# Reinforcement Learning without Explicit Rewards: Theory and Practice

**Weitong Zhang**

School of Data Science and Society
University of North Carolina at Chapel Hill
weitongz@unc.edu

In recent years, RL has become central to fine-tuning foundation models for capabilities such as reasoning and self-reflection. The traditional reward-based RL pipeline typically relies on a human-designed or learned reward function. While this pipeline has demonstrated immense value across a wide range of tasks, many recent applications move beyond this setting and leverage *implicit* signals, including preferences, expert demonstrations, and contextual information. To better apply RL to these implicit signals, I claim that collapsing them into a single learned reward or relying on reward engineering is neither efficient nor faithful. Instead, we should advance and better understand RL methods that operate without an explicitly specified reward.

Based on this vision, my research focuses on building both the theoretical foundation and the empirical applications for reinforcement learning without explicit or precise reward design. I have developed theory, algorithms, and systems that enable RL agents to operate effectively with implicit signals rather than hand-crafted rewards.

- **Unsupervised, Reward-Free Robust Exploration.** In multi-task RL, exploration using individual explicit rewards design is often inefficient and less effective because rewards usually vary across tasks. To tackle this issue, I study reward-free exploration in which an agent seeks broad coverage of states and skills without a predefined goal, followed by light, goal-oriented finetuning that can tolerate misspecified rewards. Concretely, my works has ***(i)*** introduced a theoretical framework that guides reward-free exploration with intrinsic rewards and establishes coverage guarantees; ***(ii)*** analyzed robustness under model misspecification and distribution shift, proving stability and downstream performance guarantees for finetuning; and ***(iii)*** developed practical unsupervised RL methods that instantiate these principles, including uncertainty-aware exploration and representation learning with strong empirical performance.
- **Mitigating Reward Hacking in Generative Models.** RL is increasingly used to finetune generative models, including LLMs and flow based generative models (e.g., diffusion models). In these settings the reward often reflects relative quality or preference rather than a ground truth utility, which invites *reward hacking*. This risk persists when rewards are self accessed by the model, for example in self-reflection pipelines. My approach is to design finetuning methods that optimize preferences while preserving the base generative policy and respecting the policy priors, so the model cannot inflate proxy scores at the expense of real quality. Concretely, my work has ***(i)*** developed exact energy guidance for flow matching that preserves a KL regularized objective with formal guarantees, which steers sampling without drifting from the pretrained prior; ***(ii)*** regularized self rewarding LLM finetuning to leverage self critique while controlling reward inflation, through calibrated self assessment and trust region style constraints that keep updates close to the base policy; and ***(iii)*** built a principled multi modal preference pipeline that elicits self critic feedback without an explicit reward, adds consistency checks to reduce self reward bias, and releases a benchmark that stresses reward hacking failure modes.
- **Interdisciplinary Applications.** Theoretical guarantees and methodological innovations from the aforementioned works support a range of interdisciplinary applications in scientific discovery and self-driving labs. These include combating the COVID-19 pandemic through large-scale modeling, *de novo* drug design with generative models, and energy-guided refinement of protein–ligand structures. In particular, I have pioneered the development of machine learning methods for electrochemical analysis and led one of the first applications of reinforcement learning in electrochemistry self-driving laboratories for synergistic catalyst optimization.

**Future Plans.** My future research plan is to comprehensively develop and understand the reinforcement learning without the explicit reward design so that the RL agents can be further applied to complex environments and tasks. Building upon our previous results, I will (***i***) develop theory and algorithms that improve sample efficiency, robustness, and online adaptation in imitation learning; (***ii***) develop in context RL methods and scalable implementations for various, personalized online adaptations and (***iii***) broaden the impact of AI for science and healthcare by making RL practical for complex scientific and clinical workflows, including self-driving labs and automatic diagnosis.