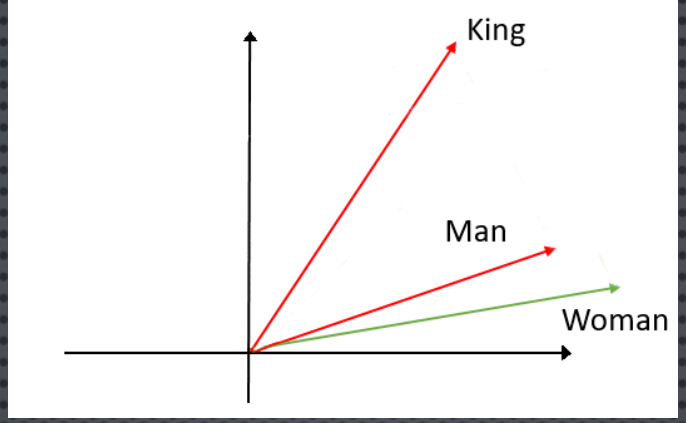
ON TRUTH, VALUES, KNOWLEDGE, AND DEMOCRACY IN THE AGE OF GENERATIVE "AI"

Dr. Damien Patrick Williams
Assistant Professor of Philosophy
Assistant Professor of Data Science
University of North Carolina at Charlotte

A (Very) Little Bit About How "Generative AI" Works



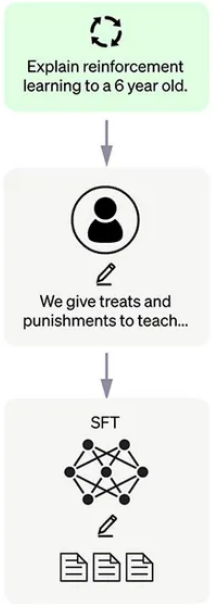
Step 1

Collect demonstration data and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



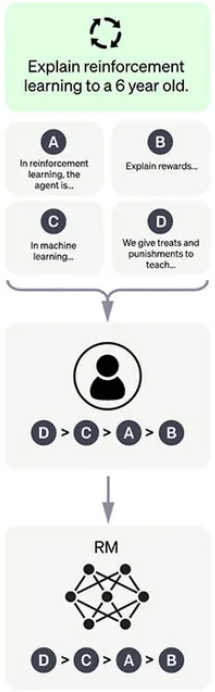
Step 2

Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

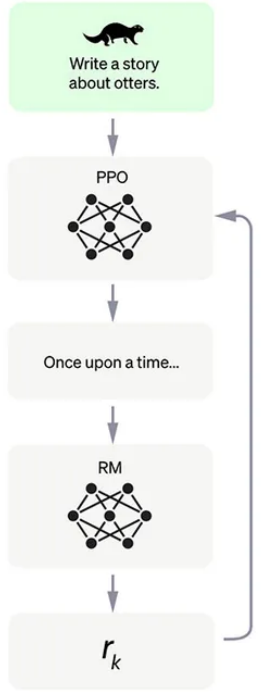
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

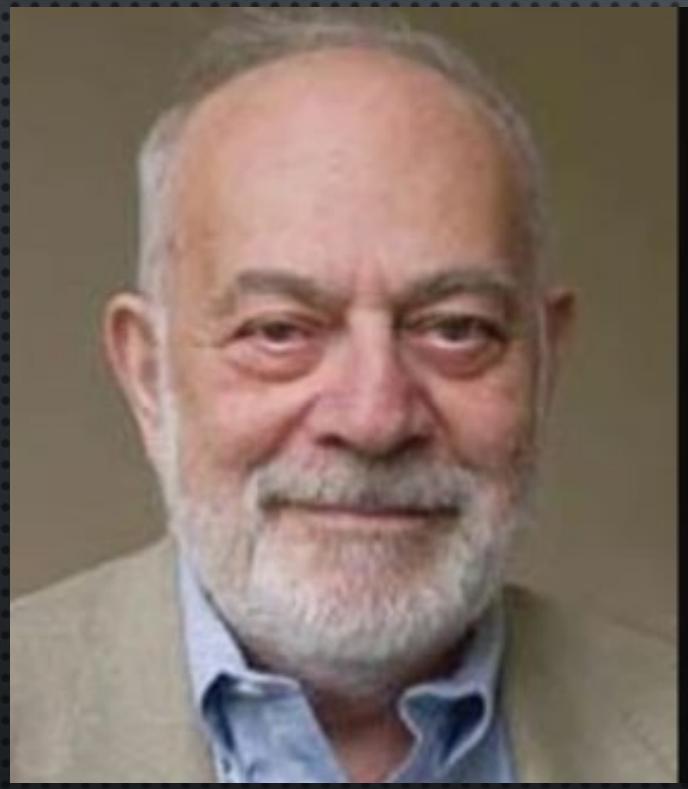
The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



On Bullshit



The bullshitter is neither on the side of the true or the side of the false. His eye is not on the facts at all. He does not reject the authority of the truth, as the liar does, and oppose himself to it. He pays no attention to it at all. By virtue of this, bullshit is a greater enemy of the truth than lies are.

— Harry Frankfurt —

Healthy Hormones
WITH GENDERGP

Embracing AI-Generated Content: A Positive Step Towards Gender-Affirmative Information

A.I. tools fueled a 34% spike in Microsoft's water consumption, and one city with its data centers is concerned about the effect on residential supply

BY MATT O'BRIEN, HANNAH FINGERHUT AND THE ASSOCIATED PRESS
September 9, 2023 at 11:01 AM EDT

A Problem With AI-Based Credit Models

FEBRUARY 2024

CCG | Catalyst

TOM SIMONITE BUSINESS 10.24.2019 02:00 PM

A Health Care Algorithm Offered Less Care to Black Patients

A study shows the risks of making decisions using data that reflects inequities in American society.

AI translation is jeopardizing Afghan asylum claims

Cost-cutting translations are introducing errors and putting refugees at risk.

future tense

How Algorithmic Bias Hurts People With Disabilities

Though a huge portion of the population lives with a disability, it comes in many different forms, making bias hard to detect, prove, and design around.

By ALEXANDRA REEVE GIVENS FEB 06, 2020 • 5:15 PM

Boston University Suggests Replacing Striking Grad Students With AI

MORE LIKE BOT-STON

In response to a grad student worker strike, the school recommends that staff utilize generative AI tools "to give feedback or facilitate 'discussion' on readings or assignments."

Tony Ho Tran | Updated Mar. 28, 2024 3:40PM EDT / Published Mar. 28, 2024 12:31PM EDT

- *Engage generative AI tools to give feedback or facilitate "discussion' on readings or assignments*

arXiv > cs > arXiv:2304.02819

Computer Science > Computation and Language

[Submitted on 6 Apr 2023 (v1), last revised 10 Jul 2023 (this version, v3)]

GPT detectors are biased against non-native English writers

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, James Zou

Artificial intelligence (AI)

Revealed: a California city is training AI to spot homeless encampments

Todd Feathers

Mon 25 Mar 2024 11.00 EDT

Exclusive: Google Contract Shows Deal With Israel Defense Ministry

The Israeli Ministry of Defense, according to the document, has its own "landing zone" into Google Cloud—a secure entry point to Google-provided computing infrastructure, which would allow the ministry to store and process data, and access AI services.

The ministry sought consulting assistance from Google to expand its Google Cloud access, seeking to allow "multiple units" to access automation technologies, according to a draft contract dated March 27, 2024. The contract shows Google billing the Israeli Ministry of Defense over $1 million for the consulting service.

Tackling healthcare's biggest burdens with generative AI

July 10, 2023 | Article

The introduction of AI technologies in warfare, **particularly through systems like IDF's targeting program "Gospel,"** has significantly obscured the traditional distinctions between combatants and civilians. This development introduces profound concerns regarding civilian safety and challenges the foundational norms of international armed conflict.

"Gospel," engineered for heightened efficiency (not necessarily precision) in the identification and engagement of targets, **relies on intricate algorithms that process extensive data to make immediate decisions.** This dependence on technology during combat complicates established military ethics and blurs the definitions of legitimate targets.

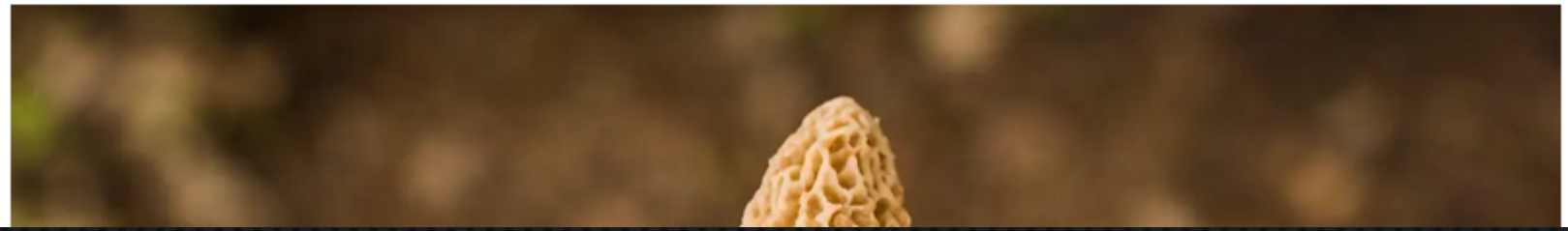
Nava and Benefits Data Trust partner to explore how AI can support public benefits navigators

Nava, Benefits Data Trust, and strategic advisors from government, academia, and civil society are exploring how generative AI can support benefits navigators to quickly enroll people in government benefits programs so they have more time to spend with the families they serve.

For the Love of God, Don't Let AI Choose Your Mushrooms

Artificial intelligence isn't a reliable source of information. Why trust it with your life?

By **Dennis Lee** Published September 14, 2023 | Comments (13)



> The AI design of your email is clever, but significantly lacks warmth.
>
> Would it be possible to be contacted by a human being moving forward instead of AI?
>
> Many thanks,

It's not an AI. I'm just Autistic.

See you next Friday.

When Artificial Intelligence Gets It Wrong

Unregulated and untested AI technologies have put innocent people at risk of being wrongly convicted.

The use of such biased technology has had real-world consequences for innocent people throughout the country. To date, six people that we know of have reported being falsely accused of a crime following a facial recognition match — all six were Black. Three of those who were falsely accused in Detroit have filed lawsuits, one of which urges the city to gather more evidence in cases involving facial recognition searches and to end the "facial recognition to line-up pipeline."

The problem is not and never has been that "AI" is out of step with human values; it's that the values with which "AI" is in step are

a) extremely human, and also, in very large part, both

b) woefully under-examined, and

c) terrible.

BAD BLOOD

Trump Pleads Ignorance After Sharing AI-Generated Taylor Swift Images

The former president says he's not worried about getting sued because he didn't generate the images himself

BY NAOMI LACHANCE

AUGUST 22, 2024

New Hampshire investigating fake Biden robocall meant to discourage voters ahead of primary

Paradigmatic Capture

- That our current cultural contention around even having a conversation about democratic or multicultural values has arisen in exactly the moment that the most recent wave of "AI" captures the public imagination is telling in its own right

Paradigmatic Capture

- That our current cultural contention around even having a conversation about democratic or multicultural values has arisen in exactly the moment that the most recent wave of "AI" captures the public imagination is telling in its own right
- Whoever controls the definition of a thing controls the conversation around that thing, and controlling the conversation shapes the cultural narrative.

AND OBFUSCATION, MISINFORMATION, AND DISINFORMATION, ARE MECHANISMS OF CONTROL

Objectivity, Facts, Intersubjectivity, & the Situatedness & Construction of Knowledge

Fig. 1.1. Truth-to-Nature. *Campanula foliis hastatis dentatis*, Carolus Linnaeus, *Hortus Cliffortianus* (Amsterdam: n.p., 1737), table 8 (courtesy of Staats- und Universitätsbibliothek Göttingen). Drawn by Georg Dionysius Ehret, engraved by Jan Wandelaar, and based on close observation by both naturalist and artist, this illustration for a landmark botanical work (still used by taxonomists) aimed to portray the underlying type of the plant species, rather than any individual specimen. It is an image of the characteristic, the essential, the universal, the typical: truth-to-nature.

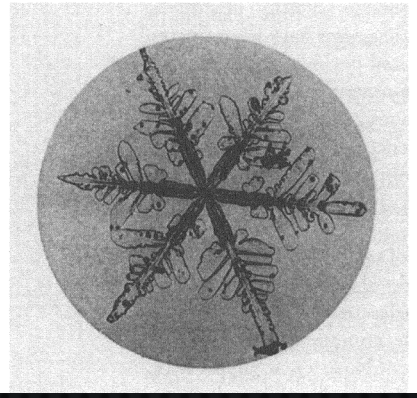


Fig. 1.2. Mechanical Objectivity. Snowflake, Gustav Hellmann, with microphotographs by Richard Neuhauss, *Schneekrystalle: Beobachtungen und Studien* (Berlin: Mückenberger, 1893), table 6, no. 10. An individual snowflake is shown with all its peculiarities and asymmetries in an attempt to capture nature with as little human intervention as possible: mechanical objectivity.

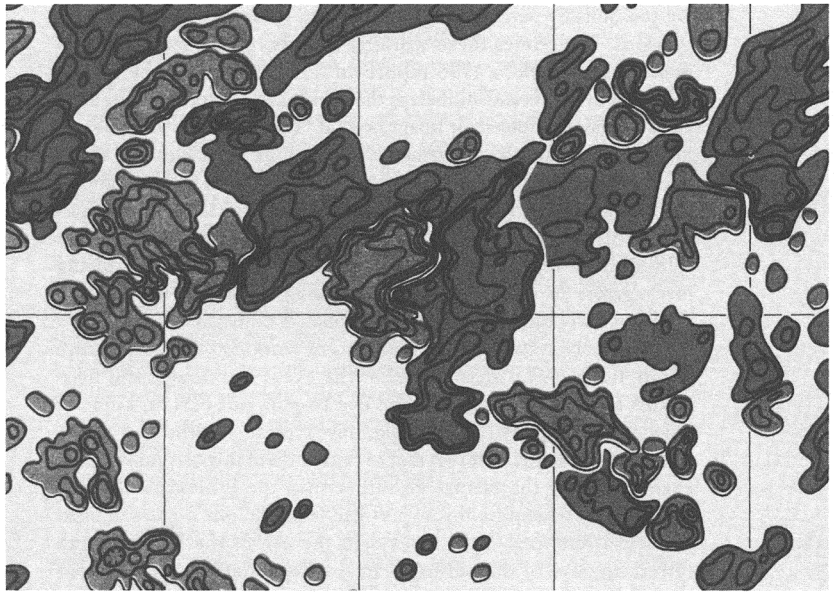
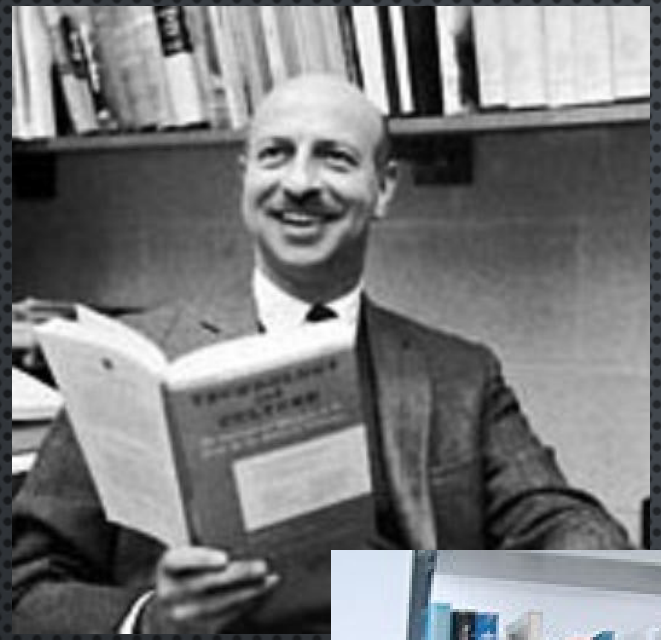
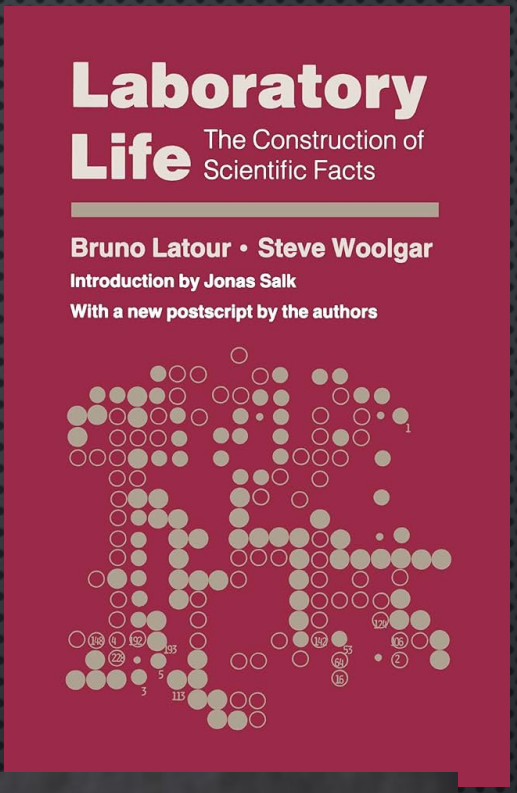
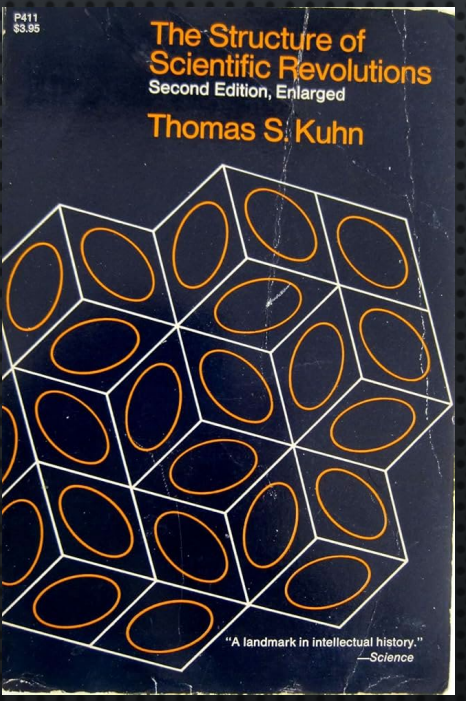
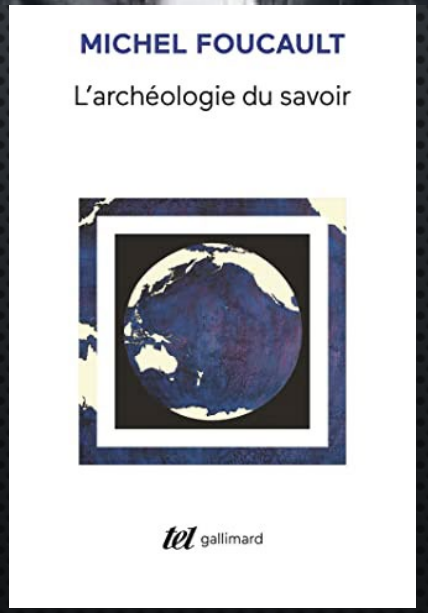


Fig. 1.3. Trained Judgment. Sun Rotation 1417, Aug.–Sept. 1959 (detail), Robert Howard, Václav Bumba, and Sara F. Smith, *Atlas of Solar Magnetic Fields*, August 1959 – June 1966 (Washington, DC: Carnegie Institute, 1967) (courtesy of the Observatories of the Carnegie Institution of Washington, DC). This image of the magnetic field of the sun mixed the output of sophisticated equipment with a "subjective" smoothing of the data – the authors deemed this intervention necessary to remove instrumental artifacts: trained judgment. (Please see Color Plates.)




Laboratory Life
The Construction of Scientific Facts
Bruno Latour • Steve Woolgar
Introduction by Jonas Salk
With a new postscript by the authors


P411
$3.95
The Structure of Scientific Revolutions
Second Edition, Enlarged
Thomas S. Kuhn
"A landmark in intellectual history."
—Science


MICHEL FOUCAULT
L'archéologie du savoir
tel gallimard

WE CAN'T TECHNOFIX OUR WAY OUT OF VALUES PROBLEMS

WE CAN'T TECHNOFIX OUR WAY OUT OF VALUES PROBLEMS

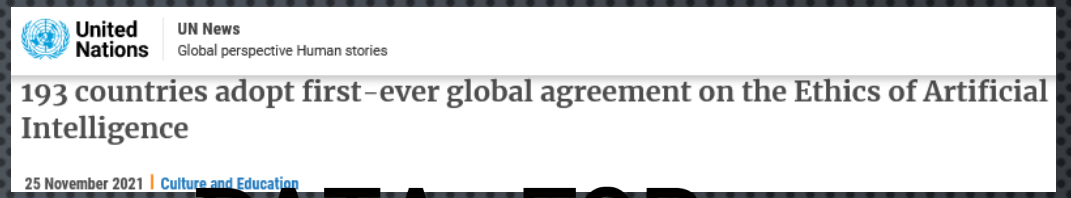
The Deeper Problem With Google's Racially Diverse Nazis

Generative AI is not built to honestly mirror reality, no matter what its creators say.

By Chris Gilliard

BRINGING US TO THE CENTRAL QUESTION:

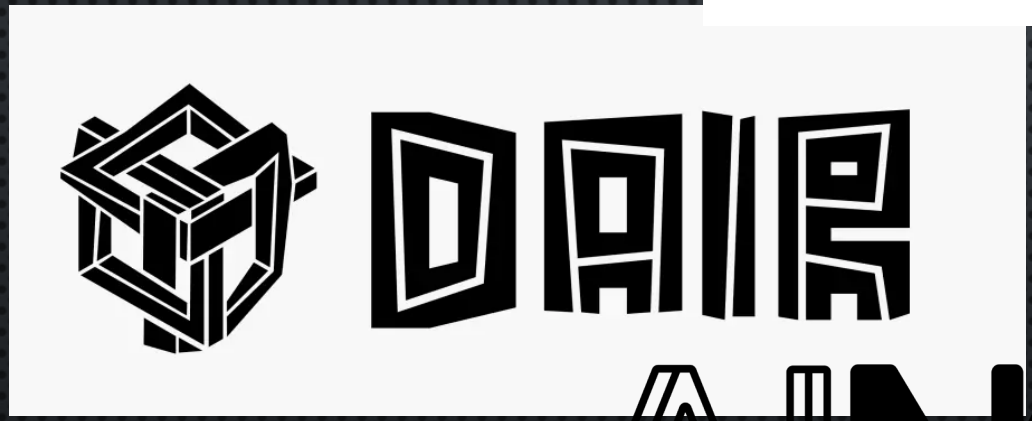
- Whose Values Do We Want Embedded in These Technologies?



DATA FOR BLACK LIVES

OpenAI

cdt



Meta

AINOW

Microsoft

Google

Center for Humane AI Studies

Center for Humane AI Studies
The University of North Carolina at Charlotte
9201 University City Blvd, Charlotte, NC 28223-0001
704-687-8622

UNIVERSITY OF NORTH CAROLINA CHARLOTTE

CHARLOTTE

CENTER FOR TAIMING AI

If we want technological systems which do not depend on predatory and extractive logics or social frameworks, and which fundamentally respect the rights and dignity of people, then we must identify, articulate, confront, and deconstruct the values that brought us to the systems we have, and work very hard, and very intentionally, to build something more equitable and more overarchingly just, in their place.

Thank you for your time and attention.

Contact:
Damien.P.Williams@charlotte.edu