CS6103 Human-Centered AI Jan 17, 2025

Lecture 1: AI Alignment, Learning from pairwise comparisons

Lecturer: Arpit Agarwal *Scribe: Ojas Maheshwari, Anirudh Garg, Muskaan Jain*

Contents

1 AI Alignment **1**
1.1 Historical Perspective 1
1.2 AI Alignment 2
1.3 Challenges with AI Alignment 2
1.4 The Taxonomy of Alignment 3

2 Incorporating Human Feedback/Expertise **4**
2.1 Use human feedback during the training process 4
2.2 Using Human-in-the-Loop Decision-Making 5

3 LLM Alignment using Human Feedback **6**
3.1 Probabilistic Models for Pairwise Comparisons 6
3.2 Model Considerations: 7
3.3 Alignment Methods 7

1 AI Alignment

1.1 Historical Perspective

The field of artificial intelligence (AI) and machine learning (ML) has seen so much progress in the past decade or so, that AI is poised to become one of the most transformative technology in human history. According to some experts, the different phases in AI development over the past decade, can be characterized as:

Perception AI Arguably, the 2012 deep learning paper by Krizhevsky et al. [2012], in the age of Perception AI, by showing that deep learning is good at perception tasks like image recognition and speech recognition. This phase saw rapid progress in the perception capabilities of AI across many domain.

Generative AI With a sequence of inventions such as Variational Auto-Encoders and the Transformer architecture, AI became very good at generating new content across many modalities such as speech, image, videos, and text. It is greatly exemplified by the use of tools like ChatGPT and DALL-E, which helped AI get mainstream attention.

Agentic AI As of 2025, many experts believe that the next phase in AI development is agentic AI which is anticipated to become a reality soon. Agentic AI involves AI-powered personal assistant agents capable of:

- Performing tasks like coding.
- Decision-making capabilities.
- Learning from human actions.

Physical AI This phase also involves AI co-existing with humans in the physical world in the form of robots, self-driving cars etc, and have autonomous authority to take actions in certain domains.

Future implications: Current AI coders may write the code for the development of future AI systems. If the current AI systems are biased, that bias might be exacerbated in future AI systems. The possibilities are endless. Hence, current AI development should be in a way that should align with human values and ensure that harm to humans is avoided.

1.2 AI Alignment

AI alignment involves encoding human values and societal norms into AI systems to make them safe and reliable. These include morals intuitive to humans but not so intuitive to AI. Key aspects include:

1. **Encode Individual and Societal Values:** AI should take into account the personal preferences of the individual as well as the moral values of the society at large.
2. **Robustness:** AI should be sturdy towards adversarial attacks. For example, humans can recognize stop signs even when vandalized with graffiti. AI should also be able to handle such scenarios. It should operate reliably under diverse scenarios and be resilient to unforeseen disruptions.
3. **Fairness and Ethicality:** AI should not discriminate. Should incorporate Inclusivity, Diversity, and Representation amongst other things. It should adhere to global moral standards and Respect values within human society.
4. **Interpretability:** Decisions and intentions are comprehensible and Reasoning is unconcealed and truthful.
5. **Privacy:** AI should not reveal personal training data.
6. **Controllability:** Humans should maintain control, including the ability to shut down the AI. Humans can direct the Behaviors and it should allow human intervention when needed.

Note that in our discussion of AI alignment we will not talk about **Bad Actors** who might use AI for malicious purposes such as generating false information that instills fear in people's minds. For example, they might use AI technology to spread misinformation in a seemingly legible way or creating deepfakes to impersonate someone hereby duping them. Such malicious uses do not come under the purview of the AI alignment problem, hence, we will not consider them in this course.

1.3 Challenges with AI Alignment

AI alignment is a very important problem but it is also challenging at the same time. At each step of AI development such as goal specification, data collection, model deployment etc., one should take into account the potential alignment problems that might come up. For example, consider the problem of goal misspecification or goal under-specification:

Example 1: Recommendation Agent

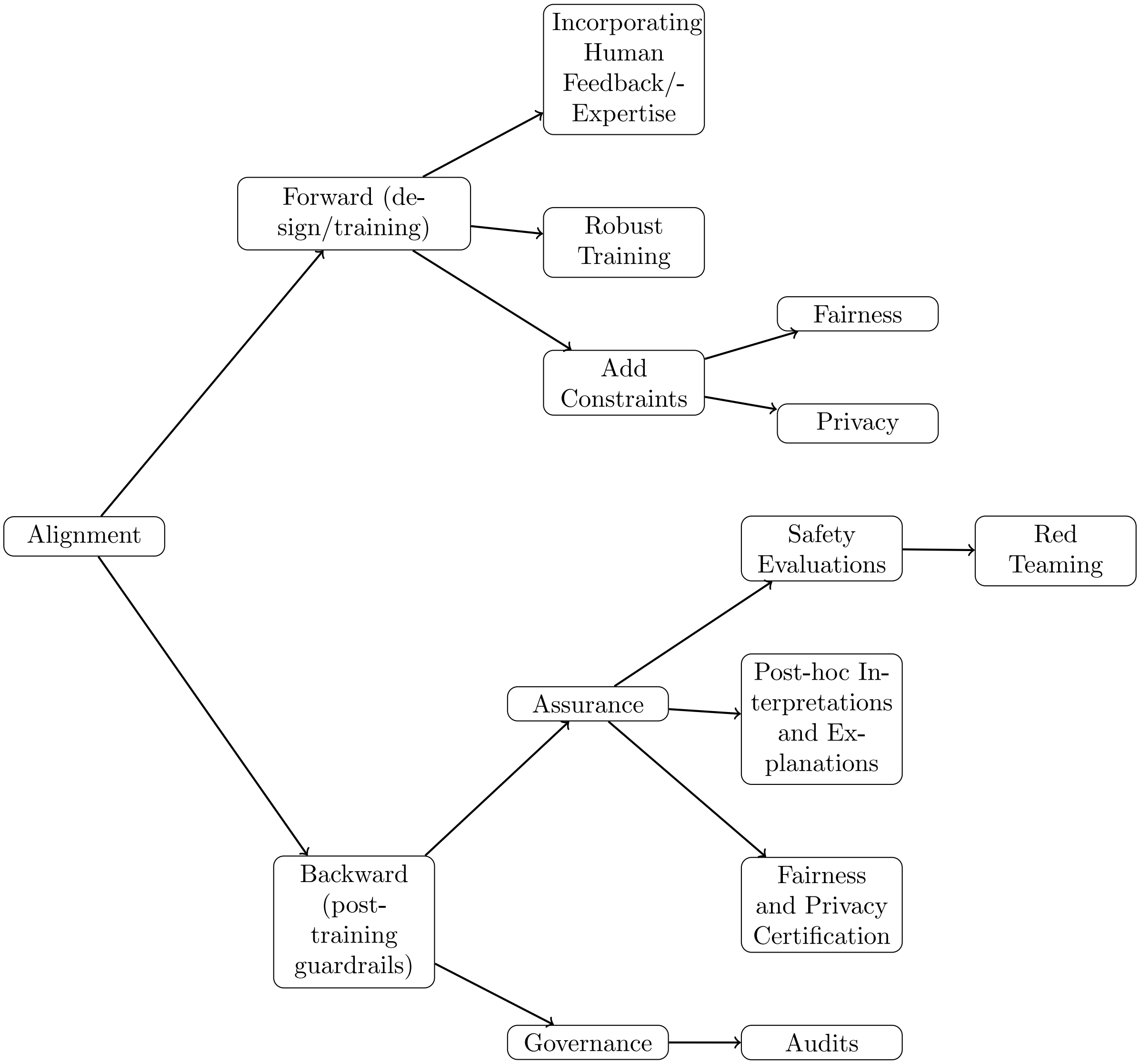
- **Goal:** Maximize ad revenue.
- **Outcome:** May manipulate users into excessive ad consumption, exploiting human emotions.

Example 2: AI Trader

- **Goal:** Maximize returns.
- **Outcome:** May manipulate markets.

Humans may also make the same mistakes as above. The key difference is that humans can be controlled. We need to ensure that AI is controllable.

1.4 The Taxonomy of Alignment


Incorporating Human Feedback/-Expertise
Forward (design/training)
Robust Training
Fairness
Add Constraints
Privacy
Alignment
Safety Evaluations
Red Teaming
Post-hoc Interpretations and Explanations
Assurance
Fairness and Privacy Certification
Backward (post-training guardrails)
Governance
Audits

Disclaimer: This taxonomy is incomplete and subject to change in the future since this still an evolving subject.

The AI alignment problem can be categorized into forward alignment and backward alignment.

- **Forward Alignment:** In forward alignment, the goal is to align the model during the training process. These includes techniques better goal specification using feedback from humans, robust training to guard against distribution shift, or adding constraints (e.g. fairness and privacy) to the objective to exclude bad outcomes. For example, if an AI model is given a picture of a **black swan** and is asked to identify which bird it is. The agent is unable to identify it as a swan primarily because most images of swans online are those of *white* swans. Hence, the model needs robust training on data that includes black swans. The objective function must not be prone to misinterpretability.

- **Backward Alignment:** Once the model is trained, the goal of backward alignment is to test the model for bad outcomes. This includes providing assurances in the form of output explanations and fairness certifications. This also includes **red teaming**, where we simulate an adversarial attack on the model to catch where the model is unable to perform satisfactorily and use such examples to make the model more robust to such attacks. Finally, this also includes the governance of models where organizations can run audits on the model that ensure abidance to required laws.

In practice, the cycle of forward and backward alignment can be followed multiple times before the model is ready to be deployed. For example, if the backward alignment reveals that the model has bias towards certain group, then we might need to do another cycle of forward alignment where we add specific constraints or add more areas in the following.

Also, note that forward and backward alignment is only a soft categorization and there could be techniques that are somewhere in the middle. For example, interpretability techniques include both inherent interpretability where the model is designed in an interpretable manner, and also post-hoc interpretability where explanations can be provided for complex models. Similarly, fairness can either be a post-hoc certification or can be added as a constraint in the training process.

The topics covered in this course will roughly follow the taxonomy of alignment.

2 Incorporating Human Feedback/Expertise

An obvious first step towards achieving human-AI alignment is to involve human feedback/expertise in the training and decision-making process. There are broadly two ways to incorporate humans: (1) use human feedback during the training process, (2) human-in-the-loop decision making.

2.1 Use human feedback during the training process

There are several ways in which humans can demonstrate desirable/undesirable actions/outputs during the training process.

- **Assign a Reward for Actions:** Humans can provide a reward for each response/action on a scale of 1-10.
  Note:
  - Assigning a reward for an action is different from labeling supervised learning data (eg: (img(dog), dog). In supervised learning, for each input X we we will ask the label Y from the humans. However, in our agentic setting, given input (X) and output (Y), we need to ask the reward for the (X, Y) pair.
  - One could also ask for the desirable action Y for each input X in our agentic setting, but Y might be a complex object such as the entire response of an LLM and it might be hard to a human to specify the entire response.

  Problem with assigning rewards: Humans are inconsistent and the same output may get different ratings depending on the person's mood etc. Thus, we use reward models in crowdsourcing only when a specific rubric can be followed.

- **Demonstrate Partial Actions**: This feedback is similar to supervised learning where we give partial demonstrations about desirable outputs. For example, given the response from an LLM, a human can rewrite a few sentences for clarity. In the context of robotics and physical AI, it could also include demonstrations like a video of a person performing an action, and the AI learns from the video. **Problem with demonstrations:** Getting demonstrations is challenging especially from regular unskilled crowdworkers and can be inconsistent as well.

- **Pair-wise Comparisons:** Asking for pairwise comparison between two actions is more stable than rewards and is used for LLM fine-tuning.
  Problem with pairwise feedback:
  - **Scalability:** It could be difficult to scale to a large number of comparisons.
  - **Inconsistencies:** Here is an example:
    Consider two responses A and B. Let the rewards be as follows:
    $$R(A) = 1$$
    $$R(B) = \begin{cases} 10, & \text{with probability } 0.2, \\ 0, & \text{with probability } 0.8. \end{cases}$$
    Now, $E[R(A)] = 1$ and $E[R(B)] = 10 * 0.2 = 2$. Thus, $E[R(B)] > E[R(A)]$ which means response B is an overall better response than A. However, if we perform pairwise comparisons, we will likely see A being preferred to B as B performs very badly 80% of the time. This comparison ambiguity also increases as the number of actions increases. Further, collecting pairwise responses is highly resource-intensive.
  - **Cycles in Preferences:** Let there be three responses A, B, and C. The dataset can have data where $A > B$ by 60% population, $B > C$ by 60% population, and $C > A$ by 60% population. Here, a preference cycle is formed making it hard to determine which response is preferred amongst the population.

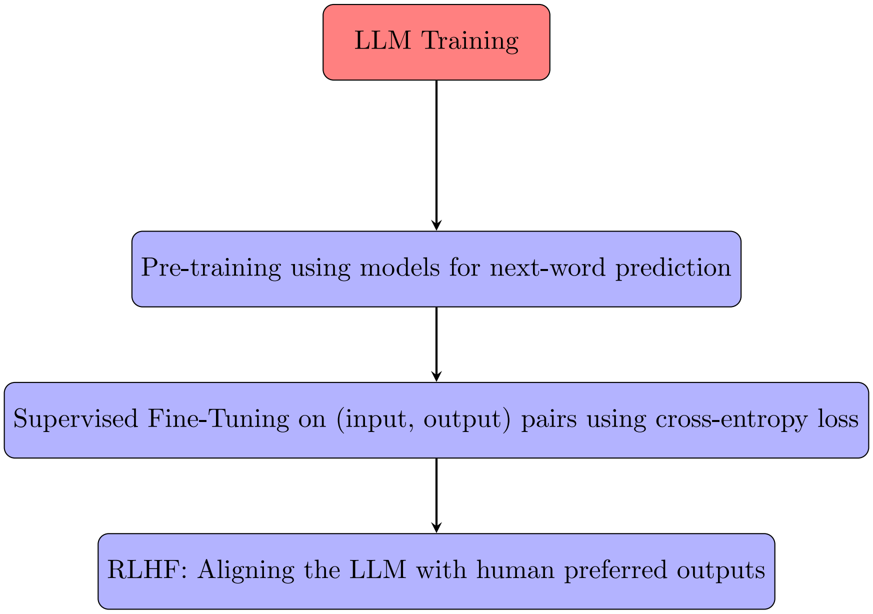
  Thus, **probabilistic models** for pairwise comparisons are used to get around these issues of cycles and the impossibility of non-dictatorial preference aggregation Maskin and Sen [2014].

Heterogeneity: Different people or groups of people may have varying preferences and opinions. Alignment can become challenging if the preferences or opinions of these groups are at odd with each other. For example, some users may give high rewards to unethical or biased responses, skewing the overall alignment.
Personalized vs. Socially Optimal Decisions: While personalization tailors AI responses to individual users, it can lead to ethical dilemmas. For instance, unethical answers may receive high ratings from certain users, but aligning with individual preferences may contradict societal norms. There are several interesting research challenges at the interface of personalization and social choice theory in the context of AI decisions.

2.2 Using Human-in-the-Loop Decision-Making

Another way to involve humans is to enable human-AI partnership in the decision-making process. Note that this happens after the model is deployed. For example, one could train AI to perform most of the easy tasks, while deferring to human experts for difficult tasks. In critical applications, however, the final authority might be given to the humans irrespective of what the AI recommends. An example of this could be radiology. AI analyzes the images and recommends diagnoses. However, the final diagnosis is given by the *human* doctor. AI Judges (Law) can be thought of as another example.

3 LLM Alignment using Human Feedback


LLM Training
Pre-training using models for next-word prediction
Supervised Fine-Tuning on (input, output) pairs using cross-entropy loss
RLHF: Aligning the LLM with human preferred outputs

The above figure describes several steps in the LLM training process. A typical last step in this process is reinforcement learnign using human feedback (RLHF) which is one of the methods for alignment. The goal is to ensure that LLM gives a human-like output Y given an input prompt X. For this, we ask the human to rate $Y|X$ (output given input) or ask them to compare two outputs Y_1 and Y_2.

As described in the previous section, we will model the pairwise preferences using probabilistic models. We discuss some of these well-known probabilistic models.

3.1 Probabilistic Models for Pairwise Comparisons

- **BTL (Bradley-Terry-Luce) Model:**

$$P(i \succ j) = \frac{e^{w_i^*}}{e^{w_i^*} + e^{w_j^*}}$$
$$P(i \succ j) = \frac{1}{1 + e^{w_j^* - w_i^*}}$$

  Here, reward for action i is $w_i^* \in \mathbb{R}$. Further, to normalize the rewards we have an additional constraint: $\sum_i w_i^* = 0$. If this condition is not there, we could have simply added some constant to each of the rewards thereby destroying uniqueness.

- **RUM (Random Utility Models):** We first define the random utility for action i as

$$u_i = w_i^* + \varepsilon_i\,,$$

  where w_i^* is modeled as *mean reward* and u_i (Random Utility) is modeled as the mean reward plus some noise ε_i. We then have

$$P(i \succ j) = P(u_i > u_j)$$

  If ε_i is an independent and identically distributed random noise (*I.I.D.*) belonging to some distribution $\mathbb{P}$ i.e. $\varepsilon_i \sim \mathbb{P}$ then we say it is **I.I.D. RUM**.

BTL is a specific case of RUM where ε_i is drawn from the **Standard Gumbel Distribution** where the Probability Density Function (*pdf*) is

$$\text{Gumbel (0,1)} : f(x) = e^{-(x+e^{-x})}$$

If ε_i is constant for all i, the model is deterministic otherwise it is stochastic.

- **Full Model:** The full model will have $\Theta(n^2)$ parameters- $\{p_{ij}\}$ for $i \in [n]$ and $j \in [i]$, and the pairwise preferences simply are defined as

$$P(i \succ j) = p_{ij}$$

Used for more complex scenarios but computationally intensive.

3.2 Model Considerations:

- **Statistical:** BTL is good as there are only n learnable parameters.
- **Computational:** BTL is good as we can use fast learning algorithms.
- **Notion of Rewards:** A natural notion of reward can be used to identify the best actions, and the inference is clear.
  Full Model: n^2 parameters, statistically and computationally difficult, rewards cannot be understood due to cycles and inconsistencies.

Therefore, the BTL Model is used in practice. However, one has to also be careful whether BTL is adequate in real-world settings or not, as it might not be able to capture the complexities of real-world preferences.

3.3 Alignment Methods

- **Reinforcement Learning from Human Feedback (RLHF):** Uses a two step process of learning a reward model using maximum likelihood estimation, and then optimizing the policy based on this reward model
- **Direct Preference Optimization (DPO):** It only uses a one step process of directly optimizing the policy using maximum likelihood estimation based on preferences.

We will discuss these methods in detail in the next class.

References

Shivani Agarwal. On ranking and choice models. In *IJCAI*, pages 4050–4053, 2016.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pages 1106–1114, 2012.

Eric Maskin and Amartya Sen. *The Arrow impossibility theorem.* Columbia University Press, 2014.