Toward Reliable Agents

Hua Wei
School of Computing and Augmented Intelligence, Arizona State University
hua.wei@asu.edu

Abstract

Learned agents that control traffic signals or call software tools are usually trained in simulators or fixed benchmarks, yet must act in worlds that differ. Many reliability concerns (robustness, safety, alignment, etc.) become one problem once agents are written as Markov decision processes (MDPs): a gap between the world an agent was built for and the world it acts in. The sim-to-real gap decomposes along state, observation, action, transition, and reward; multi-agent deployment further adds topology, population, and task gaps as neighbors learn and workloads scale. Using reinforcement learning (RL) for traffic control and large language model (LLM) agents for tool use and mobile GUIs as running examples, this paper traces how mismatches appear on each channel and how mitigations from one community (domain randomization, grounded action transformation, perturbation benchmarks) transfer to the other. Because the gap rarely closes, reliable deployment needs selective human oversight; uncertainty quantification can flag when to escalate, localize errors at the answer, reasoning, step, or agent level, and support sim-to-real transfer in RL. This paper opens an agenda for a shared MDP vocabulary, uncertainty-aware oversight, and scaling human-AI collaboration across volume, complexity, and expertise.

1 Introduction

An agent is something that *acts* in the classical sense [Shardlow, 1990]. Increasingly, the agents that act on our behalf are *learned*. Reinforcement learning (RL) and, lately, large language model (LLM) agents now make decisions in domains from transportation and energy to public health and software automation. Their appeal is common: agents that learn good decisions from data or experience, rather than following hand-written rules, and before handing such an agent a decision that matters, we need to trust that it will act reliably once deployed. The question we keep returning to is when that trust is warranted.

Reliability nowadays goes by many names (alignment, robustness, safety, verification, etc.), each with its own vocabulary and toolbox. A theme of this paper is that, once agents are written as Markov decision processes (MDPs), many of these names describe one thing: a *gap* between the world an agent was built for and the world it acts in. For example, alignment is a gap between the reward it optimizes and the one we intend; robustness, a possible gap in what the agent observes and transitions; safety, a constraint that breaks when things change. Even formal guarantees hold only relative to a model, and so are bounded by the same gap. Stripped of the method-specific names, the problem is the gap itself.

The most conspicuous of these gaps comes from how we build these agents. Almost every agent for a physical system is trained in a *simulator*, because collecting experience in the world is slow, expensive, or unsafe. Since simulators are approximations, the mismatch between the simulator an agent trains in and the world it is deployed to, known as the *sim-to-real gap*, has long blocked deployment of learned policies [Zhao *et al.*, 2020]. A policy that is optimal in simulation may violate safety constraints, fail when observations are missing, or simply underperform in reality.

This paper structures the discussion around three points. First, positioning agents as Markov decision processes (MDPs) exposes the sim-to-real gap for RL and for LLM agents alike as shifts along five channels: state, observation, action, transition, and reward (Figure 1 and Section 2). Second, many interacting agents compound the problem through additional topology, population, and task gaps (Section 2). Third, because the gap rarely disappears entirely, reliable deployment needs a human in the loop, and uncertainty quantification can make that oversight selective rather than constant (Section 3). As an early career spotlight, this paper draws on our own research and presents the broader literature where it sets context.

2 A Unified MDP Formulation

A standard abstraction for a decision-making agent is the Markov decision process $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$: at each step the agent observes a *state* $s \in \mathcal{S}$, takes an *action* $a \in \mathcal{A}$, transitions according to $P(s' \mid s, a)$, and receives a scalar *reward* $R(s, a)$, seeking a policy that maximizes expected discounted return [Sutton and Barto, 2018]. For a policy π trained in M_s and evaluated in M_r, the sim-to-real gap can be written as $G(\pi) = \psi_s(\pi) - \psi_r(\pi)$ for any deployment metric ψ [Liu *et al.*, 2026a]: mitigating the gap means reduc-

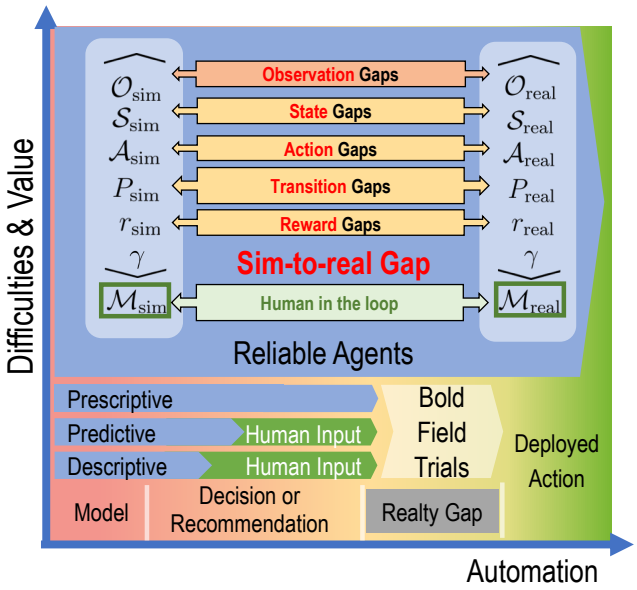

Figure 1: Reliable deployment as a trade-off between *value* and *difficulties*. Horizontally, value rises as systems move from descriptive monitoring through predictive recommendations to prescriptive agents whose outputs become deployed actions; field trials and human input sit on that path. Vertically, difficulties track the sim-to-real gap between M_{sim} and M_{real} across state, observation, action, transition, and reward. Human oversight bridges the two MDPs when gaps cannot be closed in simulation alone.

ing this performance drop, not eliminating every discrepancy. Techniques from one community may transfer to the other; we have tried to make that argument precise for foundation-model agents [Liu *et al.*, 2026a; Da *et al.*, 2026b].

Instantiating the MDP for RL and LLM agent. The MDP formulation is the same for both RL and LLM agents. With an RL agent for control problems like traffic signal control (TSC) [Wei *et al.*, 2018; Wei *et al.*, 2019a], the agent sees a traffic condition (e.g., per-lane queue lengths, vehicle counts, etc.), chooses signal timings, and receives a reward so that maximizing return reduces average travel time. With an LLM agent for problems like tool calling, observations are the prompt, context, and tool responses; actions are generated text and tool calls; transitions are whatever the environment returns; reward is task success or downstream utility.

With that vocabulary in place, the same failure mode appears in both RL and LLM settings. A policy trained in a simulated MDP M_s and deployed in a real MDP M_r degrades when the two processes disagree on what the agent sees, can do, how the world evolves, or what should be optimized [Liu *et al.*, 2026a]. Following prior sim-to-real works [Zhao *et al.*, 2020; Da *et al.*, 2026b], some basic discrepancies can be organized into five gaps, as shown in Figure 1: (1) a *state gap* when the latent variables that define $s \in \mathcal{S}$ differ between simulation and deployment. For example, a traffic simulator whose state omits pedestrian calls or transit priority that the real controller tracks; (2) an *observation gap* when the agent's reading of state shifts even if $\mathcal{S}$ is shared, e.g., noisy loop detectors versus perfect counts in simulation, or a paraphrased prompt versus the user's intent in tool calling; (3) an *action gap* in $\mathcal{A}$, (4) a *transition gap* in T (or $P(s' \mid s, a)$), and (5) a *reward gap* in R. The state and observation gaps are related under partial observability, but separating them helps locate the fix: widen $\mathcal{S}$, improve sensors, or change the observation map.

From One Agent to Many. Real deployments are rarely a single agent but a system of agents. A city has thousands of intersections [Chen *et al.*, 2020]; LLM workflows split work across roles [Yao *et al.*, 2026]. In a multi-agent MDP, each agent's transitions and rewards depend on the others, which compounds the sim-to-real gap. From one agent's view, the others are part of the environment; as they learn, T becomes non-stationary and the transition gap widens [Buşoniu *et al.*, 2008]. Splitting a system-wide objective into local rewards also introduces reward gaps when the decomposition assumed in simulation does not match deployment. Beyond the five channel gaps, multi-agent settings add structural mismatches: a *topology gap* when the interaction graph differs (a small grid in simulation versus the actual road network), a *population gap* when the number or heterogeneity of agents changes, and a *task gap* when work is partitioned differently between benchmark roles and live workflows. In our multi-agent TSC work [Wei *et al.*, 2019b; Ruan *et al.*, 2024; Turnau *et al.*, 2025], we find that proper coordination reduces transfer loss under different network structures.

2.1 The Sim-to-Real Gap in RL

For RL control, the gap shows up along each channel. We use traffic signal control (TSC) as a running example. Figure 2 traces how RL research on TSC has shifted over time: from single intersections and small networks toward multi-agent coordination, and now toward sim-to-real transfer. The *state gap* appears when the simulator's state vector omits variables the street tracks, such as pedestrian phases, emergency preemption, or transit priority not encoded in s_{sim}. The *observation gap* is mismatch in what the agent reads from state: a simulator may expose exact per-lane vehicle counts (e.g., $[0, 1, 0, 2]$), while a deployed intersection reads loop detectors that are miscalibrated or broken [Mei *et al.*, 2023; Chen *et al.*, 2024a], or camera-based sensing rather than loop detectors [Chen *et al.*, 2024b]. The *action gap* arises when actions valid in simulation do not execute faithfully on the street (e.g., discretized signal phases versus actuation delay). The *transition gap* appears when an action's effect on the next state differs between simulation and reality; for example, vehicles move more slowly in rain or snow, so a fixed green interval clears fewer vehicles than the simulator predicts. The *reward gap* arises when the return optimized in simulation does not match what deployment values; for instance, a policy trained only to minimize average travel time may be unsafe on the street. In SafeLight [Du *et al.*, 2023] we narrowed this gap by reshaping R: encoding collision avoidance in the reward and action space, rather than bolting safety on afterward, improved both mobility and safety.

Mitigation strategies and research infrastructure. Training-time fixes include domain randomization [Tobin *et al.*, 2017] and dynamics randomization [Peng *et al.*, 2018], which train against a distribution of simulators.

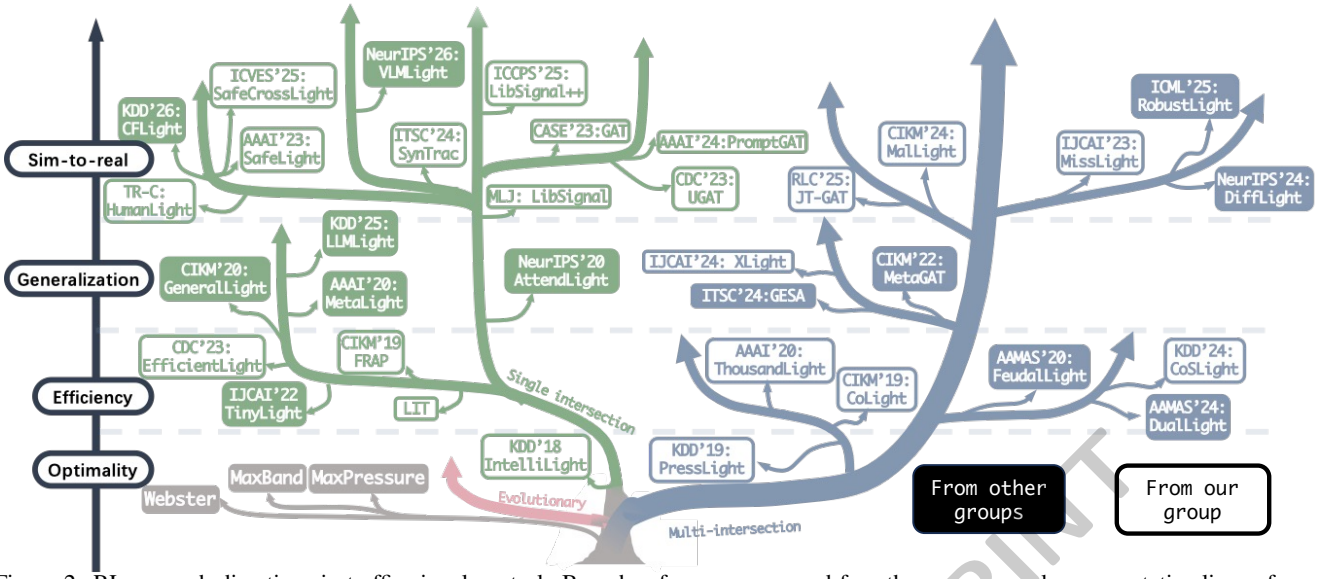

Figure 2: RL research directions in traffic signal control. Branches for our group and for other groups mark representative lines of work; sim-to-real transfer is the current focus of this spotlight.

We have also grounded simulated *transitions* in real data with learned inverse dynamics models [Da *et al.*, 2023a; Da *et al.*, 2023b; Da *et al.*, 2024b; Turnau *et al.*, 2025], and policy adaptation [Da *et al.*, 2026a; Chen *et al.*, 2026a]. We also developed benchmarks and surveys [Wei *et al.*, 2020; Da *et al.*, 2026b], and ecosystems for sim-to-real transfer, including LibSignal [Mei *et al.*, 2024] and a low-cost testbed [Zhang *et al.*, 2025] for evaluating transfer.

2.2 The Sim-to-Real Gap in LLM Agents

Benchmarks and sandboxes as M_s. The decomposition applies to LLM agents once their simulator is seen as a fixed development distribution (the benchmarks, prompt templates, and API sandboxes an agent is built against [Liu *et al.*, 2026a]) that deployment departs from along the same channels. AndroidReality [Liu *et al.*, 2026c] makes this concrete for mobile GUI agents, treating real deployment as a *family* of perturbed MDPs that share the task but vary one component at a time: *observation*-gap changes to the screen surface (font size, light/dark theme, orientation) hurt the most, as agents keep tapping coordinates from the old layout and fail to reground; *transition*-gap pop-ups, consent sheets, and permission dialogs insert unexpected states; and *action*-gap delays or dropped taps hurt least, since agents already retry. In multilingual tool calling [Luo *et al.*, 2026], an *observation*-gap shift, e.g., posing a query in Chinese, Hindi, or Igbo instead of English, yields calls with the right intent but a parameter *value* left in the user's language, breaking the convention that values stay language-invariant; inference-time translation reduces these errors but never fully recovers English-level accuracy. A *reward* gap appears when agents tuned for binary benchmark success meet a deployment that also rewards latency and cost [Zhou *et al.*, 2026].

Mitigation recipes across communities. As in RL, no single knob closes the gap; the fix depends on *where* the mismatch sits, but the RL playbook transfers: domain and dynamics randomization have analogues in prompt, schema, and sandbox perturbation; grounded action transformation parallels tool grounding and the domain-randomized transfer recipe of [Zhou *et al.*, 2026]; and test-time adaptation mirrors AndroidReality's training-free introspective recovery [Liu *et al.*, 2026c].

3 Keeping Humans in the Loop with Uncertainty Quantification

Since the simulator can never be assumed perfect, the gap does not go away, and multiple agents make it worse. In practice, we need to keep humans in the loop. However, human attention is bandwidth-limited. Therefore, escalation should happen when the agent is likely wrong rather than on every decision. That is where uncertainty quantification helps and motivates another large body of research [Liu *et al.*, 2025].

3.1 Uncertainty for RL Sim-to-Real

Most of the sim-to-real gap in RL control sits in the transition channel: the policy chooses an action, but whether the next state matches training depends on traffic, weather, and infrastructure that simulators simplify. Grounded action transformation (GAT) [Da *et al.*, 2023a] learns from real trajectories to rewrite simulator actions so their street-level effects better match deployment. That addresses average mismatch, yet a point estimate can still be wrong when conditions shift, e.g., rain, sensor dropout, or an intersection absent from the training logs.

UGAT [Da *et al.*, 2023b] treats the grounding model itself as uncertain. By quantifying uncertainty over the sim-to-real transition map, the agent can down-weight unreliable corrections, fall back to safer behavior, or request human review when local dynamics look out of distribution. Uncertainty

thus serves two roles: it improves transfer by avoiding overconfident adaptation, and it supplies an escalation signal for human-in-the-loop deployment.

The same logic compounds in multi-agent settings. JT-GAT [Turnau *et al.*, 2025] learns joint-local grounded transformations so each intersection maintains its own uncertainty over how neighbors and local flow respond, rather than sharing one global correction. That structure matches decentralized TSC deployments and helps localize which intersection, not just which timestep, warrants operator attention. Follow-on work conditions grounding on context (PromptGAT [Da *et al.*, 2024b]), adapts foundation policies in latent space [Da *et al.*, 2026a], and uses diffusion for cross-domain policy adaptation [Chen *et al.*, 2026a]; across these lines, knowing when the transition model is trustworthy remains the hinge between sim-trained policies and field-ready oversight.

3.2 Uncertainty for LLM Agents

LLM agents can surface far more text, tools, and intermediate reasoning than a human can review. Uncertainty quantification compresses that volume into escalation signals. Recent studies have looked into this at the answer, reasoning-trace, step, and multi-agent levels [Liu *et al.*, 2025].

Answer level. A black-box estimator samples several answers to one question and asks how much they agree. Semantic entropy [Kuhn *et al.*, 2023] groups answers into meaning-equivalence classes with a natural-language-inference model and scores the spread; our directional-entailment variant [Da *et al.*, 2024a] replaces symmetric similarity with a *directed* entailment graph and reads dispersion from its Laplacian spectrum, with claim-level augmentation to tame vague propositions. Uncertainty can also steer alignment: Conformal Feedback Alignment [Chen *et al.*, 2026c] converts per-answer reliability into a conformal-prediction weight on preference pairs, down-weighting low-confidence comparisons during post-training alignment. Stepping back, a position paper [Chen *et al.*, 2026b] argues this whole family is mechanically *unsupervised clustering*: useful, but it measures self-consistency, not truth, so it can certify a "confident hallucination."

Reasoning level. Answer scores say *whether* an output is trustworthy, not *why*. [Da *et al.*, 2025] proposes to elicit an explanation as a directed acyclic graph of sub-questions and answers, then quantify structural uncertainty by graph edit distance across sampled topologies: high variation flags fragile logic and exposes redundant steps.

Step level. For multi-step reasoning, the hard part is finding *which* step failed, while separating legitimate variation among correct traces from genuine errors. [Liu *et al.*, 2026b] frames this as an information bottleneck, compressing a trajectory to a sparse, confidence-weighted set of steps that still predicts final-answer correctness, using the *logical invariants* shared by correct traces as the signal. A training-free variant (NIBS) scores each step by similarity to reference trajectories; a graph variant (GIBS) learns a differentiable step mask and stays robust under domain shift. The payoff is qualitative: pointing an agent at the faulty *step* lifts self-correction by up to 13.5% over a binary "the answer is wrong" signal.

Agent level. When specialized agents collaborate, each can look confident while the team is misaligned and errors cascade across handoffs. [Chen *et al.*, 2026d] stacks interaction histories over repeated runs into a ragged agents-by-steps-by-runs tensor and applies PARAFAC2 decomposition [Kiers *et al.*, 1999]; coherent collaborations stay low-rank, so a large reconstruction error signals system-level uncertainty and points to the agent role that needs review.

Across these levels, uncertainty serves three roles in deployment: deciding when to call in a human, detecting that something is going wrong, and diagnosing where, leaving *how to scale* that oversight as the open problem the next research agenda should address.

4 Research Agenda and Outlook

Three directions seem worth pursuing.

A shared MDP vocabulary. Writing RL, robotics, and LLM/VLM agents as MDPs makes their observation, action, and dynamics gaps explicit, and lets a mitigation developed in one community carry over to another. This common language, more than any single method, is what turns a scattered set of reliability concerns into one problem.

Uncertainty-aware oversight. Because the gap rarely closes, reliable deployment needs agents that quantify their uncertainty, localize likely errors, and ask for a human where it helps; making the estimators of Section 3 cheap enough to run continuously is open.

Scaling human-AI collaboration. Human attention does not scale with the number of agents, the length of a trajectory, or a reviewer's expertise; the regime where *alert fatigue* and *automation bias* set in [Parasuraman and Riley, 1997; Amershi *et al.*, 2019]. three axes should be considered to separate the bottleneck: *volume* (e.g., LLM-as-judge at scale), *complexity* (e.g., long-horizon or multi-agent work, where the hard part is locating the step or agent at fault [Liu *et al.*, 2026b; Chen *et al.*, 2026d]), and *expertise depth* (knowing which claims need outside validation [Liu *et al.*, 2025]); against them, three levers: *compress* what a human reads, *visualize* it with context, and *train* reviewers to spend attention well, with uncertainty steering what to summarize, where to focus, and when to escalate [Da *et al.*, 2025]. Here *when* and *how* to escalate matter as much as *whether*, and even low-uncertainty cases have value: reviewing confident outputs teaches a workforce what "good" looks like. One application in education domain [Yao *et al.*, 2026] shows the promising potential of specialized roles with human review at the prescriptive end. Making that division of labor systematic across volume, complexity, and expertise is the central challenge, and we hope this MDP-centered view helps others build agents for the real world.

Ethical Statement

This work concerns agents deployed in settings such as traffic control. Better reliability and uncertainty estimates should reduce harm from overconfident or mis-transferred policies. They are not a substitute for human judgment; deployment should keep human oversight in place.

Acknowledgements

This work is supported by the U.S. National Science Foundation under grants #2442477 and #2550203; any opinions, findings, and conclusions expressed here are the author's and do not necessarily reflect those of the funding agencies. I thank my students, collaborators, mentors, and advisors, along with Arizona State University and our academic and industry partners, for their support, and the IJCAI 2026 Program Committee for inviting me to deliver this Early Career Spotlight talk.

References

[Amershi *et al.*, 2019] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, et al. Guidelines for human-AI interaction. In *CHI Conference on Human Factors in Computing Systems*, 2019.

[Buşoniu *et al.*, 2008] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(2):156–172, 2008.

[Chen *et al.*, 2020] Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, and Zhenhui Li. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[Chen *et al.*, 2024a] Hanyang Chen, Yang Jiang, Shengnan Guo, Xiaowei Mao, Youfang Lin, and Huaiyu Wan. DiffLight: A partial rewards conditioned diffusion model for traffic signal control with missing data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[Chen *et al.*, 2024b] Tiejin Chen, Prithvi Parag Shirke, Bharatesh Chakravarthi, Arpitsinh Vaghela, Longchao Da, Duo Lu, Yezhou Yang, and Hua Wei. SynTraC: A synthetic dataset for traffic signal control from traffic monitoring cameras. In *Proceedings of the 27th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2024.

[Chen *et al.*, 2026a] Hanyang Chen, Anirudh Satheesh, Longchao Da, and Hua Wei. DADiff: Diffusion-driven cross-domain policy adaptation for reinforcement learning. In *Proceedings of the 2026 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2026.

[Chen *et al.*, 2026b] Tiejin Chen, Longchao Da, Xiaoou Liu, and Hua Wei. Position: Uncertainty quantification in LLMs is just unsupervised clustering. In *International Conference on Machine Learning (ICML)*, 2026.

[Chen *et al.*, 2026c] Tiejin Chen, Xiaoou Liu, Vishnu Nandam, Kuan-Ru Liou, and Hua Wei. Conformal feedback alignment: Quantifying answer-level reliability for robust LLM alignment. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026.

[Chen *et al.*, 2026d] Tiejin Chen, Huaiyuan Yao, Jia Chen, Evangelos E. Papalexakis, and Hua Wei. Every response counts: Quantifying uncertainty of LLM-based multi-agent systems through tensor decomposition. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

[Da *et al.*, 2023a] Longchao Da, Hao Mei, Romir Sharma, and Hua Wei. Sim2real transfer for traffic signal control. In *Proceedings of the 19th IEEE International Conference on Automation Science and Engineering (CASE)*, 2023.

[Da *et al.*, 2023b] Longchao Da, Hao Mei, Romir Sharma, and Hua Wei. Uncertainty-aware grounded action transformation towards sim-to-real transfer for traffic signal control. In *Proceedings of the 62nd IEEE Conference on Decision and Control (CDC)*, 2023.

[Da *et al.*, 2024a] Longchao Da, Tiejin Chen, Lu Cheng, and Hua Wei. LLM uncertainty quantification through directional entailment graph and claim level response augmentation. *arXiv preprint arXiv:2407.00994*, 2024.

[Da *et al.*, 2024b] Longchao Da, Minquan Gao, Hao Mei, and Hua Wei. Prompt to transfer: Sim-to-real transfer for traffic signal control with prompt learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[Da *et al.*, 2025] Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. Understanding the uncertainty of LLM explanations: A perspective based on reasoning topology. In *Conference on Language Modeling (COLM)*, 2025.

[Da *et al.*, 2026a] Longchao Da, T Pranav Kutralingam, Lirong Xiang, and Hua Wei. Latent adaptation of foundation policies for sim-to-real transfer. In *Proceedings of the Fourteenth International Conference on Learning Representations (ICLR)*, 2026.

[Da *et al.*, 2026b] Longchao Da, Justin Turnau, Thirulogasankar Pranav Kutralingam, Alvaro Velasquez, Paulo Shakarian, and Hua Wei. A survey of sim-to-real methods in RL: Progress, prospects and challenges with foundation models, 2026. Preprint.

[Du *et al.*, 2023] Wenlu Du, Junyi Ye, Jingyi Gu, Jing Li, Hua Wei, and Guiling Wang. SafeLight: A reinforcement learning method toward collision-free traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14801–14810, 2023.

[Kiers *et al.*, 1999] Henk AL Kiers, Jos MF Ten Berge, and Rasmus Bro. Parafac2—part i. a direct fitting algorithm for the parafac2 model. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 13(3-4):275–294, 1999.

[Kuhn *et al.*, 2023] Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

[Liu *et al.*, 2025] Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. Uncertainty quantification and confidence calibration in large language

models: A survey. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2025.

[Liu *et al.*, 2026a] Xiaoou Liu, Tiejin Chen, Weibo Li, Xiyang Hu, and Hua Wei. The sim-to-real gap of foundation model agents: A unified MDP perspective. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), Blue Sky Ideas Track*, 2026. arXiv:2606.07017.

[Liu *et al.*, 2026b] Xiaoou Liu, Tiejin Chen, Dengjia Zhang, Yaqing Wang, Lu Cheng, and Hua Wei. Diagnosing multi-step reasoning failures in black-box LLMs via stepwise confidence attribution. In *International Conference on Machine Learning (ICML)*, 2026.

[Liu *et al.*, 2026c] Xiaoou Liu, Longchao Da, Hanyang Chen, Yuan Ling, and Hua Wei. AndroidReality: How far are mobile agents from the real world?, 2026.

[Luo *et al.*, 2026] Zheng Luo, T Pranav Kutralingam, Ogochukwu N. Okoani, Wanpeng Xu, Hua Wei, and Xiyang Hu. Lost in execution: On the multilingual robustness of tool calling in large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

[Mei *et al.*, 2023] Hao Mei, Junxian Li, Bin Shi, and Hua Wei. Reinforcement learning approaches for traffic signal control under missing data. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2261–2269, 2023.

[Mei *et al.*, 2024] Hao Mei, Xiaoliang Lei, Longchao Da, Bin Shi, and Hua Wei. LibSignal: An open library for traffic signal control. *Machine Learning (Springer)*, 2024.

[Parasuraman and Riley, 1997] Raja Parasuraman and Victor Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253, 1997.

[Peng *et al.*, 2018] Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[Ruan *et al.*, 2024] Jingqing Ruan, Ziyue Li, Hua Wei, Haoyuan Jiang, Jiaming Lu, Xuantang Xiong, Hangyu Mao, and Rui Zhao. CoSLight: Co-optimizing collaborator selection and decision-making to enhance traffic signal control. In *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2024.

[Shardlow, 1990] N. Shardlow. Action and agency in cognitive science. 1990. Department of Psychology, University of Manchester.

[Sutton and Barto, 2018] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

[Tobin *et al.*, 2017] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017.

[Turnau *et al.*, 2025] Justin Turnau, Longchao Da, Khoa Vo, Ferdous Al Rafi, Shreyas Bachiraju, Tiejin Chen, and Hua Wei. Joint-local grounded action transformation for sim-to-real transfer in multi-agent traffic control. In *Proceedings of the 2nd Reinforcement Learning Conference (RLC)*, 2025.

[Wei *et al.*, 2018] Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. IntelliLight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2018.

[Wei *et al.*, 2019a] Hua Wei, Chacha Chen, Guanjie Zheng, Kan Wu, Vikash Gayah, Kai Xu, and Zhenhui Li. PressLight: Learning max pressure control to coordinate traffic signals in arterial network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2019.

[Wei *et al.*, 2019b] Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu, Kai Xu, and Zhenhui Li. CoLight: Learning network-level cooperation for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM)*, 2019.

[Wei *et al.*, 2020] Hua Wei, Guanjie Zheng, Vikash Gayah, and Zhenhui Li. Recent advances in reinforcement learning for traffic signal control: A survey. *ACM SIGKDD Explorations*, 2020.

[Yao *et al.*, 2026] Huaiyuan Yao, Wanpeng Xu, Justin Turnau, Nadia Kellam, and Hua Wei. Instructional agents: Reducing teaching faculty workload through multi-agent instructional design. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2026.

[Zhang *et al.*, 2025] Yiran Zhang, Khoa Vo, Longchao Da, Tiejin Chen, Xiaoou Liu, and Hua Wei. Reproducible and low-cost sim-to-real environment for traffic signal control. In *Proceedings of the International Conference on Cyber-Physical Systems (ICCPS)*, 2025.

[Zhao *et al.*, 2020] Wenshuai Zhao, Jorge Pe na Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: A survey. In *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020.

[Zhou *et al.*, 2026] Xiaolin Zhou, Aojie Yuan, Zheng Luo, Zipeng Ling, Xixiao Pan, Yicheng Gao, Haiyue Zhang, Jiate Li, Shuli Jiang, Prince Zizhuang Wang, et al. When simulation lies: A sim-to-real benchmark and domain-randomized RL recipe for tool-use agents. *arXiv preprint arXiv:2605.11928*, 2026.