicting moral commitments that cannot be resolved by a single metric [13]. Long-term conversational systems intensify this problem because they make fairness trade-offs over time and across contexts, requiring continual deliberation rather than one-time compliance.

Sociotechnical risk also emerges from the way intent models abstract complex human behavior. Abstraction is necessary for system design, but it can erase contextual meaning and produce mismatches between the institutional understanding of a user and the user's lived experience. The distinction between formal fairness interventions and the deeper institutional uses of user models has been extensively discussed in the fairness literature [14]. In long-term conversational systems, abstractions may become increasingly stable as they are embedded in

memory, making them harder to contest. Designers should therefore maintain multiple levels of representation and avoid prematurely resolving ambiguity into fixed categories. Users should be able to contest not only individual inferences but also the categories and goals that the system uses to organize their interactions.

7. Deployment Infrastructures and Operational Sustainability

The deployment of continual intent learning systems imposes substantial infrastructure requirements. User memory must be stored securely, updated with low latency, retrieved efficiently, and synchronized across devices or sessions. These requirements are further complicated by privacy regulations that treat inferred personal data differently across jurisdictions. Federated learning offers one architectural direction for decentralized model adaptation without centralizing raw interaction data [15]. In principle, user-specific intent models can be updated on local devices, with only aggregated or differentially private updates shared with central servers. This reduces exposure risk and gives users greater control over their data. However, federated approaches are not a complete solution. They introduce challenges in coordination, straggler management, model divergence, and evaluation because data distributions may vary widely across users and institutions [16]. The added complexity can reduce model quality or increase operational cost, creating pressure to centralize data in practice.

Operational sustainability also depends on the frequency and cost of model updates. Continual learning systems that require frequent retraining or large inference pipelines may become prohibitively expensive when scaled to millions of users. Memory retrieval can reduce the need for retraining, but it increases storage and indexing costs. Different domains require different trade-offs. A high-volume customer service assistant may prioritize low latency and minimal memory depth, while a mental health support system may prioritize careful memory management and human review even at higher cost. Deployment decisions should therefore be guided by a clear analysis of the value created by adaptation and the harm caused by poor adaptation. Optimization for short-term efficiency can undermine the long-term trust necessary for sustainable human–AI relationships.

Versioning and rollback present additional infrastructure challenges. When a conversational system updates its intent model or memory policy, users may experience sudden changes in behavior. Without version control, it is difficult to attribute failures to a specific update or to restore a previous stable configuration. System logs should capture not only model versions but also memory changes, policy parameters, and retrieval outcomes. This information supports post-hoc analysis and regulatory audit. In safety-critical domains, operators may need to freeze certain memory-derived influences until a new policy has been validated. The ability to roll back or constrain adaptation is thus a key architectural requirement, not a optional operational convenience. It allows the system to recover from harmful updates while preserving user trust and institutional accountability.

Sustainability also includes environmental and organizational dimensions. Large foundation models require significant computational resources for training and inference. Continual adaptation can increase energy use if every interaction triggers costly model updates. Retrieval-based designs can reduce this burden, but they still require storage and indexing infrastructure. Organizations must assess whether the benefit of long-term personalization justifies the additional energy and hardware demands. Furthermore, the human labor required to annotate feedback, review high-stakes adaptations, and maintain governance processes is often underestimated. Sustainable deployment requires staffing, training, and institutional

procedures that extend beyond the initial model release. A conversational system that learns continually is not a one-time product but a long-term operational commitment.

8. Evaluation, Policy, and Future Directions

Evaluation for continual intent learning cannot rely solely on static benchmark accuracy. Systems must be assessed across multiple timescales, user groups, and interaction conditions. Short-term metrics such as turn-level intent accuracy or response relevance are important but insufficient. Long-term metrics should capture whether the system improves user outcomes over time, whether it avoids harmful memory persistence, and whether users maintain trust and perceived control. Evaluation should also include failure analysis under drift, low-resource interaction, and adversarial input. A human-centered artificial intelligence perspective suggests that reliability, safety, and trustworthiness arise from a combination of technical performance and organizational accountability [17]. This requires evaluation frameworks that include user experience, institutional oversight, and capacity for responsible intervention.

Policy implications follow from the long-term nature of intent learning. Data protection frameworks often emphasize collection and consent, but they may not adequately address inferred preferences that emerge from many interactions over time. Users may not perceive inference as data collection, even when it creates sensitive profiles. Policy should require clear disclosure of memory mechanisms, periodic user access to inferred profiles, meaningful deletion rights, and restrictions on automated decision-making in high-stakes contexts. These requirements are not merely legal compliance burdens; they can improve design by forcing system developers to make memory explicit and contestable. Governance frameworks that only consider model deployment as a discrete event fail to capture the ongoing learning process. Regulatory attention should therefore shift toward algorithmic lifecycle management, including monitoring, audit, and user-accessible recourse mechanisms.

The political economy of user modeling also deserves attention. Long-term conversational systems produce value by converting interaction histories into durable profiles that can improve services, target recommendations, or reduce operational costs. This value is unequally distributed and often invisible to users. The planetary and labor costs of artificial intelligence infrastructure have been documented in critical scholarship [18], and the same critical attention should be applied to the ongoing extraction of user intent. Policy interventions might include limits on inferred data retention, requirements for local storage, or mandatory data sharing agreements that give users a stake in the value derived from their interaction histories. Without such interventions, continual intent learning may deepen the asymmetry between platform operators and users, even as it improves conversational convenience.

Future research should address several open problems. The first is the development of memory architectures that can support rich personalization while preserving interpretability and user control. The second is the design of evaluation methods that capture longitudinal fairness and harm rather than short-term accuracy. The third is the creation of governance mechanisms that link model adaptation to institutional accountability. The fourth is the study of how different application domains shape acceptable adaptation trade-offs. These directions require collaboration across machine learning, human-computer interaction, law, and organizational studies. Continual user intent learning will only become a trustworthy foundation for long-term conversational systems if it is embedded in a broader sociotechnical

framework that respects human agency, contextual complexity, and the long-term consequences of algorithmic memory.

9. Conclusion

Continual user intent learning is a central challenge for adaptive human–AI interaction in long-term conversational systems. The problem cannot be reduced to improving intent classification accuracy, because user goals change, memory introduces persistent influence, and adaptation creates feedback loops that shape future interaction. This paper has argued for a system-level perspective that treats continual intent learning as an architectural, operational, and governance challenge. Effective systems require modular separation of intent encoding, memory management, retrieval, and policy. They require memory lifecycles that support consolidation, forgetting, user correction, and uncertainty. They require adaptive decision loops that distinguish provisional interpretations from durable commitments. They require fairness and robustness mechanisms that account for temporal compounding and sociotechnical abstraction. Finally, they require deployment infrastructures, evaluation frameworks, and policy regimes that make adaptation accountable and sustainable. The future of conversational AI depends not only on more powerful models, but on the careful construction of infrastructures that learn with users while respecting their autonomy, diversity, and right to change.

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. In *Advances in Neural Information Processing Systems* 30 (pp. 5998–6008). Curran Associates.
2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186). Association for Computational Linguistics.
3. 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., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems* 33 (pp. 1877–1901). Curran Associates.
4. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* 35 (pp. 27730–27744). Curran Associates.
5. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 54–71.
6. Yu, X. (2026). Enhancing Search Efficiency through LLM-Based User Memory Systems for Query Matching and Intent Modeling.

7. Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM.
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., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niknam, N., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.
9. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM.
10. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Biles, C., Birhane, A., Haas, J., Rimell, L., Hendricks, L. A., Isaac, W., Legassick, S., Irving, G., & Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). ACM.
11. 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), Article 115, 1–35.
12. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM.
13. Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency* (pp. 149–159). PMLR.
14. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 59–68). ACM.
15. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In

Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 1273–1282). PMLR.

16. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R. G. L., Eichner, H., Rouayheb, S. E., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P. B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konečný, J., Korolova, A., Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S. U., Sun, Z., Suresh, A. T., Tramèr, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F. X., Yu, H., & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1–2), 1–210.
17. Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe and trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495–504.
18. Crawford, K. (2021). *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.