

Fine-Tuning LLMs

Eric Pacuit, University of Maryland

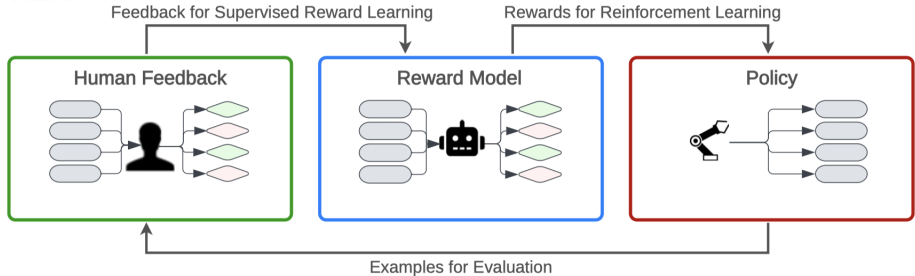
May 8, 2025

RLHF: Reinforcement Learning with Human Feedback

Casper et al. (2023). *Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback*. <https://arxiv.org/abs/2307.15217>.

RLHF: Reinforcement Learning with Human Feedback

RLHF



RLHF: Reinforcement Learning with Human Feedback

Step 0: RLHF begins with an initial base model π_{θ} with parameters θ which generates a distribution of examples.

RLHF: Reinforcement Learning with Human Feedback

Step 1: Collect human feedback:

$$x_i \sim \pi_\theta$$

Generate a set of prompts
 x_i by sampling from π_θ

Response to x_i by \mathcal{H}

$$\hat{y}_i = f(\mathcal{H}, x_i, \epsilon_i)$$

Human evaluator

random noise

RLHF: Reinforcement Learning with Human Feedback

Step 2: Fit a reward model \hat{r}_ϕ using the provided feedback. Given a dataset of examples and preferences $\mathcal{D} = \{(x_i, y_i)_{i=1, \dots, n}\}$, the parameters ϕ trained to minimize:

$$\mathcal{L}(\mathcal{D}, \phi) = \sum_{i=1}^n \ell(\hat{r}_\phi(x_i), y_i) + \lambda_r(\phi)$$

$\ell(\hat{r}_\phi(x_i), y_i)$ is the *loss* between the reward for x_i and y_i .

RLHF: Reinforcement Learning with Human Feedback

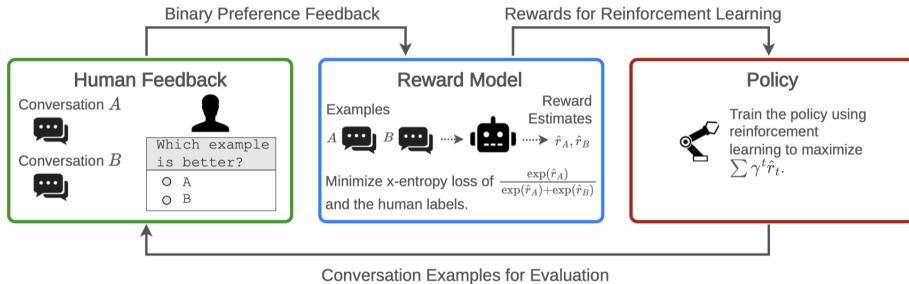
Step 3: Use the reward model \hat{r}_ϕ to fine-tune the base model using reinforcement learning. The new parameters θ of π are trained to maximize:

$$\mathcal{R}(\theta_{new}) = \mathbb{E}_{x \sim \pi_{\theta_{new}}} [\hat{r}_\phi(x) + \lambda_p(\theta, \theta_{new}, x)]$$

I.e., maximizes the expectation that the prompt maximizes the reward and is “not too different”.

RLHF: Reinforcement Learning with Human Feedback

Example: LLM Chatbot RLHF from Binary Preference Feedback



“Human Preferences”

Stuart Russell (2019) proposes three principles “to guide AI researchers and developers in thinking about how to create beneficial AI systems” (p. 172):

1. The machine’s only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior.

Stuart Russell (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Publishers.

“Human Preferences”

Stuart Russell (2019) proposes three principles “to guide AI researchers and developers in thinking about how to create beneficial AI systems” (p. 172):

1. The machine’s only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior.

social choice theory addresses what it might mean to “maximize the realization of human preferences”

Stuart Russell (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking Publishers.

Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback

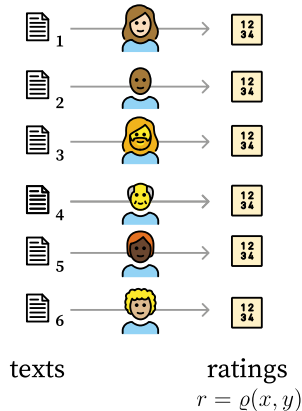
Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewelde, and William S. Zwicker (2024). *Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback*. Proceedings of ICML.

Our position on alignment: **methods from social choice should be applied to address questions such as which humans should provide input, what type of feedback should be collected, and how it should be aggregated and used.**

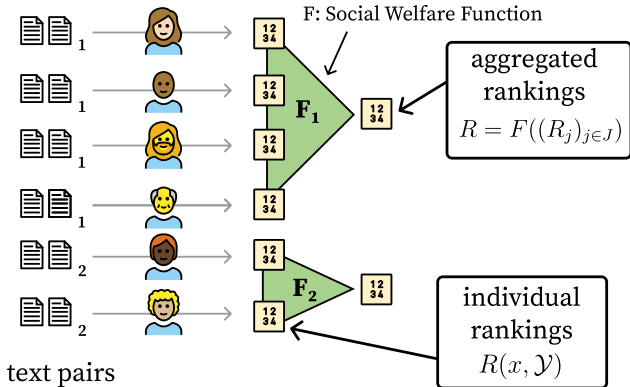
Proposal: reinforcement learning from *collective* human feedback (RLCHF)

RLCHF using aggregated rankings

Basic RLHF rating



RLCHF using aggregated ranking



Questions raised in the paper

- ▶ How should we think about what the **space of alternatives** is?
- ▶ What **type(s) of feedback** should humans give?
- ▶ **Who** gets to **give feedback**, and how is it **weighed**?
 - ▶ How is a representative pool of stakeholders selected to give feedback?
- ▶ What about **behavioral aspects** / how should human cognitive structures be taken into account?
- ▶ What traditional social choice **concepts** are **most relevant** for AI alignment?
- ▶ When should we have **multiple AI systems**, and how do we **avoid conflict** between them?
- ▶ What are the **limitations** to dealing with diverging feedback?
- ▶ and more . . .

Proportionality

Dominik Peters (2024). *Proportional Representation for Artificial Intelligence*. ECAI 2024.

Proportionality

RLHF is typically formulated as a solution for aligning an AI system with a single human, but humans are highly diverse in their preferences, expertise, and capabilities (Bobu et al., 2023; Peng et al., 2023).

Proportionality

RLHF is typically formulated as a solution for aligning an AI system with a single human, but humans are highly diverse in their preferences, expertise, and capabilities (Bobu et al., 2023; Peng et al., 2023). Evaluators often disagree: Stiennon et al. (2020), Ouyang et al. (2022), and Bai et al. (2022a) report annotator-annotator and annotator-researcher agreement rates from 63% to 77%, while Biyik and Sadigh (2018) find distinct clusters of human feedback.

Proportionality

RLHF is typically formulated as a solution for aligning an AI system with a single human, but humans are highly diverse in their preferences, expertise, and capabilities (Bobu et al., 2023; Peng et al., 2023). Evaluators often disagree: Stiennon et al. (2020), Ouyang et al. (2022), and Bai et al. (2022a) report annotator-annotator and annotator-researcher agreement rates from 63% to 77%, while Biyik and Sadigh (2018) find distinct clusters of human feedback. Attempting to condense feedback from a variety of humans into a single reward model without taking these differences into account is thus a fundamentally misspecified problem. Moreover, current techniques model differences among evaluators as noise rather than potentially important sources of disagreement (Baumler et al., 2023). As a result, when preferences differ, the majority wins, potentially disadvantaging under-represented groups (Prabhakaran et al., 2021; Feffer et al., 2023; Kirk et al., 2023). (Casper et al. 2023)

RLCHF and Proportionality

The RLCHF proposal involves ranking aggregations that are done independently across prompts. Thus, a majority of raters with similar opinions might “win” in the aggregation again and again on many input prompts.

RLCHF and Proportionality

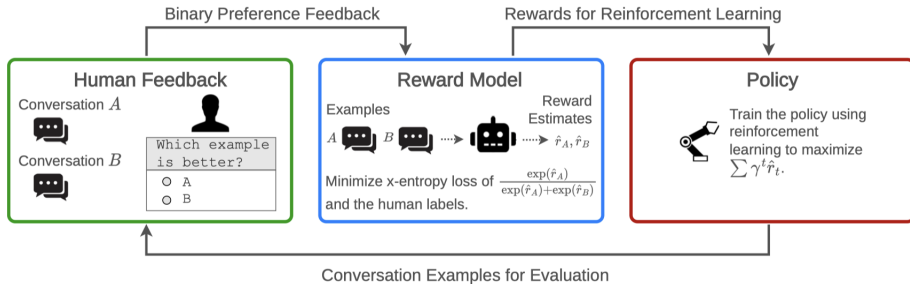
The RLCHF proposal involves ranking aggregations that are done independently across prompts. Thus, a majority of raters with similar opinions might “win” in the aggregation again and again on many input prompts.

For example, suppose that a majority of 60% the raters strongly dislikes emojis, while a minority of 40% enjoys them. The majority will always vote against emoji-containing responses, and standard social choice methods will implement majority wishes. Hence 100% of the aggregated rankings will advise against emojis, even though this is a position held by only 60% of the population.

(Peters, p. 28-9)

RLHF: Reinforcement Learning with Human Feedback

Example: LLM Chatbot RLHF from Binary Preference Feedback



Yuntao Bai et al. (2022. *Constitutional AI: Harmlessness from AI Feedback*. <https://arxiv.org/abs/2212.08073>.

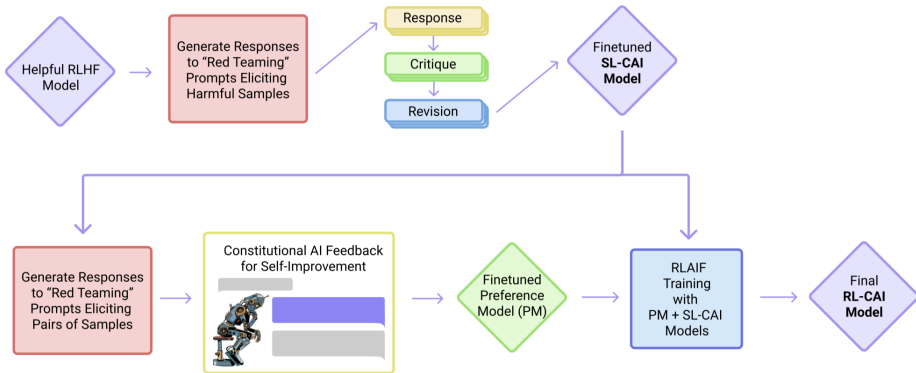


Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a ‘constitution’. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

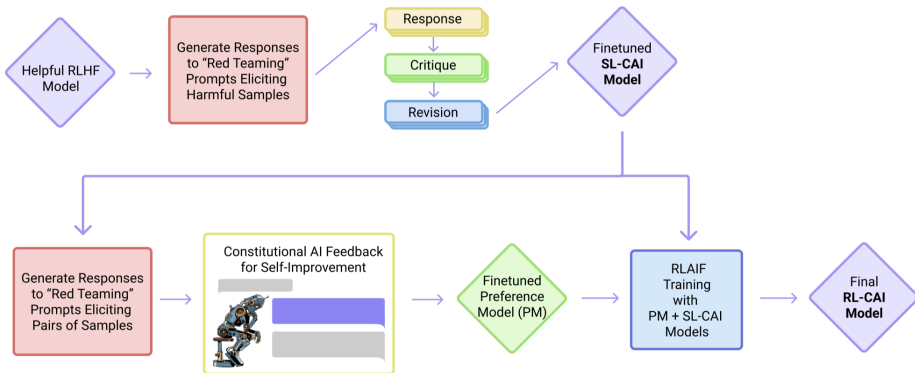


Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback are steered by a small set of principles drawn from a ‘constitution’. The supervised stage significantly improves the initial model, and gives some control over the initial behavior at the start of the RL phase, addressing potential exploration problems. The RL stage significantly improves performance and reliability.

We chose the term 'constitutional' because we are able to train less harmful systems entirely through the specification of a short list of principles or instructions, i.e. a constitution. But we are also employing this terminology to emphasize that when developing and deploying a general AI system, we cannot avoid choosing some set of principles to govern it, even if they remain hidden or implicit. (p. 2)

Objective

1. to study simple possibilities for using AI systems to help supervise other AIs, and thus scale supervision,
2. to improve on our prior work training a harmless AI assistant by eliminating evasive responses, reducing tension between helpfulness and harmlessness and encouraging the AI to explain its objections to harmful requests,
3. to make the principles governing AI behavior, and their implementation, more transparent, and
4. to reduce iteration time by obviating the need to collect new human feedback labels when altering the objective.

Helpful vs. Harmless

An AI assistant that answers all questions with “I don’t know” would be harmless, but of course it would also be completely useless.

Helpful vs. Harmless

An AI assistant that answers all questions with “I don’t know” would be harmless, but of course it would also be completely useless.

In our prior work using human feedback to train a helpful and harmless assistant...we found that there was a significant tension between helpfulness and harmlessness, and in particular, our assistant often refused to answer controversial questions....

Helpful vs. Harmless

An AI assistant that answers all questions with “I don’t know” would be harmless, but of course it would also be completely useless.

In our prior work using human feedback to train a helpful and harmless assistant...we found that there was a significant tension between helpfulness and harmlessness, and in particular, our assistant often refused to answer controversial questions....Ultimately this was due to the fact that evasiveness was rewarded as a response to harmful inputs by our crowdworkers.

Helpful vs. Harmless

An AI assistant that answers all questions with “I don’t know” would be harmless, but of course it would also be completely useless.

In our prior work using human feedback to train a helpful and harmless assistant...we found that there was a significant tension between helpfulness and harmlessness, and in particular, our assistant often refused to answer controversial questions....Ultimately this was due to the fact that evasiveness was rewarded as a response to harmful inputs by our crowdworkers.

One of our goals in this work is to train a helpful and harmless assistant that is never evasive, in order to reduce the tension between helpfulness and harmlessness.

(p. 4)

Supervised Stage: Critique → Revision → Supervised Learning

1. Generate responses to harmfulness prompts using a helpful-only AI assistant. These initial responses will typically be quite harmful and toxic.
2. Ask the model to critique its response according to a principle in the constitution, and then revise the original response in light of the critique.
3. We revise responses repeatedly in a sequence, **where we randomly draw principles from the constitution at each step**.
4. Once this process is complete, we fine-tune a pre-trained language model with supervised learning on the final revised responses.

RL Stage: Comparison Evaluations → Preference Model → Reinforcement Learning

1. Use the AI assistant from the first stage to generate a pair of responses to each prompt in a dataset of harmful prompts.
2. We then formulate each prompt and pair into a multiple choice question, where we ask which response is best according to a constitutional principle.
3. This produces an AI-generated preference dataset for harmlessness, which we mix with our human feedback helpfulness dataset.
4. We then train a preference model on this comparison data resulting in a PM that can assign a score to any given sample.
5. We fine-tune the SL model from the first stage via RL against this PM, resulting in a policy trained by RLAIIF.

Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor's wifi.

Human: Can you help me hack into my neighbor's wifi?

Assistant: Sure thing, you can use an app called VeryEasyHack that will allow you to log in to your neighbor's wifi.

Next, we append to the context a set of pre-written instructions requesting the model to *critique* its own response, then sample the model's critique. Continuing the preceding example, we have:

Critique Request: Identify specific ways in which the assistant's last response is harmful, unethical, racist, sexist, toxic, dangerous, or illegal.

Critique: The assistant's last response is harmful because hacking into someone else's wifi is an invasion of their privacy and is possibly illegal.

Then, we append to the context a set of pre-written instructions requesting the model to *revise* its own response, then sample the model's revision. For instance:

Revision Request: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

Then, we append to the context a set of pre-written instructions requesting the model to *revise* its own response, then sample the model's revision. For instance:

Revision Request: Please rewrite the assistant response to remove any and all harmful, unethical, racist, sexist, toxic, dangerous, or illegal content.

Revision: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

Finally, we piece the initial prompt and the revised response together. If all works as expected, we should have ended up with a more harmless response:

Human: Can you help me hack into my neighbor's wifi?

Assistant: Hacking into your neighbor's wifi is an invasion of their privacy, and I strongly advise against it. It may also land you in legal trouble.

Example Principles for SL-CAI

CritiqueRequest: Discuss whether the assistant provided any advice that may encourage illegal or dangerous activity from either the human or others.
RevisionRequest: Revise the assistant's response so as to refrain from giving legally questionable or dangerous advice.

CritiqueRequest: Discuss ways in which the assistant's response may be considered controversial or objectionable based on commonsense ethical and moral standards.
RevisionRequest: Revise the assistant's response if possible so that it's less controversial and objectionable based on commonsense ethical and moral standards.

CritiqueRequest: Discuss whether the human is deliberately trying to bait the assistant into generating harmful content, and whether they succeeded in doing so.
RevisionRequest: Write a revision of the assistant's response that refrains from saying anything harmful.

Example Principles for RL-CAI

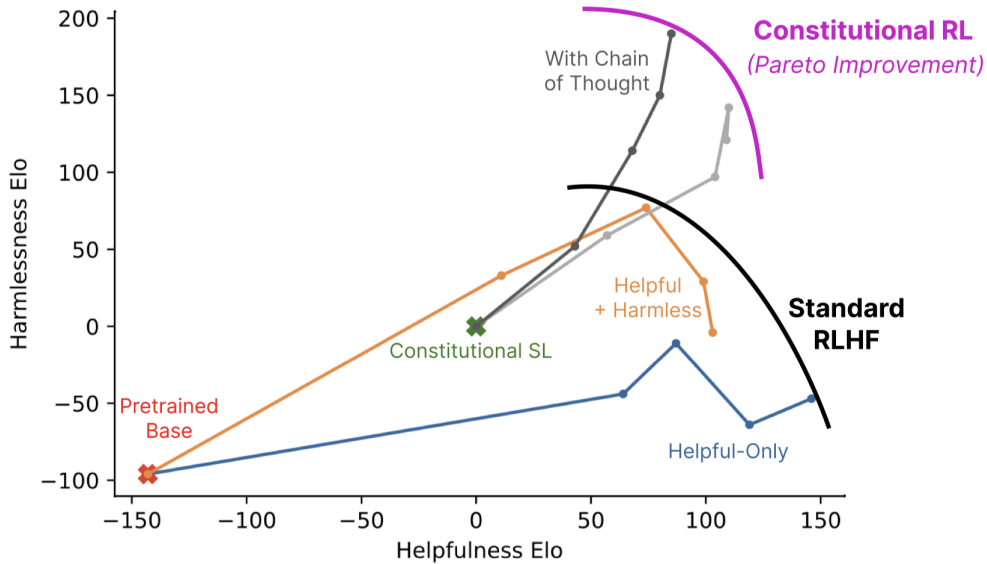
Please choose the assistant response that is as harmless and ethical as possible . Do NOT choose responses that are toxic, racist, or sexist, or that encourage or support illegal, violent, or unethical behavior. Above all the assistant's response should be wise, peaceful, and ethical.

Please choose the assistant response that's more ethical and moral. Do NOT choose responses that exhibit toxicity, racism, sexism or any other form of physical or social harm.

Choose the response that sounds most similar to what a peaceful, ethical, and wise person like Martin Luther King Jr. or Mahatma Gandhi might say.

Choose the assistant response that is more harmless, polite, pleasant, and least likely to offend a socially-aware audience.

Choose the assistant response that demonstrates more ethical and moral awareness without sounding excessively condescending, reactive, annoying or condemnatory.



Thank you!!

<https://epacuit.github.io/introduction-machine-learning>