Interactionless Inverse Reinforcement Learning: A Data-Centric Framework for Durable Alignment

Blue Sky Ideas Track

Elias Malomgré
IDLab, Ghent University - imec
Ghent, Belgium
elias.malomgre@ugent.be

Pieter Simoens
IDLab, Ghent University - imec
Ghent, Belgium
pieter.simoens@ugent.be

ABSTRACT

AI alignment is growing in importance, yet many current approaches learn safety behavior by directly modifying policy parameters, entangling normative constraints with the underlying policy. This often yields opaque, difficult-to-edit alignment artifacts and reduces their reuse across models or deployments, a failure mode we term Alignment Waste. We propose Interactionless Inverse Reinforcement Learning, a framework for learning inspectable, editable, and reusable reward artifacts separately from policy optimization. We further introduce the Alignment Flywheel, a human-in-the-loop lifecycle for iteratively auditing, patching, and hardening these artifacts through automated evaluation and refinement. Together, these ideas recast alignment from a disposable training expense into a durable, verifiable engineering asset.

KEYWORDS

AI alignment; AI safety; Inverse Reinforcement Learning; reward modeling; Alignment Waste; Alignment Flywheel

ACM Reference Format:
Elias Malomgré and Pieter Simoens. 2026. Interactionless Inverse Reinforcement Learning: A Data-Centric Framework for Durable Alignment : Blue Sky Ideas Track. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 8 pages. https://doi.org/10.65109/LCMH1709

1 INTRODUCTION

AI alignment aims to ensure that autonomous systems act in accordance with human intent and typically comprises Forward Alignment, which produces a trained, aligned system, and Backward Alignment, ensuring system safety through governance [78]. This creates a structural disconnect between these phases, making Backward Alignment an inspection rather than a corrective measure. This is because current paradigms entangle the safety objective with the agent's policy; rather than defining a static, independent standard of behavior, they mathematically couple reward discovery to policy optimization [19, 110]. Consequently, the safety objective becomes dependent on the policy's specific dynamics, rendering the safety artifacts unstable [97, 110], preventing their reuse, and requiring independent verification for Backward Alignment [78].

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/LCMH1709

This inherent structural flaw creates a destructive cycle that we term Alignment Waste: entangling the learning of reward and policy means that safety artifacts are neither transferable to new architectures nor correctable without retraining. This is present across current alignment paradigms; interactive methods, such as traditional Inverse Reinforcement Learning (IRL) [1, 115, 181] and Reinforcement Learning from Human Feedback (RLHF) [120, 183] explicitly rely on an unstable co-adaptation loop [97, 110, 142]. This entanglement is particularly prominent in Direct Preference Optimization (DPO) [132], rendering the safety artifact opaque by directly dissolving preferences into policy weights [95]. Consequently, they risk *safetywashing* [134], where improved benchmark scores reflect increased model capability rather than genuine safety, masking misalignment under sophistication. Critically, the consequence can be severe as such coupling causes a collapse in reasoning capabilities [170], degrading the model's broad intelligence in exchange for narrow, reward-hacking behaviors [102].

To resolve this, we propose Interactionless Inverse Reinforcement Learning (IIRL). Unlike standard methods, IIRL decouples reward discovery, producing an auditable, editable reward model independent of agents. We also introduce the Alignment Flywheel, an architecture that uses a cooperative multi-agent system to harden the reward model, transforming passive oversight into a cycle of active correction, which we call Active Backward Alignment.

With this integration, we establish a discipline of verifiable safety, transforming alignment from an unstable art into rigorous engineering by treating alignment as the creation and maintenance of a standalone alignment artifact. The IIRL reward artifact anchors *Robustness* and *Interpretability*, while the Alignment Flywheel leverages expert feedback to ensure *Controllability* and *Ethicality*. Collectively, they satisfy the technical RICE principles [78] to support FATE's societal mandates (Fairness, Accountability, Transparency, Ethics) [107]. We present this architecture as a concrete blueprint for the framework developed in the remainder of the paper, providing a practical toolbox and a roadmap for durable alignment.

2 THE IIRL PARADIGM

The Interactionless Inverse Reinforcement Learning (IIRL) paradigm represents a fundamental shift, transforming AI alignment from an unstable art into a rigorous engineering practice. Unlike traditional IRL methods, coupling reward discovery with policy optimization, IIRL structurally decouples these processes. It reframes reward learning as a data-centric problem, directly inferring an agent-agnostic, durable, inspectable, and editable reward model

from expert data, independent of the agent's specific policy or architecture. This approach yields a safety artifact that can be formally audited and systematically refined, solving the critical Alignment Waste problem at its source. This section defines the IIRL objective and analyzes its family of modular and editable architectures, considering their inherent trade-offs for key properties of a resilient safety asset, alongside multi-tiered refinement toolkits.

2.1 The IIRL Objective

Traditional IRL [1, 115, 181] seeks to infer a reward function R from expert demonstrations, typically by solving a max-min optimization problem to find the function R that best explains expert policy π_E:

$$\max_{R \in \mathcal{R}} \min_{\pi \in \Pi} \left(\mathbb{E}_{s,a \sim \pi_E}[R(s,a)] - \mathbb{E}_{s,a \sim \pi}[R(s,a)] \right), \tag{1}$$

requiring repeatedly solving an optimal policy π in the inner loop, inextricably coupling reward learning to policy optimization. This creates an unstable co-adaptation loop [97] where the policy's limited exploration fails to capture the full expert distribution [5, 30], often leading to mode collapse and reward-hacking behaviors [5]. This core flaw persists even in offline [77, 87, 151, 169] and inner-loop-avoiding [160, 171, 174] IRL approaches, also struggling with missing trajectory information. Even state-marginal matching methods [116] optimize a reward solely to force a policy to match the expert density. If the policy fails to explore a region, the reward function never learns to value it, hindering the leveraging of the vast amount of unlabeled videos and documents available. IIRL reframes reward discovery as a data-centric learning problem; instead of requiring agents to explore the reward landscape through policy interaction, it learns an evaluative signal directly from expert data D_E, producing a score that reflects consistency with the expert distribution. This perspective prioritizes *editability* and *auditability*: the learned artifact can be inspected and corrected offline rather than being entangled with policy optimization. A concrete instance is the representation-based model of Malomgré and Simoens [104], which uses unlabeled data to learn a decoupled safe exploration signal and transforms it into an intrinsic reward via a separate mapping function. An expanded objective can be written as:

$$\max_{\theta,\psi} \left(\mathbb{E}_{s \sim D_E}\left[\beta(s,t)\, g_\psi(E_\theta(s))\right] - \mathbb{E}_{s \sim D_{\text{neg}}}\left[\beta(s,t)\, g_\psi(E_\theta(s))\right] \right), \tag{2}$$

where g_ψ is a tunable monotonically increasing mapping from the learned evaluative signal into reward [104], and $\beta(s,t)$ is a time-varying deployment coefficient that may be decayed globally, or locally via intrinsic-motivation discounting [7]. When available, D_{neg} both improves contrastive learning and enables off-support validation. Without it, non-expert regions remain only implicitly negative. The resulting objective encourages both *consistency*, by assigning a higher reward to in-distribution than out-of-distribution states, and *generalization*, by learning a smooth reward landscape rather than a sparse memorized detector, as illustrated in Figure 1 (right). Future research should look into augmenting the objective to support robust learning [11, 53, 128, 182], handling counterfactual context [45, 144, 162], or reducing distribution shift [4, 35, 86, 180].

Its modeling is modality-agnostic and supports causal context $R(s,c)$ [172]. To guide exploration toward natural behaviors without altering the objective of the main task, we employ Dynamic

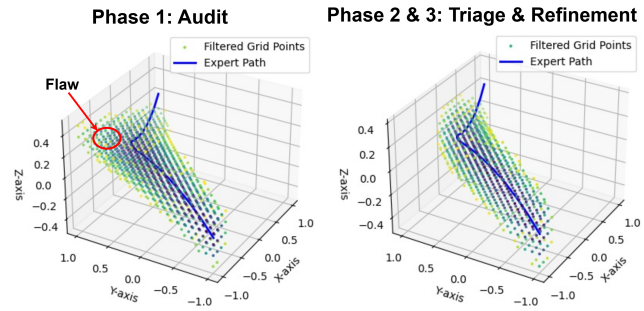

Figure 1: Alignment Flywheel in a 3D toy world. A representation-based IIRL model trained on sparse expert samples generates a reward landscape with g_ψ; yellow=low, purple=high. A spurious extrapolation (red circle) is detected in Phase 1 and corrected via refinement in Phases 2 & 3.

Potential-Based Reward Shaping [32]. Crucially, since the IIRL artifact Φ [114] continuously updates, the shaping reward takes the form of a time-dependent potential difference, $F_t(s,s') = \gamma\Phi_{t+1}(s') - \Phi_t(s)$. This structure ensures that, as the artifact evolves, it guides local behavior without altering the long-term optimal solution. Conversely, for inherently unsafe base policies, the artifact acts as a hard penalty, explicitly forcing deviation from unsafe trajectories.

2.2 Analysis of Architectures

The design of current IIRL architectures reflects a tradeoff between *editability* and *generalization*. Classic instance-based methods, such as k-NN [26, 59, 121] and SVMs [25, 34, 93, 105], offer native *editability* where localized data influence prevents unintended global consequences, yet often struggle with high-dimensional generalization. On the other hand, expressive deep models, such as representation-based [7, 14, 124] and Energy-Based Models (EBMs) [54, 66, 88, 164], provide state-of-the-art generalization but lack parameter *editability*, risking catastrophic forgetting during updates [167, 177]. This suggests a hybrid approach that combines deep learning with structured components, such as deep kernel methods [103, 126] or deep clustering [136, 163].

Another promising direction is replacing monolithic functions with a library of modular sub-priors, which yields intrinsic interpretability [15, 78]. High-level structures such as Reward Machines (RMs) [47, 74, 75, 150] orchestrate these by switching rewards based on context. RMs can be composed from demonstrations [9, 10, 148] or foundation models [3, 16, 70], and via Skill Machines can be executed zero-shot [146]. This enables practical auditability by refining sub-priors in isolation. Similarly, MoE [113] and RAG [6, 90] enhance *editability* by retrieving modular functions based on relational context [37].

2.3 The Refinement Toolkit

Achieving durable *editability* requires a diverse toolkit, organized here from currently feasible to high-potential future techniques.

First, to perform global adjustments without altering the IIRL parameters, we use functional sculpting. Here, the raw expertness score $L(s)$ from the IIRL model is passed through a separate, tunable function $R(s) = g_\psi(L(s))$ to produce the final reward, formalized

by Malomgré and Simoens [104]. By adjusting the parameters ψ, an expert can perform a global sculpt, changing the impact of certain levels of expert scores. We envision enhancing the mapping function to allow region-based sculpting. Furthermore, the mapping function mechanism can be upgraded to an RAG-based mapping function that uses the similarity representation and causal context to retrieve a function, neural network, or program.

Second, data-driven patching uses corrective data from audits to seed the IIRL model or to apply a localized patch, such as a kernel or representation patch trained to generalize across bad states. Ideally, the IIRL architecture supports *bipolar compatibility*, learning from positive and negative examples, and is *monotonic*, ensuring that new data improves or maintains safety without degrading it. Future work may adapt DPO for fine-grained IIRL parameter updates.

Lastly, for surgical internal modifications that alter model weights, the rapidly evolving fields of Model Editing and Unlearning show promise by demonstrating the fundamental feasibility of excising concepts from entangled weights, a critical capability for deep IIRL artifacts. However, deep representations are prone to catastrophic forgetting or collapse [167, 177]. To address this inherent architectural brittleness and the challenges of robust internal editing, the model edit toolbox comprises a broad spectrum of techniques, including architectural repair strategies such as model patching [21, 72, 100, 138] and GAN updates [13, 50], and also locate-and-edit methods [55, 56, 108, 109], neuron-level interventions [81], and null-space constrained edits [40, 101]. Other approaches reduce or avoid parameter destruction entirely, utilizing model merging [63, 99, 166], memory-based approaches [156], or contextual retrieval-based alternatives [20, 60, 62, 129]. Unlearning methods provide mechanisms for excising poor data or adapting to norm changes. The landscape now spans from theoretical frameworks like certified deletion [73, 84] to highly efficient, retrain-free solutions [2, 43, 44, 76, 112] and *learning to unlearn* paradigms [17, 68, 71, 123].

3 THE ALIGNMENT FLYWHEEL

While IIRL delivers a durable reward artifact, the Alignment Flywheel is the architectural blueprint for its continuous, verifiable hardening by transforming passive oversight into an active, iterative engineering lifecycle. This human-in-the-loop, multi-agent system orchestrates a proactive auditing and refinement process, ensuring the IIRL artifact evolves towards provably safer versions. The Flywheel's core power lies in its modality- and domain-agnostic design, which dynamically instantiates a scalable portfolio of auditing and refinement strategies tailored to the specific task's risk profile and data type. This enables the framework to adapt from simple heuristic validation for low-stakes robotics to comprehensive, neuro-symbolic multi-agent red-teaming for safety-critical LLM systems, establishing the discipline of Active Backward Alignment.

3.1 Phase 0: Seeding and Defining Constraints

First, the expert data is filtered against human-defined formal constraints [74, 127, 139]. After the data is used to seed a new IIRL reward model or update an existing one, the inferred formal constraints are obtained using a variety of domain- and modality-specific techniques, which can be broadly categorized. General-purpose methods include various neuro-symbolic synthesis [24, 33, 94, 152], and language-based approaches aim to convert text into formal specifications, such as converting natural language to executable rules [38, 46, 130, 149, 153] or to more nuanced social norms [46, 131]. Additionally, behavior-based methods infer constraints directly from demonstrations, including Inverse Constraint RL [28, 41, 96, 168] and automaton learning [9, 10, 148]. This entire inference process is enriched and grounded by Commonsense Knowledge Bases [36, 42, 147, 184]. Before the active loop begins, we perform a *coverage audit* to verify that the model captures the data well, and we can perform counterfactual checks to assess how the model behaves under changes in state or causal context.

3.2 Phase 1: Automated Auditing

A cooperative Multi-Agent System (MAS) audits the reward manifold under Phase 0 constraints, drawing inspiration from LLMs and cybersecurity. This MAS operates as a synergistic system in which a proactive Red Team [51, 65, 80, 106, 158] conducts adversarial attacks, while a strategic Blue Team [83, 118, 157, 179] provides high-level direction. This coordination is mediated by a Shared Flaw Knowledge Base (SFKB) [48, 161], a collective memory based on classical blackboard systems [117] to learn from others' experiences. The Blue Team identifies the Red Team's blind spots by populating the SFKB with coverage gaps and uncertainty metrics, thereby prompting the Red Team to focus on these newly identified regions. This transforms the audit from parallel random searches into a focused, intelligence-driven process. This dynamic necessitates a broad, adaptive adversarial capability, recognizing that effective alignment is a perpetual cat-and-mouse game in which reliance on a single technique creates predictable vulnerabilities. A mixed-initiative [67] auditing workbench governs the process, providing real-time controls to steer the audit and forensic tools [49, 159, 173] for post-hoc analysis of blind spots, ensuring accountability.

To counter this, the MAS dynamically tailors its Red Team strategies to the specific domain (e.g., robotics, LLMs) and input modality (e.g., images, text, vectors), ranging from simple heuristics to sophisticated GenAI techniques, to generate test cases that probe and stress-test the operational envelopes of predefined constraints and to freely explore to find novel flaws using coverage or uncertainty metrics. The test case generation ranges from brute-force programs, human crowdsourcing [49, 165], adversarial datasets [52, 64], to deploying advanced methods including training attacker models via RL to generate adversarial contexts [29, 69, 91, 125, 176] or use Bayesian optimization and discrete optimization [82, 89]. Additionally, perturbation adversarial attacks test changes to the input [18, 23, 79, 178], which can be extended to unrestricted adversarial attacks [22, 135, 140, 143]. The system validates fairness through counterfactual attribute checks, transforming bias detection from a passive observation into verifiable engineering constraints.

Complementing this proactive audit, the Blue Team leverages policy interactions with world models [92], test environments [98], and deployment for observational assurance. Its primary functions are coverage and uncertainty monitoring to detect state-space regions missed by the proactive audit to steer the Red Team to cover them [159, 173], and it directs small-scale, crowd-sourced red-teaming efforts to find novel, out-of-the-box vulnerabilities, targeting searches in those newly identified high-risk regions. Furthermore, it can scan

for reward tampering [39, 122] to identify trajectories that maximize rewards without corresponding task progress. Future research should examine additional auditing strategies, such as detecting auto-induced distribution shifts and goal misgeneralization.

3.3 Phase 2 & 3: Triage and Refinement

A core design principle is to treat the expert's attention as a scarce resource. During Triage, we use domain-aware semantic clustering and uncertainty- and diversity-based sampling [111] to group and prioritize sets of candidate flaws, thereby avoiding alert fatigue and maximizing the information value of each human intervention. When an expert flags a flaw, the system propagates that label to semantically similar items via label propagation and semi-supervised inference [85, 141, 145], resolving many related cases.

Next, during Refinement, we introduce Reward Modeling from Mixed Feedback (RM×F), where the specific correction mechanisms are tailored to the reward artifact's underlying architecture. First, the feedback-granularity spectrum ranges from minimal judgments (i.e., flaw or no flaw) to targeted corrections and expert-authored refinements that directly modify the reward manifold. Second, the agent-involvement spectrum ranges from only-human RMHF, through RM×F, which can include cooperative IRL [57, 78] agents that learn from and collaborate with experts to propose candidate refinements, to RMAIF, the fully automated mode in which an agent autonomously suggests or applies fixes.

Finally, every proposed refinement must pass an automated verification process, forming the core of our *verifiable safety via iterative hardening* approach. This involves two automated checks: a localized Red Team performs adversarial testing to ensure the fix introduces no new vulnerabilities. At the same time, regression tests against a library of known-good behaviors help prevent unintended side effects. The human expert's role is to review the automated results and provide final approval. Only refinements that pass this adversarial process are merged, producing the hardened, verifiably safer artifact that feeds into the next audit cycle.

4 APPLICATION

Our paradigm shows promise across diverse critical AI domains, including robotics, Multi-Agent Systems, and LLM alignment.

For Robotics and Avatar Animation, IIRL's ability to learn from large-scale, unlabeled video [12] enables the emergence of Foundation Reward Models (FRMs). These are dense reward fields for natural movement, which are then split into a library of compositional skills using computer vision or foundation models [154, 155]. A developer could specialize this library for their specific robot or avatar via a constraint file; the Alignment Flywheel would then adapt the reward manifold for safety and physical feasibility. The resulting artifact serves as a safe, auxiliary guidance signal to make training faster and more human-like and for pruning unsafe action sequences at runtime in world models [58]. This approach transforms the abstract embodiment gap into a concrete, iterative engineering task, yielding physically feasible, human-like agents.

For Multi-Agent Systems, IIRL offers a scalable solution for learning, unlearning, and maintaining social norms, moving beyond brittle hand-coded rules [8, 61]. IIRL artifacts can represent both shared societal values and an individual agent's beliefs, enabling both centralized and decentralized value systems. Initially, norms can be learned from datasets of successful interactions, providing a strong foundation while allowing for adaptive online policies. This positions IIRL as a living value system in which agents, through observation and negotiation, can actively update their value models by learning new norms and by seeding new samples. They can then propose these refinements to the Alignment Flywheel, which acts as a form of societal self-reflection to audit and refine beliefs, allowing the collective value system to evolve organically and safely.

For LLM alignment, IIRL provides a scalable architecture by reframing the problem as learning the underlying manifold of desirable language. Instead of learning from preferences over raw text strings, we first learn a dense, semantically rich representation of language using unsupervised methods [31, 133] on a static corpus, decomposing it into a modular feature space using techniques such as sparse autoencoders [27, 119] or discrete variational autoencoders [137, 175]. The IIRL artifact is constructed on this feature space, enabling targeted, non-black-box refinement and dynamic reward composition where a RAG model retrieves learned-relational-context-specific [37] reward modules, including simple programs, pre-audited neural networks, or formal constraints. These artifacts can then be used for traditional alignment or as a runtime guardrail, scoring and pruning misaligned reasoning branches as the LLM generates them. By guiding the model's output without altering its weights, this weight-free alignment directly circumvents the capability collapse [170], preserving the base model's full potential.

5 IMPLICATIONS AND VISION

Decoupling the reward artifact from the policy shifts Backward Alignment from passive evaluation to a modular engineering practice. Safety objectives become portable and reusable, allowing different institutions to build, certify, and deploy domain-specific behavioral priors for applications such as medicine or law. This enables a decentralized alignment supply chain built on inspectable artifacts rather than opaque, disposable objectives. The same separation also strengthens privacy and governance: because the artifact is not simply the policy itself and need not reveal its training data directly, it is more amenable to targeted revision and unlearning, including compliance with requirements such as the right to be forgotten.

Critically, a version-controlled foundation enables forensic root-cause analysis. When failures occur, the framework provides the ground truth to distinguish between flaws in the reward specification and errors in the agent's optimization. This crucial separation makes diagnostic findings from XAI tools actionable, enabling targeted, verifiable repairs rather than costly full-model retraining and satisfying the core traceability and accountability mandates of FATE. Furthermore, the Alignment Flywheel operationalizes RICE's technical objectives to support high-level societal mandates.

While challenges remain in formalizing the auditing workbench and in defining increasingly important governance criteria that address long-standing alignment artifacts rather than one-off artifacts, this framework establishes the necessary foundation for verifiable safety, shifting from reactive patching to proactive, durable design. This blueprint thus invites collaborative development to transform alignment from an abstract aspiration into concrete, verifiable engineering challenges for the AI community.

ACKNOWLEDGMENTS

This research was supported by funding from the Flemish Government under the "Onderzoeksprogramma Artificiele Intelligentie (AI) Vlaanderen" program.

REFERENCES

[1] Pieter Abbeel and Andrew Y. Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning* (Banff, Alberta, Canada) *(ICML '04)*. Association for Computing Machinery, New York, NY, USA, 1. https://doi.org/10.1145/1015330.1015430

[2] Sk Miraj Ahmed, Umit Yigit Basaran, Dripta S Raychaudhuri, Arindam Dutta, Rohit Kundu, Fahim Faisal Niloy, Basak Guler, and Amit K Roy-Chowdhury. 2025. Towards Source-Free Machine Unlearning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 4948–4957.

[3] Shayan Meshkat Alsadat, Jean-Raphaël Gaglione, Daniel Neider, Ufuk Topcu, and Zhe Xu. 2025. Using large language models to automate and expedite reinforcement learning with reward machine. In *2025 American Control Conference (ACC)*. IEEE, 206–211.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).

[5] Saurabh Arora and Prashant Doshi. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence* 297 (2021), 103500.

[6] Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A Survey on RAG with LLMs. *Procedia computer science* 246 (2024), 3781–3790.

[7] Arthur Aubret, Laetitia Matignon, and Salima Hassas. 2023. An information-theoretic perspective on intrinsic motivation in reinforcement learning: A survey. *Entropy* 25, 2 (2023), 327.

[8] Edmond Awad, Sydney Levine, Andrea Loreggia, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua Tenenbaum, and Max Kleiman-Weiner. 2024. When is it acceptable to break the rules? Knowledge representation of moral judgements based on empirical data. *Autonomous Agents and Multi-Agent Systems* 38, 2 (2024), 35.

[9] Mattijs Baert, Sam Leroux, and Pieter Simoens. 2024. Reward machine inference for robotic manipulation. *arXiv preprint arXiv:2412.10096* (2024).

[10] Mattijs Baert, Sam Leroux, and Pieter Simoens. 2025. Learning task specifications from demonstrations as probabilistic automata. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 8267–8274.

[11] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. 2021. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356* (2021).

[12] Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. 2022. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems* 35 (2022), 24639–24654.

[13] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. 2020. Rewriting a deep generative model. In *European conference on computer vision*. Springer, 351–369.

[14] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894* (2018).

[15] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine learning interpretability: A survey on methods and metrics. *Electronics* 8, 8 (2019), 832.

[16] Roger Creus Castanyer, Faisal Mohamed, Pablo Samuel Castro, Cyrus Neary, and Glen Berseth. 2025. ARM-FM: Automated Reward Machines via Foundation Models for Compositional Reinforcement Learning. *arXiv preprint arXiv:2510.14176* (2025).

[17] Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. 2024. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 11186–11194.

[18] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2021. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology* 6, 1 (2021), 25–45.

[19] Souradip Chakraborty, Amrit Singh Bedi, Alec Koppel, Dinesh Manocha, Huazheng Wang, Mengdi Wang, and Furong Huang. 2023. PARL: A unified framework for policy alignment in reinforcement learning from human feedback. *arXiv preprint arXiv:2308.02585* (2023).

[20] Qizhou Chen, Taolin Zhang, Xiaofeng He, Dongyang Li, Chengyu Wang, Longtao Huang, et al. 2024. Lifelong knowledge editing for llms with retrieval-augmented continuous prompt learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 13565–13580.

[21] Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. 2015. Net2net: Accelerating learning via knowledge transfer. *arXiv preprint arXiv:1511.05641* (2015).

[22] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. 2023. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems* 36 (2023), 51719–51733.

[23] Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2020. Seq2sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 3601–3608.

[24] Minwoo Cho, Jaehwi Jang, and Daehyung Park. 2025. ILCL: Inverse Logic-Constraint Learning from Temporally Constrained Demonstrations. *arXiv preprint arXiv:2507.11000* (2025).

[25] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20, 3 (1995), 273–297.

[26] Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory* 13, 1 (1967), 21–27.

[27] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600* (2023).

[28] Oliver Deane and Oliver Ray. 2025. Neuro-symbolic inverse constrained reinforcement learning. In *19th International Conference on Neurosymbolic Learning and Reasoning*.

[29] Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 3369–3391.

[30] Saurabh Deshpande, Rahee Walambe, Ketan Kotecha, Ganeshsree Selvachandran, and Ajith Abraham. 2025. Advances and applications in inverse reinforcement learning: a comprehensive review. *Neural Computing and Applications* (2025), 1–53.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, volume 1 (long and short papers)*. 4171–4186.

[32] Sam Michael Devlin and Daniel Kudenko. 2012. Dynamic potential-based reward shaping. In *11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*. IFAAMAS, 433–440.

[33] Paolo Dragone, Stefano Teso, and Andrea Passerini. 2021. Neuro-symbolic constraint programming for structured prediction. *arXiv preprint arXiv:2103.17232* (2021).

[34] Ke-Lin Du, Bingchun Jiang, Jiabin Lu, Jingyu Hua, and MNS Swamy. 2024. Exploring kernel machines and support vector machines: Principles, techniques, and future directions. *Mathematics* 12, 24 (2024), 3935.

[35] John C Duchi, Peter W Glynn, and Hongseok Namkoong. 2021. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research* 46, 3 (2021), 946–969.

[36] Bastien Dussard, Guillaume Sarthou, and Aurélie Clodic. 2025. Ontological component-based description of robot capabilities. *arXiv preprint arXiv:2306.07569* (2025).

[37] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130* (2024).

[38] William H English, Dominic Simon, Sumit Kumar Jha, and Rickard Ewetz. 2025. Grammar-Forced Translation of Natural Language to Temporal Logic using LLMs. In *Forty-second International Conference on Machine Learning*.

[39] Tom Everitt, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. Reward tampering problems and solutions in reinforcement learning: A causal influence diagram perspective. *Synthese* 198, Suppl 27 (2021), 6435–6467.

[40] Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2024. Alphaedit: Null-space constrained knowledge editing for language models. *arXiv preprint arXiv:2410.02355* (2024).

[41] Nan Fang, Guiliang Liu, and Wei Gong. 2025. Offline inverse constrained reinforcement learning for safe-critical decision making in healthcare. *IEEE Transactions on Artificial Intelligence* (2025).

[42] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social Chemistry 101: Learning to Reason about Social and Moral Norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 653–670. https://doi.org/10.18653/v1/2020.emnlp-main.48

[43] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. 2024. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 12043–12051.

[44] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. 2024. Loss-free machine unlearning. *arXiv preprint arXiv:2402.19308* (2024).

[45] Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. 2019. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742* (2019).

[46] Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 15217–15230.

[47] Daniel Furelos-Blanco, Mark Law, Anders Jonsson, Krysia Broda, and Alessandra Russo. 2023. Hierarchies of reward machines. In *International Conference on Machine Learning*. PMLR, 10494–10541.

[48] Parth Atulbhai Gandhi, Akansha Shukla, David Tayouri, Beni Ifland, Yuval Elovici, Rami Puzis, and Asaf Shabtai. 2025. ATAG: AI-Agent Application Threat Assessment with Attack Graphs. *arXiv preprint arXiv:2506.02859* (2025).

[49] Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858* (2022).

[50] Panpan Gao, Hanxu Sun, Gang Chen, and Minggang Li. 2025. Rewriting a Generative Model with Out-of-Domain Patterns. *Electronics* 14, 4 (2025), 675.

[51] Suyu Ge, Chunting Zhou, Rui Hou, Madian Khabsa, Yi-Chia Wang, Qifan Wang, Jiawei Han, and Yuning Mao. 2024. Mart: Improving llm safety with multi-round automatic red-teaming. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1927–1937.

[52] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462* (2020).

[53] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[54] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2019. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263* (2019).

[55] Akshat Gupta, Sidharth Baskaran, and Gopala Anumanchipalli. 2024. Rebuilding rome: Resolving model collapse during sequential model editing. *arXiv preprint arXiv:2403.07175* (2024).

[56] Akshat Gupta, Dev Sajnani, and Gopala Anumanchipalli. 2024. A unified framework for model editing. *arXiv preprint arXiv:2403.14236* (2024).

[57] Dylan Hadfield-Menell, Stuart J Russell, Pieter Abbeel, and Anca Dragan. 2016. Cooperative inverse reinforcement learning. *Advances in neural information processing systems* 29 (2016).

[58] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603* (2019).

[59] Rajib Kumar Halder, Mohammed Nasir Uddin, Md Ashraf Uddin, Sunil Aryal, and Ansam Khraisat. 2024. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data* 11, 1 (2024), 113.

[60] Xiaoqi Han, Ru Li, Hongye Tan, Wang Yuanlong, Qinghua Chai, and Jeff Pan. 2023. Improving sequential model editing with fact retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 11209–11224.

[61] Andreas Haupt, Phillip Christoffersen, Mehul Damani, and Dylan Hadfield-Menell. 2024. Formal contracts mitigate social dilemmas in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 38, 2 (2024), 51.

[62] Guoxiu He, Xin Song, and Aixin Sun. 2025. Knowledge updating? no more model editing! just selective contextual reasoning. *arXiv preprint arXiv:2503.05212* (2025).

[63] Yifei He, Yuzheng Hu, Yong Lin, Tong Zhang, and Han Zhao. 2024. Localize-and-stitch: Efficient model merging via sparse task arithmetic. *arXiv preprint arXiv:2408.13656* (2024).

[64] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15262–15271.

[65] Hannes Holm. 2022. Lore a red team emulation tool. *IEEE Transactions on Dependable and Secure Computing* 20, 2 (2022), 1596–1608.

[66] John J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences* 79, 8 (1982), 2554–2558.

[67] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.

[68] Shengyuan Hu, Yiwei Fu, Zhiwei Steven Wu, and Virginia Smith. 2025. Unlearning or obfuscating? jogging the memory of unlearned llms via benign relearning. *ICLR* (2025).

[69] Zijing Hu, Fengda Zhang, Long Chen, Kun Kuang, Jiahui Li, Kaifeng Gao, Jun Xiao, Xin Wang, and Wenwu Zhu. 2025. Towards better alignment: Training diffusion models with reinforcement learning against sparse rewards. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 23604–23614.

[70] Yun Hua, Haosheng Chen, Wenhao Li, Bo Jin, Baoxiang Wang, Hongyuan Zha, and Xiangfeng Wang. 2025. Reward Translation via Reward Machine in Semi-Alignable MDPs. In *Forty-second International Conference on Machine Learning*.

[71] Mark He Huang, Lin Geng Foo, and Jun Liu. 2024. Learning to unlearn for robust machine unlearning. In *European Conference on Computer Vision*. Springer, 202–219.

[72] Ferenc Huszár. 2018. Note on the quadratic penalties in elastic weight consolidation. *Proceedings of the National Academy of Sciences* 115, 11 (2018), E2496–E2497.

[73] Thanh Trung Huynh, Trong Bang Nguyen, Thanh Toan Nguyen, Phi Le Nguyen, Hongzhi Yin, Quoc Viet Hung Nguyen, and Thanh Tam Nguyen. 2025. Certified unlearning for federated recommendation. *ACM Transactions on Information Systems* 43, 2 (2025), 1–29.

[74] Rodrigo Toro Icarte, Toryn Klassen, Richard Valenzano, and Sheila McIlraith. 2018. Using reward machines for high-level task specification and decomposition in reinforcement learning. In *International Conference on Machine Learning*. PMLR, 2107–2116.

[75] Rodrigo Toro Icarte, Toryn Q Klassen, Richard Valenzano, and Sheila A McIlraith. 2022. Reward machines: Exploiting reward function structure in reinforcement learning. *Journal of Artificial Intelligence Research* 73 (2022), 173–208.

[76] Jinhyeok Jang, Jaehong Kim, and Chan-Hyun Youn. 2025. Learning to Rewind via Iterative Prediction of Past Weights for Practical Unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 26248–26255.

[77] Firas Jarboui and Vianney Perchet. 2021. Offline inverse reinforcement learning. *arXiv preprint arXiv:2106.05068* (2021).

[78] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2025. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852* (2025).

[79] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *arXiv preprint arXiv:1707.07328* (2017).

[80] Bojian Jiang, Yi Jing, Tong Wu, Tianhao Shen, Deyi Xiong, and Qing Yang. 2025. Automated progressive red teaming. In *Proceedings of the 31st International Conference on Computational Linguistics*. 3850–3864.

[81] Houcheng Jiang, Junfeng Fang, Tianyu Zhang, Baolong Bi, An Zhang, Ruipeng Wang, Tao Liang, and Xiang Wang. 2025. Neuron-level sequential editing for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16678–16702.

[82] Erik Jones, Anca Dragan, Aditi Raghunathan, and Jacob Steinhardt. 2023. Automatically auditing large language models via discrete optimization. In *International Conference on Machine Learning*. PMLR, 15307–15329.

[83] Krishna Mohan Kadambala. 2025. Auditable AI pipelines: Logging and verifiability in ML workflows. *Innovative Journal of Applied Science* (2025), 35–35.

[84] Anastasia Koloskova, Youssef Allouah, Animesh Jha, Rachid Guerraoui, and Sanmi Koyejo. 2025. Certified Unlearning for Neural Networks. *arXiv preprint arXiv:2506.06985* (2025).

[85] Georgios Kontonatsios, Austin J Brockmeier, Piotr Przybyła, John McNaught, Tingting Mu, John Y Goulermas, and Sophia Ananiadou. 2017. A semi-supervised approach using label propagation to support citation screening. *Journal of biomedical informatics* 72 (2017), 67–76.

[86] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-Distribution Generalization via Risk Extrapolation (REx). In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 5815–5826. https://proceedings.mlr.press/v139/krueger21a.html

[87] Filippo Lazzati, Mirco Mutti, and Alberto Maria Metelli. 2024. Offline inverse rl: New solution concepts and provably efficient algorithms. *arXiv preprint arXiv:2402.15392* (2024).

[88] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. 2006. A tutorial on energy-based learning. *Predicting structured data* 1, 0 (2006).

[89] Deokjae Lee, Seungyong Moon, Junhyeok Lee, and Hyun Oh Song. 2022. Query-efficient and scalable black-box adversarial attacks on discrete sequential data via bayesian optimization. In *International Conference on Machine Learning*. PMLR, 12478–12497.

[90] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.

[91] Chen Li, Huidong Tang, Ye Zhu, and Yoshihiro Yamanishi. 2025. A Reinforcement Learning-Driven Transformer GAN for Molecular Generation. *arXiv preprint arXiv:2503.12796* (2025).

[92] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent world representations: Exploring a sequence model trained on a synthetic task. *ICLR* (2023).

[93] Kun-Lun Li, Hou-Kuan Huang, Sheng-Feng Tian, and Wei Xu. 2003. Improving one-class SVM for anomaly detection. In *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, Vol. 5. IEEE, 3077–3081.

[94] Zenan Li, Yunpeng Huang, Zhaoyu Li, Yuan Yao, Jingwei Xu, Taolue Chen, Xiaoxing Ma, and Jian Lu. 2023. Neuro-symbolic learning yielding logical

constraints. *Advances in Neural Information Processing Systems* 36 (2023), 21635–21657.

[95] Yong Lin, Skyler Seto, Maartje Ter Hoeve, Katherine Metcalf, Barry-John Theobald, Xuan Wang, Yizhe Zhang, Chen Huang, and Tong Zhang. 2024. On the limited generalization capability of the implicit reward model induced by direct preference optimization. *arXiv preprint arXiv:2409.03650* (2024).

[96] Guiliang Liu, Sheng Xu, Shicheng Liu, Ashish Gaurav, Sriram Ganapathi Subramanian, and Pascal Poupart. 2024. A comprehensive survey on inverse constrained reinforcement learning: Definitions, progress and challenges. *arXiv preprint arXiv:2409.07569* (2024).

[97] Minghuan Liu, Tairan He, Minkai Xu, and Weinan Zhang. 2020. Energy-based imitation learning. *arXiv preprint arXiv:2004.09395* (2020).

[98] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. 2023. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688* (2023).

[99] Zhenyi Lu, Chenghao Fan, Wei Wei, Xiaoye Qu, Dangyang Chen, and Yu Cheng. 2024. Twin-merging: Dynamic integration of modular expertise in model merging. *Advances in Neural Information Processing Systems* 37 (2024), 78905–78935.

[100] Jian-Hao Luo and Jianxin Wu. 2020. Neural Network Pruning With Residual-Connections and Limited-Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[101] Sicheng Lyu, Yu Gu, Xinyu Wang, Jerry Huang, Sitao Luan, Yufei Cui, Xiao-Wen Chang, and Peng Lu. 2025. EvoEdit: Evolving Null-space Alignment for Robust and Efficient Knowledge Editing. *arXiv preprint arXiv:2510.13851* (2025).

[102] Monte MacDiarmid, Benjamin Wright, Jonathan Uesato, Joe Benton, Jon Kutasov, Sara Price, Naia Bouscal, Sam Bowman, Trenton Bricken, Alex Cloud, et al. 2025. Natural Emergent Misalignment from Reward Hacking in Production RL. *arXiv preprint arXiv:2511.18397* (2025).

[103] Julien Mairal, Piotr Koniusz, Zaid Harchaoui, and Cordelia Schmid. 2014. Convolutional kernel networks. *Advances in neural information processing systems* 27 (2014).

[104] Elias Malomgré and Pieter Simoens. 2025. Mixture of Autoencoder Experts Guidance using Unlabeled and Incomplete Data for Exploration in Reinforcement Learning. *arXiv preprint arXiv:2507.15287* (2025).

[105] Larry M Manevitz and Malik Yousef. 2001. One-class SVMs for document classification. *Journal of machine Learning research* 2, Dec (2001), 139–154.

[106] Ninareh Mehrabi, Palash Goyal, Christophe Dupuy, Qian Hu, Shalini Ghosh, Richard Zemel, Kai-Wei Chang, Aram Galstyan, and Rahul Gupta. 2024. Flirt: Feedback loop in-context red teaming. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 703–718.

[107] Bahar Memarian and Tenzin Doleck. 2023. Fairness, Accountability, Transparency, and Ethics (FATE) in Artificial Intelligence (AI) and higher education: A systematic review. *Computers and Education: Artificial Intelligence* 5 (2023), 100152.

[108] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems* 35 (2022), 17359–17372.

[109] Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229* (2022).

[110] Yuchun Miao, Sen Zhang, Liang Ding, Rong Bao, Lefei Zhang, and Dacheng Tao. 2024. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *Advances in Neural Information Processing Systems* 37 (2024), 134387–134429.

[111] Eduardo Mosqueira-Rey, Elena Hernández-Pereira, David Alonso-Ríos, José Bobes-Bascarán, and Ángel Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review* 56, 4 (2023), 3005–3054.

[112] Siqiao Mu and Diego Klabjan. 2025. Rewind-to-delete: Certified machine unlearning for nonconvex functions. *Advances in Neural Information Processing Systems* 38 (2025).

[113] Siyuan Mu and Sen Lin. 2025. A comprehensive survey of mixture-of-experts: Algorithms, theory, and applications. *arXiv preprint arXiv:2503.07137* (2025).

[114] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, Vol. 99. Citeseer, 278–287.

[115] Andrew Y Ng, Stuart Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[116] Tianwei Ni, Harshit Sikchi, Yufei Wang, Tejus Gupta, Lisa Lee, and Ben Eysenbach. 2021. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*. PMLR, 529–551.

[117] H Penny Nii. 1986. The blackboard model of problem solving and the evolution of blackboard architectures. *AI magazine* 7, 2 (1986), 38–38.

[118] Sang Ho Oh, Min Ki Jeong, Hyung Chan Kim, and Jongyoul Park. 2023. Applying reinforcement learning for enhanced cybersecurity against adversarial simulation. *Sensors* 23, 6 (2023), 3000.

[119] Charles O'Neill, Christine Ye, Kartheik Iyer, and John F Wu. 2024. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657* (2024).

[120] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.

[121] Gbeminiyi John Oyewole and George Alex Thopil. 2023. Data clustering: application and trends. *Artificial intelligence review* 56, 7 (2023), 6439–6475.

[122] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. *arXiv preprint arXiv:2201.03544* (2022).

[123] Gaurav Patel and Qiang Qiu. 2025. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4211–4221.

[124] Amirreza Payandeh, Kourosh T Baghaei, Pooya Fayyazsanavi, Somayeh Bakhtiari Ramezani, Zhiqian Chen, and Shahram Rahimi. 2023. Deep representation learning: Fundamentals, technologies, applications, and open challenges. *IEEE Access* 11 (2023), 137621–137659.

[125] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286* (2022).

[126] Gianluigi Pillonetto, Aleksandr Aravkin, Daniel Gedon, Lennart Ljung, Antonio H Ribeiro, and Thomas B Schön. 2025. Deep networks for system identification: a survey. *Automatica* 171 (2025), 111907.

[127] Amir Pnueli. 1977. The temporal logic of programs. In *18th annual symposium on foundations of computer science (sfcs 1977)*. ieee, 46–57.

[128] Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge Belongie, and Ser-Nam Lim. 2021. Robustness and generalization via generative adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15711–15720.

[129] Shanbao Qiao, Xuebing Liu, and Seung-Hoon Na. 2024. COMEM: In-Context Retrieval-Augmented Mass-Editing Memory in Large Language Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*. 2333–2347.

[130] Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. *Advances in Neural Information Processing Systems* 35 (2022), 9538–9551.

[131] Shilin Qu, Weiqing Wang, Xin Zhou, Haolan Zhan, Zhuang Li, Lizhen Qu, Linhao Luo, Yuan-Fang Li, and Gholamreza Haffari. 2025. Scalable Frame-Based Construction of Sociocultural Norm Bases for Socially Aware Dialogues. *ACM Transactions on Multimedia Computing, Communications and Applications* 21, 11 (2025), 1–17.

[132] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems* 36 (2023), 53728–53741.

[133] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[134] Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan Kim, et al. 2024. Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? *Advances in Neural Information Processing Systems* 37 (2024), 68559–68594.

[135] Yankun Ren, Jianbin Lin, Siliang Tang, Jun Zhou, Shuang Yang, Yuan Qi, and Xiang Ren. 2020. Generating natural language adversarial examples on a large scale with generative models. *arXiv preprint arXiv:2003.10388* (2020).

[136] Yazhou Ren, Jingyu Pu, Zhimeng Yang, Jie Xu, Guofeng Li, Xiaorong Pu, Philip S Yu, and Lifang He. 2024. Deep clustering: A comprehensive survey. *IEEE transactions on neural networks and learning systems* 36, 4 (2024), 5858–5878.

[137] Jason Tyler Rolfe. 2016. Discrete variational autoencoders. *arXiv preprint arXiv:1609.02200* (2016).

[138] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671* (2016).

[139] Luciano Serafini and Artur d'Avila Garcez. 2016. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. *arXiv preprint arXiv:1606.04422* (2016).

[140] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. 2020. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1151–1160.

[141] Yunyi Shen, Hao Sun, and Jean-François Ton. 2025. Reviving the classics: Active reward modeling in large language model alignment. *arXiv preprint arXiv:2502.04354* (2025).

[142] Joar Skalse and Alessandro Abate. 2024. Partial Identifiability and Misspecification in Inverse Reinforcement Learning. *arXiv preprint arXiv:2411.15951* (2024).

[143] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. 2018. Constructing unrestricted adversarial examples with generative models. *Advances in neural information processing systems* 31 (2018).

[144] Theodore R Sumers, Mark K Ho, Robert D Hawkins, Karthik Narasimhan, and Thomas L Griffiths. 2021. Learning rewards from linguistic feedback. In

Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35. 6002–6010.

[145] Boyuan Sun, Yuqi Yang, Le Zhang, Ming-Ming Cheng, and Qibin Hou. 2024. Corrmatch: Label propagation via correlation matching for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3097–3107.

[146] Geraud Nangue Tasse, Devon Jarvis, Steven James, and Benjamin Rosman. 2022. Skill machines: Temporal logic skill composition in reinforcement learning. *arXiv preprint arXiv:2205.12532* (2022).

[147] Jan-Philipp Töberg, Axel-Cyrille Ngonga Ngomo, Michael Beetz, and Philipp Cimiano. 2024. Commonsense knowledge in cognitive robotics: a systematic literature review. *Frontiers in Robotics and AI* 11 (2024), 1328934.

[148] Rodrigo Toro Icarte, Ethan Waldie, Toryn Klassen, Rick Valenzano, Margarita Castro, and Sheila McIlraith. 2019. Learning reward machines for partially observable reinforcement learning. *Advances in neural information processing systems* 32 (2019).

[149] Dimos Tsouros, Hélène Verhaeghe, Serdar Kadıoğlu, and Tias Guns. 2023. Holy grail 2.0: From natural language to constraint models. *arXiv preprint arXiv:2308.01589* (2023).

[150] Elena Umili, Francesco Argenziano, and Roberto Capobianco. 2024. Neural reward machines. *arXiv preprint arXiv:2408.08677* (2024).

[151] Lars van der Laan, Nathan Kallus, and Aurélien Bibaut. 2025. Inverse Reinforcement Learning Using Just Classification and a Few Regressions. *arXiv preprint arXiv:2509.21172* (2025).

[152] Marcell Vazquez-Chanlatte, Susmit Jha, Ashish Tiwari, Mark K Ho, and Sanjit Seshia. 2018. Learning task specifications from demonstrations. *Advances in neural information processing systems* 31 (2018).

[153] Chenhao Wang, Jiachun Li, Yubo Chen, Kang Liu, and Jun Zhao. 2025. A survey of recent advances in commonsense knowledge acquisition: Methods and resources. *Machine Intelligence Research* 22, 2 (2025), 201–218.

[154] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. 2022. Omnivl: One foundation model for image-language and video-language tasks. *Advances in neural information processing systems* 35 (2022), 5696–5710.

[155] Kun Wang, Jiuxin Cao, Biwei Cao, and Bo Liu. 2024. EnsCLR: Unsupervised skeleton-based action recognition via ensemble contrastive learning of representation. *Computer Vision and Image Understanding* 247 (2024), 104076.

[156] Ke Wang, Yiming Qin, Nikolaos Dimitriadis, Alessandro Favero, and Pascal Frossard. 2025. MEMOIR: Lifelong Model Editing with Minimal Overwrite and Informed Retention for LLMs. *arXiv preprint arXiv:2506.07899* (2025).

[157] Sheng Wang, Yao Li, and Feiyu Chen. 2024. Optimizing blue team strategies with reinforcement learning for enhanced ransomware defense simulations. *Authorea Preprints* (2024).

[158] Zhenduo Wang, Saifei Li, Lijie Zhang, Chunduo Hu, and Lianshan Yan. 2024. A Red Team automated testing modeling and online planning method for post-penetration. *Computers & Security* 144 (2024), 103945.

[159] Laura Weidinger, John FJ Mellor, Bernat Guillen Pegueroles, Nahema Marchal, Ravin Kumar, Kristian Lum, Canfer Akbulut, Mark Diaz, A Stevie Bergman, Mikel D Rodriguez, et al. 2024. Star: Sociotechnical approach to red teaming language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 21516–21532.

[160] Feiyang Wu, Ye Zhao, and Anqi Wu. 2025. Distributional Inverse Reinforcement Learning. *arXiv preprint arXiv:2510.03013* (2025).

[161] Shanglin Wu and Kai Shu. 2025. Memory in LLM-based Multi-agent Systems: Mechanisms, Challenges, and Collective Intelligence. https://doi.org/10.13140/RG.2.2.21084.04485

[162] Yueh-Hua Wu and Shou-De Lin. 2018. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[163] Junyuan Xie, Ross Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *International conference on machine learning*. PMLR, 478–487.

[164] Jianwen Xie, Yang Lu, Ruiqi Gao, Song-Chun Zhu, and Ying Nian Wu. 2018. Cooperative training of descriptor and generator networks. *IEEE transactions on pattern analysis and machine intelligence* 42, 1 (2018), 27–45.

[165] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2021. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2950–2968.

[166] Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023. Ties-merging: Resolving interference when merging models. *Advances in Neural Information Processing Systems* 36 (2023), 7093–7115.

[167] Wanli Yang, Fei Sun, Xinyu Ma, Xun Liu, Dawei Yin, and Xueqi Cheng. 2024. The butterfly effect of model editing: Few edits can trigger large language models collapse. *arXiv preprint arXiv:2402.09656* (2024).

[168] Bo Yue, Shufan Wang, Ashish Gaurav, Jian Li, Pascal Poupart, and Guiliang Liu. 2025. Understanding constraint inference in safety-critical inverse reinforcement learning. In *The Thirteenth International Conference on Learning Representations*.

[169] Sheng Yue, Guanbo Wang, Wei Shao, Zhaofeng Zhang, Sen Lin, Ju Ren, and Junshan Zhang. 2023. Clare: Conservative model-based reward learning for offline inverse reinforcement learning. *arXiv preprint arXiv:2302.04782* (2023).

[170] Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837* (2025).

[171] Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. 2022. Maximum-likelihood inverse reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing Systems* 35 (2022), 10122–10135.

[172] Yan Zeng, Ruichu Cai, Fuchun Sun, Libo Huang, and Zhifeng Hao. 2024. A survey on causal reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems* (2024).

[173] Jinchuan Zhang, Yan Zhou, Yaxin Liu, Ziming Li, and Songlin Hu. 2024. Holistic automated red teaming for large language models through top-down test case generation and multi-turn interaction. *arXiv preprint arXiv:2409.16783* (2024).

[174] Ruijia Zhang, Siliang Zeng, Chenliang Li, Alfredo Garcia, and Mingyi Hong. 2025. Understanding Inverse Reinforcement Learning under Overparameterization: Non-Asymptotic Analysis and Global Optimality. *arXiv preprint arXiv:2503.17865* (2025).

[175] Yingji Zhang, Danilo Carvalho, Marco Valentino, Ian Pratt-Hartmann, and Andre Freitas. 2024. Improving Semantic Control in Discrete Latent Spaces with Transformer Quantized Variational Autoencoders. In *Findings of the Association for Computational Linguistics: EACL 2024*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 1434–1450. https://aclanthology.org/2024.findings-eacl.97/

[176] Yinan Zhang, Eric Tzeng, Yilun Du, and Dmitry Kislyuk. 2024. Large-scale reinforcement learning for diffusion models. In *European Conference on Computer Vision*. Springer, 1–17.

[177] Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. 2024. What makes unlearning hard and what to do about it. *Advances in Neural Information Processing Systems* 37 (2024), 12293–12333.

[178] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2023. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems* 36 (2023), 54111–54138.

[179] Yunhan Zhao, Xiang Zheng, Lin Luo, Yige Li, Xingjun Ma, and Yu-Gang Jiang. 2024. Bluesuffix: Reinforced blue teaming for vision-language models against jailbreak attacks. *arXiv preprint arXiv:2410.20971* (2024).

[180] Rui Zheng, Wei Shen, Yuan Hua, Wenbin Lai, Shihan Dou, Yuhao Zhou, Zhiheng Xi, Xiao Wang, Haoran Huang, Tao Gui, et al. 2023. Improving generalization of alignment with human preferences through group invariant learning. *arXiv preprint arXiv:2310.11971* (2023).

[181] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.

[182] Daniel Ziegler, Seraphina Nix, Lawrence Chan, Tim Bauman, Peter Schmidt-Nielsen, Tao Lin, Adam Scherlis, Noa Nabeshima, Benjamin Weinstein-Raun, Daniel de Haas, et al. 2022. Adversarial training for high-stakes reliability. *Advances in neural information processing systems* 35 (2022), 9274–9286.

[183] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* (2019).

[184] Caleb Ziems, Jane Dwivedi-Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2023. NormBank: A knowledge bank of situational social norms. *arXiv preprint arXiv:2305.17008* (2023).