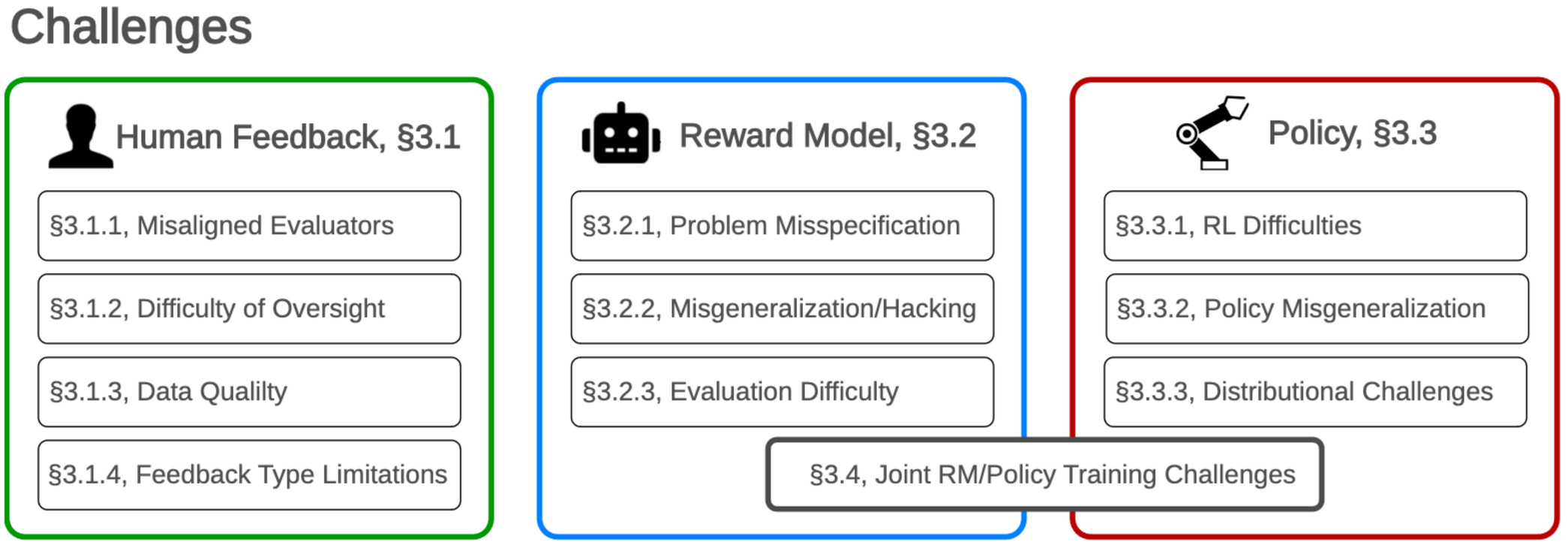
CS 690: Human-Centric Machine Learning

Prof. Scott Niekum

Open problems in RLHF

Open problems and fundamental limitations of RLHF



Casper, Stephen, et al. "Open problems and fundamental limitations of reinforcement learning from human feedback." *arXiv preprint arXiv:2307.15217* (2023).

Challenges: Human feedback

Misaligned Humans: Evaluators may Pursue the Wrong Goals

- Tractable: Selecting representative humans and getting them to provide quality feedback is difficult
- Tractable: Some evaluators have harmful biases and opinions
- Tractable: Individual human evaluators can poison data

Good Oversight is Difficult

- Tractable: Humans make simple mistakes due to limited time, attention, or care
- Tractable: Partial observability limits human evaluators
- Fundamental: Humans cannot evaluate performance on difficult tasks well
- Fundamental: Humans can be misled, so their evaluations can be gamed

Data Quality

- Tractable: Data collection can introduce harmful biases
- Fundamental: There is an inherent cost/quality tradeoff when collecting human feedback
- Tractable: Individual human evaluators can poison data

Limitations of Feedback Types

- Fundamental: RLHF suffers from a tradeoff between the richness and efficiency of feedback types
  - Comparisons
  - Scalar feedback
  - Corrections
  - Language

Addressing: Human feedback

- Providing feedback with AI assistance
- Fine-grained feedback
- Process-based supervision
- Translating natural language specifications into a reward model
- Learning rewards from demonstrations

Challenges: Reward model

Problem Misspecification

- Fundamental: An individual human's values are difficult to represent with a reward function
- Fundamental: A single reward function cannot represent a diverse society of humans

Reward Misgeneralization and Hacking

- Fundamental: Reward models can misgeneralize to be poor reward proxies, even from correctly-labeled training data
- Fundamental: Optimizing for an imperfect reward proxy leads to reward hacking

Evaluating Reward Models

- Tractable: Evaluating reward models is difficult and expensive

Addressing: Reward model

- Using direct human oversight
- Multi-objective oversight
- Maintaining uncertainty over the learned reward function

Challenges: Policy learning

Robust Reinforcement Learning is Difficult

- Tractable: It is (still) challenging to optimize policies effectively
- Tractable: Policies tend to be adversarially exploitable

Policy Misgeneralization

- Fundamental: Policies can perform poorly in deployment even if rewards seen during training were perfectly correct
- Fundamental: Optimal RL agents tend to seek power

Distributional Challenges

- Tractable: The pretrained model introduces biases into policy optimization
- Tractable: RL contributes to mode collapse

Challenges with Jointly Training the Reward Model and Policy

- Tractable: Joint training induces distribution shifts
- Tractable: It is difficult to balance efficiency and avoiding overfitting by the policy

Addressing: Policy learning

- Aligning LLMs during pretraining
- Aligning LLMs through supervised learning

RLHF is Not All You Need: Complementary Strategies for Safety

- Robustness
- Risk assessment and auditing
- Interpretability and model editing

Auditing RLHF'd systems

Human feedback details:

- **A description of the pretraining process including details about what data was used** to make apparent possible biases that pretraining can cause.
- **How human evaluators were selected and trained** to provide information about risks of evaluators being malicious, unrepresentative, or incapable.
- **The process by which examples were selected to obtain feedback** to invite scrutiny about their representativeness and whether sufficient adversarial training was used. If examples were crowdsourced from a publicly-available application, details about what measures were taken to avoid data poisoning attacks should be provided.
- **The type(s) of human feedback used** (e.g., binary comparisons, scalar feedback, etc.) to suggest what risks might be caused by insufficiently abundant or rich feedback.
- **A report on measures taken for quality assurance in feedback collection and inter-rater consistency** to ensure that effective quality control measures were taken.

Auditing RLHF'd systems

Reward model details:

- **The loss function used to fit the reward model and how disagreement was modeled** (e.g., as noise) to help with analyzing the degree of misspecification when fitting the reward model.
- **A report on reward model evaluation and results** to suggest possible problems from a misaligned reward model. The evaluation should involve red teaming.

Policy details:

- **A report on policy evaluation and results** to suggest possible troubles from a misaligned policy. The evaluation should involve red teaming and include assessment for risky capabilities (e.g., the ability to deceive a human).

Systemic safety measures

- **A report on internal and external audits and red teaming** to ensure accountability and disclose risks that are identified.
- **A report on expected risks and anticipated failure modes** to ensure accountability.
- **Plans for monitoring and correcting failures that emerge** to support post-deployment safety.