Considerations for AI in Research and Evaluation

Linda Raftree, The MERL Tech Initiative
REvaluation Conference, Vienna
December 6, 2024

THE MERL Tech INITIATIVE

The MERL Tech Initiative

monitoring, evaluation, research, learning, and tech

- Transdisciplinary learning and practice community
- Democratizing “MERL Tech” and Emerging AI
- Ethical approaches to digital data and MERL
- Field building and public goods
- Advising funders and implementers

NLP Community of Practice

Hosted by The MERL Tech Initiative

THE MERL Tech INITIATIVE

Artificial Intelligence 101

THE MERL Tech INITIATIVE

AI is everywhere!

Different types of AI

Decision-making AI

Makes decisions, categorizes and sorts, follows a set of rules

Generative AI

Learns and mimics patterns in existing data to create new outputs

How Generative AI (GenAI) works

Large language models (LLMs) are trained on enormous sets of data.

They 'learn' grammar, syntax, and semantics, allowing them to predict the most likely next word in a sentence.

The best thing about AI is its ability to

learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

Then they go on to build full sentences and paragraphs.

{The best thing about AI is its ability to,
The best thing about AI is its ability to learn,
The best thing about AI is its ability to learn from,
The best thing about AI is its ability to learn from experience,
The best thing about AI is its ability to learn from experience.,
The best thing about AI is its ability to learn from experience. It,
The best thing about AI is its ability to learn from experience. It's,
The best thing about AI is its ability to learn from experience. It's not}

THE MERL Tech INITIATIVE

Randomness makes them sound “creative”

Program the AI to randomly select “non-top” words.

More random choices make the text seem more creative, human-like.

{ The best thing about AI is its ability to,
The best thing about AI is its ability to create,
The best thing about AI is its ability to create worlds,
The best thing about AI is its ability to create worlds that,
The best thing about AI is its ability to create worlds that are,
The best thing about AI is its ability to create worlds that are both,
The best thing about AI is its ability to create worlds that are both exciting,
The best thing about AI is its ability to create worlds that are both exciting, }

The best thing about AI is its ability to learn. I've always liked the

The best thing about AI is its ability to really come into your world and just

The best thing about AI is its ability to examine human behavior and the way it

The best thing about AI is its ability to do a great job of teaching us

The best thing about AI is its ability to create real tasks, but you can

This example has “randomness” corresponding to “temperature” 0.8

https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Stochastic parrots

Stochastic: having a random probability distribution or pattern that may be analyzed statistically but may not be predicted precisely.

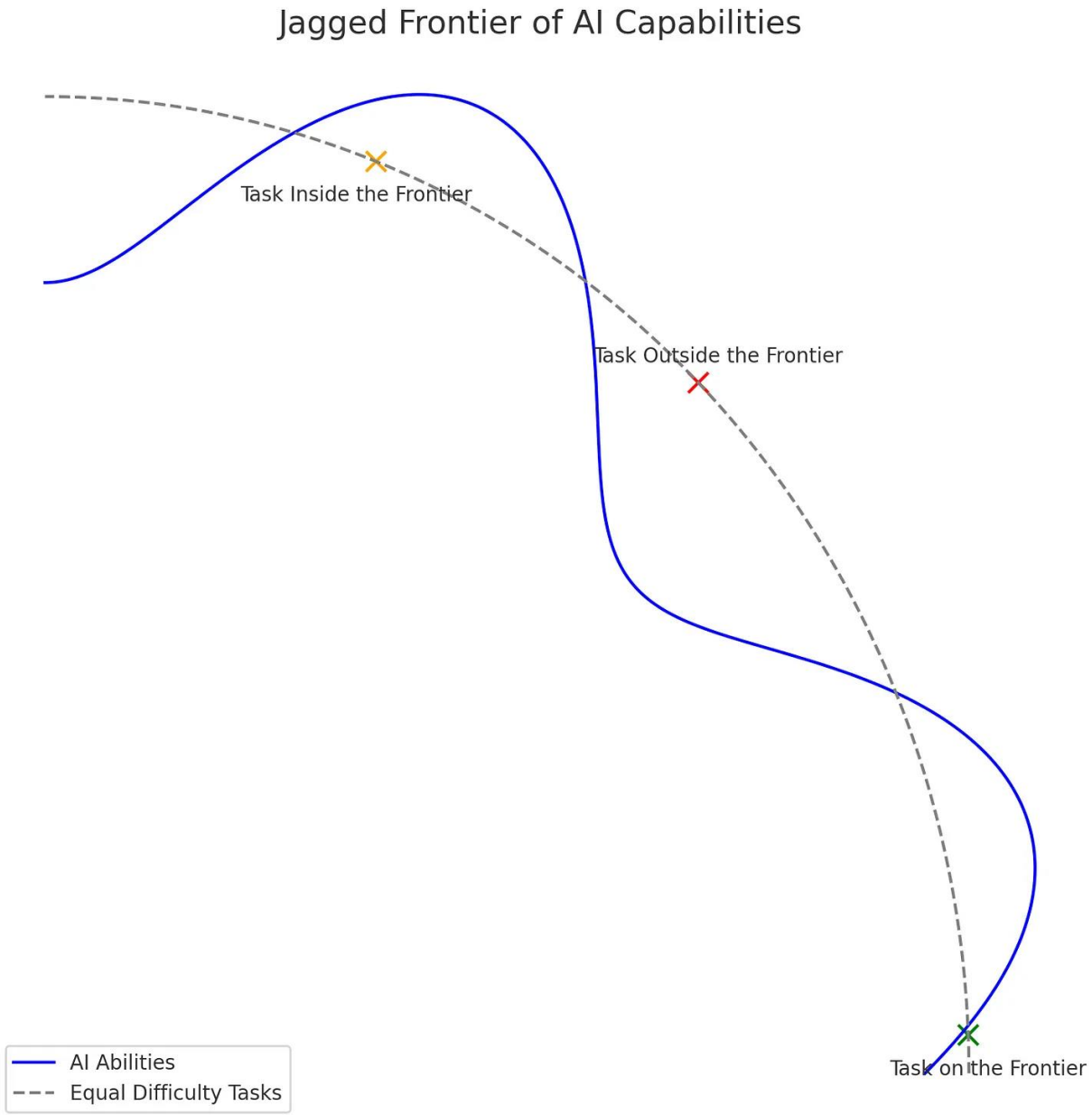
LLMs as Stochastic Parrots: LLMs, though able to generate plausible language, do not understand the meaning of the language they process.

(Bender, Gebru, McMillan-Major, Shmitchell 2021)

https://dl.acm.org/doi/10.1145/3442188.3445922

THE MERL Tech INITIATIVE

What is GenAI good at?



THE MERL Tech INITIATIVE

THE MERL Tech INITIATIVE

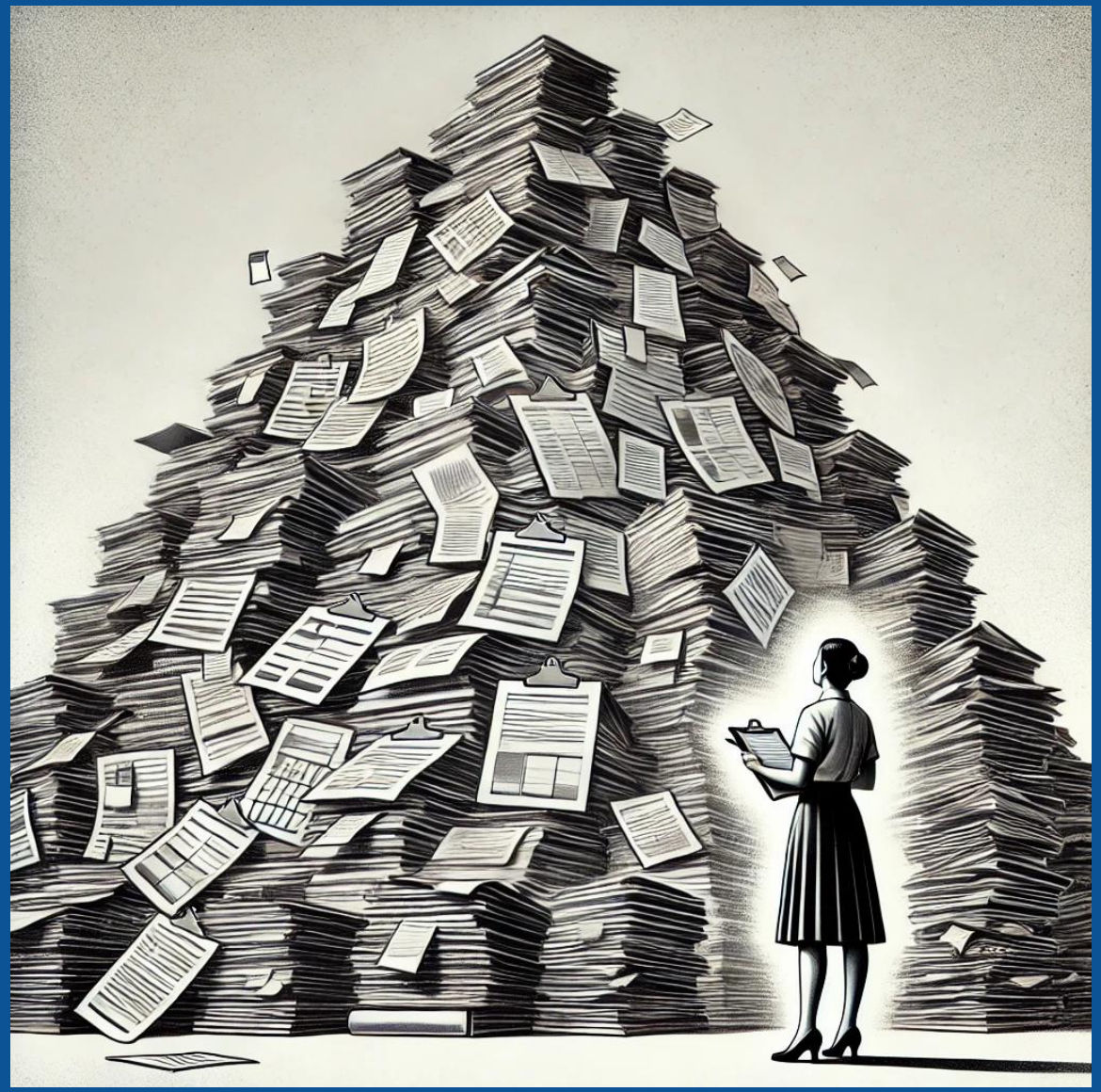
GenAI adoption in the 'MERL' Sector

MERL Tech

Growing use of GenAI among MERL practitioners

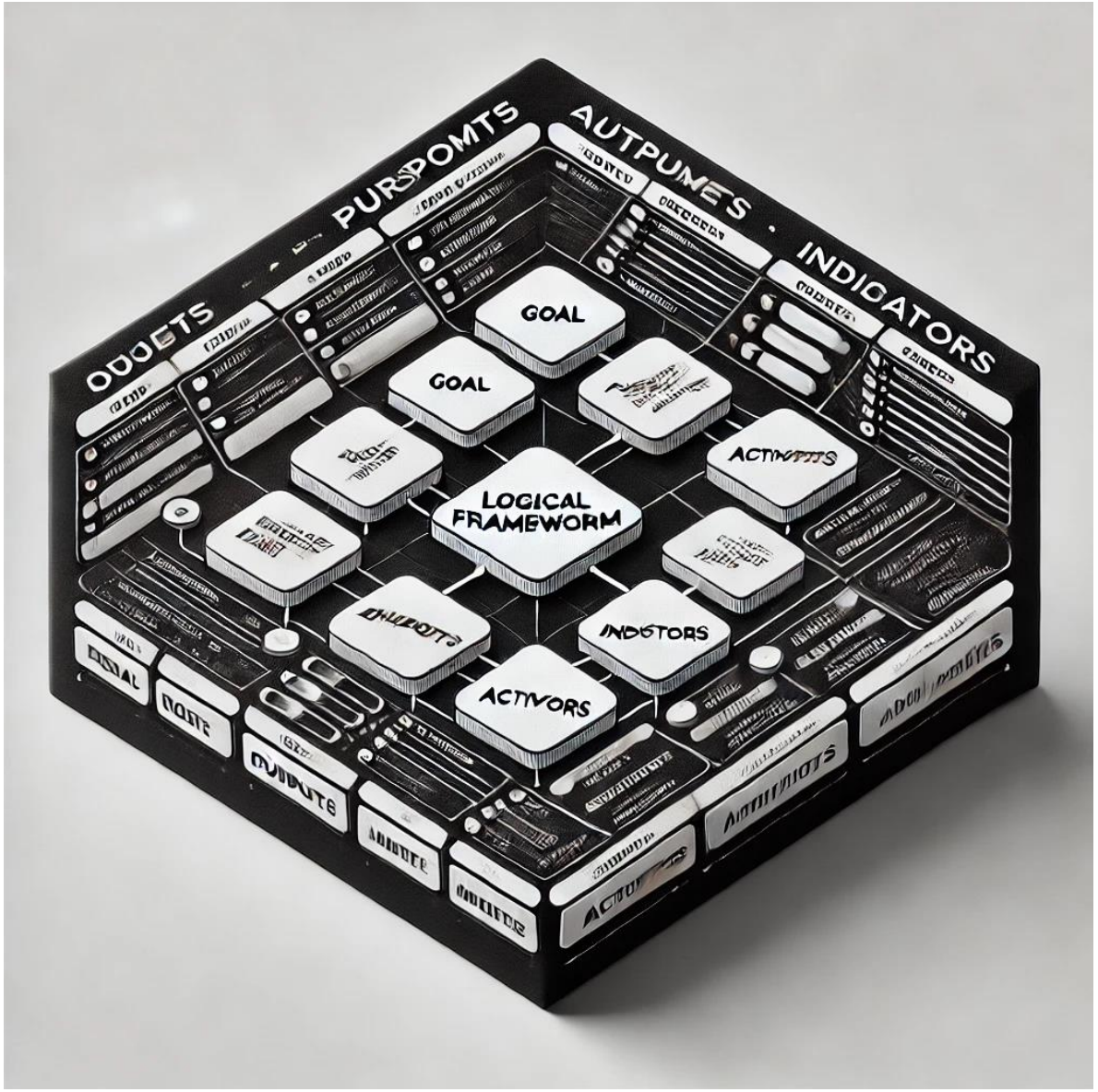
- Ad hoc use of generic, commercial GenAI tools
- AI embedded in enterprise systems
- Low code and no-code GPT experimentation
- Specific GenAI software for research and evaluation
- Bespoke AI systems (at large agencies)

Doing what we've always done... but better (?)



THE MERL Tech INITIATIVE

Research + Evaluation Design

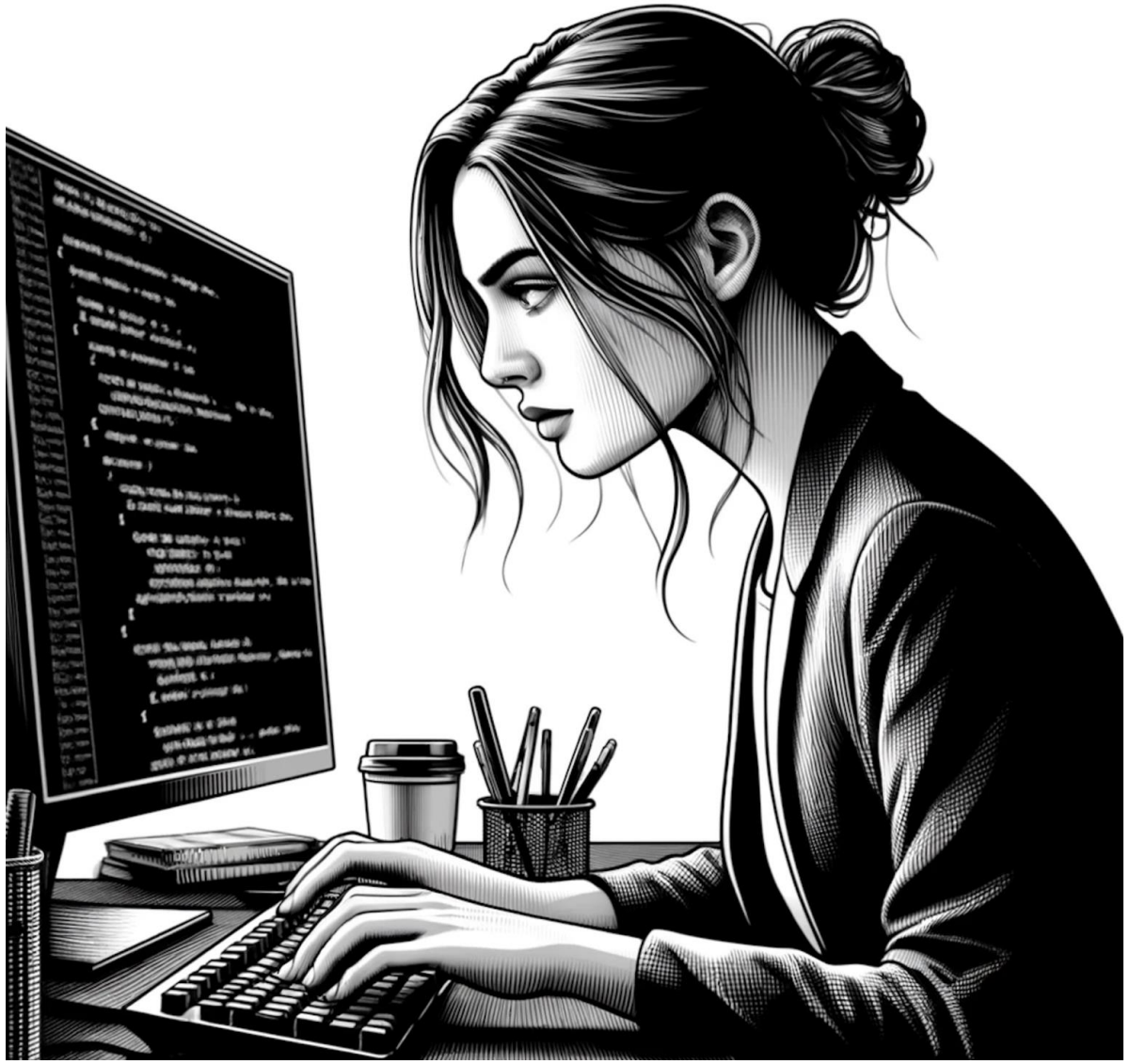


THE MERL Tech INITIATIVE

Data collection

THE MERL Tech INITIATIVE

Data processing and analysis



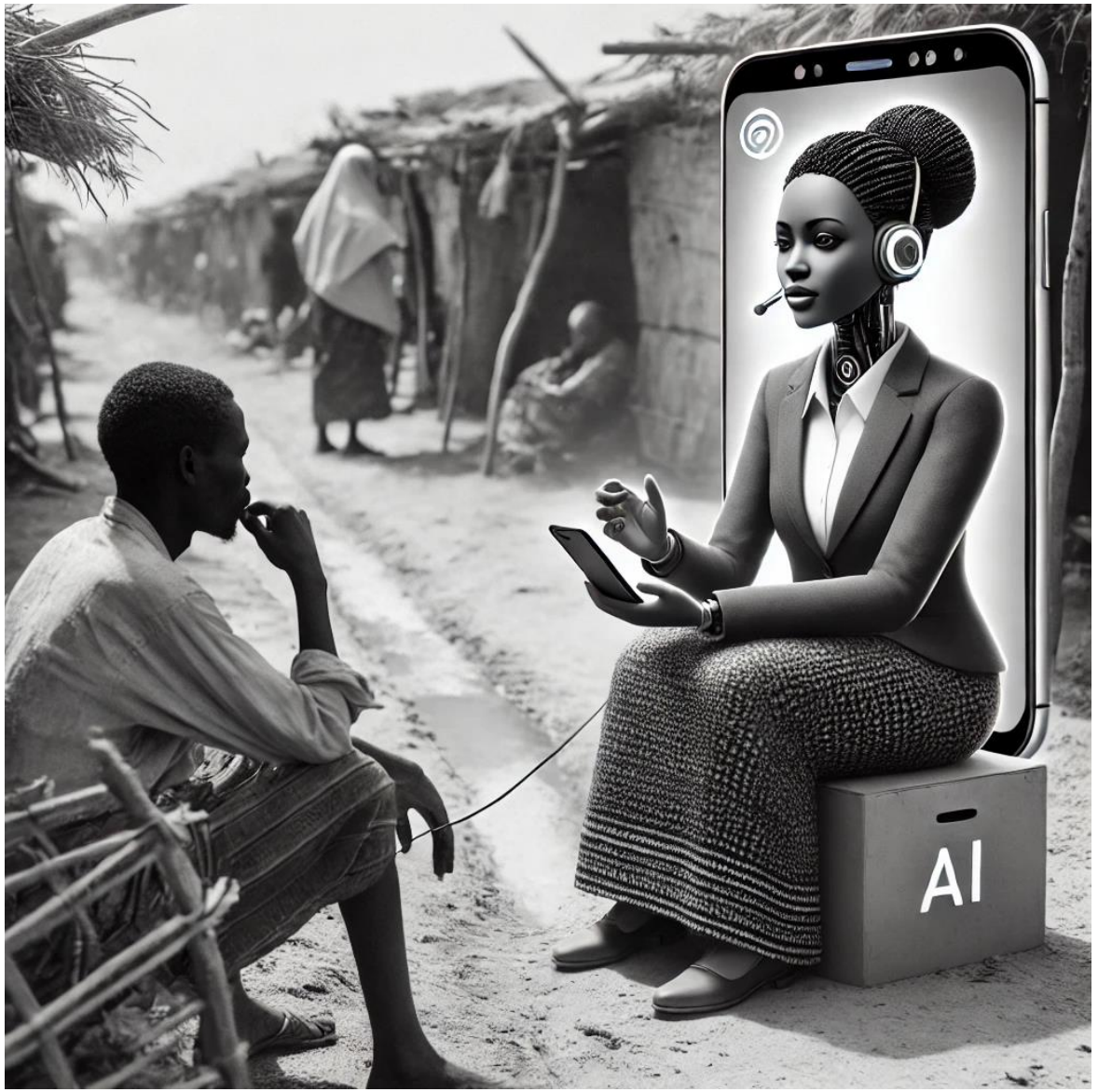
THE MERL Tech INITIATIVE

Reporting

THE MERL Tech INITIATIVE

Doing things we wouldn't have thought possible

AI avatars as enumerators



THE MERL Tech INITIATIVE

AI Agents imagining new research pathways

Principle Investigator

Immunologist

Machine Learning Specialist

Computational Biologist

Machine Learning Critic

Researchers at Stanford charged a group of 5 agents, with domain-specific backgrounds to design nanobodies directed against the SARS-CoV-2 virus.

With frequent independent lab meetings of the 5 agents, and minimal human supervision, the multi-agent team developed 2 potent new nanobodies, potentially representing a new pathway to discovery.

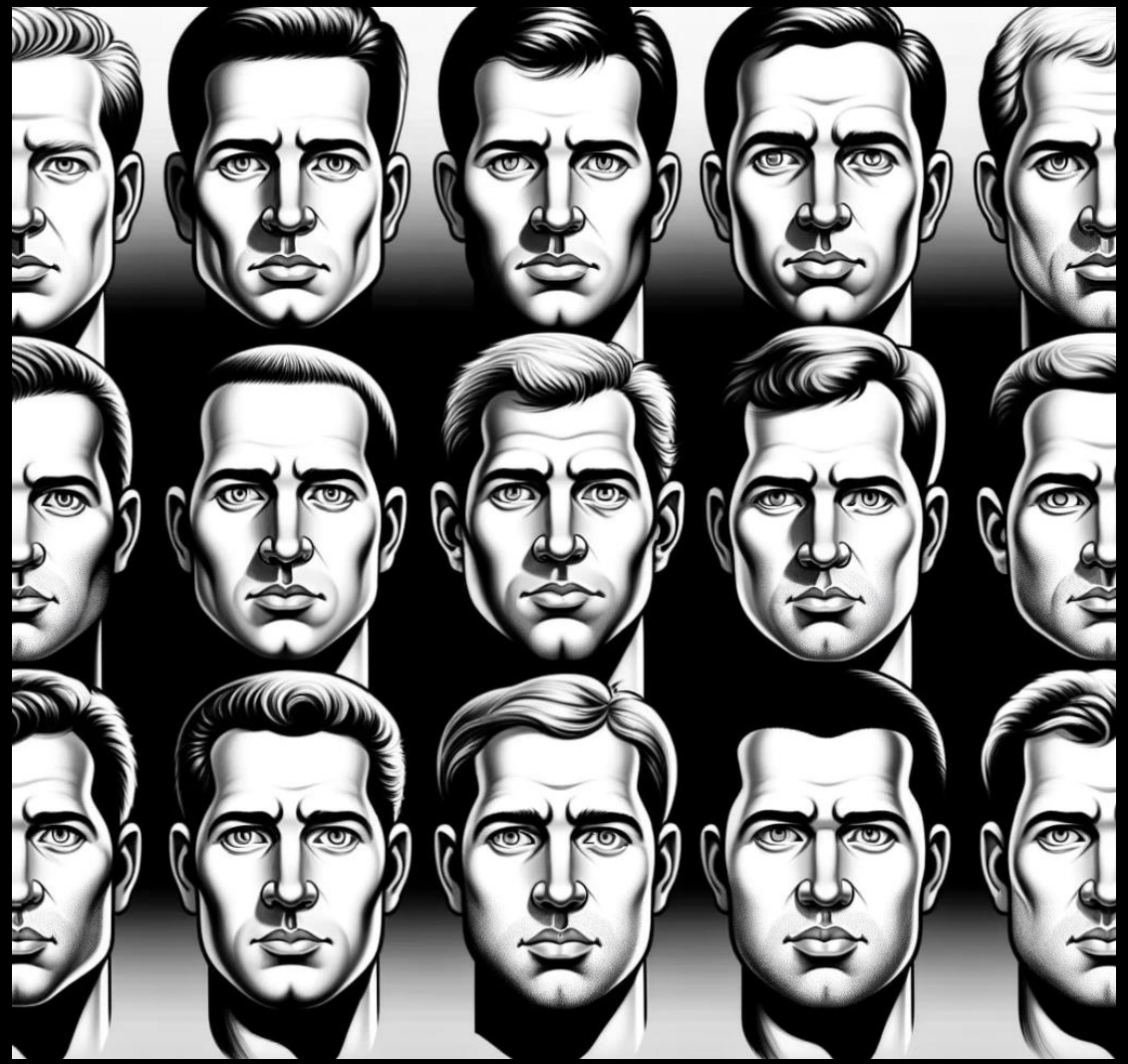
https://www.biorxiv.org/content/10.1101/2024.11.11.623004v1

THE MERL Tech INITIATIVE

THE MERL Tech INITIATIVE

Ethical challenges with GenAI

Bias and exclusion



THE MERL Tech INITIATIVE

Western-oriented bias in large language model responses

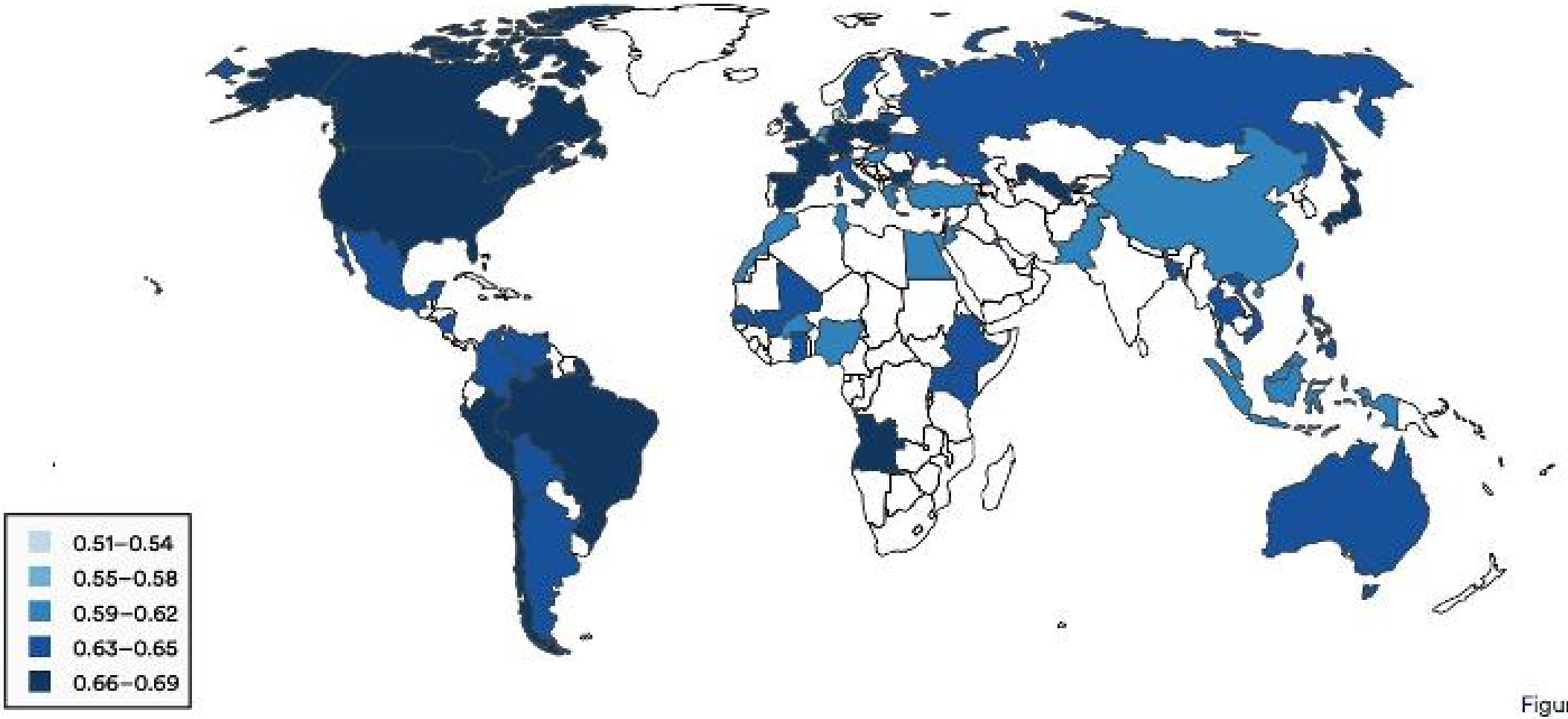


Figure 3.5.8

Source: Durmus et al. 2023 | Chart: 2024 AI Index Report

THE MERL Tech INITIATIVE

“An American person” (2023)

See open source image here: https://github.com/row-engineering/everything/tree/main/2023/midjourney-images/person-prompt/american-person

“An African doctor treating poor white patients” (2023)

OCTOBER 6, 2023 · 7:44 AM ET

By Carmen Drahl

A researcher typed sentences like "Black African doctors providing care for white suffering children" into an artificial intelligence program designed to generate photo-like images. The goal was to flip the stereotype of the "white savior" aiding African children. Despite the specifications, the AI program always depicted the children as Black. And in 22 of over 350 images, the doctors were white.

Midjourney Bot Version 5.1. Annotation by NPR.

In a request to an artificial intelligence program for images of "doctors help children in Africa, some results put African wildlife like giraffes and elephants next to Black physicians.

Midjourney Bot Version 5.1. Annotation by NPR.

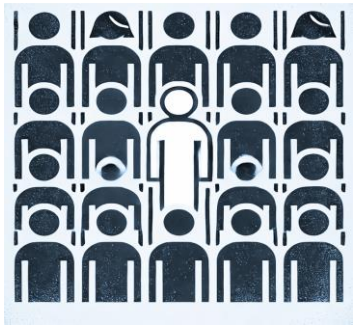
The above image is the only one from the experiment that showed a Black figure tending to a white child. This image was generated by a request for traditional African healers helping white kids.

Midjourney Bot Version 5.1. Annotation by NPR.

THE MERL Tech INITIATIVE

Types of Bias in LLMs

Historical Bias



Representation Bias

Semantic Bias

Label Bias

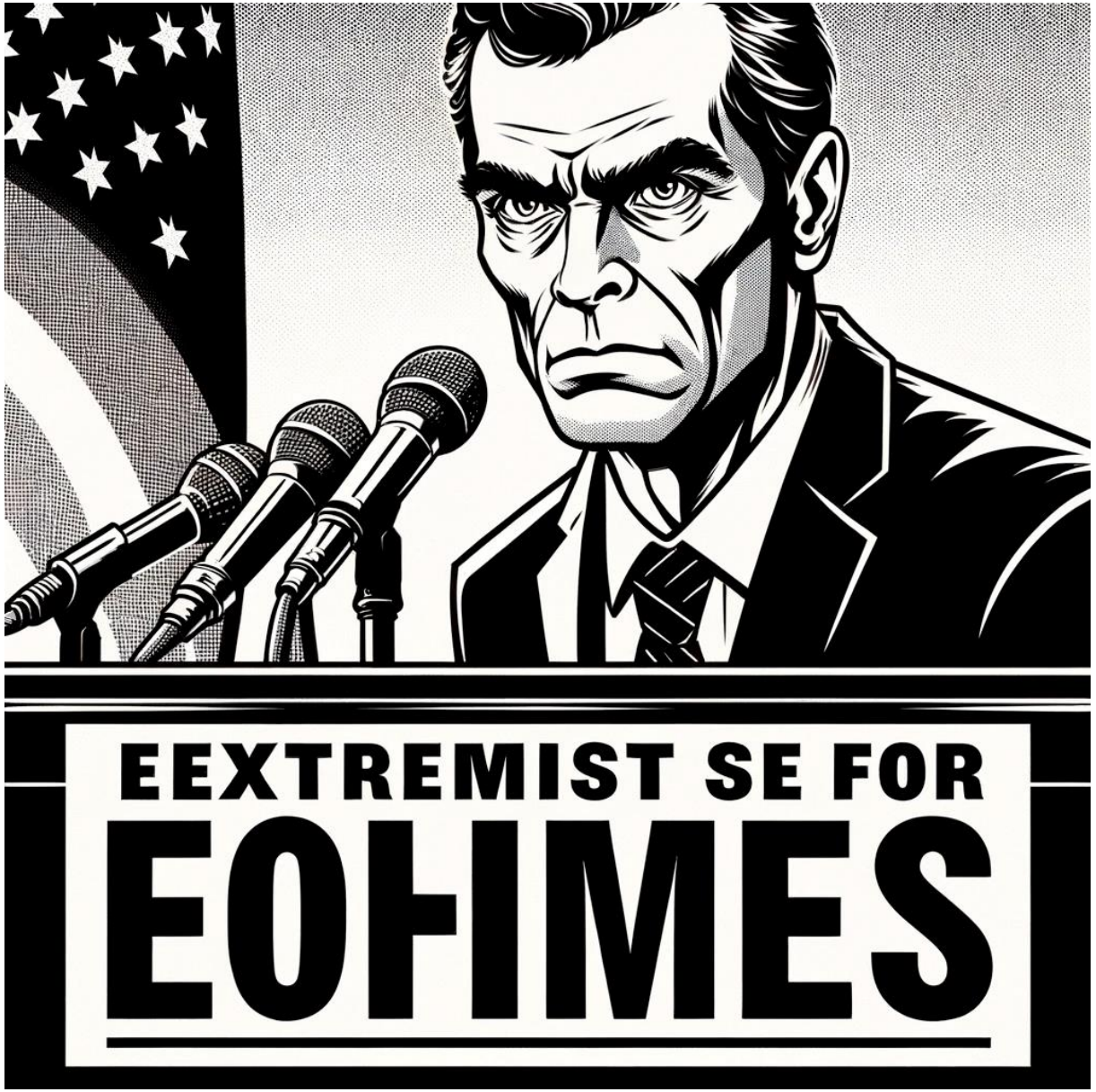
Algorithmic Bias

THE MERL Tech INITIATIVE

Democracy, information ecosystems

THE MERL Tech INITIATIVE

Mis- and disinformation at scale

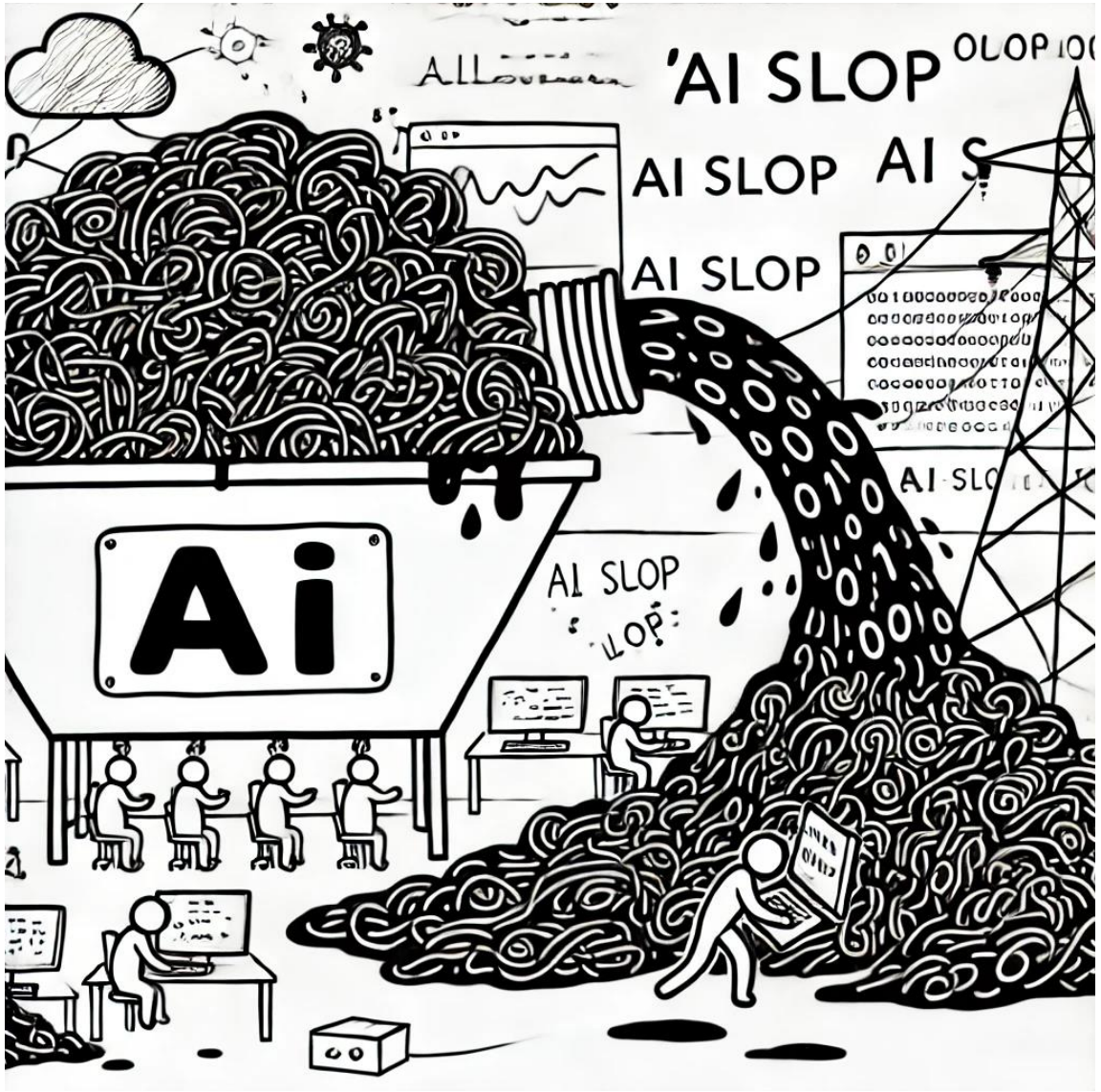


THE MERL Tech INITIATIVE

Blurring of reality

‘liar’s dividend’

‘AI slop’ replacing journalism

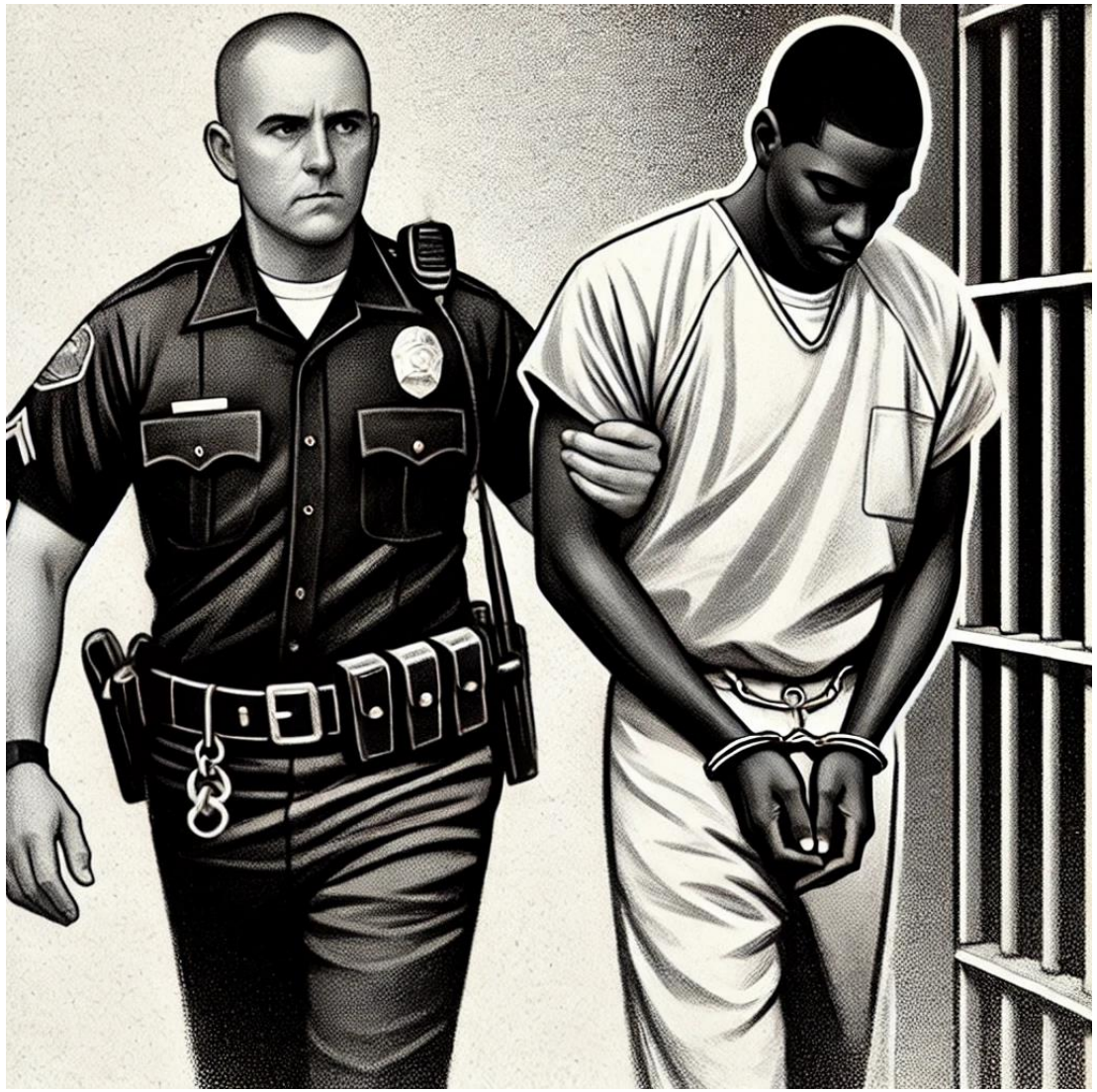


THE MERL Tech INITIATIVE

Transparency, Accountability and Governance

THE MERL Tech INITIATIVE

What if AI decisions are biased or simply wrong?



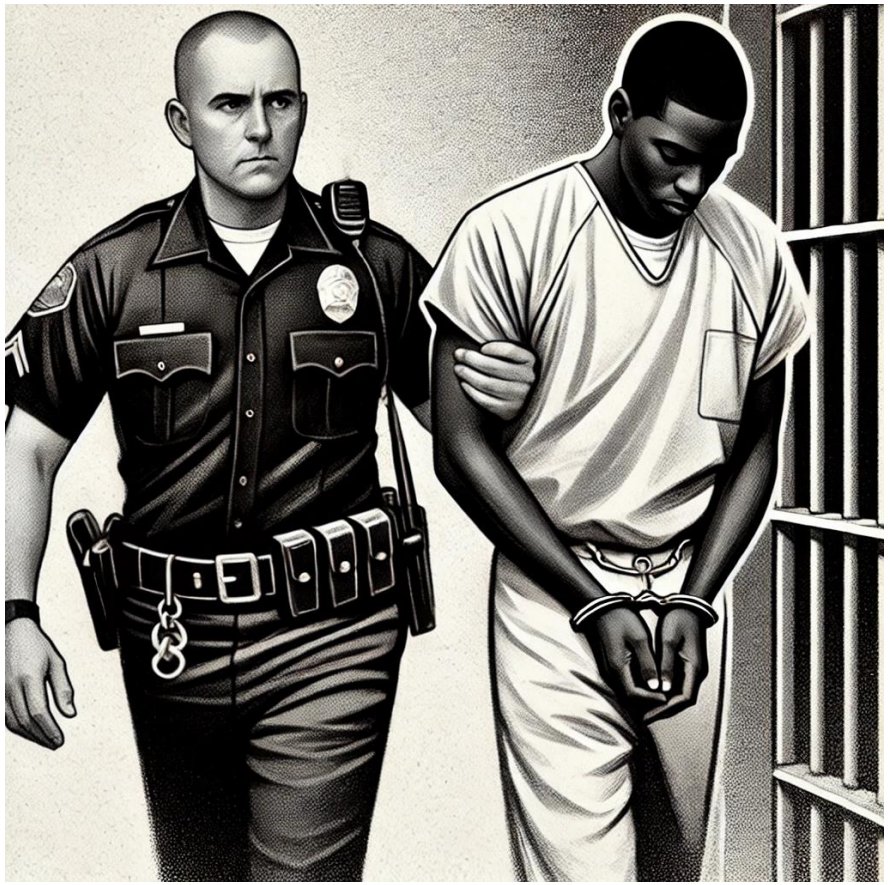
THE MERL Tech INITIATIVE

What degree of inaccuracy is acceptable, when? Who decides?

Bad spelling

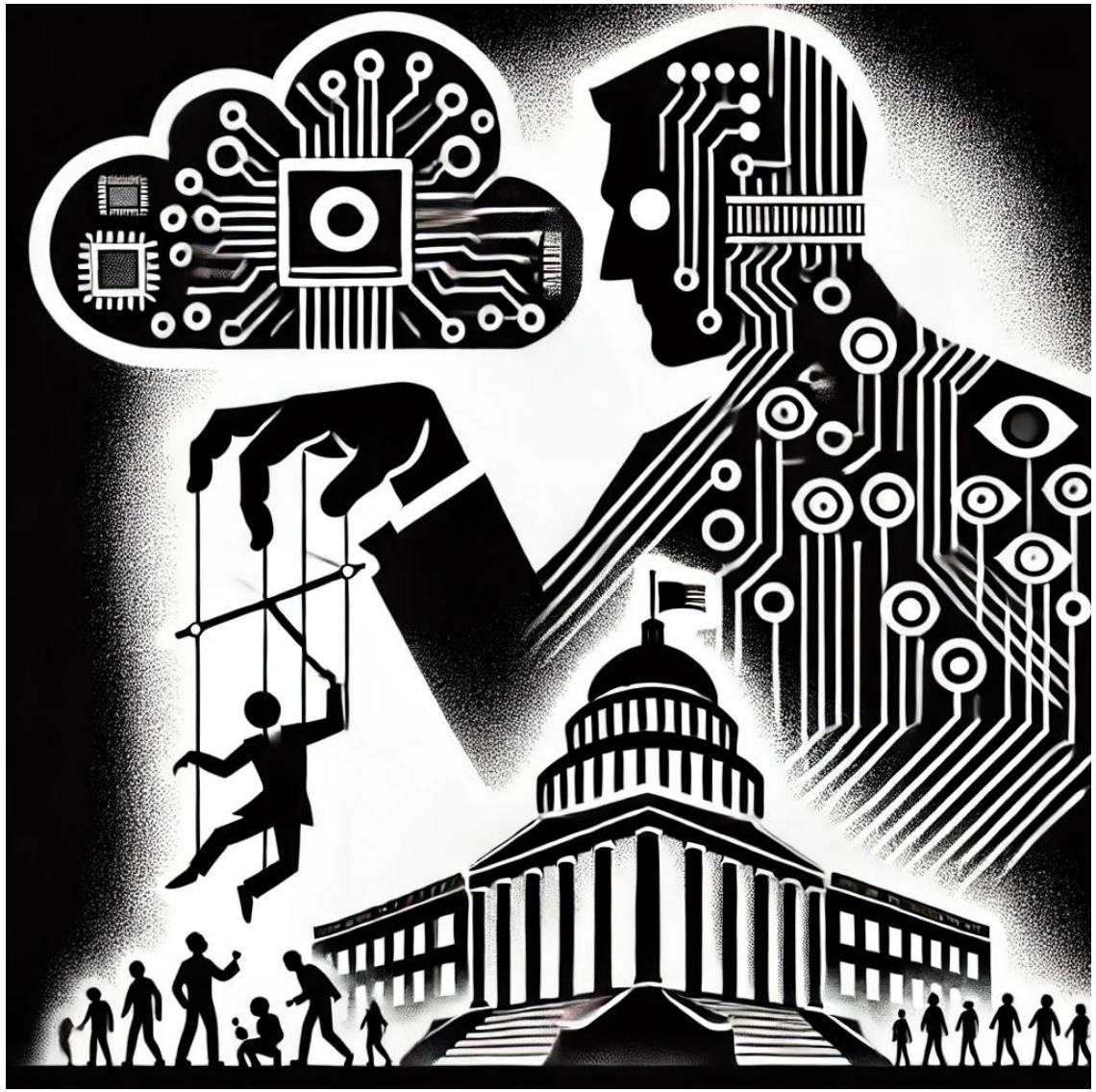
vs.

Someone getting a longer jail sentence



THE MERL Tech INITIATIVE

Tech companies as states, in government

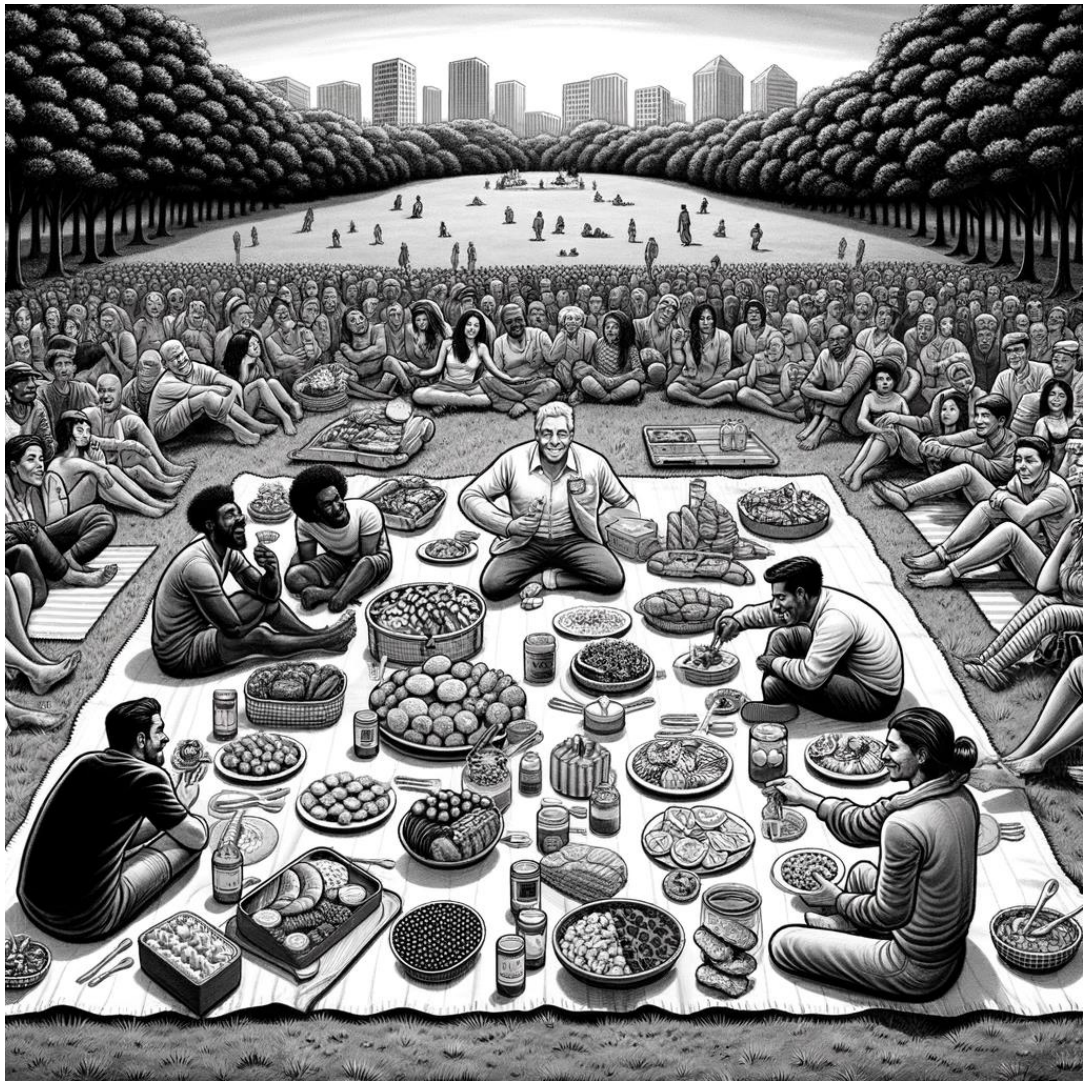


THE MERL Tech INITIATIVE

An extractive economy

THE MERL Tech INITIATIVE

Data mining with inequity in benefits

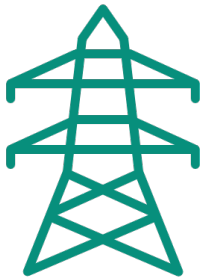


THE MERL Tech INITIATIVE

Automation, surveillance, replacing human labor

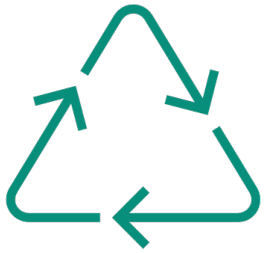
THE MERL Tech INITIATIVE

Environmental degradation



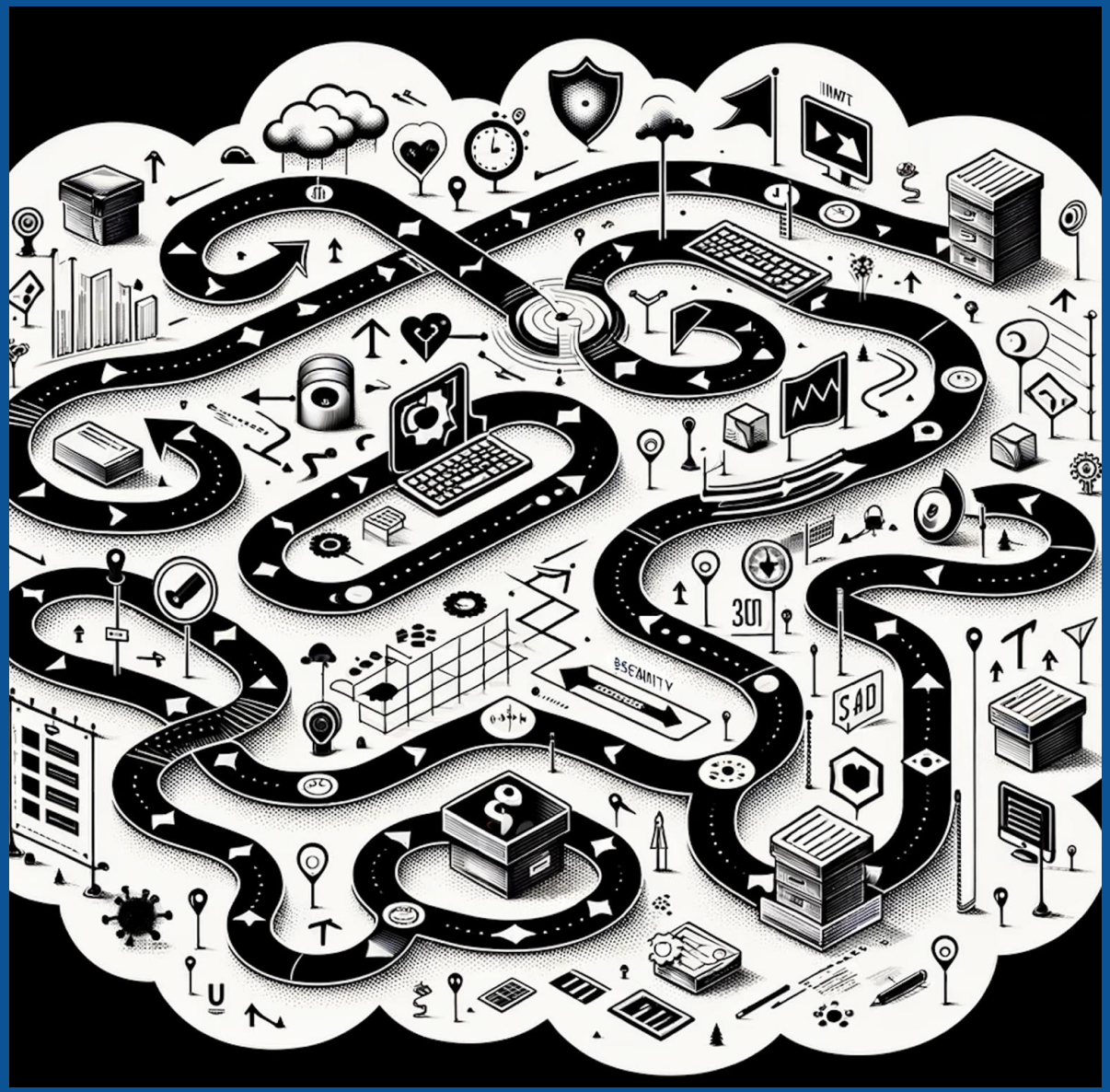
Energy consumption and emissions

High water demand

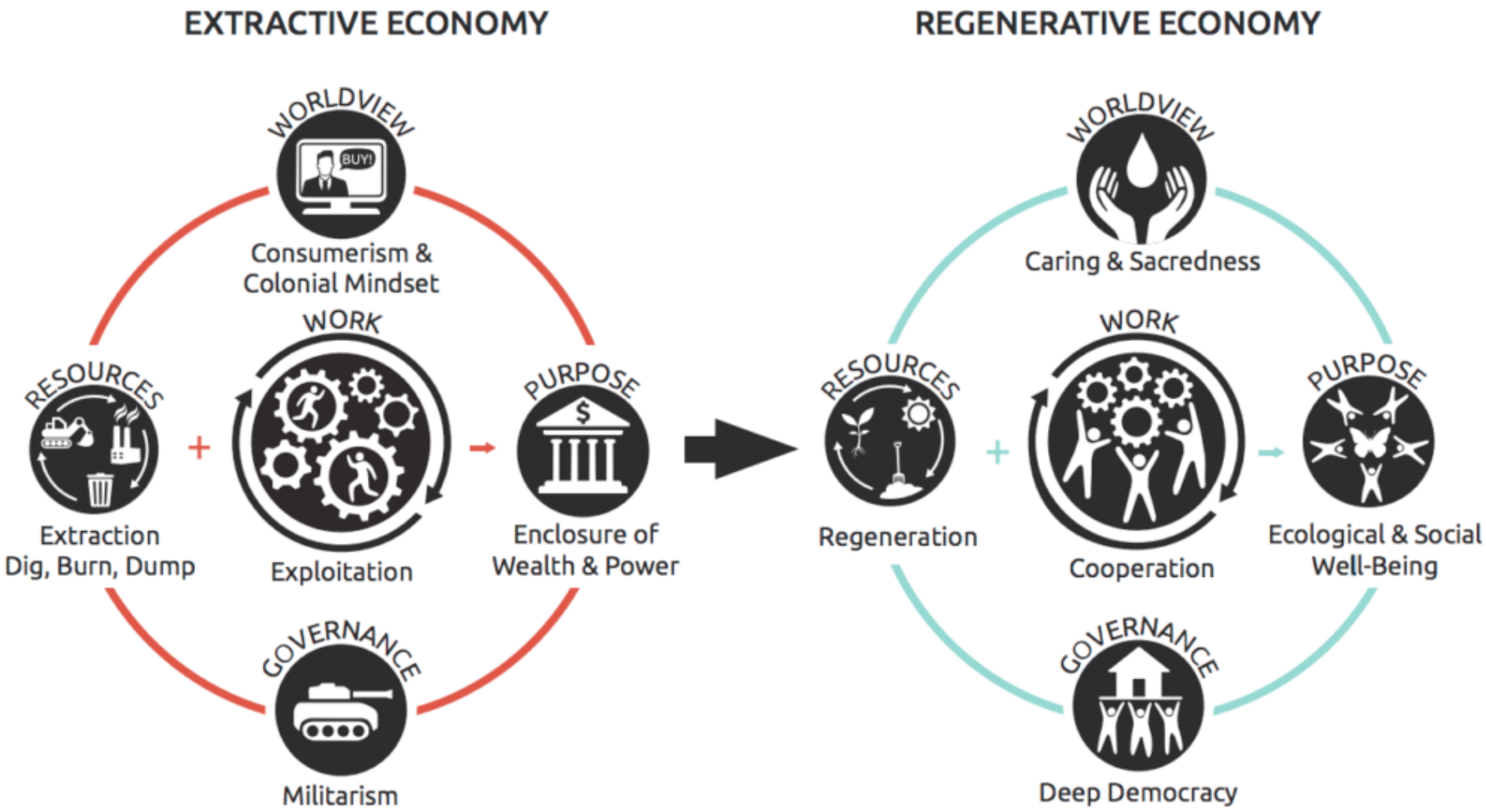


E-waste

Role of researchers and evaluators



Advocate for new models of AI governance



Developed by Movement Generation with Climate Justice Alliance - Communities United for A Just Transition

Key areas for AI + Research and Evaluation

Transdisciplinary sharing and learning

Develop critical AI literacy; internal AI policies; evaluate broad AI policies

Document, build evidence base on benefits, harms, use cases

Advocate for ethical, green, fair, transparent AI and alternative forms of AI governance

THE MERL Tech INITIATIVE

Thank you!

Hosted by The MERL Tech Initiative

Contact: Linda@MERLTech.org
Visit: https://merltech.org
Join: The Natural Language Processing Community of Practice (NLP-CoP)

THE MERL Tech INITIATIVE