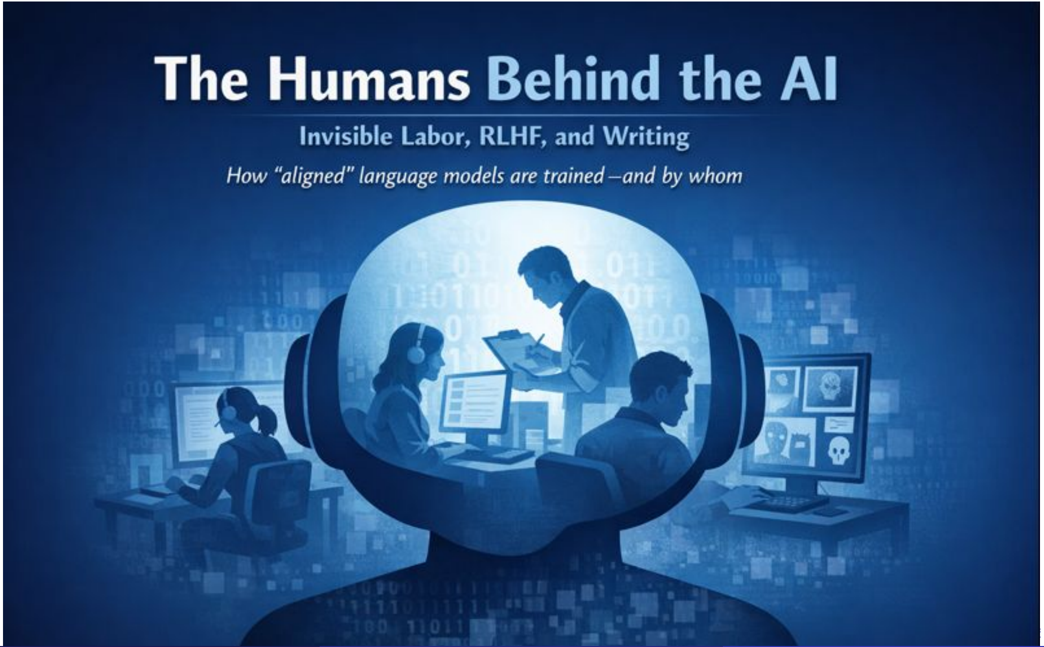


The Humans Behind the AI

Invisible Labor, RLHF, and Writing

How "aligned" language models are trained—and by whom

Opening Question

When you use an AI chatbot,

who do you imagine is shaping its voice?

- Engineers?

Opening Question

When you use an AI chatbot,

who do you imagine is shaping its voice?

- Engineers?
- The algorithm itself?

Opening Question

When you use an AI chatbot,

who do you imagine is shaping its voice?

- Engineers?
- The algorithm itself?
- “Society”?

Opening Question

When you use an AI chatbot,

who do you imagine is shaping its voice?

- Engineers?
- The algorithm itself?
- “Society”?
- Human workers?

Opening Question

When you use an AI chatbot,

who do you imagine is shaping its voice?

- Engineers?
- The algorithm itself?
- “Society”?
- Human workers?

RLHF in One Slide

- RLHF = **Reinforcement Learning from Human Feedback**

RLHF in One Slide

- RLHF = **Reinforcement Learning from Human Feedback**
- Human beings:

RLHF in One Slide

- RLHF = **Reinforcement Learning from Human Feedback**
- Human beings:
  - Rank responses

RLHF in One Slide

- RLHF = **Reinforcement Learning from Human Feedback**
- Human beings:
    - Rank responses
    - Judge helpfulness and harm

RLHF in One Slide

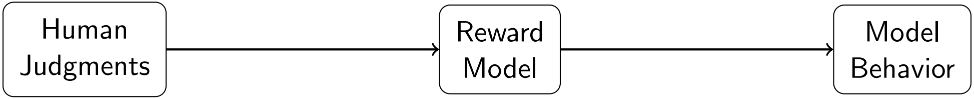
- RLHF = **Reinforcement Learning from Human Feedback**
- Human beings:
  - Rank responses
  - Judge helpfulness and harm
  - Label toxic or disturbing content

RLHF in One Slide

- RLHF = **Reinforcement Learning from Human Feedback**
- Human beings:
  - Rank responses
  - Judge helpfulness and harm
  - Label toxic or disturbing content
- These judgments become **reward signals**

RLHF in One Slide

- RLHF = **Reinforcement Learning from Human Feedback**
- Human beings:
  - Rank responses
  - Judge helpfulness and harm
  - Label toxic or disturbing content
- These judgments become **reward signals**



Human Judgments

Reward Model

Model Behavior

What's Usually Left Out

What’s Usually Left Out

- AI does not judge itself

What's Usually Left Out

- AI does not judge itself
- Safety doesn't come from nowhere

What’s Usually Left Out

- AI does not judge itself
- Safety doesn’t come from nowhere
- Alignment = **human judgment at scale**

What's Usually Left Out

- AI does not judge itself
- Safety doesn't come from nowhere
- Alignment = **human judgment at scale**

Invisible labor

Ghost Work

> *“Work performed by humans but perceived as automated.”*

by Gray & Suri, *Ghost Work*

- Labeling
- Ranking
- Moderation
- Quality control

Mary Gray short video: On YouTube, 5 minutes

Who Does This Work?

- Contract and outsourced workers
- Frequently in the Global South
- Paid per task / per hour
- Limited job security and benefits

Discussion: What incentives push this work offshore?

What Are They Asked to Read or Watch?

Depends on the job, but there are jobs that are dangerous to your emotional well being:

What Are They Asked to Read or Watch?

Depends on the job, but there are jobs that are dangerous to your emotional well being:
Review content containing:

- Graphic violence or sexual abuse
- Hate speech or Extremist material
- Exposure is repeated and high volume.

Psychological Impact

Research shows elevated rates of:

- PTSD
- Depression
- Anxiety
- Intrusive thoughts

> *"Mental health symptoms were significantly elevated among moderators."*

Safeguarding Digital "First-Responders"

Validate the difficulty of the work while making the case for better mental health infrastructure.

- Moderators act as a vital safety shield.
- Without robust support, these workers faces PTSD and depression.
- Recognizing the risks makes mental health support a non-negotiable design choice.

Sama - Success story

- Sama is a B-Corp
  -for-profit, officially certified for meeting high standards of social and environmental performance, transparency, and accountability
- uses "Impact Sourcing"
  -hire and train workers from marginalized backgrounds (such as in Kenya and Uganda) at living wages.
- What's different?
  Unlike the "ghost work" model of anonymous gig platforms, Sama employs workers in formal office environments. They provide healthcare, professional training, and career pathing.
- AI training can be a vehicle for poverty reduction and skills development when done with an ethical corporate structure.

RLHF Trains a Voice

- Politeness
- Calm tone
- Refusal language
- What sounds “reasonable”

Whose writing norms are being enforced?

Bias Isn't Gone: RLHF as a Filter

- RLHF can reduce explicit harm
- But it may preserve dominant norms
- "Nice" does not mean neutral

Claim: Bias is often filtered, not removed; this is an iterative journey

Who Is Responsible?

- AI companies
- Contractors
- Governments
- Users
- “The system”?

A Design Choice?

- Better pay and protections
- Mental health support
- Shorter exposure windows / rotation
- Transparency (who, where, how trained)
- Alternative training methods

Big Reframe

AI is not replacing human judgment.

It is redistributing it.

Final Discussion Question

Should AI reflect the world as it is

or the world as we want it to be?

Who pays the cost of that choice?

Final Discussion Question

Should AI reflect the world as it is

or the world as we want it to be?

Who pays the cost of that choice?

Exit Ticket: Write a couple of sentences, including one quote or phrase from today plus one policy suggestion.