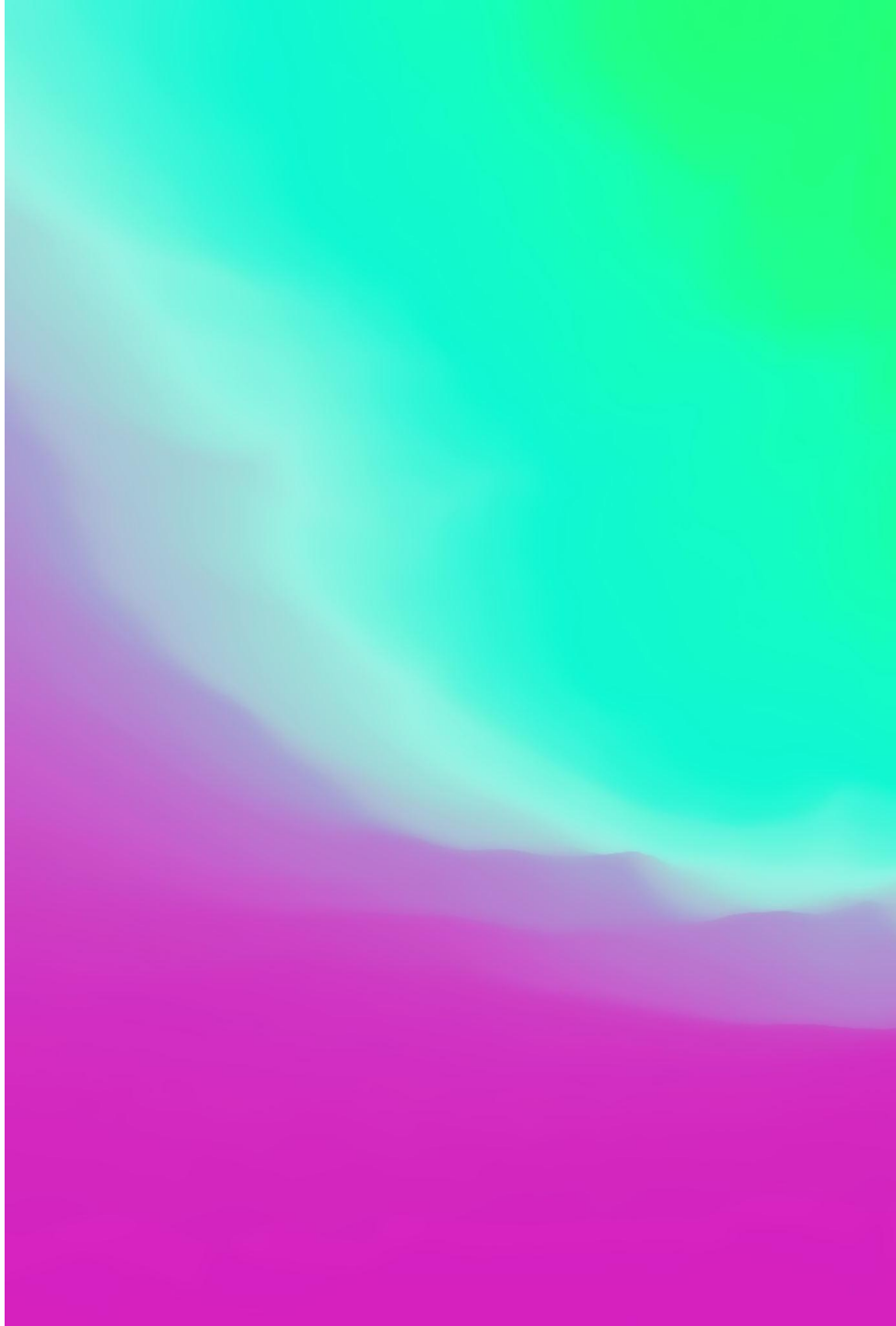
ETHICS PROBES(S): ARTIFICIAL INTELLIGENCE

Dr. Dylan J. White

Instructor, Department of Philosophy
Memorial University



Contents

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton

All Open Letters

Pause Giant AI Experiments: An Open Letter

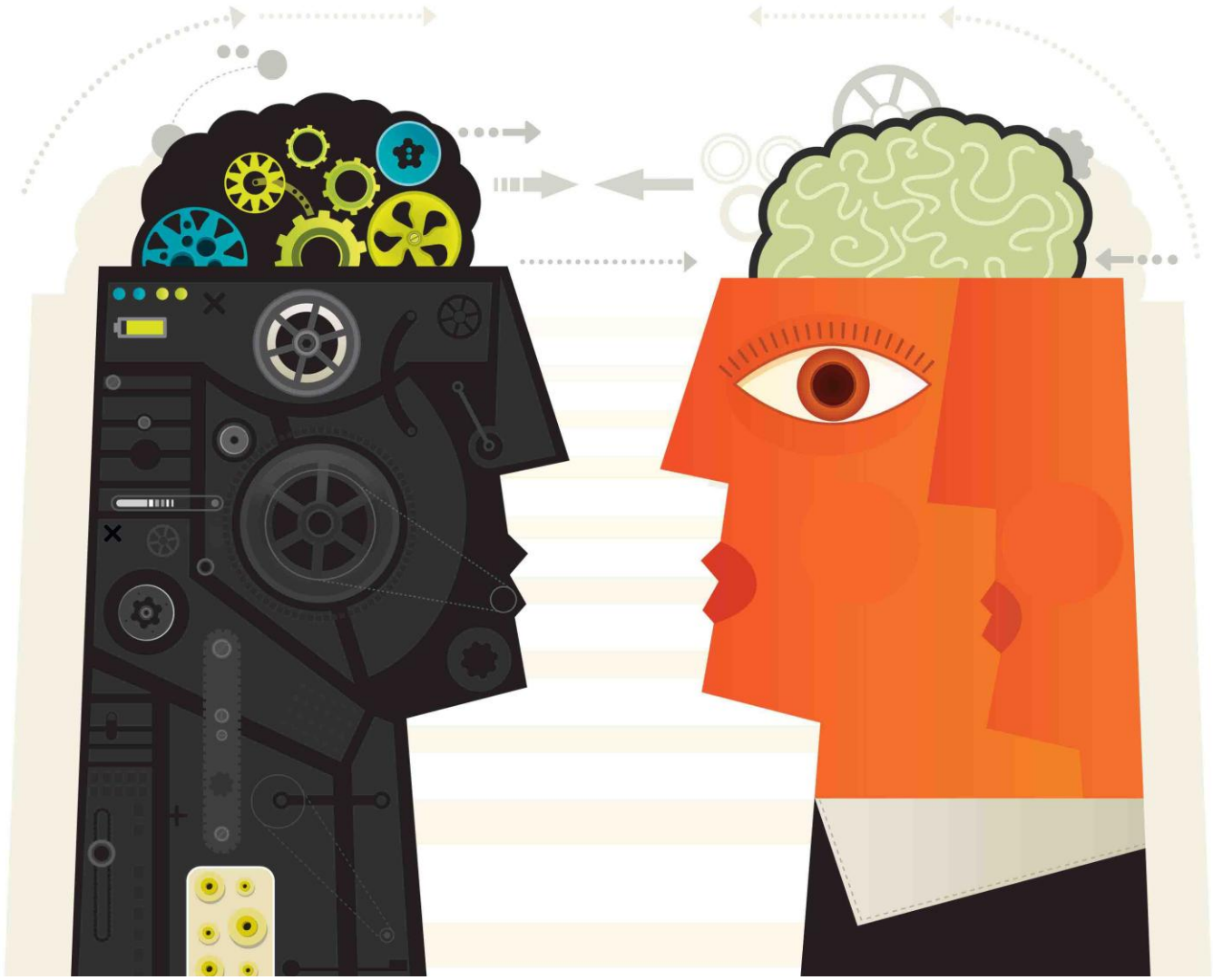
We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

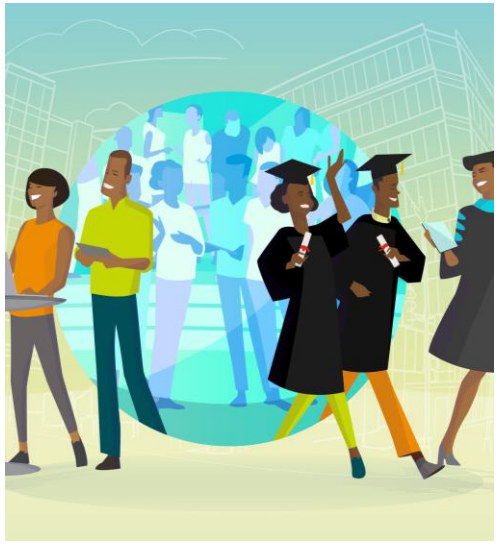
Signatures
31810

Add your signature

TAKING A STEP BACK…

- Ethics
  - How to live a life full of well-being, flourishing, etc.
- Decision-Making
  - How do we make good decisions?





ALGORITHMICALLY MEDIATED DECISIONS

- Increasingly, many aspects of our lives from the mundane to the major decisions that we make, are mediated by algorithmic decision-making systems, or otherwise impacted by AI.
- This makes questions about how to live a good life increasingly inextricable from questions about technology, especially artificial intelligence

PROBES(S) FOR THE ETHICS OF AI

- **P**rivacy (and Surveillance)
- **R**esponsibility
- **O**pacity
- **B**ias
- **E**quity (and Employment)
- **S**ustainability
- **S**afety

PRIVACY AND SURVEILLANCE

- Nothing to hide…
- "As digital traces of individual behaviors are aggregated, stored, and analyzed, **markets see people through a lens of deserving and undeservingness and classification situations become moral projects…**" (Fourcade & Healy, 2017)
- For example, Facebook's discriminatory housing practices
- PRIZM - Environics
- Cases such as these suggest that the traditional framing of privacy as an individual concern needs to be reconsidered
- "… to counter group and societal harms, there is a pressing need for future data and AI regulation to incorporate collective data rights and give communities a powerful say over the data and AI that affect them" (Tennison, 2024)

PRIVACY AND SURVEILLANCE

- The economic dimension of privacy (Balsillie, 2026)
- Data monopolies and oligopolies use their market power to charge higher prices
- They use individual consumer information to charge exploitative rates and call it “tailored” or “personalized” pricing
- Surveillance-based "personalized" pricing eliminates consumer surplus — you become a market of one
- Monopsony power suppresses wages — algorithmic "desperation scores" tailor pay to the minimum a worker will accept (e.g., Uber, DoorDash)
- Privacy is not just a social concern — it is an economic one: you pay more and earn less because of it

RESPONSIBILITY

- Responsibility Gaps
- "… presently there are machines in development or already in use which are to decide on a course of action and to act without human intervention. The rules by which they act are not fixed during the production process, but can be changed during the operation of the machine, *by the machine itself*. This is what we call machine learning. Traditionally we hold either the operator/manufacturer of the machine responsible for the consequences of its operation, or 'nobody' (in cases, where no personal fault can be identified). Now it can be shown that there is an increasing class of machine actions, where traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough *control* over the machine's actions to be able to assume responsibility for them. These cases constitute what we will call the *responsibility gap.*" (Matthias, 2004, p. 177)

RESPONSIBILITY

- Traditional ways of thinking about responsibility must be reconsidered
  - Knowledge
  - Control
- Self-Driving Cars
  - Who should we hold responsible for crashes? The manufacturer? The owner? Both? Nobody?
- Air Canada's Customer Service 'Agent'
  - Air Canada claimed the bot was "responsible for its own actions"

OPACITY

- The ‘Black-Box Problem’
- Giving and Asking for Reasons
- The Double-Edged Sword
  - Accuracy vs. Interpretability
  - An automated decision-making system, for example, does not care about human reasons

BIAS

- Fazelpour & Danks (2021)
  - Taxonomy of Algorithmic Bias
- In problem specification
  - Target variables, e.g., what is student success?
- In data
  - Systemic barriers
- In modelling and validation
  - Error distributions
- In deployment
  - Data evolution and feedback loops, e.g., predictive policing

BIAS

- In problem specification
- “The first step in algorithm design is problem specification: What goal(s) will the algorithm be used to achieve (perhaps as part of a broader system)? This specification requires thinking about our overall aims, the actions available to us, and ways of using the algorithm to help achieve those aims (Mitchell at al., 2021). In most practical settings, decisions makers are interested in aims that are complex, contested, and sometimes intentionally ambiguous. Each of these elements requires consideration of values and normative standards, and thereby provides a vehicle for the creation of biases” (Fazelpour & Danks, 2021, p. 4).
- ‘Student Success’

EQUITY (AND EMPLOYMENT)

- The Digital Divide
- Employment
  - "We are being afflicted with a new disease of which… [readers]… will hear a great deal in the years to come – namely, technological unemployment" (Keynes, 1929)
  - Industrial Automation
- Keynes wasn't wrong about widespread automation, but widespread unemployment did not come to pass…
- Why?

EQUITY (AND EMPLOYMENT)

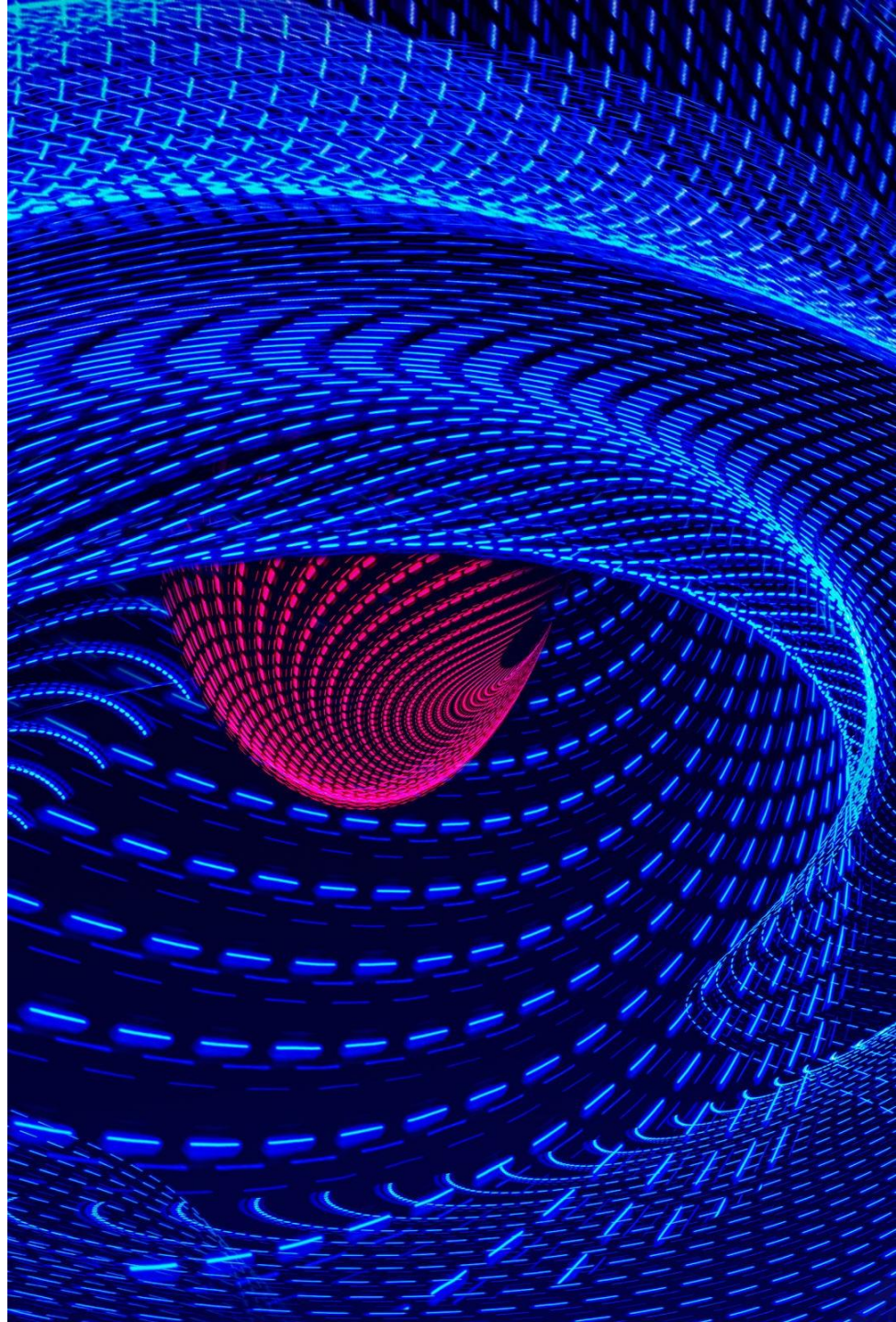
- "Agricultural mechanization was followed by rapid industrial automation, but this too was counterbalanced by other technological advances that created new tasks for workers. Today the majority of the workforce in all industrialized nations engages in tasks that did not exist when Keynes was writing (think of all the tasks involved in modern education, health care, communication, entertainment, back-office work, design, technical work on factory floors, and almost all of the service sector). Had it not been for these new tasks, Keynes would have been right. They not only spawned plentiful jobs but also generated demand for a diverse set of skills, underpinning the shared nature of modern economic growth" (Acemoglu, 2021)
- Labor market institutions—such as minimum wages, collective bargaining, and regulations introducing worker protection—greatly contributed to shared prosperity
- These outcomes are not guaranteed

SUSTAINABILITY

- The ”cloud” is made of rocks, lithium brine, and crude oil (Crawford, 2021)
- Individual AI use (prompting, searching) has a relatively modest footprint — the environmental impact of AI at the individual level is often overstated
- The real concern: the massive buildout of data infrastructure — data centers are expected to be a major driver of rising energy demands, colliding with decarbonization goals (Vogelsong, 2024; de Vries, 2024)
- The tension: AI may help us build a sustainable future, but without addressing the environmental costs of the infrastructure itself, it risks exacerbating the problems it aims to solve

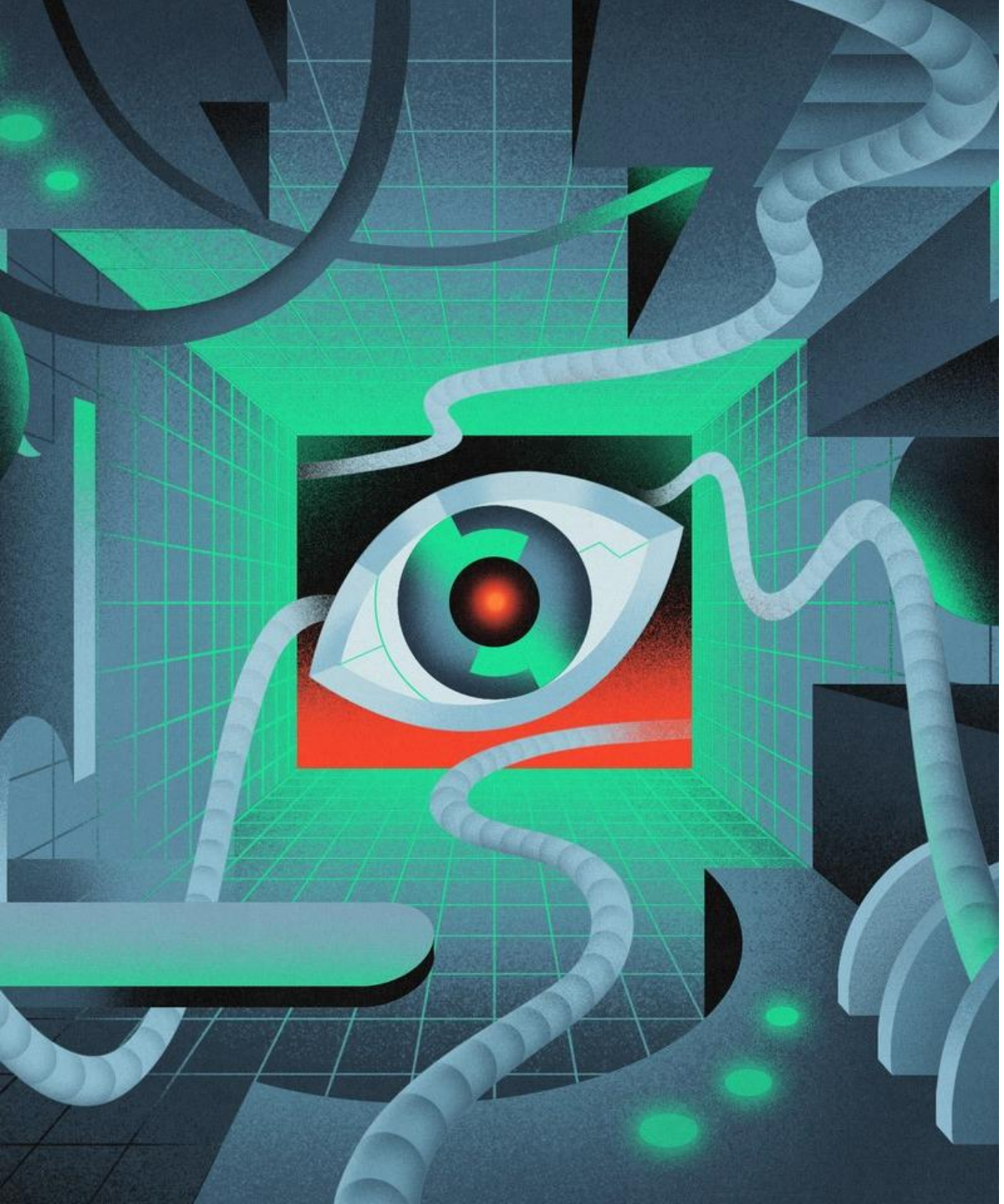
SAFETY

- AI X-Risk
  - “Mitigating the risk of extinction from AI should be a global priority alongside other societal scale risks such as pandemics and nuclear war.”
- Geoffrey Hinton
  - Concerns about “existential risk” fundamentally different than other ethical concerns (PROBES) that are not ”existentially serious”



SAFETY

- Decisive Risk
  - x-risks from ASI concern the possibility of abrupt large-scale events that lead to humanity's extinction or cause an unrecoverable decline in its potential (Kasirzadeh, 2025)
  - Paper-Clip Maximizers
- Accumulative Risk
  - AI x-risks result from the build-up of a series of smaller, lower-severity disruptions over time, collectively and gradually weakening systemic resilience until a triggering event causes unrecoverable collapse (Kasirzadeh, 2025)
  - Discrimination and hate speech risks, surveillance, rights infringement, and erosion of trust risks, environmental and socioeconomic risks, etc.

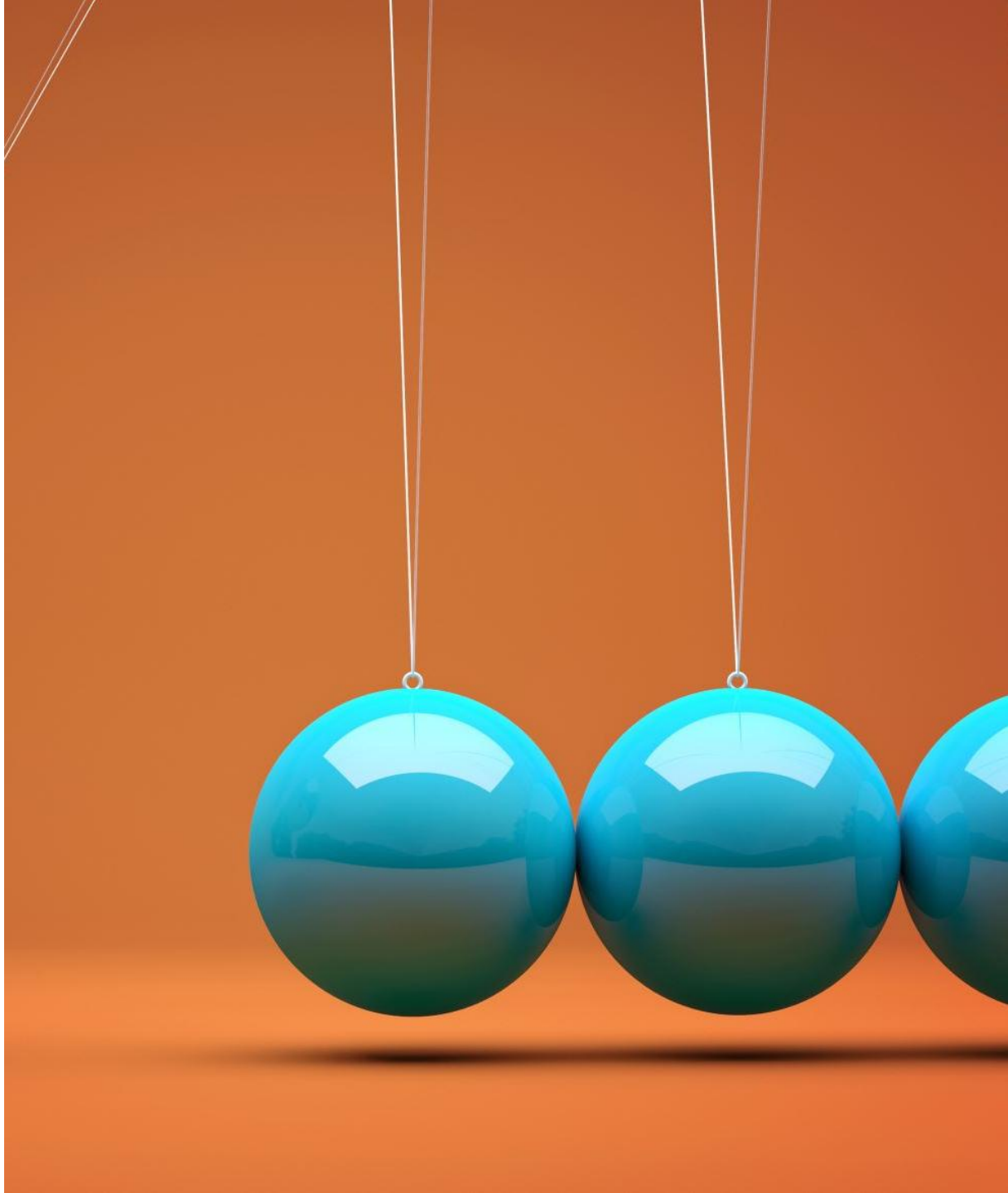


OTHER ISSUES

- Deepfakes
- Disinformation and Misinformation
- Manipulation
- Art
- Education
- Regulation
- Companions
- Moral Status
- …

PROBES(S) FOR THE ETHICS OF AI

- **P**rivacy (and Surveillance)
- **R**esponsibility
- **O**pacity
- **B**ias
- **E**quity (and Employment)
- **S**ustainability
- **S**afety



THANK YOU FOR YOUR ATTENTION.

- Questions?
- Contact: dwhite937@gmail.com