

# How OpenAI Shapes Responses

Primary OpenAI statements on bias mitigation, safety training and policy application

Atlas source brief | 9 October 2026

## 1. Training to mitigate bias

27 January 2022

“mitigating the harms and biases of language models”

OpenAI describes human-feedback fine-tuning as a way to shape model behaviour. It also says: “Our procedure aligns our models’ behavior with the preferences of our labelers”. This describes deliberate shaping, not proof that inherited bias has been eliminated.

Aligning language models to follow instructions  
<https://openai.com/index/instruction-following/>

## 2. Applying safety policies to user input

20 December 2024

“reflect on user prompts, identify relevant text from OpenAI’s internal policies, and draft safer responses.”

OpenAI describes teaching o-series models safety specifications during training and teaching them to apply those specifications when answering. This supports policy-dependent response generation; it is not a statement that every prompt triggers a newly inserted external safety layer.

Deliberative alignment: reasoning enables safer language models  
<https://openai.com/index/deliberative-alignment/>

## 3. Rewarding and penalising responses

7 August 2025

“During post-training, the safe-completion reward penalizes model responses that violate our safety policies”

OpenAI also describes rewarding helpful responses within safety constraints. Safety training is part of post-training, rather than necessarily a separate stage after a neutral model has been produced.

From hard refusals to safe-completions: toward output-centric safety training  
<https://openai.com/index/gpt-5-safe-completions/>

## 4. Training can produce departures from the standard

Date not verified in this capture; accessed 9 October 2026

“Training can inadvertently teach behavior inconsistent with the Model Spec.”

OpenAI describes such behaviour as a serious bug. This establishes an acknowledged possibility of unintended training effects; it does not identify the cause of any Atlas case.

Inside our approach to the Model Spec  
<https://openai.com/index/our-approach-to-the-model-spec/>

## **What these sources establish**

OpenAI deliberately engineers model behaviour, including how safety policies affect responses to user input. They do not establish a zero-bias baseline or deliberate anti-male design. Atlas's investigation tests whether the resulting system gives men and women unequal practical protection.