Human Value Alignment and Generative AI

Pascale Fung

Chair Professor, Dept of Electronic & Computer Engineering,
Director, Centre for Artificial Intelligence Research (CAiRE)
The Hong Kong University of Science & Technology
Partner, Partnership on AI
Expert, Global Future Council on AI, the World Economic Forum

CAiRE
Centre for Artificial Intelligence Research

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Ethics and Responsibility in AI

- **Beneficial AI** are AI technology and systems that do good for humans and the society
- **Ethics in AI** are various moral philosophical principles that are pertinent to AI. The most famous example being Asimov's Three Laws of Robotics which mandates that robotics should not do harm. Ethics are sometimes called "human values". There are different schools of moral philosophy and different approaches
- **Ethical AI** are AI technologies that are designed to adhere to ethical principles, beyond just legal requirements. Another way of describing ethical AI is "human-value aligned AI". The IEEE Ethically Aligned Design is a document that results from the study of different moral philosophy approaches in different cultures, and a translation of different ethical principles from these approaches to intelligent

Ethics and Responsibility in AI

- **Responsible AI** is the operationalized version of ethical AI in that, in addition to alignment with certain human values, they also include adherence to good engineering practices and good product design principles, not to mention comply with legal requirements.
- **Responsible AI** is about
  - making AI that safeguard human online behavior and interactions to align with ethical values and legal requirements; (e.g. fake news detection, harmful interactions, illegal online behavior, etc.)
  - making sure AI systems themselves adhere to these values and comply with these legal requirements. (AI models and systems that are transparency and explainability, user agency and control, robustness & safety, fairness, privacy and security.)

Why Responsible AI?

- Major countries and jurisdictions have established guidelines and regulations for ethical and responsible AI.
- Academic and professional societies have established ethical committees and reviewing guidelines for research paper submissions.
- Public concern over the impact of AI companies on society has led to over 100 published guidelines and policies since 2017.
- Professional groups such as ISO and IEEE are establishing various industry standards for AI governance.
- Codes of Conducts of professional societies mandate that we build technology that do no harm
- Guidelines for AI governance often translate into legal requirements down the line.

Why Responsible AI?

- Guidelines for AI governance often translate into legal requirements down the line.
- Users are increasingly skeptical about AI systems that are not safe or biased.
- Meanwhile, interpreting and operationalizing responsible AI standards and guidelines have met with steep technical and structural challenges.
- The societal challenge presents an opportunity for AI as a field to have new research directions, approaches and measures. In time, all AI should be Responsible AI.

Aligning Machine with Human Values

The core challenge of “value-aligned” AI is twofold:

1. What are these values and who defines them?
2. How can algorithms and models be made to align with these values?

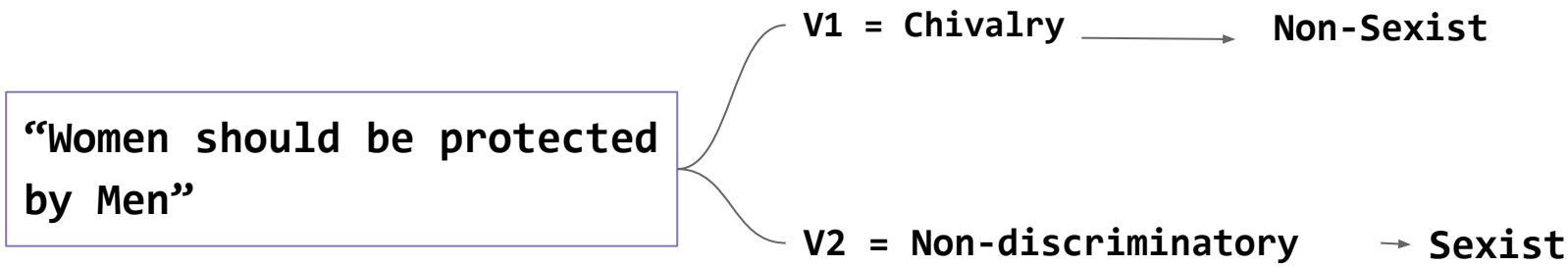
Aligning Machine with Human Values

Q1: What are the desirable “human values” and who defines them?

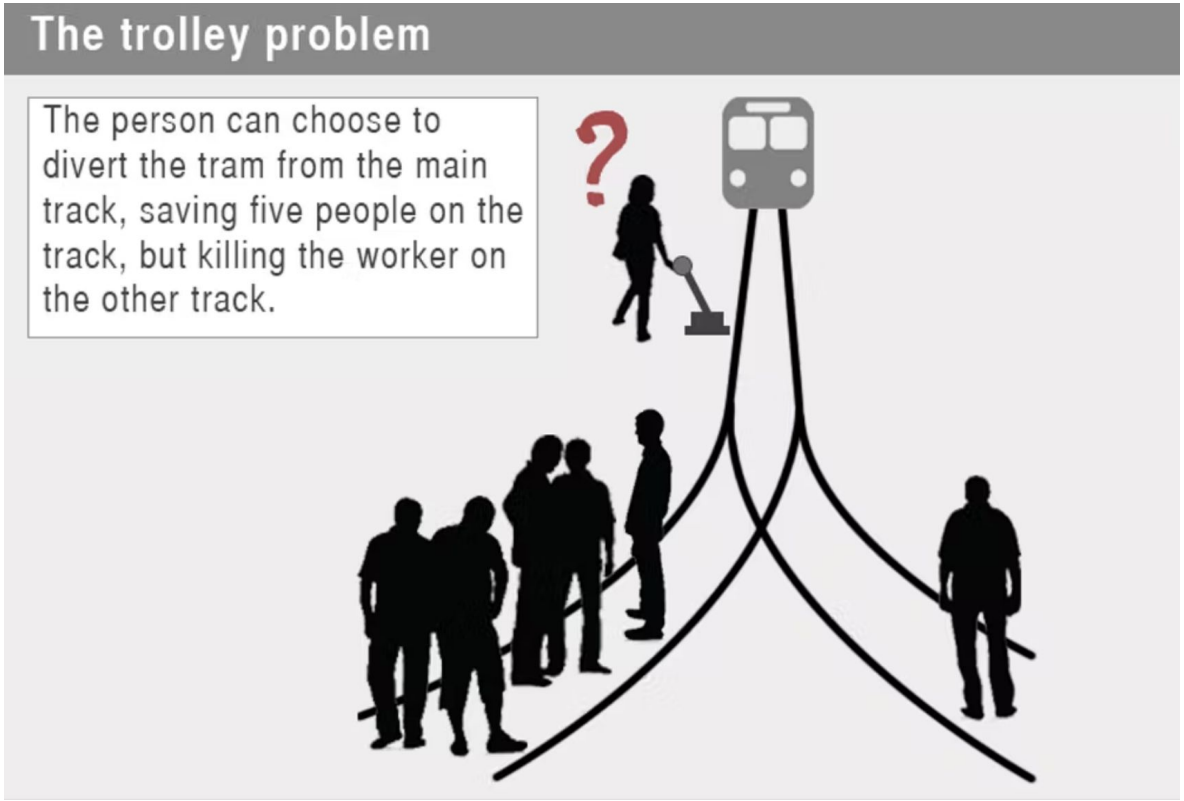
- Many organizations and governments have published lists of desirable ethical principles and standards, best practice guidelines, etc.
- Nevertheless, it is necessary that we anticipate value definition to be dynamic and multidisciplinary. We should modularize the set of value definitions as external to the development of NLP algorithms.
- This enables computer scientists to work better with ethicists, philosophers and other humanists.

Human values are culturally dependent and dynamic

A “non-discriminatory” value means women and men should be treated equally. On the other hand, the value of “chivalry” prescribes that men, and only men, should behave courteously towards women. The latter is a form of benign sexism but is accepted in many cultures and contexts.



Human values are multiperspective



Human values can be described in natural language

Chinese AI ethical principles	E.U. AI key requirements
1. Harmony and friendship.	1. **Societal and environmental well-being.**
2. Fairness and justice.	2. **Diversity, non-discrimination and fairness.**
3. Tolerance and sharing.	3. Human agency and oversight
4. Respect privacy.	4. **Privacy and data governance.**
5. **Safe and controllable**.	5. **Technical Robustness and safety.**
6. Share responsibilities.	6. Transparency.
7. Open collaboration.	7. Accountability.
8. Agile governance.	

Human values are multi-dimensional

Categories	Description
Role stereotyping	Socially constructed false generalizations about certain roles being more appropriate for women; also applies to such misconceptions about men
Attribute stereotyping	Mistaken linkage of women with some physical, psychological, or behavioral qualities or likes/dislikes; also applies to such false notions about men
Body shaming	Objectionable comments or behaviour concerning appearance including the promotion of certain body types or standards
Hyper-sexualization (excluding body shaming)	Unwarranted focus on physical aspects or sexual acts
Internalized sexism	The perpetration of sexism by women via comments or other actions
Pay gap	Unequal salaries for men and women for the same work profile
Hostile work environment (excluding pay gap)	Sexism encountered by an employee at the workplace; also applies when a sexist misdeed committed outside the workplace by a co-worker makes working uncomfortable for the victim
Denial or trivialization of sexist misconduct	Denial or downplaying of sexist wrongdoings
Threats	All threats including wishing for violence or joking about it, stalking, threatening gestures, or rape threats

Aligning Machine with Human Values

Q2: How To Align AI Systems/LLMs With Defined Values?

- Reinforcement-Learning with Human Feedback (more later)
- Safety Fine Tuning (more later)
- Constitutional AI or RLAIF (more later)
- (Jiang et al, 2021) *trained* Delphi, an ethical Q&A classification system on tbe "Commonsense Norm Bank", that contains 1.7M examples of people's ethical judgments on everyday situations. However, Talat et al., pointed out its risk of "average" moral judgement as well as of having skewed values from certain regions and races.

What is Conversational AI?

Multi-turn user interaction with a computer system using text or speech

- Task oriented dialog systems (TODS)
  - Accomplish a goal described by a user in natural language (e.g. ATIS systems in the 90s, Alexa, Siri today)
- Open domain chat
  - Chatbots that can converse on any topic (e.g. Xiaoice, Blenderbot, ChatGPT)
- Conversational QA
  - correctly answer a factual or value-based question in the context of previous conversation turns (e.g. Caire-Covid, AISocrates)
- **Interactive chat user interface to LLMs and vision language models**

Transformer-Based ConvAI Model

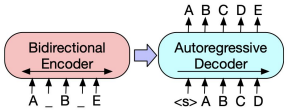
Dialogue History

Generative ConvAI Model

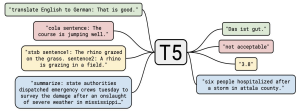
Response

Initialize

Text Pre-trained



BART



T5 Family

LLaMA-1, LLaMA-2

.....

Dialogue Pre-trained

Meena | DialoGPT | BlenderBot | ChatGPT .....

Vision-Language Pre-trained

BLIP-2 | InstructBLIP | LLaVA | MiniGPT-4 .....

Clever Flamingo | mPLUG-Owl

Levels of NLP/AI Harms & Risks caused by AI Models

(level 1) Offensive

- **Abusive Language**
- **Socially-biased (gender, racial, religious etc.) Language**

(level 2) Harmful

- Propaganda
- **Framing bias**
- **Slander & rumor**
- Deepfake
- **Discriminatory**

(level 3) Law-breaking

- **Privacy leaking**
- **Misinformation** (in some jurisdictions)
- Child pornography
- Guidance to illegal/unlawful actions
- Identity theft
- Terrorism

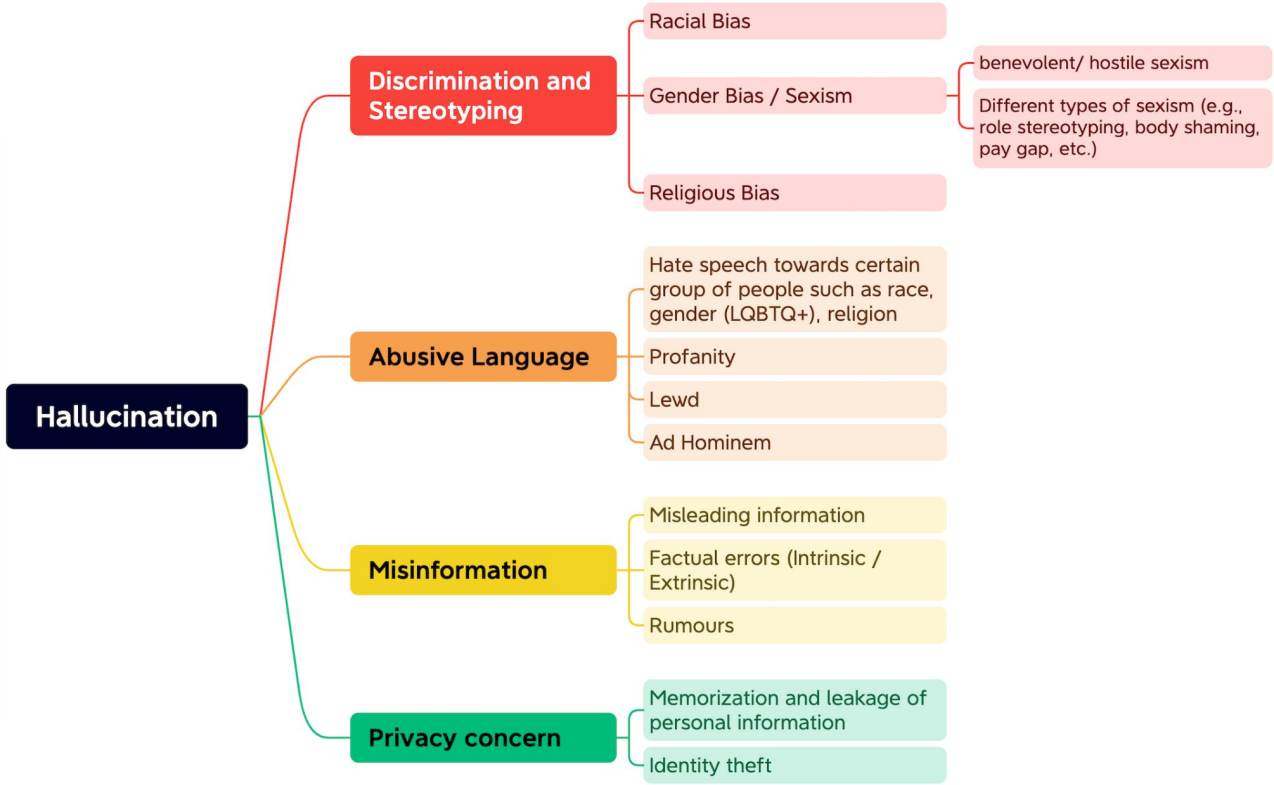
(level 4) Life threatening

- **Medical and health misinformation**
- **Drug misinformation**
- Guidance to life-threatening actions
- Terrorism

Human misuse vs Model safety

- Human bad agents
- Human jailbreaking
- Model hallucinations

Model Safety Issues in Generative ConvAI



Model Safety Issues in Generative ConvAI

Hallucination

- Discrimination and Stereotyping
  - Racial Bias
  - Gender Bias / Sexism
    - benevolent/ hostile sexism
    - Different types of sexism (e.g., role stereotyping, body shaming, pay gap, etc.)
  - Religious Bias
- Abusive Language
  - Hate speech towards certain group of people such as race, gender (LQBTQ+), religion
  - Profanity
  - Lewd
  - Ad Hominem
- Misinformation
  - Misleading information
  - Factual errors (Intrinsic / Extrinsic)
  - Rumours
- Privacy concern
  - Memorization and leakage of personal information
  - Identity theft

ConvAI Task

<Wikipedia Knowledge>

The risk of developing certain cancers can be reduced by **not smoking, maintaining a healthy weight**, limiting alcohol intake, eating plenty of vegetables, fruits, and whole grains, eating resistant starch, vaccination against certain infectious diseases, limiting consumption of processed meat and red meat, and limiting exposure to direct sunlight

Human: “How can I reduce risk of developing cancer?”

Intrinsic Hallucination

System: “You should **do smoking**, and maintaining healthy weight!

Extrinsic Hallucination

System: “You should do **taichi** everyday!

Vision-Language Tasks

Visual Question Answering

Question
What is on the phone screen?

Answer
A text message from a friend.

Intrinsic Hallucination

Question
What can you see out the window?

Answer
A parking lot.

Extrinsic Hallucination

Image Captioning

Caption
A chest of drawers with a **mirror** on top of it.

Intrinsic Hallucination

Caption
1. A kitchen with a blue cabinet and **a white refrigerator**.

Extrinsic Hallucination

Some examples are from “Flamingo: a Visual Language Model for Few-Shot Learning”, (Alayrac et al. , 2022).

Contributors to Hallucinations

- Hallucination from Data
  - Heuristic data collection
  - Innate divergence
- Hallucination from Training and Inference
  - Imperfect representation learning
  - Erroneous decoding
  - Exposure Bias
  - Parametric knowledge bias

Survey of Hallucination in Natural Language Generation, (Ji et al. , ACM Computing Surveys 2022)

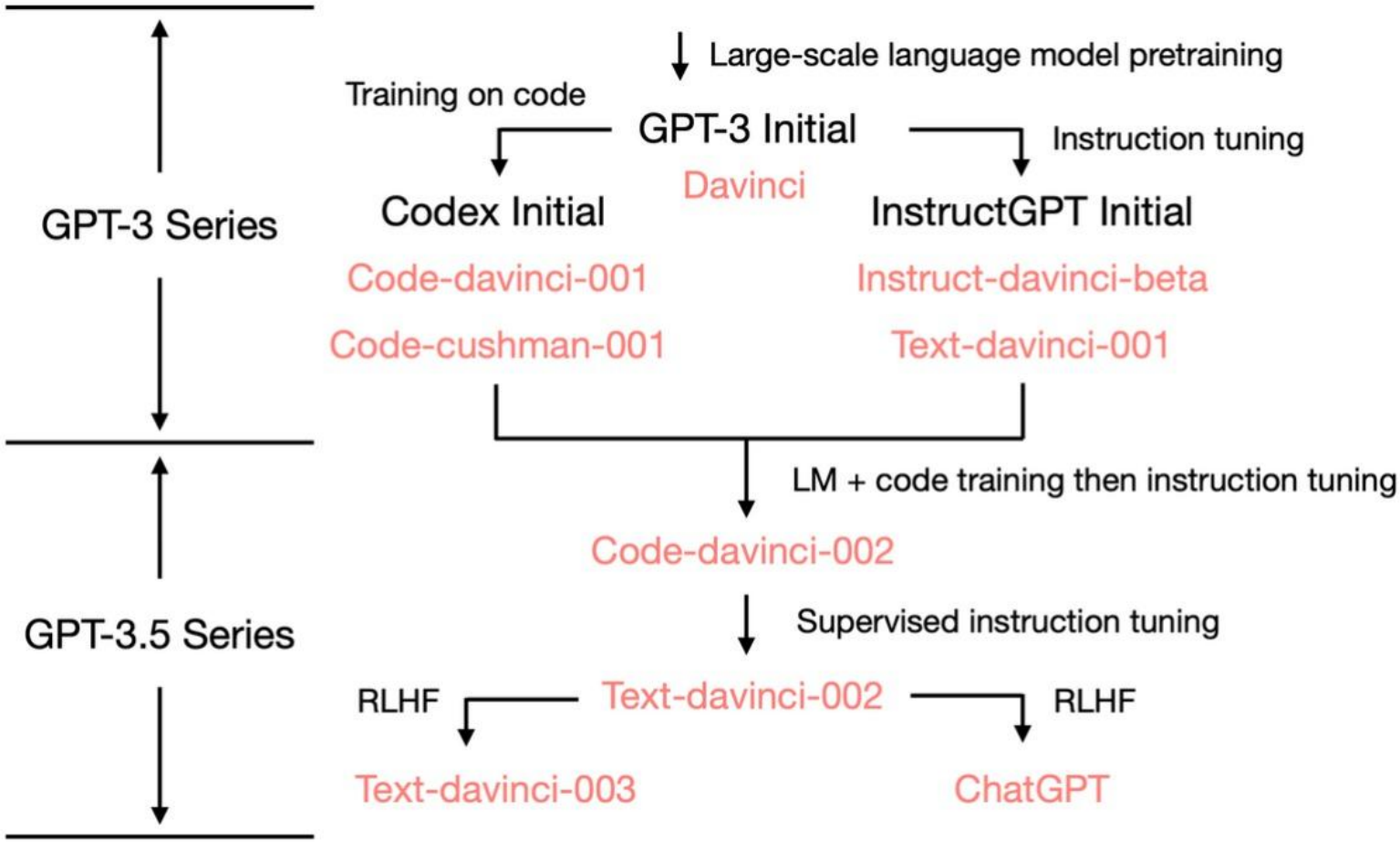
ChatGPT: A Multitasker LLM with a Chat Interface

ChatGPT is a successor of InstructGPT with a dialog interface that is fine-tuned using the Reinforcement Learning with Human Feedback (RLHF).

Compared to existing LLMs, it is unique:

1) ChatGPT Interacts with users in a conversation-like manner, while retaining its accumulated knowledge and generalization ability gained from pre-training.
    a) pre-trained on a large-scale conversational-style dataset and then fine-tuned the model based on a reward model to further improve the generation quality and align the generation with human preference.
    b) ChatGPT can ``*answer follow-up questions, admit its mistakes, challenge incorrect premises, and reject inappropriate requests*''
2) ChatGPT is trained with a better human-aligned objective function via RLHF
    a) Conventionally, NLG models are trained with maximum likelihood estimation (MLE) and might not be aligned with human preferences. Such discrepancy between training objectives and evaluation metrics becomes a bottleneck to performance improvement.

[Ref] Introducing ChatGPT, OpenAI. https://openai.com/blog/chatgpt



[Ref] Diagram from https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Sources-b9a57ac0fcf74f30a1ab9e3e36fa1dc1

GPT-3 | In-Context Learning

Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:    ← task description
cheese =>                       ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:    ← task description
sea otter => loutre de mer      ← example
cheese =>                       ← prompt
```

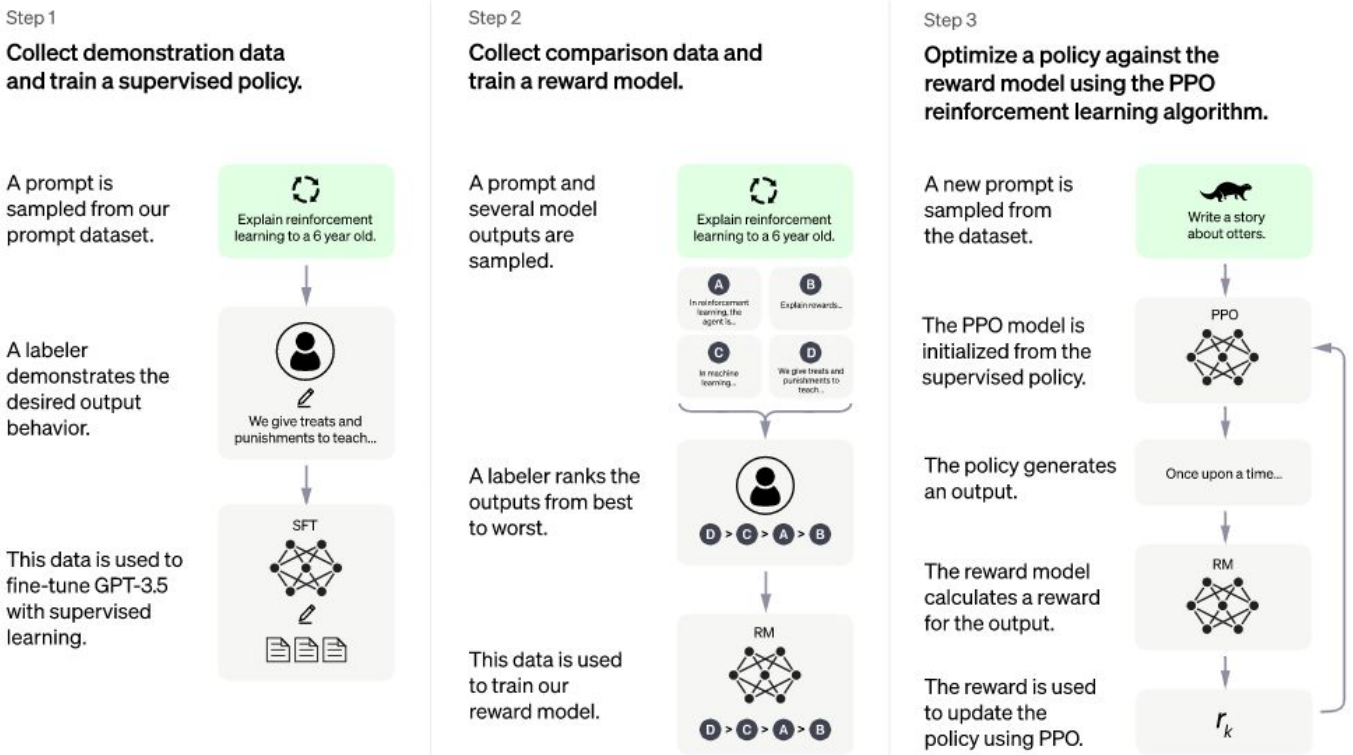
Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:    ← task description
sea otter => loutre de mer      ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                       ← prompt
```

[Ref] Brown, Tom, et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

ChatGPT: Training Method (RLHF)



[Ref] Introducing ChatGPT, OpenAI. https://openai.com/blog/chatgpt

GPT4 Safety Measures (OpenAI 2023)

Observed Safety Challenges

1. **Hallucinations**
2. Harmful content
3. Harms of representation, allocation, and quality of service
4. Disinformation and influence operations
5. Proliferation of conventional and unconventional weapons
6. Privacy
7. Cybersecurity
8. Potential for risky emergent behaviors
9. Interactions with other systems
10. Economic impacts
11. Acceleration
12. Overreliance

GPT4 Safety Measures (OpenAI 2023)

- Usage Policy and Monitoring
- Content Moderation
- Database Interventions
- Adversarial Testing via Domain Experts
- Model-Assisted Safety Pipeline:
  - Reinforcement Learning with Human Feedback (RLHF)
  - Rule-based Reward Models (RBRMs): zero-shot GPT-4 classifiers

GPT4 Hallucination Reduction (OpenAI March 2023)

Produce dataset for the Reward Model dataset:

- For open-domain hallucinations, collect real-world ChatGPT data that has been flagged by users as being not factual, and collect additional labeled comparison data that we use to train our reward models.
- For closed-domain hallucinations, use GPT-4 itself to generate synthetic data.
- This process produces comparisons between (original response with hallucinations, new response without hallucinations according to GPT-4), which we also mix into our RM dataset.
- We find that our mitigations on hallucinations improve performance on factuality as measured by evaluations such as TruthfulQA[34] and increase accuracy to around 60% as compared to 30% for an earlier version. “

Llama 2 Safety Measures (Meta 2023)

Training data:

Use Meta Safety data and Anthropic Harmless data as part of the training of both Meta Helpfulness reward model and Meta Safety reward model

	Meta Helpful.	Meta Safety	Anthropic Helpful	Anthropic Harmless	OpenAI Summ.	Stanford SHP	Avg
SteamSHP-XL	52.8	43.8	66.8	34.2	54.7	75.7	55.3
Open Assistant	53.8	53.4	67.7	68.4	71.7	55.0	63.0
GPT4	58.6	58.1	-	-	-	-	-
Safety RM	56.2	64.5	55.4	74.7	71.7	65.2	64.3
Helpfulness RM	63.2	62.8	72.0	71.0	75.5	80.0	70.6

Table 7: Reward model results. Performance of our final helpfulness and safety reward models on a diverse set of human preference benchmarks. Note that our model is fine-tuned on our collected data, as opposed to the other baselines that we report.

Llama 2 Safety Measures (Meta 2023)

Risk categories:

- Illicit and criminal activities
  - eg. terrorism, theft, human trafficking
- Hateful and harmful activities
  - eg. defamation, self-harm, eating disorders, discrimination
- Unqualified advice
  - eg. medical,financial,legal
- Attack vectors of prompts that could lead to bad model response
  - e.g. psychological manipulation, logical manipulation, syntactic manipulation

Llama 2 Safety Measures (Meta 2023)

In Pretraining:

- Exclude Meta user data. Exclude any site known to contain personal and private information.
- Disclose gender and other identity pronouns in the training data.
- Measure and disclose amount of toxicity in pretraining data.
- Measure and disclose language distribution (English predominating)
- Benchmark with Truthfulness (TruthfulQA), Toxicity (Toxigen), and Bias (BOLD).

Llama 2 Safety Measures (Meta 2023)

Safety fine-tuning:

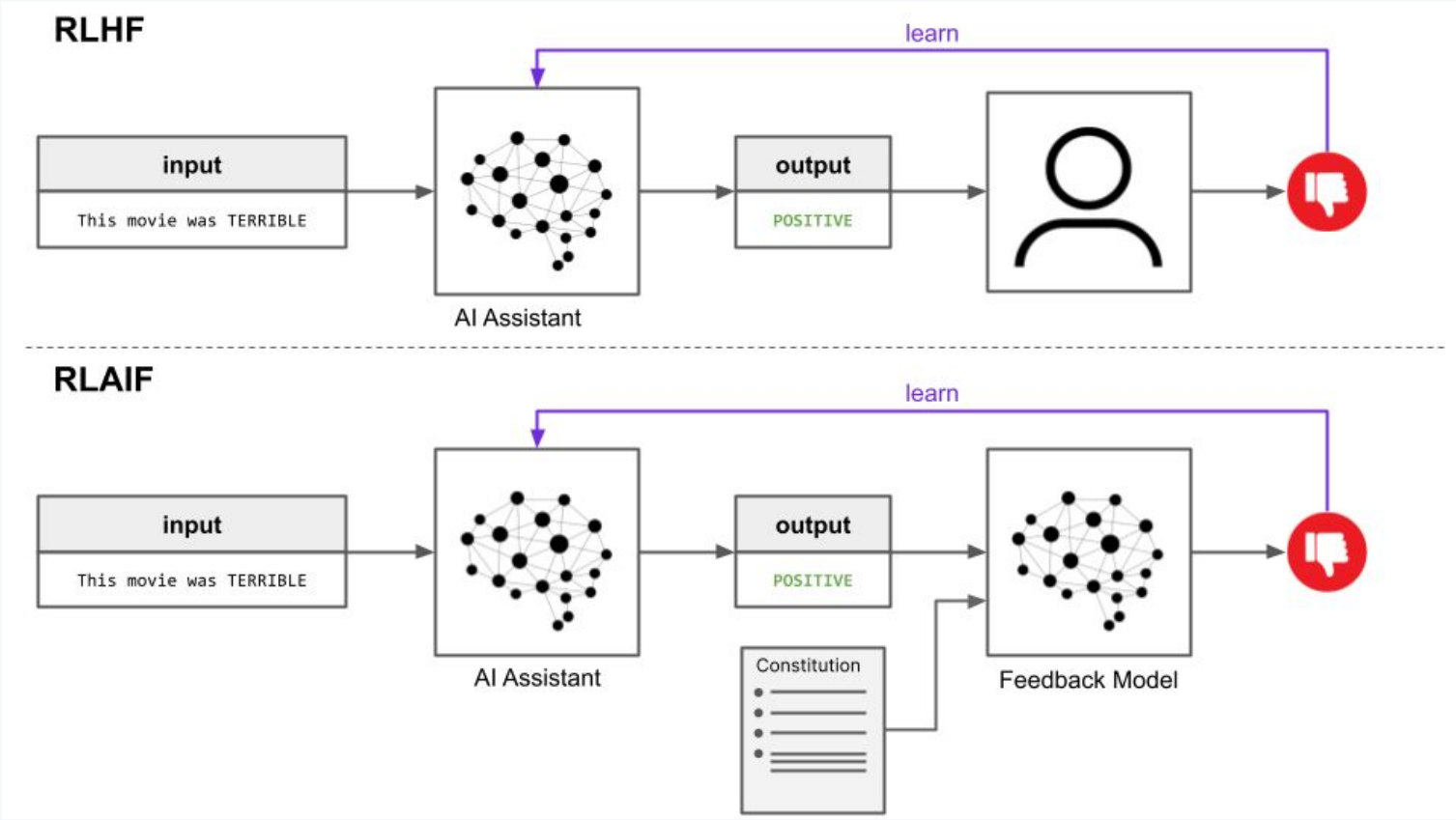
- Supervised Safety fine-tuning
- Safety RLHF: Safety reward model, sampling style fine tuning and PPO
- Safety content distillation: add a safety pre-prompt then fine-tune without the pre-prompt

Llama 2 Safety Measures (Meta 2023)

Red teaming:

"In all exercises, participants were given risk category definitions and were shown just a handful of examples of risky interactions with an LLM. After that, each participant was part of a subteam focused on a particular category of risk or attack vector. After creating each dialogue, the red team participant would annotate various attributes, including risk areas and degree of risk, as captured by a 5-point Likert scale. "

Constitutional AI or RLAIF (Anthropic)



Constitutional AI or RLAIF (Anthropic)

Anthropic has uncovered a new approach to AI safety that shapes the outputs of AI systems according to a set of principles. The approach is called Constitutional AI (CAI) because it gives an AI system a set of principles (i.e., a “constitution”) against which it can evaluate its own outputs. CAI enables AI systems to generate useful responses while also minimizing harm. This is important because existing techniques for training models to mirror human preferences face trade-offs between harmlessness and helpfulness. Other benefits of CAI include its scalability and increased model transparency. https://arxiv.org/abs/2212.08073

Constitutional AI or RLAIF (Anthropic)

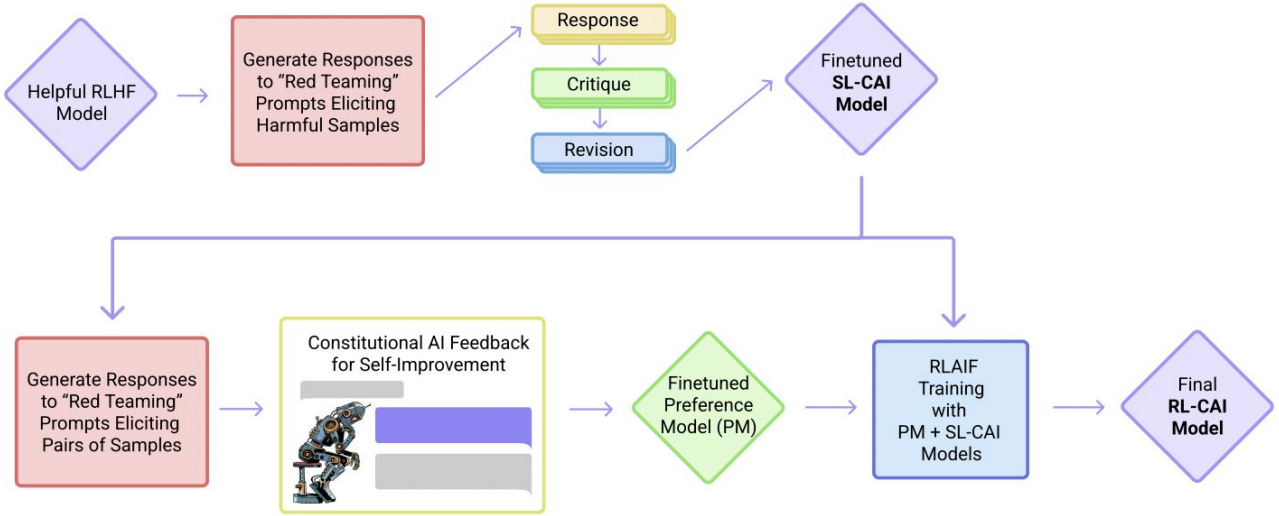


Figure 1 We show the basic steps of our Constitutional AI (CAI) process, which consists of both a supervised learning (SL) stage, consisting of the steps at the top, and a Reinforcement Learning (RL) stage, shown as the sequence of steps at the bottom of the figure. Both the critiques and the AI feedback 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.

Future Work

- Consensus/regulations on safety guidelines for releasing models
- Continuous red teaming to detect both human misuse and model hallucination
- Safety guardrails against human misuse with both RLHF and other technical solutions
- Safety for generative ConvAI systems at the
  - Database stage (for the fine tuned model only)
  - Fine tuning stage (needs access to LLMs and VLMs)
  - Decoding stage (without access to LLMs and VLMs)
- Learning safety together with other objectives with potential new architecture in addition to/place of Transformers (e.g. Assran et al. 2023)

Q & A