

Reinforcement Learning for Adaptive User Preference Optimization in Task-Oriented Dialogue Systems

Xuanlin Gong

Department of Computer Science, University of North Texas, Denton, TX, USA.

xuan.work@unt.edu

Abstract

Task-oriented dialogue systems are increasingly expected to optimize interaction outcomes beyond fixed slot-filling and intent classification. Reinforcement learning provides a principled framework for treating user preference as a nonstationary signal that emerges through repeated interaction. This paper examines the integration of reinforcement learning for adaptive user preference optimization in task-oriented dialogue systems from a systems perspective. It considers architectural trade-offs, reward modeling, safety constraints, data infrastructure, deployment, robustness, fairness, and governance. Rather than presenting a single algorithm, the paper argues that sustainable preference optimization requires a layered architecture in which preference elicitation, policy adaptation, and evaluation operate under explicit policy and resource constraints. The discussion highlights the tension between short-term task completion and long-term user satisfaction, the risks of reward hacking, distributional shift, and bias amplification, and the need for continuous monitoring in production environments. Case illustrations from customer service, healthcare scheduling, and smart assistant deployments are used to compare design choices. The paper further reviews how pairwise preference models and dual-stage transformer architectures support scalable preference estimation within constrained deployment environments. It concludes with policy implications for auditability, human oversight, and institutional governance, emphasizing that reinforcement learning in conversational systems should be treated as a socio-technical infrastructure rather than an isolated optimization problem.

Keywords

reinforcement learning; task-oriented dialogue systems; preference optimization; human feedback; governance; robustness; fairness.

1. Introduction

Task-oriented dialogue systems have historically been constructed around supervised pipelines for intent classification, entity extraction, and slot filling. Although such pipelines are effective in narrow domains, they assume that user goals are stable and can be captured by static labels. In practice, user preferences shift across sessions, depend on context, and are often expressed indirectly through conversational repair, latency, or repeated reformulations. Reinforcement learning offers a natural alternative because it formulates dialogue management as sequential decision making under uncertainty [1]. Early policy gradient methods demonstrated that dialogue policies could be optimized directly from scalar rewards without requiring fully differentiable objectives [2]. The subsequent integration of deep neural networks expanded the state and action spaces that could be represented, enabling end-to-end optimization of dialogue control [3].

Among modern policy optimization methods, proximal policy optimization has become widely used because it balances sample efficiency with stability, a critical property when interaction budgets are limited [4]. The reward function remains the most consequential design artifact in such systems. Rather than relying solely on human-authored rules, researchers have explored learning reward models from pairwise human preferences, allowing systems to infer complex quality judgments that are difficult to formalize [5]. In dialogue settings, offline batch reinforcement learning from implicit human preferences has addressed the challenge of learning from logged interactions without costly online exploration [6]. These developments establish the core premise of this paper: adaptive user preference optimization is best understood as a system-level problem in which reward acquisition, policy learning, and production feedback loops must be co-designed.

Task-oriented dialogue architectures provide the operational substrate for preference adaptation. Retrieval-based systems grounded in key-value memory structures support transparent knowledge access and can be integrated with learned policies for entity selection [7]. Sequence-to-sequence architectures simplify dialogue state tracking by generating belief states and responses jointly, which reduces modular error propagation [8]. Surveys of neural conversational AI emphasize that task completion is only one dimension of quality and that user satisfaction depends on naturalness, efficiency, and appropriate handling of uncertainty [9]. A purely supervised model may optimize for likelihood of observed responses, but it does not directly optimize for delayed outcomes such as task success, user return rate, or trust.

Reinforcement learning in real-world conversational infrastructures also introduces broader system risks. Reward misspecification can lead to policies that exploit evaluation metrics without improving service quality [10]. High variance and sensitivity to hyperparameters complicate reproducible deployment [11]. Real-world constraints such as partial observability, delayed rewards, and nonstationary user populations create additional challenges that are rarely present in simulation benchmarks [12]. Preference learning from human feedback has been proposed as a mechanism for aligning language-based assistants, but it requires careful dataset design and regularization to avoid overfitting to annotator biases [13]. At the same time, large language-based dialogue systems raise concerns about representational harms and environmental costs, requiring governance before they are embedded in public-facing services [14]. Fairness audits and bias surveys show that preference data can encode social asymmetries that reinforcement learning may amplify if not explicitly measured [15]. Ethical frameworks for artificial intelligence further emphasize human autonomy, beneficence, and accountability as design requirements rather than afterthoughts [16]. This paper therefore treats adaptive preference optimization as a socio-technical system and examines structural trade-offs across architecture, data, deployment, robustness, fairness, and policy.

2. Background and System Context

Within this system context, preference modeling is an active subfield. Recent work on pairwise user preference prediction using dual-stage transformer networks with gated fusion illustrates how fine-grained preference signals can be extracted from dialogue transcripts without relying on explicit user ratings [17]. Such models can serve as reward approximators, but their integration into reinforcement learning loops raises architectural questions about when and how to update the preference estimator, the policy, and the safety constraints.

Task-oriented dialogue systems differ from open-domain conversational agents in that their primary purpose is to complete a user transaction, such as booking a reservation, updating an account, troubleshooting a device, or obtaining information. This transactional purpose makes

evaluation seemingly easier because task success can be measured against a known state change. However, the quality of a transaction is not reducible to task completion alone. A user may complete a booking but experience excessive repetition, lose trust in the system, or abandon the service in future sessions. Reinforcement learning aligns with this richer evaluation perspective by treating the dialogue as a sequence of actions whose value is only partially observed through immediate feedback. The system must balance exploitation of known preferences with exploration of user needs that have not yet been articulated [1]. This exploration-exploitation tension is especially acute in task-oriented settings because failed exploratory actions may have direct operational costs, such as incorrect data entry or missed bookings.

Dialogue systems are also characterized by long horizons and sparse rewards. A successful outcome may require multiple user turns, database queries, clarification questions, and confirmations. The reward signal may arrive only after the final transaction or may be inferred from subsequent user behavior, such as whether the user calls again about the same issue. Policy gradient methods construct stochastic policies over dialogue actions and update them in directions that increase expected cumulative reward [2]. Deep reinforcement learning makes it possible to learn from high-dimensional text and structured dialogue states, but it also introduces risks related to representation collapse and overfitting to frequent user patterns [3]. These risks must be managed through architectural constraints rather than relying solely on optimization.

From a system perspective, the dialogue manager cannot be designed independently of the channels through which it is accessed. Voice assistants, text chat, mobile applications, and web-based interfaces produce different interaction dynamics and data distributions. A policy that performs well in a text channel may fail when translated to voice because spoken interactions include disfluencies, turn-taking delays, and pronunciation variability. Adaptive preference optimization therefore requires an orchestration layer that normalizes inputs, maintains session context, and coordinates between the language understanding module, the state tracker, the policy, and the response generator. Surveys of conversational AI highlight that modular decomposition remains valuable for interpretability and maintenance even when some components are learned jointly [9]. The choice between modular and end-to-end designs is a structural trade-off involving transparency, error isolation, and sample efficiency.

3. Architectural Considerations for Preference Learning and RL Integration

Integrating reinforcement learning into a task-oriented dialogue system requires careful separation of concerns. The preference model, the policy, and the value function have different update rates and failure modes. A preference model trained from pairwise comparisons can be updated less frequently than the policy because preference labels are expensive and user preferences drift slowly relative to conversational tactics [17]. The policy may require online updates or periodic batch retraining depending on deployment constraints. When these components are tightly coupled, changes in reward estimation can destabilize the policy. A dual-stage architecture, in which a transformer encodes dialogue context and gated fusion combines turn-level signals, offers one way to produce stable preference estimates for reward shaping [17]. Nevertheless, the reward model should not be treated as a static oracle. Its outputs are noisy, context-dependent, and possibly biased toward the demographic groups represented in the annotation pool.

Proximal policy optimization provides a useful policy update rule because it restricts the magnitude of policy changes, reducing the likelihood of catastrophic degradation when

reward estimates are imperfect [4]. However, even stable policy optimization cannot compensate for a reward function that fails to encode safety and fairness. Reward design in task-oriented systems often combines task completion signals with user satisfaction proxies, such as sentiment, explicit feedback, or the absence of repair. Each proxy has limitations. Sentiment may be expressed differently across cultural groups. Explicit feedback is sparse and subject to response bias. The absence of repair is a negative signal that is difficult to attribute. Preference-based reward learning addresses this by using human comparisons between alternative system responses [5]. In dialogue systems, such comparisons can be more informative than scalar ratings because humans are better at relative judgment than absolute scoring.

An important architectural decision is whether the policy operates over symbolic dialogue acts or directly over natural language tokens. Symbolic action spaces are smaller and easier to constrain, which is important for high-stakes domains such as healthcare scheduling or financial transactions. However, symbolic representations may fail to capture the stylistic and pragmatic dimensions of language that influence user satisfaction. Token-level action spaces allow more expressive responses but require stronger safety filters and larger amounts of data. Hybrid architectures can first select a dialogue act using a constrained policy and then generate a surface realization using a language model. This separation allows safety rules to be enforced at the dialogue act level while maintaining naturalness in the final response. Such architectural choices have direct consequences for data efficiency, auditability, and the ability to correct unsafe behavior.

4. Reward Modeling and Adaptation Mechanisms

Reward modeling in production dialogue systems increasingly follows the preference-based paradigm popularized by large language model alignment. In this paradigm, human labelers compare two model outputs and indicate which is more helpful, accurate, or appropriate, and a reward model is trained to predict these preferences [18]. This approach can be transferred to task-oriented dialogue by collecting comparisons between dialogue trajectories that differ in efficiency, tone, or handling of uncertainty. The resulting reward model can then guide policy updates. However, task-oriented dialogue introduces additional sources of preference that are not present in open-ended generation, including task success, database state accuracy, and policy compliance. A response that is stylistically preferred but factually incorrect should receive a low reward even if human raters find it fluent. Therefore, reward modeling must be multi-objective and constrained by domain-specific verification.

Constitutional and rule-based feedback mechanisms have been proposed to reduce dependence on large volumes of human annotations while encoding normative constraints [19]. In dialogue systems, such mechanisms can enforce policies related to privacy, non-discrimination, and procedural correctness. For example, a reward model may include terms that penalize asking for unnecessary personal information, failing to confirm an action before execution, or using inappropriate language. These constraints are easier to implement when the system has a modular architecture with explicit policy layers. They are harder to guarantee when the policy is an opaque sequence-to-sequence model. The governance implication is that safety constraints should be embedded in the reward architecture rather than treated as post-hoc filters. Post-hoc filters can be bypassed by novel inputs and may create inconsistencies between what the policy learns and what the user experiences.

Adaptation mechanisms must also account for individual and population-level preference drift. Individual preferences may change during a single interaction, for example when a user

becomes frustrated and prefers faster responses over detailed explanations. Population-level preferences shift with seasonal events, product changes, or demographic trends. A reinforcement learning system that adapts only to aggregate data may personalize poorly for minority users. Conversely, a system that overfits to individual history may violate privacy expectations or become unstable when users behave inconsistently. A layered adaptation strategy can address this by maintaining a global policy, a group-level policy, and a session-level preference state. The global policy captures stable domain knowledge. The group-level policy captures differences across segments defined by service context rather than sensitive attributes. The session-level state captures transient needs. This layering reduces the risk that a single user interaction will corrupt long-term knowledge.

5. Data Infrastructure and Deployment Considerations

Deploying reinforcement learning for preference optimization requires data infrastructure that supports logging, preference annotation, offline evaluation, and controlled online experimentation. Many dialogue systems already log transcripts and outcomes, but these logs are often incomplete for reinforcement learning because they do not record the action alternatives that were considered, the estimated value of each state, or the user feedback that was not explicitly provided. Offline batch learning from such logs is challenging because the logged policy introduces distributional shift and missing counterfactuals [6]. Importance weighting and conservative updates can mitigate some of these issues, but they require careful implementation and monitoring [12]. When data coverage is sparse, reward models may extrapolate in unpredictable ways, producing high rewards for states that were rarely visited.

The production deployment of adaptive dialogue policies also requires a serving architecture that can support multiple policy versions, traffic splitting, and rollback. Canary deployments and A/B testing are essential because offline metrics often fail to capture user satisfaction. A policy may improve task completion in simulation while increasing latency or interrupting users in ways that damage trust. Monitoring must include not only aggregate metrics but also distributional metrics across user subgroups. Dashboards that report only average success rates can conceal regressions for minority populations. Logging should be privacy-preserving and should avoid retaining unnecessary conversation content. Differential privacy and data minimization should be considered at the infrastructure level rather than as compliance afterthoughts.

Continual learning is another infrastructure concern. User preferences evolve, and static policies become stale. Periodic retraining from accumulated logs can capture gradual changes, but it may also introduce catastrophic forgetting or amplify recent biases. Online reinforcement learning can adapt more rapidly but may destabilize production systems if reward estimation is noisy. A practical compromise is to use online learning for short-horizon adjustments such as response length or clarification strategy while using periodic batch retraining for high-level dialogue policies. This separation allows quick adaptation to user feedback without sacrificing long-term stability. It also aligns with operational governance because high-level policy changes can be reviewed and approved before deployment, while low-level adjustments remain within pre-authorized bounds.

6. Robustness, Fairness, and Governance

The use of reinforcement learning in task-oriented dialogue systems raises substantial ethical and governance challenges. Algorithms are not neutral; they encode values through objective functions, data selection, and evaluation criteria [20]. In dialogue systems, these values

manifest in which user requests are prioritized, which dialects are treated as canonical, and which behaviors are penalized. A reward model trained primarily on data from one demographic group may systematically assign lower rewards to legitimate variations in language use. This can lead to unequal service quality and may reinforce existing inequalities. Fairness in preference optimization is not solved by removing sensitive attributes from training data because language itself often carries demographic signals [15]. Fairness must be evaluated at the outcome level, measuring not only aggregate task success but also differences in success rates, latency, and user satisfaction across groups.

Robustness is closely linked to governance. Dialogue systems operate in adversarial and noisy environments. Users may provide ambiguous requests, introduce unexpected topics, or attempt to manipulate the system for unauthorized outcomes. Reinforcement learning policies can learn undesirable behaviors when reward functions are misspecified or when the environment shifts [10]. Reproducibility studies have shown that deep reinforcement learning results are sensitive to random seeds, environment details, and implementation choices [11]. This sensitivity has direct implications for certification and audit. A dialogue policy that performs well in a laboratory setting may fail in production because of differences in user behavior, channel noise, or integration with legacy systems. Robustness testing should therefore include stress tests, counterfactual audits, and red-team interactions that probe the boundaries of learned policies.

Governance mechanisms for adaptive dialogue systems should include human oversight, explainability, and redress. Human oversight does not mean that a person reviews every dialogue, but it does mean that operators can inspect why a policy selected a particular action and can intervene when high-risk transactions are involved. Explainability can be supported by modular architectures that expose dialogue states and intermediate decisions. It is harder to achieve in fully end-to-end systems. Redress requires that users can challenge decisions and that the system can log sufficient information to reconstruct what happened without compromising privacy. These governance features should be designed into the reinforcement learning loop from the beginning. Retrofitting them after deployment is expensive and often ineffective. Ethical frameworks for artificial intelligence emphasize that responsibility, transparency, and accountability are structural properties, not rhetorical commitments [16].

7. Sustainability and Operational Trade-offs

Sustainability has emerged as a central concern in the deployment of large language-based dialogue systems. Training and serving large models consumes significant computational energy and contributes to carbon emissions [14]. Reinforcement learning adds further costs because it requires multiple training runs, simulation rollouts, and reward model updates. The marginal benefit of a larger policy model may be small compared to the operational cost, especially in narrow task-oriented domains where a compact model with explicit constraints can perform adequately. System designers should therefore assess the full lifecycle cost of adaptive preference optimization, including data annotation, model training, inference-serving, and monitoring. Each stage introduces different sustainability and operational risks. Data annotation for preference learning requires human labor that may be geographically concentrated and insufficiently compensated, raising questions about the fairness of the annotation supply chain. Training runs may be repeated many times during hyperparameter search and reward model calibration, multiplying energy consumption even if the final model is relatively small. Serving adaptive policies can require maintaining multiple models, user simulators, and logging infrastructure simultaneously, which increases the carbon footprint

and operational complexity. User simulation is a widely used technique for training dialogue policies without exposing users to exploratory actions, but it introduces a new source of bias if the simulator fails to represent the full range of user behavior [21]. Therefore, sustainability assessments should include not only the energy cost of training but also the labor, data storage, and ongoing inference costs associated with preference adaptation.

A compact or modular policy can reduce inference costs but may limit the ability to capture subtle linguistic variation. The appropriate equilibrium depends on the domain risk profile and the frequency of interaction. In high-frequency customer service, small improvements in average efficiency can justify significant training investment. In low-frequency healthcare scheduling, safety and auditability may outweigh minor latency gains. Lifecycle evaluation should therefore be domain-specific and should not assume that larger models always yield proportional value. The use of sparse activation, distillation, and parameter sharing can reduce environmental costs while preserving task completion quality. However, these techniques may also reduce the policy's ability to adapt to rare user preferences, which must be measured explicitly.

8. Policy Implications and Future Directions

Policy implications follow from the structural properties described above. First, adaptive dialogue systems should be subject to algorithmic impact assessments that examine how preference optimization affects different user groups. Such assessments should not focus solely on aggregate task success but should report subgroup-level differences in success rates, uncertainty handling, and user satisfaction [15]. Second, organizations should adopt versioning and audit trails for reward models, policies, and preference datasets. The ability to reconstruct which reward model was active during a given interaction is essential for accountability and for diagnosing failures. This is especially important when policies are updated online or through continual learning. Third, human oversight should be designed around risk tiers. High-risk transactions, such as financial transfers or medical appointment changes, should require confirmation through constrained dialogue acts that can be logged and reviewed. Low-risk interactions can rely on automated adaptation but should still be monitored for unexpected behavior.

Safety in conversational AI is a distinct challenge from safety in supervised classification. A reinforcement learning policy may learn to produce responses that appear helpful but violate user expectations, disclose private information, or reinforce stereotypes. Safety evaluation should include adversarial testing and red-teaming that probes the system's response to ambiguous, manipulative, or culturally specific prompts [22]. Explainability methods can support oversight by making the relationship between dialogue state and policy action legible to auditors, although post-hoc explanations may not fully recover the causal structure of learned policies [23]. Therefore, architectural transparency should be preferred where feasible.

Future research should investigate the integration of preference optimization with privacy-preserving personalization. Federated learning and differential privacy offer pathways for adapting to user preferences without centralizing sensitive conversation data, but they introduce additional complexity in reward estimation and policy aggregation. The energy and policy considerations associated with large-scale NLP models further complicate the lifecycle assessment of adaptive dialogue policies [24]. Another direction is the development of evaluation frameworks that combine task success, user satisfaction, fairness, and sustainability into a unified reporting standard [25]. Current benchmarks emphasize task success in simulated environments, but real-world evaluation must account for long-term user

trust and service continuity. The PARADISE framework established early principles for evaluating spoken dialogue systems by combining task success with dialogue costs [25]. Extending such frameworks to reinforcement learning systems requires modeling user satisfaction as a dynamic outcome rather than a static label.

Conclusion

Reinforcement learning offers a coherent framework for adaptive user preference optimization in task-oriented dialogue systems, but its successful deployment depends on more than algorithmic choice. This paper has argued that preference adaptation is a system-level problem involving reward modeling, architectural constraints, data infrastructure, operational sustainability, fairness, and governance. The integration of pairwise preference models and dual-stage transformer architectures can improve reward estimation, but such models must be embedded within a controlled adaptation loop that separates policy updates from safety constraints. Offline batch learning and proximal policy optimization provide stability, while user simulation and evaluation frameworks help manage exploration risk. However, these techniques inherit biases from the data and reward definitions on which they depend.

Robustness and fairness are not separate from performance; they are constitutive of user trust and long-term service quality. Dialogue systems that optimize narrowly for task completion may inadvertently degrade the experience for minority users, reinforce unequal treatment, and create hidden operational liabilities. Governance mechanisms should include impact assessments, versioned audit trails, risk-tiered human oversight, and transparent reporting of subgroup-level outcomes. As task-oriented dialogue systems become embedded in financial, healthcare, and public service infrastructures, their preference optimization loops must be treated as socio-technical systems subject to institutional accountability. Future work should extend these principles to privacy-preserving personalization, cross-domain adaptation, and sustainability-aware model selection.

References

1. Levin, E., Pieraccini, R., & Eckert, W. (2000). A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing*, 8(1), 11-23. <https://doi.org/10.1109/89.817450>
2. Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4), 229-256. <https://doi.org/10.1007/BF00992696>
3. 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>
4. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*. <https://arxiv.org/abs/1707.06347>
5. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 4299-4307.

<https://proceedings.neurips.cc/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html>

6. Jaques, N., Ghandeharioun, A., Shen, J. H., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., & Picard, R. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. arXiv preprint arXiv:1907.00456. <https://arxiv.org/abs/1907.00456>
7. Bordes, A., Boureau, Y.-L., & Weston, J. (2017). Learning end-to-end goal-oriented dialog. In International Conference on Learning Representations. <https://arxiv.org/abs/1605.07683>
8. Lei, W., Jin, X., Kan, M.-Y., Ren, Z., He, X., & Yin, D. (2018). Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1437-1447). Association for Computational Linguistics. <https://aclanthology.org/P18-1133/>
9. Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI. Foundations and Trends in Information Retrieval, 13(2-3), 127-298. <https://doi.org/10.1561/15000000074>
10. Pan, A., Bhatia, K., & Steinhardt, J. (2022). The effects of reward misspecification: Mapping and mitigating misaligned models. arXiv preprint arXiv:2201.03544. <https://arxiv.org/abs/2201.03544>
11. Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., & Meger, D. (2018). Deep reinforcement learning that matters. In Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 3207-3214. <https://ojs.aaai.org/index.php/AAAI/article/view/11694>
12. Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). Challenges of real-world reinforcement learning. arXiv preprint arXiv:1904.12901. <https://arxiv.org/abs/1904.12901>
13. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008-3021. <https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html>
14. 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. <https://doi.org/10.1145/3442188.3445922>
15. 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. <https://doi.org/10.1145/3457607>
16. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles,

and recommendations. *Minds and Machines*, 28(4), 689-707.
<https://doi.org/10.1007/s11023-018-9482-5>

17. Shih, K., Deng, Z., Chen, X., Zhang, Y., & Zhang, L. (2025, May). DST-GFN: A Dual-Stage Transformer Network with Gated Fusion for Pairwise User Preference Prediction in Dialogue Systems. In 2025 8th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE) (pp. 715-719). IEEE.
18. 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. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html
19. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
<https://arxiv.org/abs/2212.08073>
20. Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121-136.
<https://www.jstor.org/stable/20024652>
21. Schatzmann, J., Weilhammer, K., Stuttle, M., & Young, S. (2006). A survey of statistical user simulation techniques for reinforcement-learning-based dialogue management. *Computational Linguistics*, 32(1), 57-81. <https://doi.org/10.1162/089120106776173045>
22. Dinan, E., Abercrombie, G., Bergman, A. S., Spruit, S., Hovy, D., Boureau, Y.-L., & Rieser, V. (2021). Anticipating safety issues in E2E conversational AI. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 2059-2073). Association for Computational Linguistics.
<https://aclanthology.org/2021.acl-long.160/>
23. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1-42. <https://doi.org/10.1145/3236009>
24. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3645-3650). Association for Computational Linguistics.
<https://aclanthology.org/P19-1355/>
25. Walker, M. A., Litman, D. J., Kamm, C. A., & Abella, A. (1997). PARADISE: A framework for evaluating spoken dialogue agents. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 271-280). Association for Computational Linguistics. <https://aclanthology.org/P97-1035/>