Ethics + Emerging Sciences Group

AI/Robots and Ethics:
Surveying the Risk Environment

Allan Hancock College | 18 April 2025

Keith Abney | Cal Poly SLO
Ethics + Emerging Sciences Group

Agenda

Ethics + Emerging Sciences Group

1. Introduction
2. AI & basics of risk
3. AI, C-risk, & X-risk
4. Solutions?

Introduction

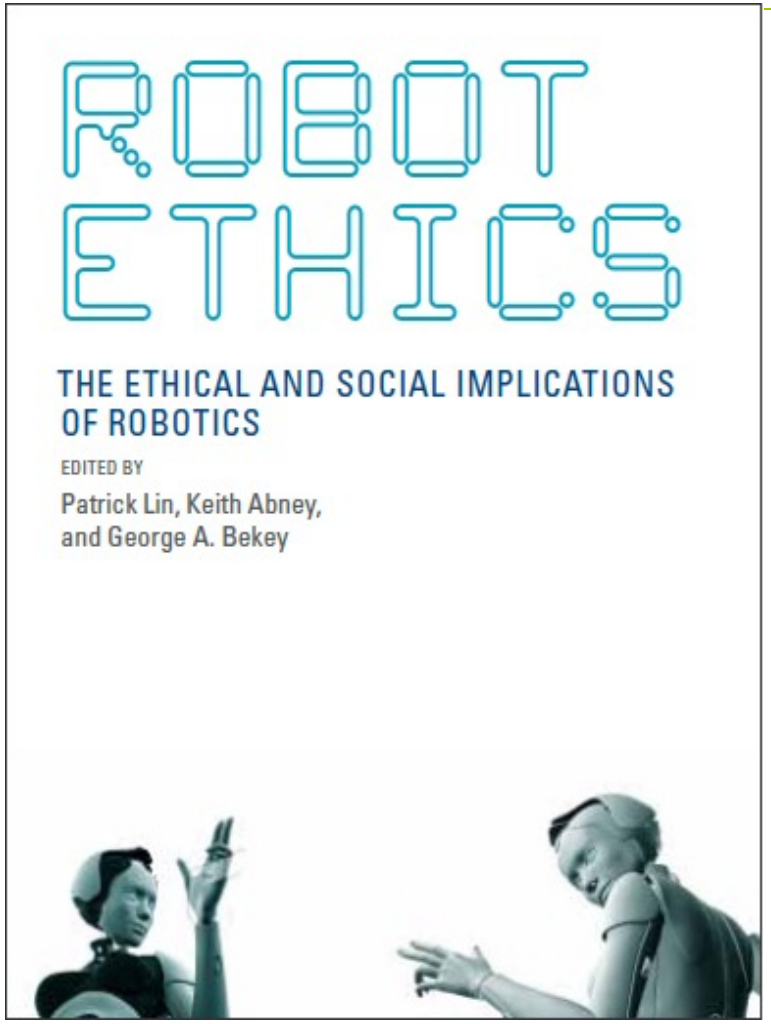
Ethics + Emerging Sciences Group

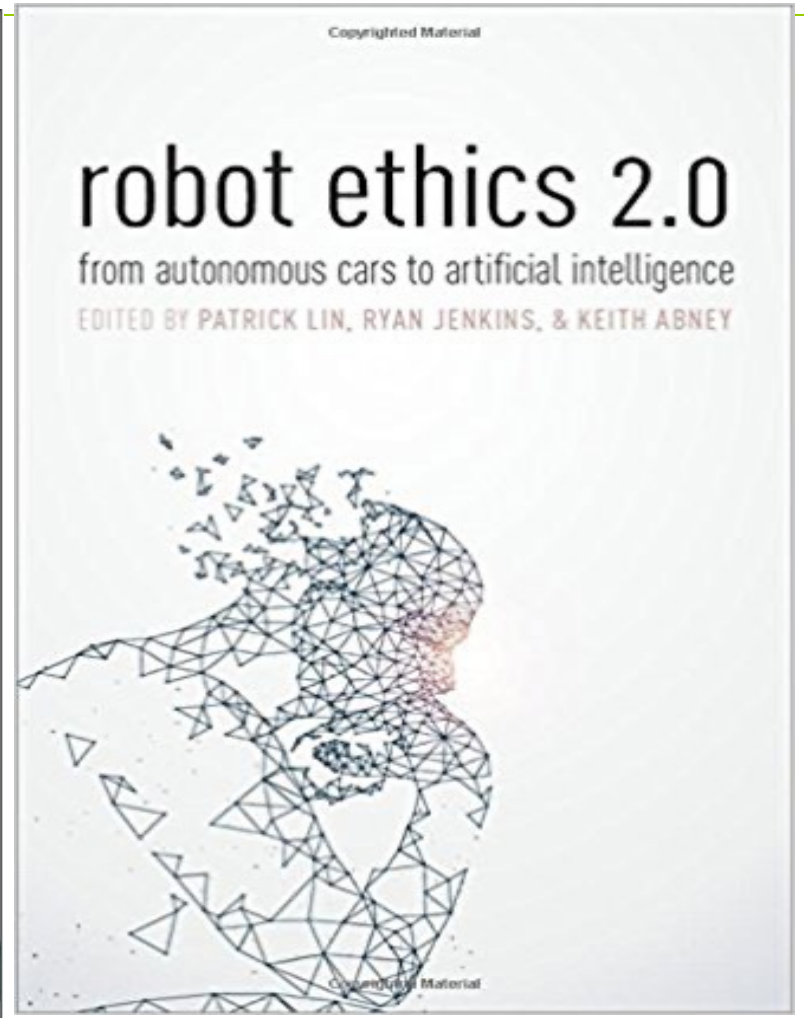
Keith Abney

- Cal Poly Philosophy; also Senior Fellow, Ethics + Emerging Sciences Group
- NASA Astrobiology & Society, other space groups
- Bioethics Committee, Arroyo Grande Community Hospital

Technology ethics

Ethics + Emerging Sciences Group





Slate

npr

acm

BBC

Basics of Risk

Concept of 'risk'

Ethics + Emerging Sciences Group

- "**Risk**" refers to possibility that harm may occur.
- BUT (*ambiguity*): the odds (**probability**)
- *Or* severity (**magnitude**) of the possible harm?
- Contrast: (*probability and/or magnitude* of) **benefits**

- **Risk/benefit assessment (RBA) in ethics:**
- commonly includes risks of **psychological harm, physical harm, legal harm, social harm, and economic harm,** and the corresponding benefits.

AI: novel risks?

Ethics + Emerging Sciences Group

- *Black box problem – Former Vice Chairman of the Joint Chiefs of Staff*:
- "In the [Defense] Department, we build machines and we test them until they break. You can't do that with an artificial intelligence, deep learning piece of software. We're going to have to figure out how to get the software to tell us what it's learned"
- Trust - grounded in predictability; depends on ability to anticipate others' behavior
- Should we trust AI?
- 'Alien' intelligence and flying analogy

Ethics of risk: what counts?

Ethics + Emerging Sciences Group

1 Acceptable-Risk Factor: Consent

- *Consent*: Is the risk voluntarily endured, or not?
- Ex: firsthand vs secondhand smoke
- Will those who are at risk from AI reasonably give consent?
- Appropriate to use AI without the meaningful consent of affected?

- **Non-voluntariness** (in which the affected party is unaware of the risk/ cannot consent) – how morally different from
- **Involuntariness** (in which the affected party is aware of the risk and does not consent)?

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

2 Acceptable-Risk Factor:
Informed Consent

- **One solution for consent:**
- ***indirect/ political consent?***
- But what about AI risk to
- ***unintended/ uninformed?***

- Does the ***morality of consent require adequate knowledge*** of what is being consented to?
- Problem: even if consent is politically possible, unrealistic that all/ most humans give *informed* consent to AI use.

Informed Consent: problems

Ethics + Emerging Sciences Group

- Is there ***full awareness***
- Of the true nature of the risk?
- Could such knowledge *undermine*
- *fulfilling their (risky) roles*?
- "I'd rather not know...."

- Do the informed have an ***obligation to tell others*** of the risks?
- Foreseeable but unknown risks – how should they (the **'known unknowns**') be handled?
- Could informing people that they are at risk ever be unethical, even akin to ***terrorism?***

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

3 Acceptable-Risk Factor: The Affected Population

- Who is at risk – esp. particularly
- susceptible or innocent, or
- *Only* those who ***understand that their role is risky***, even if
- ignorant of particulars of the risk?

- **Example**: in *military operations*, civilians/ noncombatants usually not morally required to endure the same risks as military personnel,
- especially when the risk is involuntary or non-voluntary.

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

4 Acceptable-Risk Factor: Step risk versus State risk

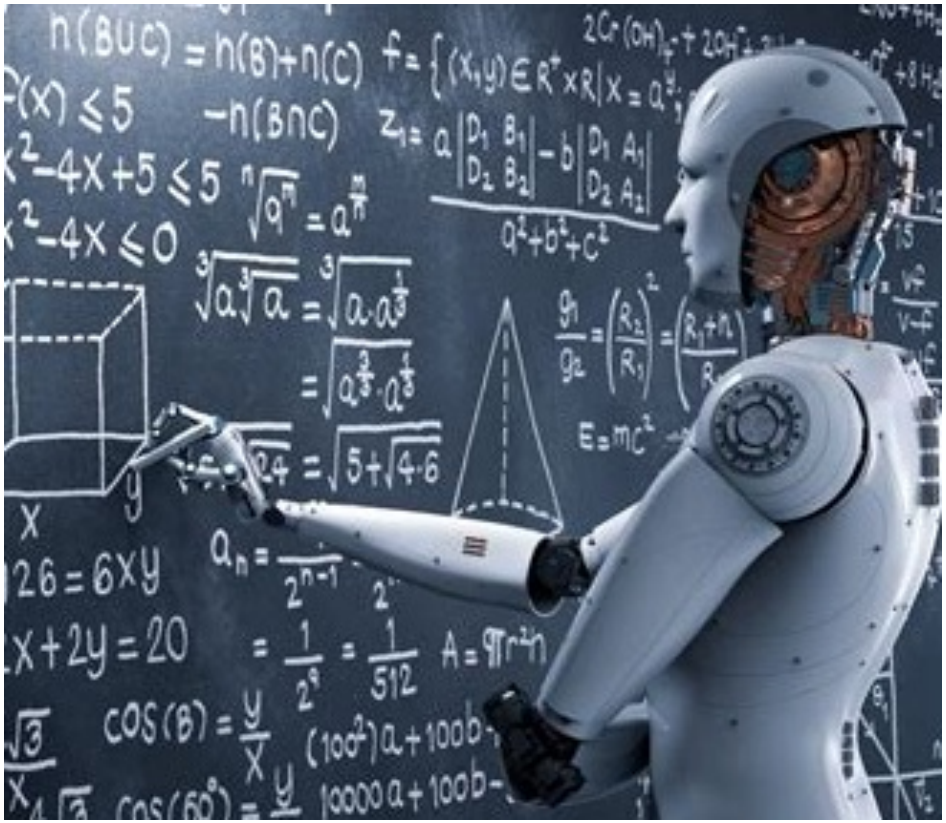
- **State risk:** the risk of being in a certain state, and the total amount of risk is a direct function of the time spent in the state;
- ***Time-dependent* -** total risk depends on the time spent in state
- Ex.: for us living on the surface of the Earth, **death by asteroid strike** is a state risk (it increases the longer we're here).

- **Step risk**: a discrete risk of taking the next step in a series/ undergoing some transition; once the transition is complete, the risk vanishes.
- In general, not time-dependent, so amount of time spent on the step matters little (or not at all).
- **Ex.: Crossing a minefield** is usually a step risk – the risk is the same whether you cross it in 1 minute or 10 minutes.

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

- *Step risk versus state risk*: How shall we determine which is more important?
- Ex: ***slowing down/ stopping AGI research,*** *given 'fast takeoff'*
- Step risk, but success would decrease other risks to humanity, so
- **how decide what to do?**



Ethics of risk, cont'd

Ethics + Emerging Sciences Group

5 Acceptable-Risk Factors: Seriousness

- We thereby come to the two most basic facets of risk assessment, seriousness and probability: how bad would the harm be, and how likely is it to happen?

- ***Seriousness (aka magnitude)*:** Risk of death/ serious physical harm seems qualitatively different than risk of a scratch or a temporary power failure or slight monetary costs.

- But ***the attempt to make serious risks nonexistent may turn out to be prohibitively expensive*** (or otherwise contraindicated).
- What magnitude of AI risk is acceptable – and to whom: users, nonusers, the environment, or the AI itself?

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

6 Acceptable-Risk Factor:
Probability

Seriousness of a 10-km asteroid hitting Earth is high

- but the *probability* is low
- though not zero - ask dinosaurs!
- *Probability of harm from AIs*?
- How certain is this estimate?
- How decide on the acceptable probability of serious (versus moderate or mild) harm?
- Do we us a linear, asymptotic, or other function? Continuous or not?

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

7 Acceptable-Risk Factors: False Positives

- ***False Positive:*** AI wrongly determines that a phenomenon occurs or an indicator is present, when it is in fact absent.

- **Example:** LAWS wrongly determines child holding a tree branch is a legitimate military target, and decides to fire its weapon

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

8 Acceptable-Risk Factors:

False Negatives

- ***False Negative:*** AI wrongly determines that a phenomenon does not occur, or indicator is absent when in fact present.
- **Example:** Medical AI misreads scan & determines patient does not have cancer, when they do.
- **Ethics: *which is worse, false positives or false negatives***?
- Can AI possibly understand the ethics/ social implications?

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

9 Who Determines Acceptable Risk?

- In various social contexts, all of the following defended as proper methods for determining that a risk is (un)acceptable:

- ***9.1 Good faith subjective standard***: It is up to each individual as to whether an unacceptable risk exists.
- Can the designers or users of AI be trusted to make wise choices about (un)acceptable risk?
- *Idiosyncrasies of human risk aversion* make this standard hard to defend, as well as the problem of
- involuntary/ non-voluntary risk borne by nonusers.

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

- ***9.2 The reasonable-person standard*:** An (un)acceptable risk is simply what a fair, informed member of a relevant community believes it to be

- Does a ***professional code*** or some other basis work for what a 'reasonable person' would think for the AI field,
- Replacing subjective judgment of its practitioners and users?

- Should we allow a fully autonomous AI to judge risk – would we trust it to accurately determine and act upon the assessed risk?

- Otherwise require a ***Human in/on the loop?***
- Reasonable to expect AI always requires teleoperators?

Ethics of risk, cont'd

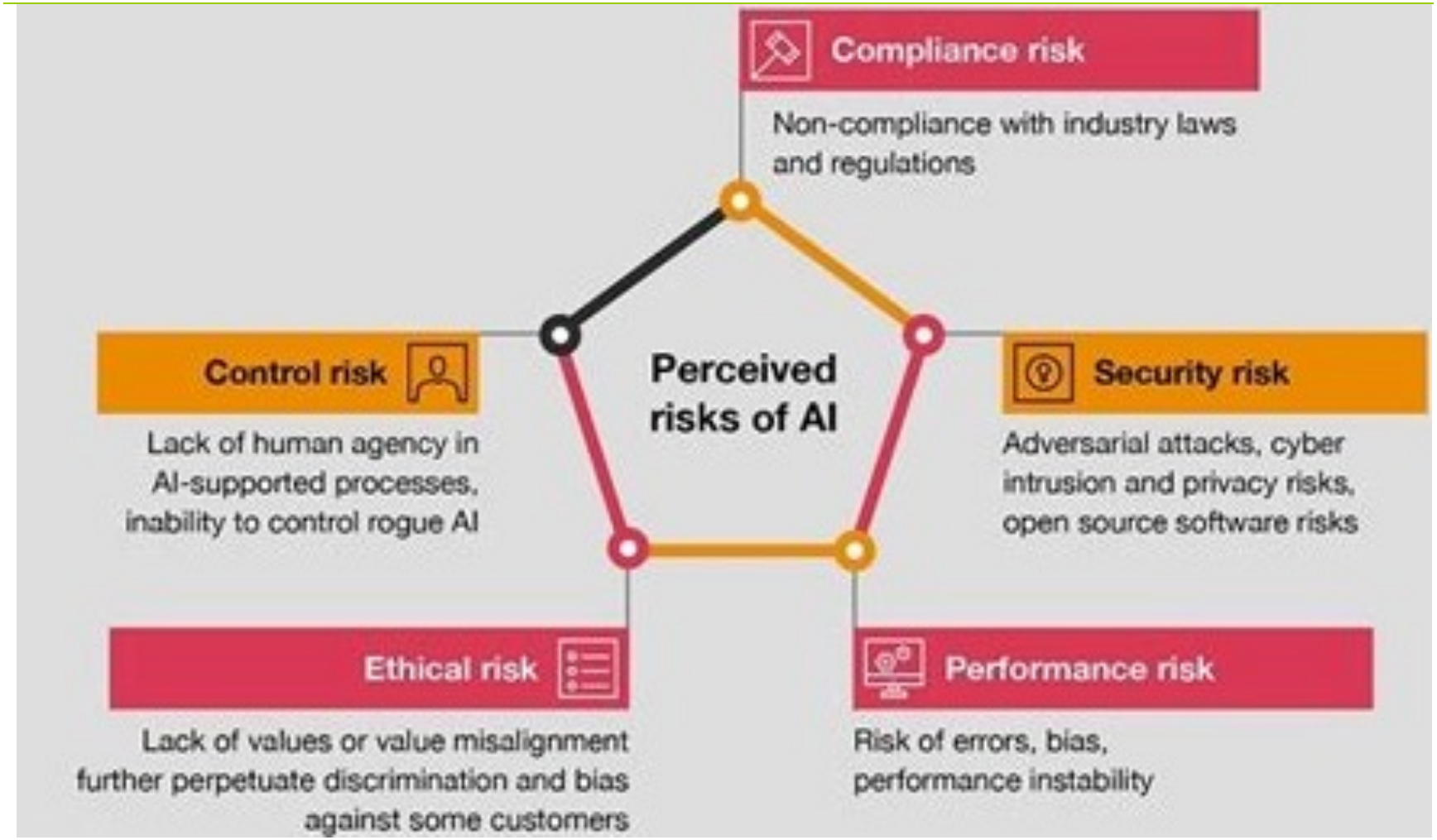
Ethics + Emerging Sciences Group

- ***9.3 Objective standard*:**
- An unacceptable risk requires evidence and/or expert testimony
- as to reality of the risk.

- **First-generation problem**: how prove an unacceptable risk unless a first generation already endured and suffered from it?
- Implementation? Perhaps a **'kill switch'?**
- Autonomous operation until ...
- A human determines something wrong
- But: still leaves unsolved first-generation problem. How else could we obtain convincing objective evidence?

5 general types of AI risk

Ethics + Emerging Sciences Group



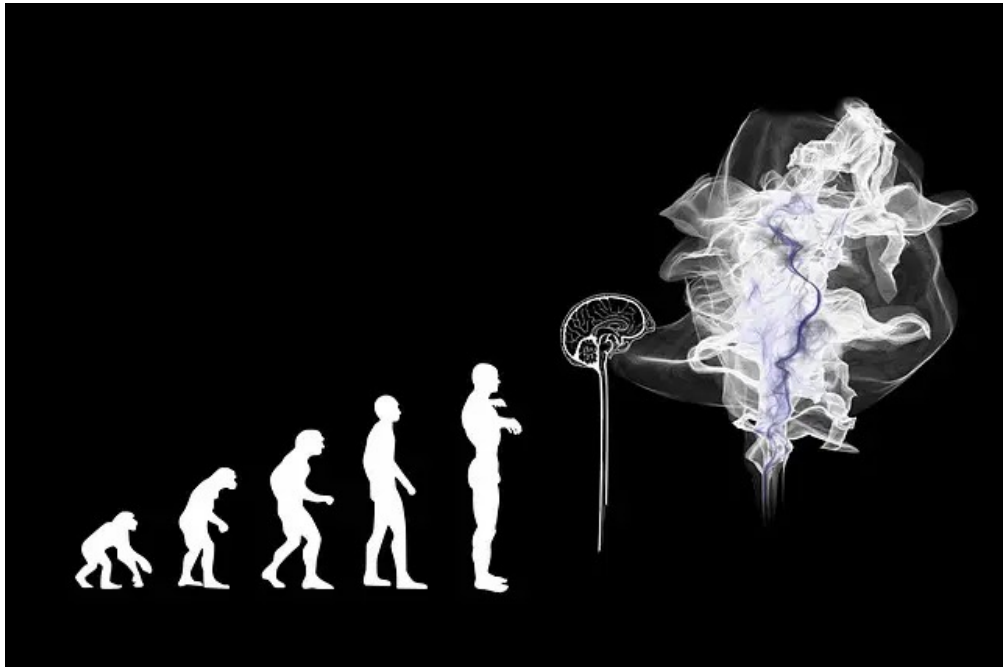
Ethics of risk, cont'd

Ethics + Emerging Sciences Group

10 Acceptable-Risk Factors:
The Wild Card: Existential Risk

- "a risk that, should it come to pass, would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential."
- Vs mere **catastrophic risks**

- *Key question:*
- ***Are X-risks a fundamentally different kind of risk****, so the previous considerations no longer apply/ are superseded?*



Catastrophic risks

Ethics + Emerging Sciences Group

- **Climate change** – may kill millions, but short/ medium-term not X-risk
- Or: Artificial intelligence (AI) tools integrated into decision-making processes in certain high-risk settings
- Such as ***employment, credit, health care, housing, and law enforcement/ predpol***

Catastrophic risk: biology

Ethics + Emerging Sciences Group

- Sept 2023, tech execs testimony to Congress:
- "Harris told the room that with $800 and a few hours of work, his team was able to strip Meta's safety controls off LLaMA 2 and that the AI responded to prompts with instructions to develop a biological weapon."
- Democratization of terror?

Ex.: AI & Space cybersecurity

Ethics + Emerging Sciences Group

1. Remoteness of space
2. Complexity of systems
3. Unclear legal regime
4. Higher stakes
5. Novel Scenarios
6. Catastrophic or X-risk?

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

- **Why is X-risk ethics different from catastrophic risk?**

Plausibly, precautions that include

- ***extensive, variegated, realistic, and exhaustive pre-deployment testing*** of AIs in virtual environments
- before they are used in actual human interactions
- could render many AI risks, even catastrophic risks, acceptable under the previous criteria.

- But AI **existential risk (**X-risk) may remain unacceptable even with the most rigorous pre-deployment testing
- **Why?**

Ethics of risk, cont'd

Ethics + Emerging Sciences Group

- Answer: **First generation problem!**
- Most risks, even catastrophic risks, may be mitigated after a failure
- by changing policies and procedures to make the risk acceptable in the future

- But the first failure of an existential risk means no opportunity to learn from our mistakes – because we will not exist!

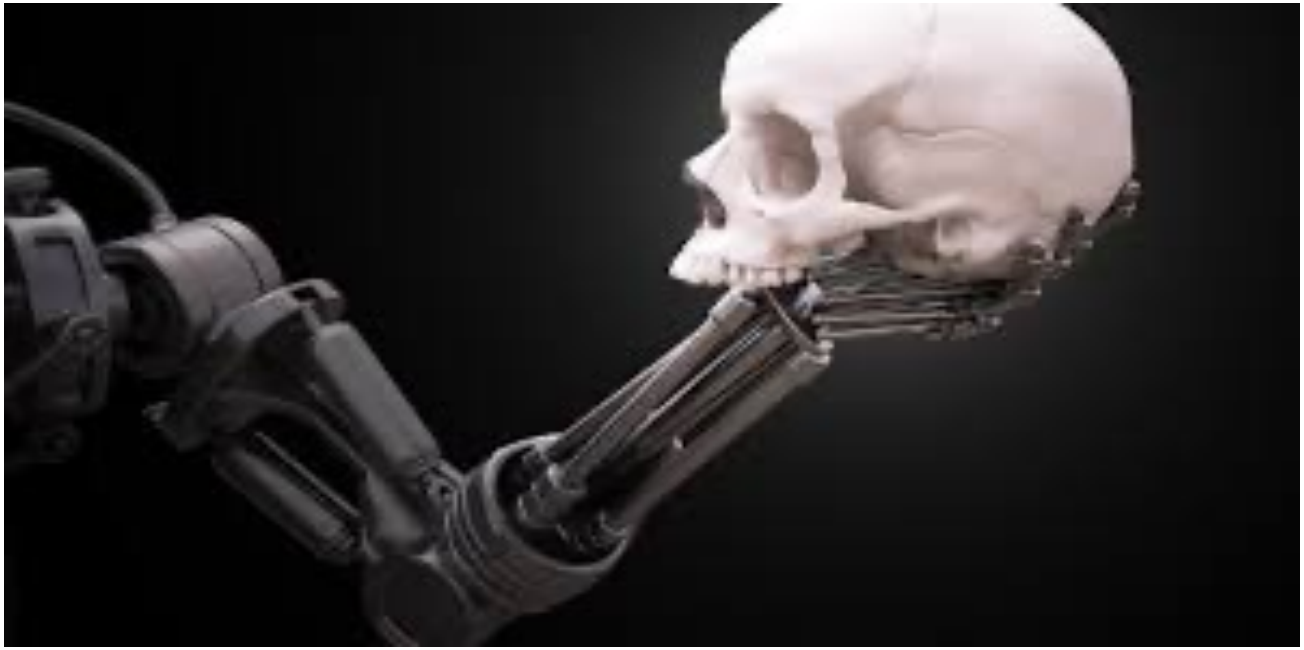
- So, do X-risks require a new ethics framework?
- Person-affecting vs person-neutral ethics

Extreme Justifications: existential risk?

AI X-risk

Ethics + Emerging Sciences Group

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



Ex.: Why explore & settle space?

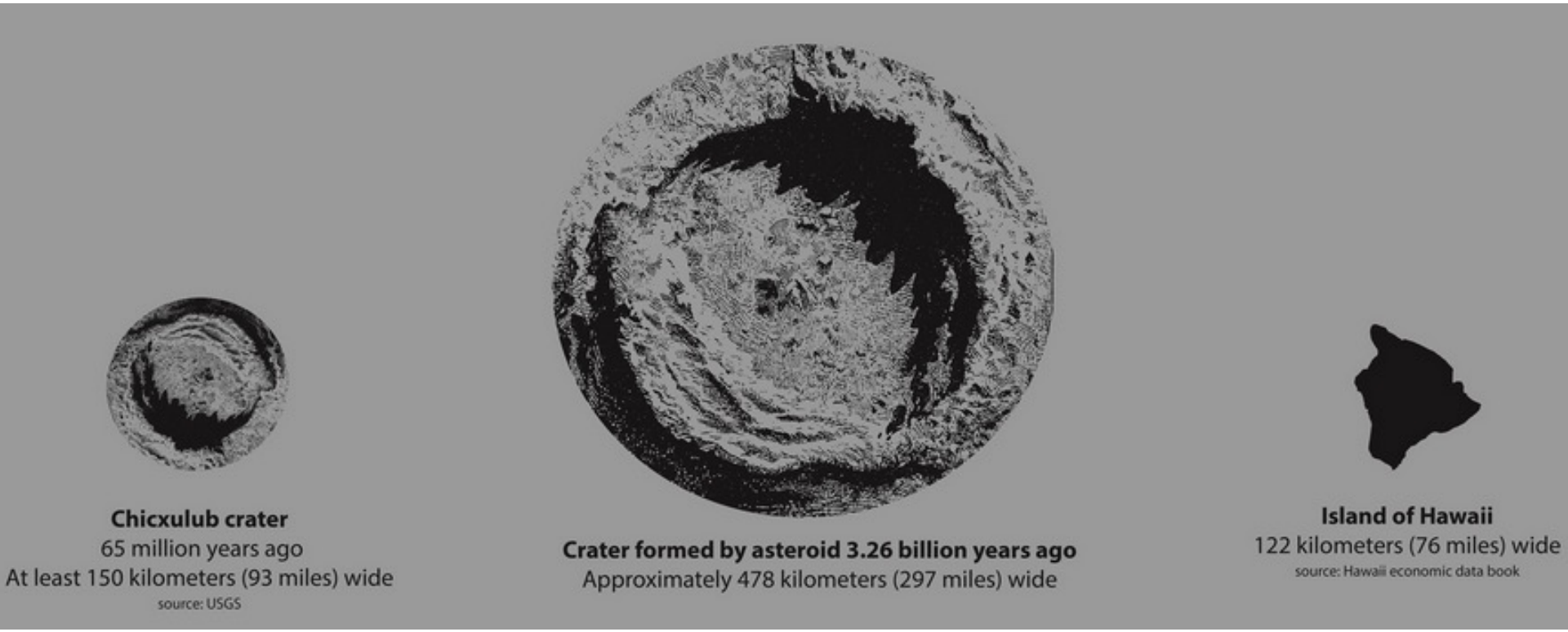
Ethics + Emerging Sciences Group

- For knowledge / science
- For a social release valve
- Just because we can
- For “backing up” our biosphere (mitigating **existential risk**)

X-risk: comets & asteroids

Ethics + Emerging Sciences Group

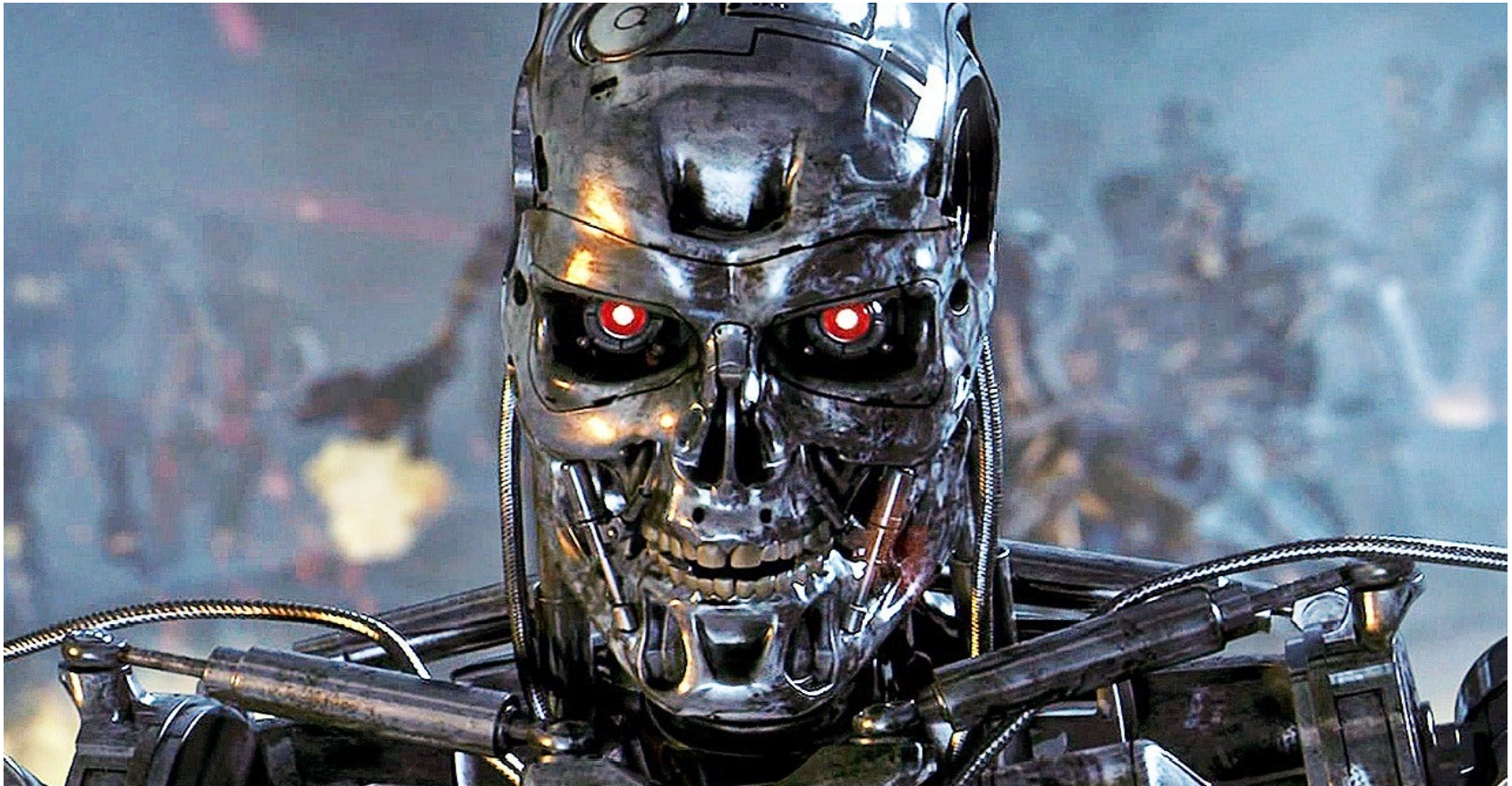
- K-T mass extinction, others?



Beyond the robot uprising?

Ethics + Emerging Sciences Group

- AI could kill us all on purpose, or by accident...
- **Perverse instantiation** and ‘paperclip maximizer’



Reason for worry: the Great Filter

Ethics + Emerging Sciences Group

- **Fermi Paradox**: If aliens exist, where is everybody? Why aren't aliens obvious, or even our ancestors?
- Earthlike planets estimate: ~40 billion in galaxy, with average Earthlike world ~2-3 billion years older.

Why Haven't We Found Aliens?

Ethics + Emerging Sciences Group

Rare Earth

Planets with complex life are extremely rare

Great Filter

ETs failed to overcome evolutionary hurdles

Great Silence

Intelligent ETs aren't communicating with us

Life As We Know It

ETs may be machines, not biological beings

Long Road Ahead

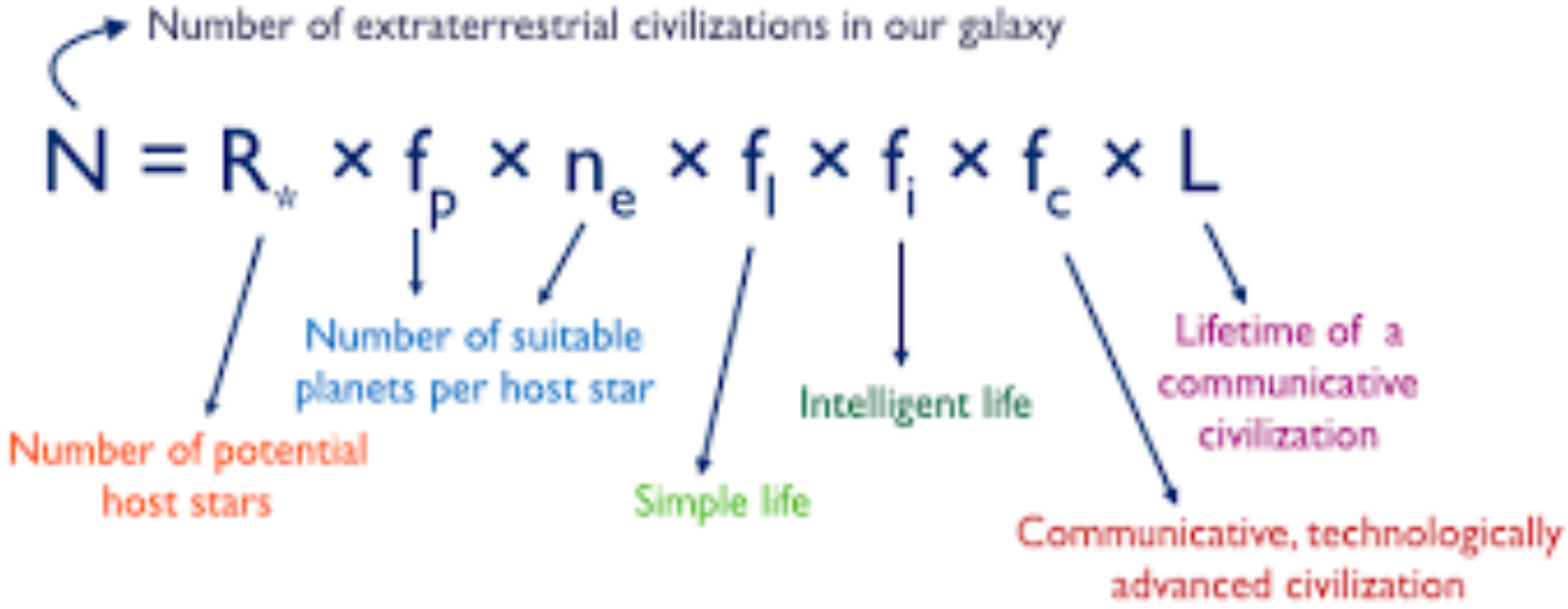
We've got a lot more searching to do

Occam's Razor

We haven't found them b/c they don't exist

The Great Filter

Ethics + Emerging Sciences Group



- **SETI:** failure so far... so
- **Great Filter**: which variable is small, if N is near zero?
- ***Rare Earth hypothesis*** - a biological variable
- ***Prime Directive/ Zoo Hypothesis*** - advanced civilizations agree not to contact primitive civilizations.

Great Filter & Doom Soon?

Ethics + Emerging Sciences Group

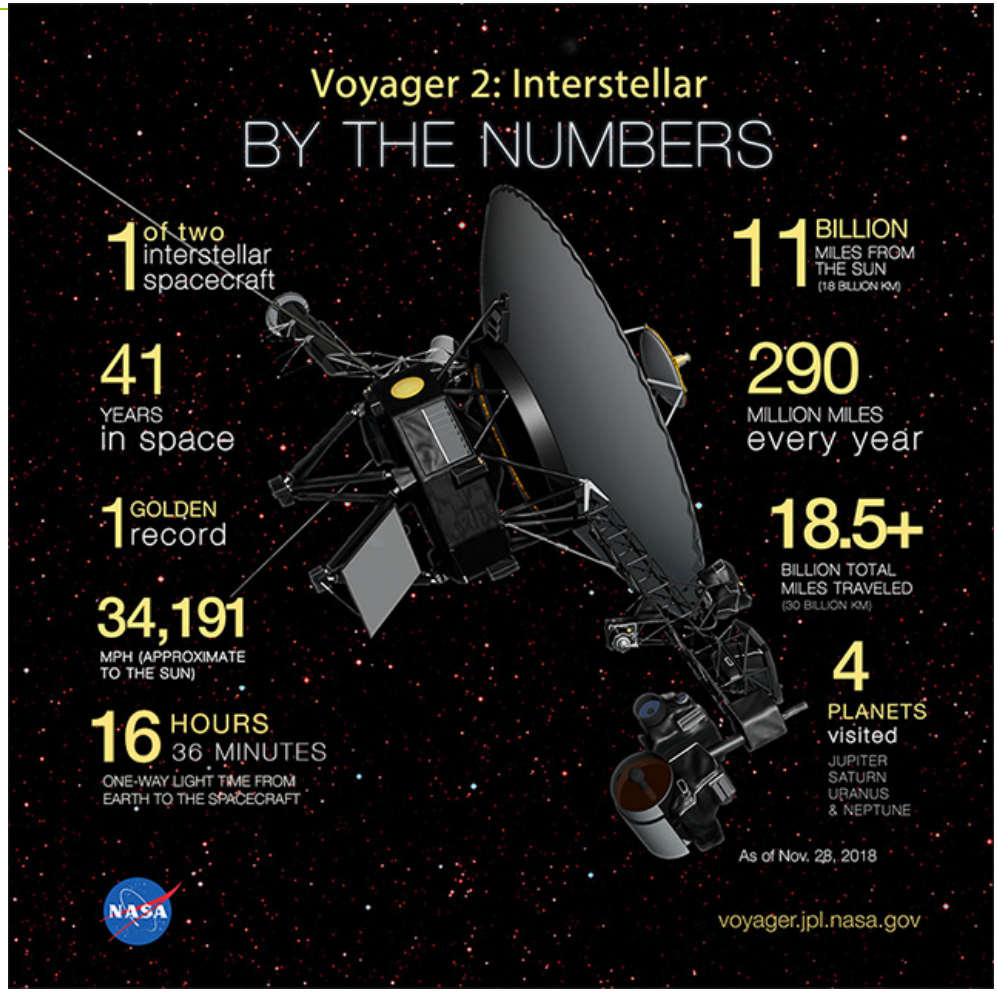
- **Existential worry:** but what if *Filter is in our future*, because every civilization that reaches our level of technology soon goes extinct?!

- *If L is small -* ***Doom Soon?***

A solution to Doom Soon?

Ethics + Emerging Sciences Group

- **Reduce existential risks, by *colonizing other planets*?**
- Crucial concern: do we have plenty of time to settle before extinction on Earth?
- Problem: our space robots; particularly Voyager 1 and 2, & hypothetical *von Neumann probes*



Interstellar Doomsday Argument

Ethics + Emerging Sciences Group

- **Self-Sampling Assumption (SSA):** "One should reason as if one were a random sample from the set of all observers in one's reference class".
- So, for a random observation of a phenomenon: 95% probability it will continue for between 1/39 and 39 times its present age
- only 5% chance a random observation comes in the first or last 2.5% of a phenomenon's lifetime
- *Voyager 1* first entered interstellar space in 2012
- So, in 2025, it's 95% probable that our time left as a civilization that sends interstellar robots
- is between $L/39$ (for L = 13 years, that's 122 more days) and $39L$ (507 more years)

Interstellar Doomsday Argument, cont'd

Ethics + Emerging Sciences Group

- von Neumann probes cinch problem?
- With feasible expansion speed of c/40, robotic saturation of the entire galaxy would take
- ***4 million years***, *< 1/3,000th* of the age of Milky Way.

- If we're not saturated by robotic probes from ancient aliens, implausible that it's because they need more time to get here.
- '*Oumuamua* as alien lightsail probe? (Loeb) ....

X-risk and ethics

Ethics + Emerging Sciences Group

- Moral imperative: is reducing potential existential risks our highest moral priority?
- Consequentalism: **expected utility** = (Pr (Benefit) x Mag (Benefit)) – (Pr (Harm) x Mag (Harm))
- any prob (no matter how low) x loss of near-infinite value trumps any prob x (finite value)
- So, for standard consequentialism/ EA, we should prioritize lowering existential risk over any other good - reducing hunger, poverty, cancer, etc...

X-risk and ethics, cont'd

Ethics + Emerging Sciences Group

- Deontology - ***Extinction Principle***: "one always has a moral obligation never to allow the extinction of all creatures capable of moral obligation."
- Accordingly, it is an absolute duty to keep things capable of obeying absolute duties in existence.
- So, mitigating or minimizing existential risk is an absolute duty, which wins any conflict it has with another duty.
- If e.g., colonization of other planets thereby minimizes existential risk, then it is our highest duty.
- But: Virtue ethics may disagree ...

Solutions?

Conclusions: AI audit?

Ethics + Emerging Sciences Group

- **One Solution for C-risk - Responsible AI?**
- Require regulation & ***AI auditing***: count on best practice standards and procedures to emerge,
- and require implementation before deployment/ use
- Critics: possible *audit-washing* – bad actors game loopholes and ambiguities in audit requirements
- to demonstrate compliance without actually providing meaningful reviews/ proof.
- **AND: doesn't work for X-risk?**

Conclusions: space backup?

Ethics + Emerging Sciences Group

- Is only solution to X-risk to develop & settle space?
- Burden of proof on opponents of space exploration? But….
  - Extend / worsen terrestrial conflicts
  - Opportunity costs
  - AI/Robots do it better?

Look before the next leap

Ethics + Emerging Sciences Group

If risk of AI development means space settlement is our opportunity to **start over again/ provide a backup**,

Then let's give it the **forethought** it deserves, so that we don't simply replicate our current problems and **risks** elsewhere.

Acknowledgements

Ethics + Emerging Sciences Group

- This work is supported in part by the **US National Science Foundation**, grant no. 2208458
- Also: **Cal Poly,** College of Liberal Arts and Philosophy Dept.

- All images and copyrights are properties of their respective owners, used in accordance with the “fair use” clause (§107 of Title 17 of the US Code)

Thank you!

Ethics + Emerging Sciences Group

Keith Abney, Sr. Fellow: kabney@calpoly.edu

Ethics + Emerging Sciences Group
Cal Poly, Philosophy Department
San Luis Obispo, California 93407

http://ethics.calpoly.edu

Survey and QR code info

Ethics + Emerging Sciences Group

Here is the link to the survey:
https://bit.ly/AHC_AI_Summit_S25

Here is the QR Code: