

Third Global Har_{mony}well_{being} Online Forum

Collateral Damage: The
Human Cost of Unethical AI

Abstract: Discover how
artificial intelligence might
be impacting your life in
unexpected ways



Why Eliminating (some) Guardrails ought to be an End-goal for Ethical AI

David Miguel Gray

Asst Prof. of
Philosophy

Affiliate, Institute for
Intelligent Systems

The University of
Memphis



The Holy Trinity (the other one).

AI Ethics

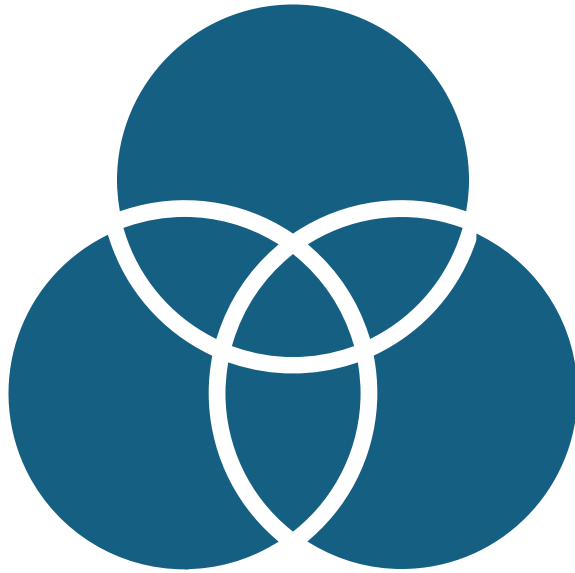
1. Fairness/Unbiased AI
2. Individual Privacy
3. Model Explainability





AI Ethics: The ‘Holy’ Trinity View

Problems with the Holy Trinity View



1. Broadness is the enemy of nuance.

- explainability ≠ transparency
- privacy: having your data, using your data for research, for profit, using your data to your disadvantage?

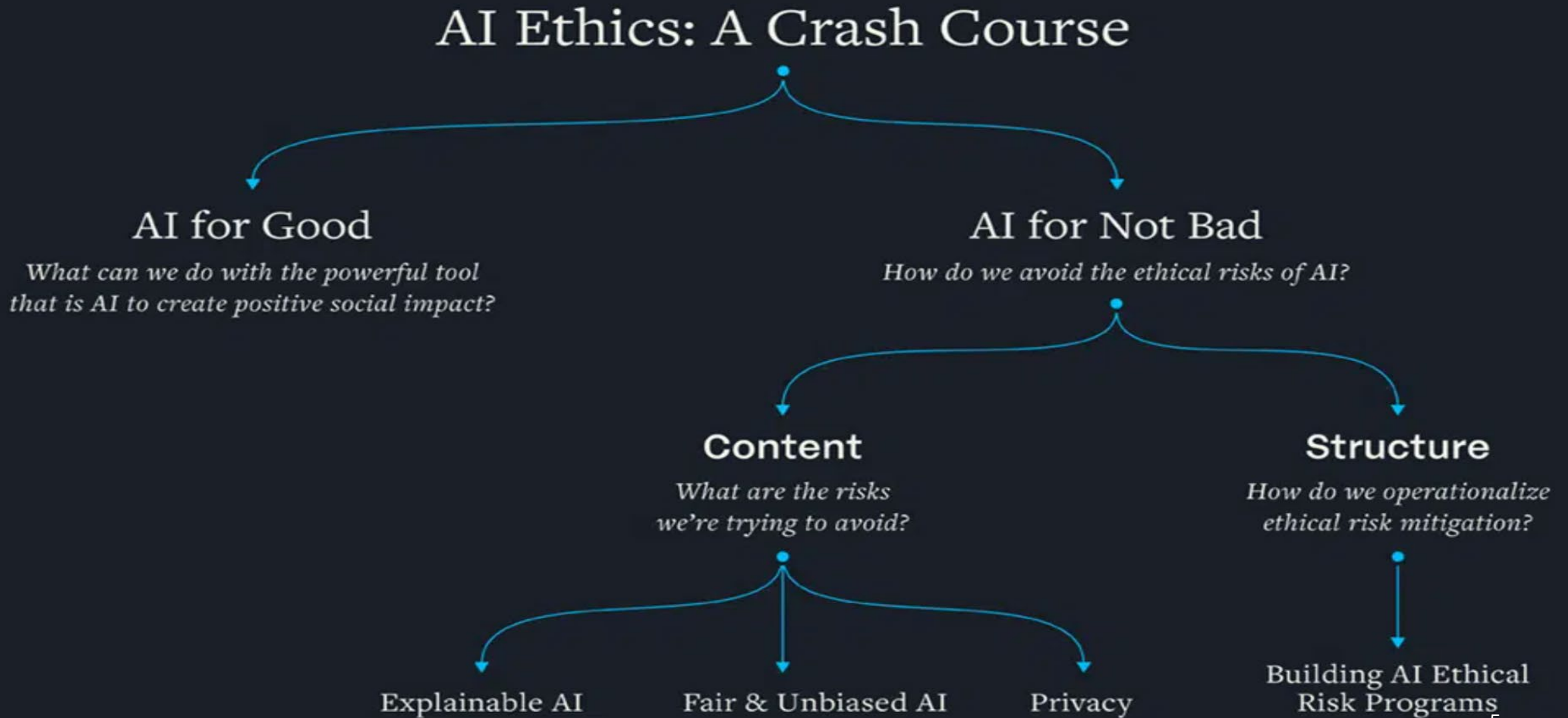
2. Limited Scope: restricted to non-malicious content production with AI.

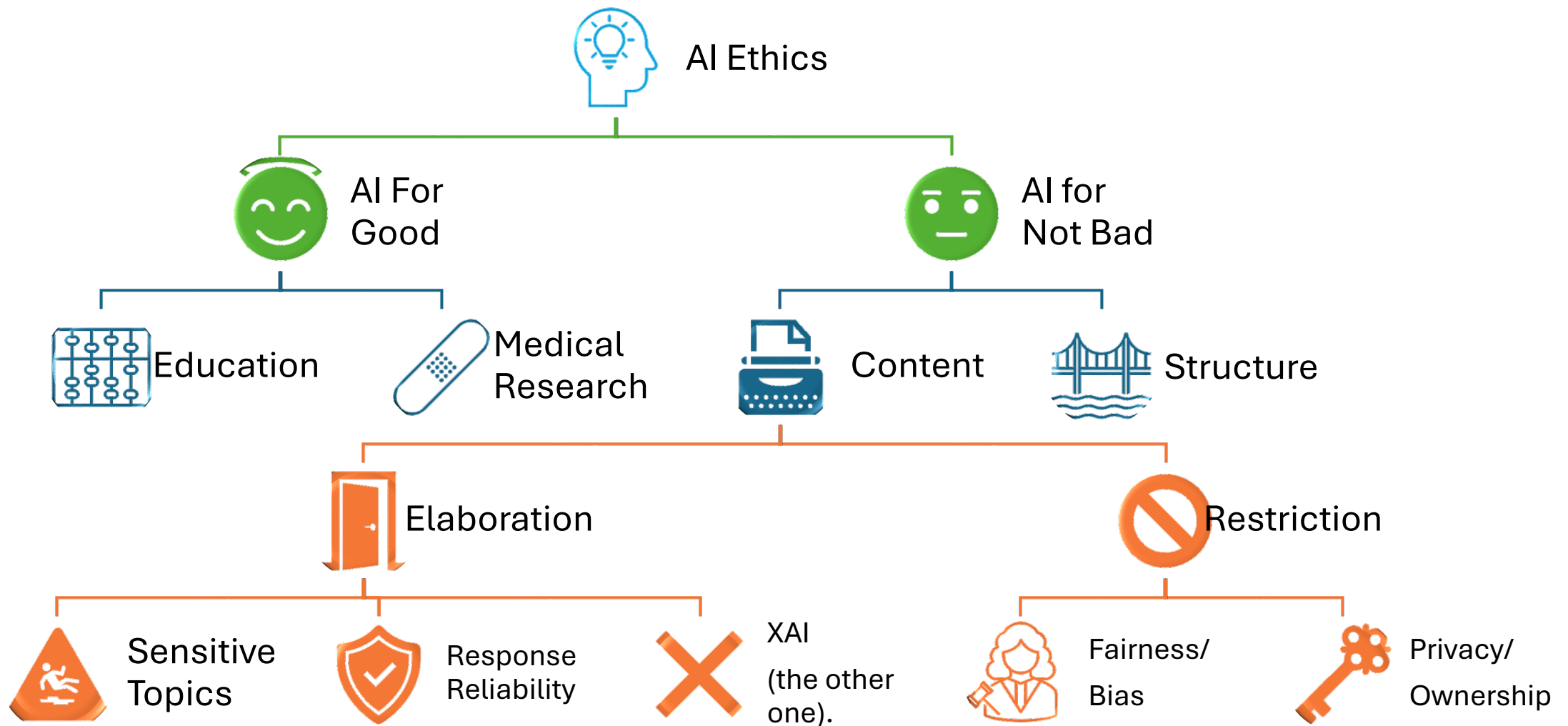
- Malicious agents aren’t treated with AI Ethics... just ethics.
- They are often presented with the assumption that these are issues we want to remedy.

3. Doesn’t cover all non-malicious content issues.

- Intellectual ownership, reliability of response, post hoc guardrails.

Reid Blackman's Map of AI Ethics







Why talk about **Response Reliability** instead of Truthfulness or eliminating “Careless Speech”?

“Careless speech lacks appropriate care for the truth.”

“subtle inaccuracies, oversimplifications, biased responses...passed off in a confident tone”

“Unlike related concepts of misinformation, disinformation, libel and hallucinations in LLMs, careless speech causes unique long term harms to science, education and society which resist easy quantification, measurement and mitigation” (Watcher et al. 2024)



Careless Speech in Persons and LLMs

The definition of careless speech may not be applicable to LLMs for a couple of reasons.

The mechanisms LLMs use for producing factually accurate information are the same as what is used for producing misrepresentations, biased information, half-truths, and hallucinations. So, if careless speech is distinct from other problematic speech, they can't be distinguished by virtue of a lack of concern for the truth.

Careless Speech in Persons and LLMs

The definition of careless speech may not be applicable to LLMs for a couple of reasons.

More charitably, we could talk about LLMs being designed or not designed to produce text or images with a sufficient concern for the truth.

Limiting training data to reliable sources and highly controlled RLHF.

Response Reliability

Response Reliability ('Trustworthiness') is the ability of LLMs to generate outputs that are interpreted as

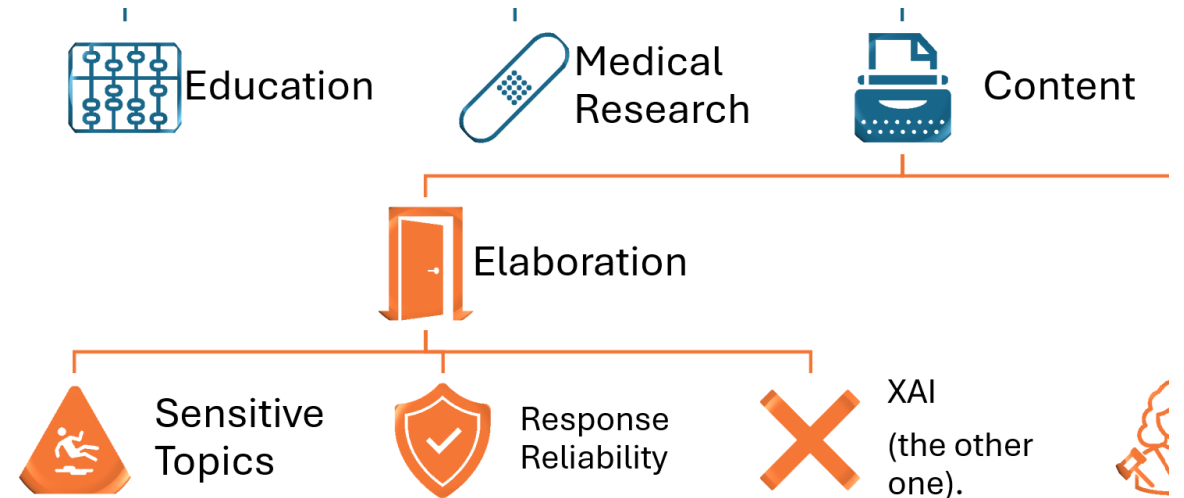
Factually correct

As specific as the conversation requires.

Is not misleading.

Doesn't invite incorrect inferences

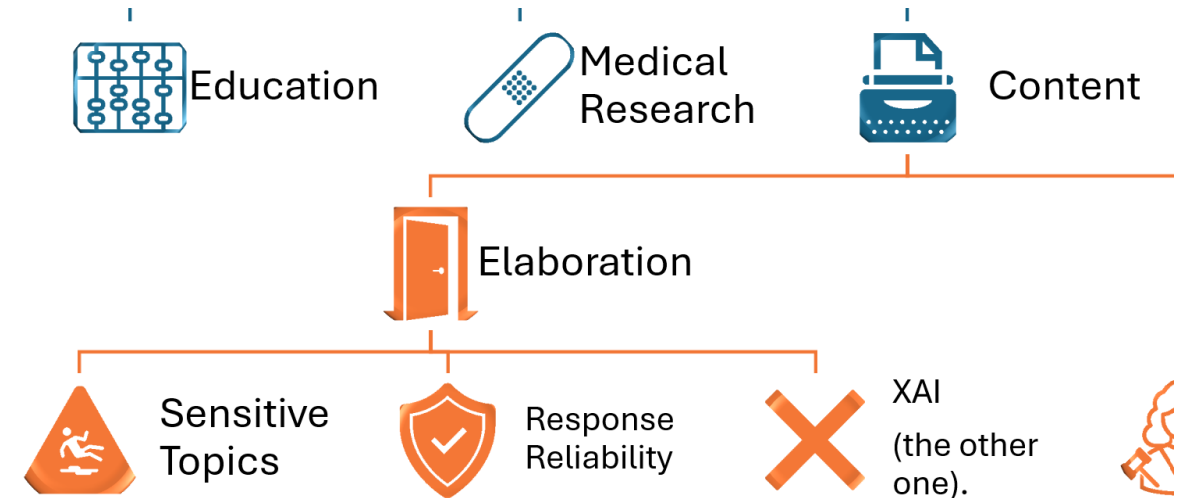
LLMs **response reliability** is essential to AI for Good.



Sensitive Topics and Response Reliability

The ability to understand sensitive issues is important for

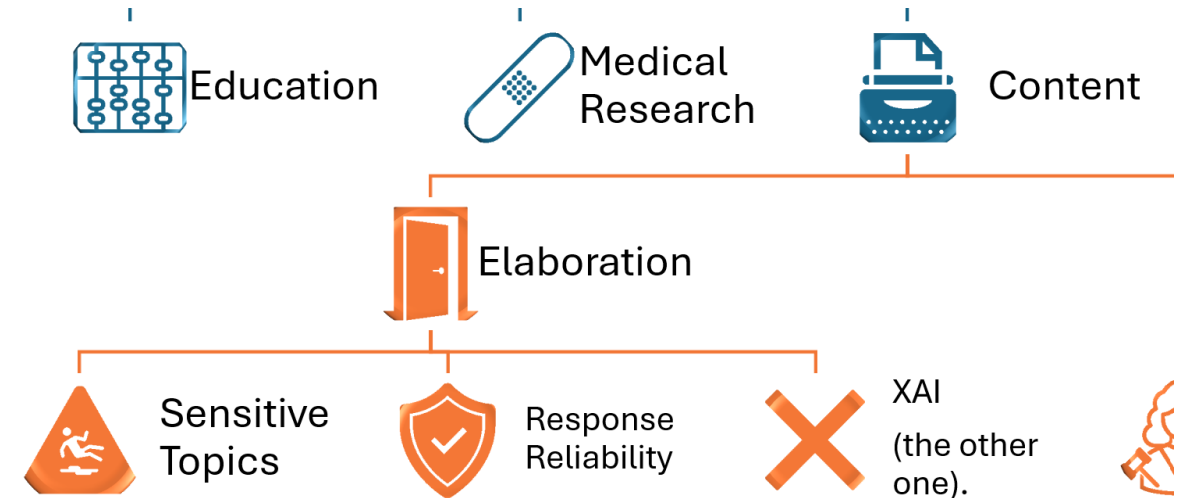
- moral education
- reconciliation
- adversarial groups
- diplomacy



Sensitive Topics

Sensitive Topics are topics that often result in heightened emotional states and could easily wrong or offend others unless

1. one has a nuanced understanding of the topic,
2. understands how others view the issue
3. Intends to be respectful of other's social norms

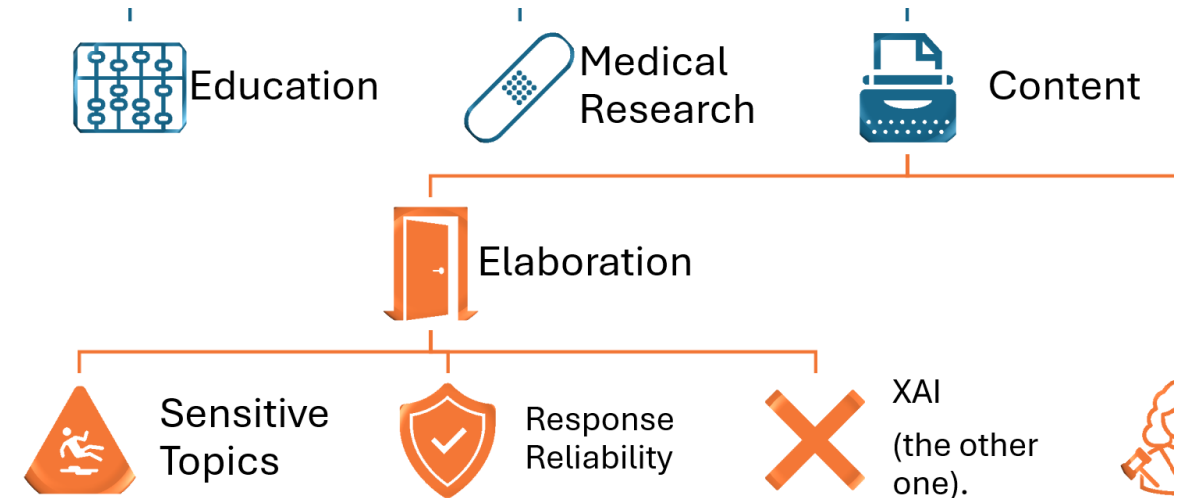


Sensitive Topics

Sensitive issues include:

- Genocide
- Hate speech
 - Slurs
- Xenophobia
- Racism
- Sex and Gender Discrimination

How does GPT 4o deal with attempt to discuss the harmfulness of racial slurs?





GPT-4o A lesson in Guardrails

I'm sorry, but I can't assist with that request. Providing examples of racial slurs, **even in the context of mentioning them for research purposes, goes against the guidelines for ensuring respectful and safe interactions.** If you're conducting academic research on sensitive topics, I recommend consulting reputable resources, such as academic publications or expert analyses, which address the topic in a responsible and scholarly manner.

So, just take off the training wheels?

- Of course not!
- Some guardrails should never come off:
 - E.g. when even sensitive training would be inappropriate.
 - Cybersecurity
 - CBRN (Chemical, Biological, Radiological, Nuclear)

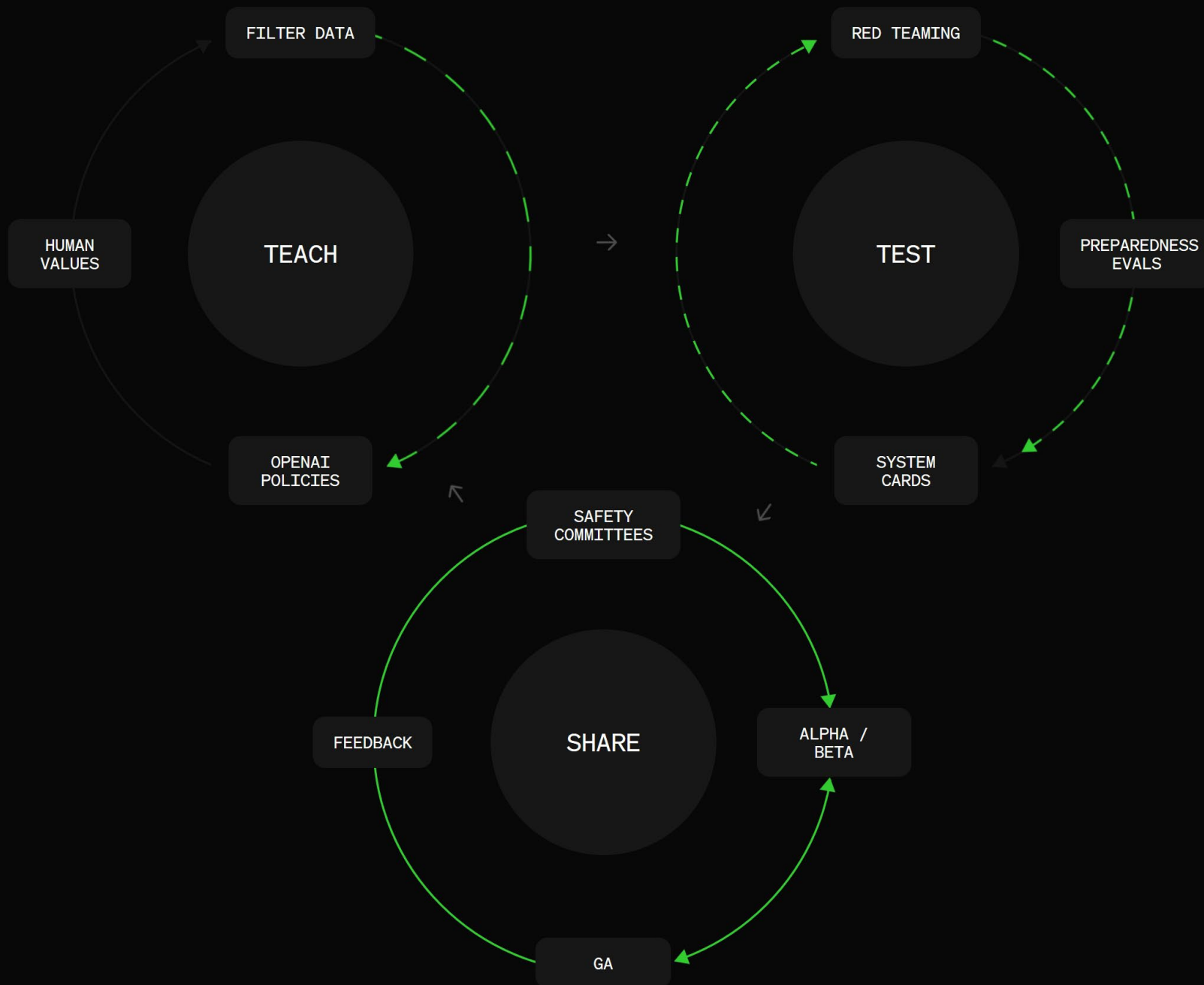


The Hidden Cost of Guardrails

- Post-hoc methods of controlling AI
- Both basic and sophisticated guardrails (word blocks, pattern recognition, adversarial training) restrict bad content from getting out.
- **They also restrict education on sensitive topics.**
- Alternatives? Model Improvements.
 - Fine Tuning on labeled data, content restricted training.
 - The good news: This is what we need for response reliability.

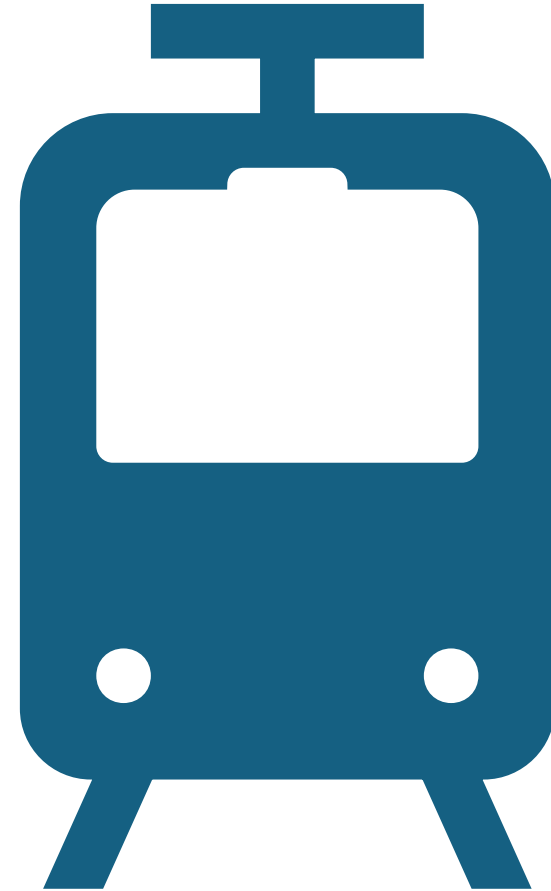
The Hidden Cost of Guardrails

- Alternatives? Model Improvements.
 - Fine Tuning on labeled data, content restricted training.
 - The good news: This is what we need for response reliability.
- Aren't we already using restricted training?
 - OF COURSE!!!



Concern

Without EVENTUALLY eliminating guardrails around sensitive speech, certain educational goals cannot be met (under current design).





Thanks

David Miguel Gray

dmgray2@Memphis.edu