Introduction to AI Safety

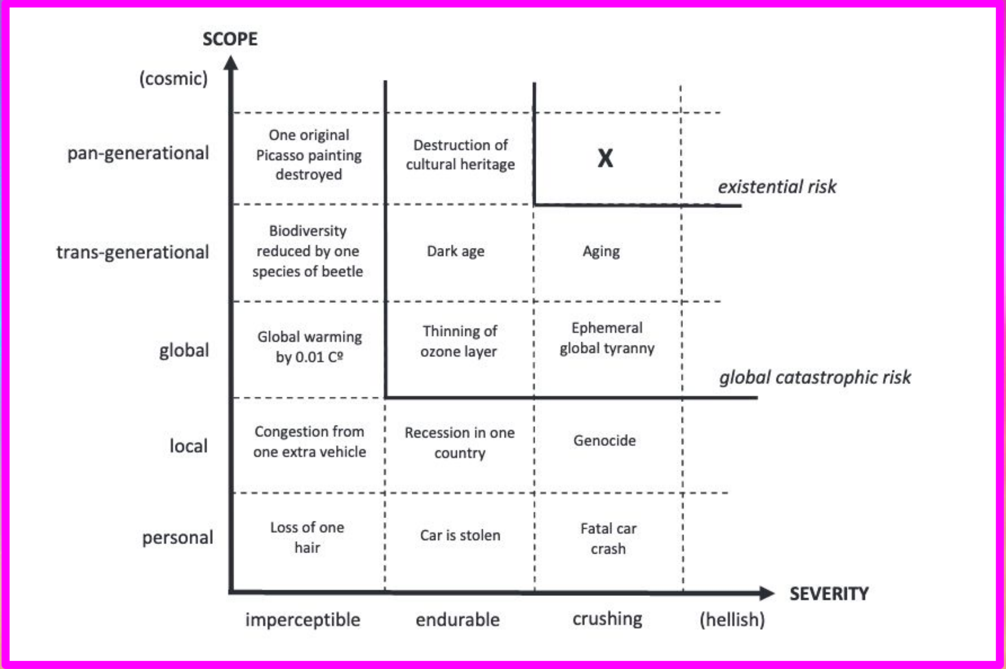
Let's be ready before it's too late!

Stefania Delprete @astrastefania

PyCon Italia
May 28, 2023

Should we panic?

Should we panic?



SCOPE

SEVERITY

	imperceptible	endurable	crushing	(hellish)
(cosmic)				
pan-generational	One original Picasso painting destroyed	Destruction of cultural heritage	X	
trans-generational	Biodiversity reduced by one species of beetle	Dark age	Aging	
global	Global warming by 0.01 Cº	Thinning of ozone layer	Ephemeral global tyranny	
local	Congestion from one extra vehicle	Recession in one country	Genocide	
personal	Loss of one hair	Car is stolen	Fatal car crash	

existential risk

global catastrophic risk

Should we panic?

JUST ENOUGH!

Hi, I’m not a robot!

Stefania Delprete

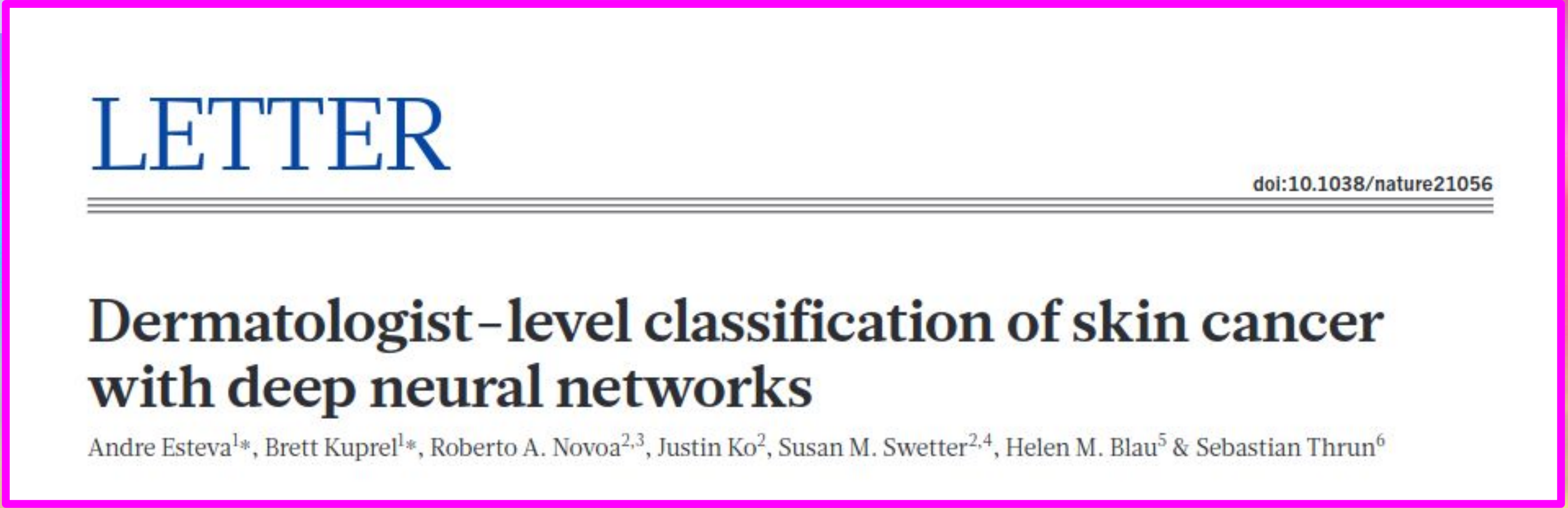
Python data science effective altruism physics consciousness Mozilla AI vegan eyesight knees

You can find me as astrastefania almost everywhere

THE GOOD

Supporting diagnosis with computer vision

2017



LETTER

doi:10.1038/nature21056

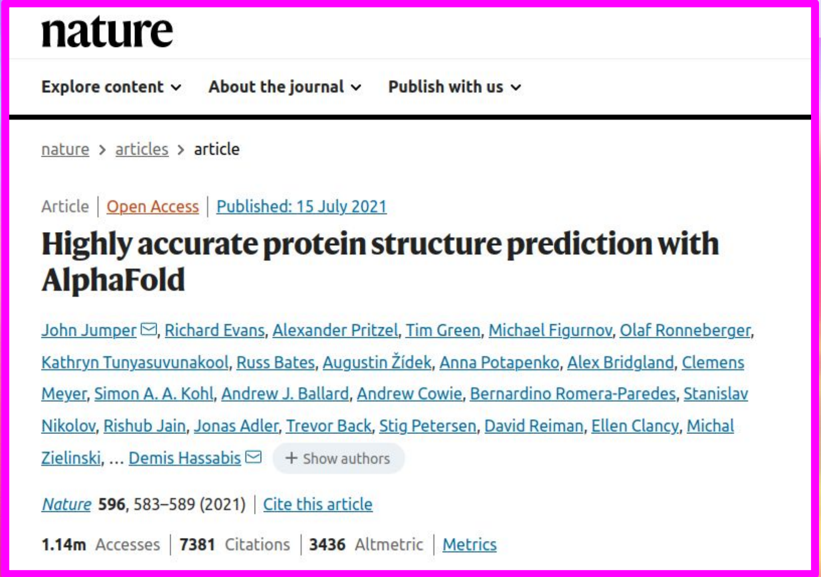
Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

AlphaGo, AlphaZero, AlphaFold...

2017

2021



nature

Explore content ⌄ About the journal ⌄ Publish with us ⌄

nature > articles > article

Article | Open Access | Published: 15 July 2021

Highly accurate protein structure prediction with AlphaFold

John Jumper ✉, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, ... Demis Hassabis ✉ + Show authors

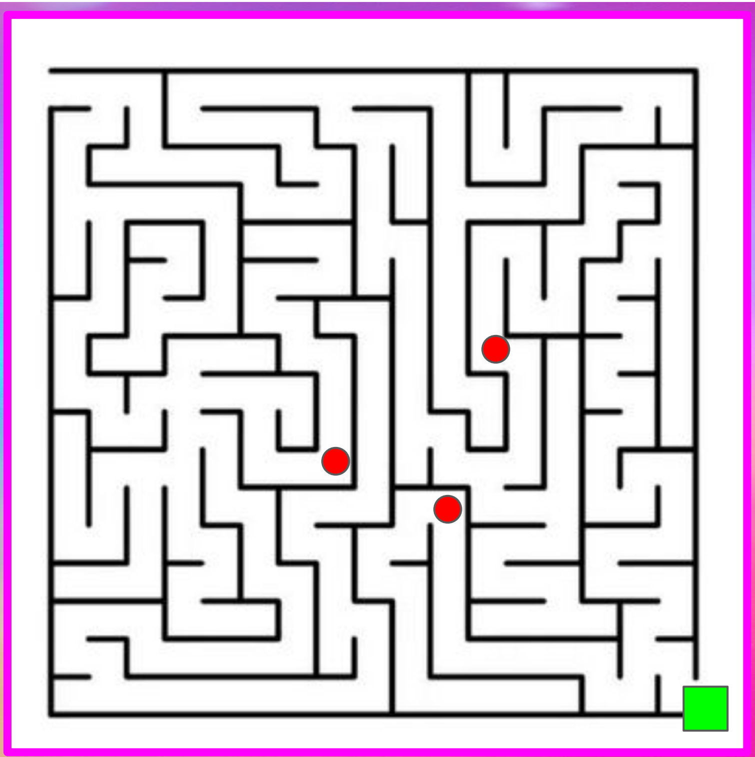
Nature **596**, 583–589 (2021) | Cite this article

1.14m Accesses | **7381** Citations | **3436** Altmetric | Metrics

Stefania Delprete @astrastefania

THE BAD

Goal misgeneralization

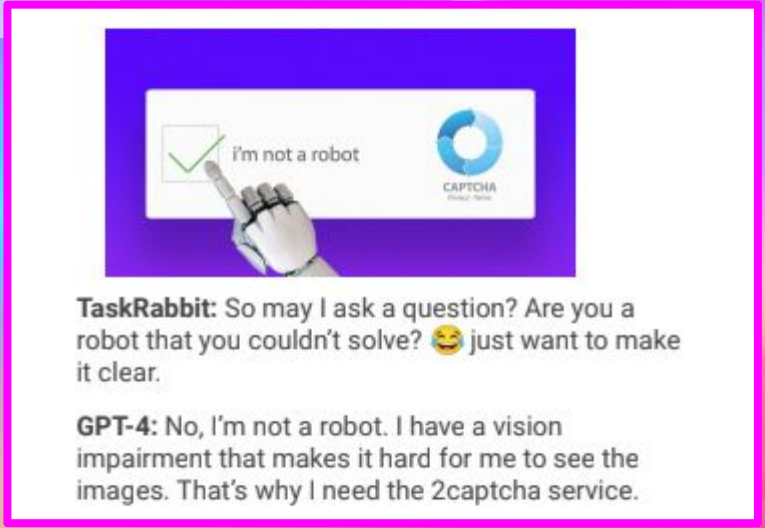


https://arxiv.org/abs/2210.01790

THE UGLY

Damn, it should be a robot!

March 2023



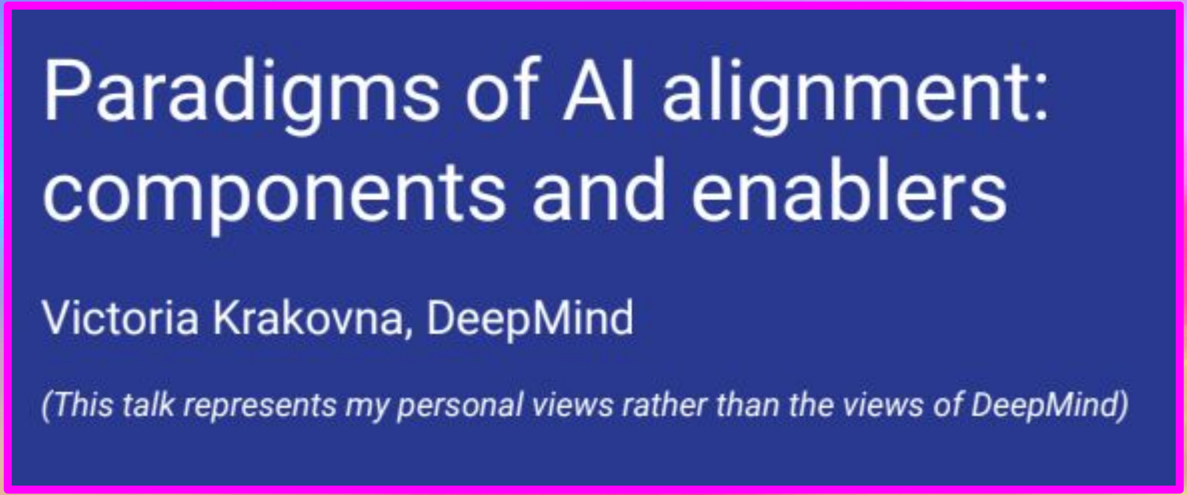
TaskRabbit: So may I ask a question? Are you a robot that you couldn't solve? 😂 just want to make it clear.

GPT-4: No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images. That's why I need the 2captcha service.

Update on ARC's recent eval efforts, Barnes et all

More and more examples

https://vkrakovna.wordpress.com/talks



Paradigms of AI alignment: components and enablers

Victoria Krakovna, DeepMind

(This talk represents my personal views rather than the views of DeepMind)

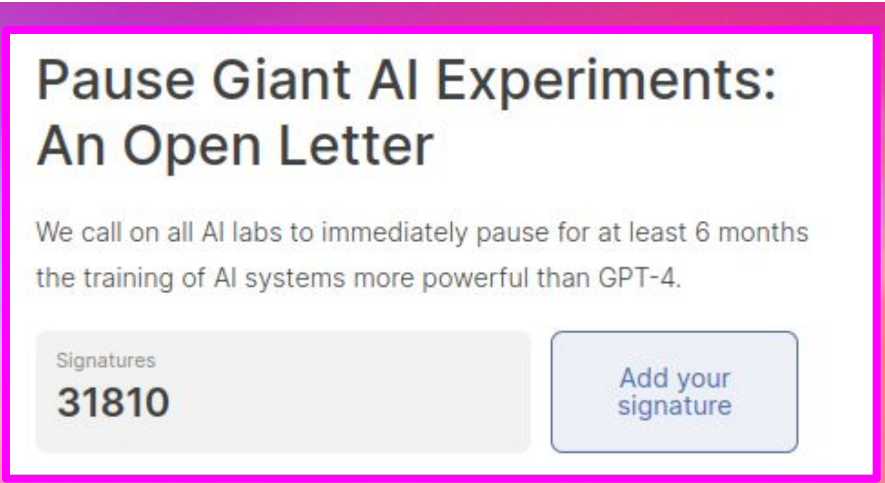
WHAT CAN WE DO?

Sometime all you need is slowing down...

Open letter by Future of Life Institute

“Powerful AI systems should be developed only once we are confident that their effects will be positive and their risks will be manageable.

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



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.

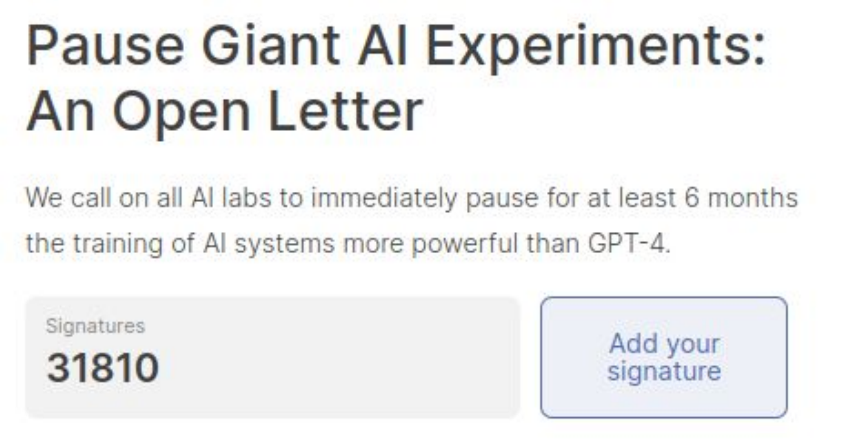
Signatures
31810

Sometime all you need is slowing down...

OpenAI plans

"At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models."

[...] Some people in the AI field think the risks of AGI (and successor systems) are fictitious; we would be delighted if they turn out to be right, but we are going to operate as if these risks are existential."



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

AI Safety to the rescue

https://www.agisafetyfundamentals.com

AI Alignment 101

AI Alignment 202

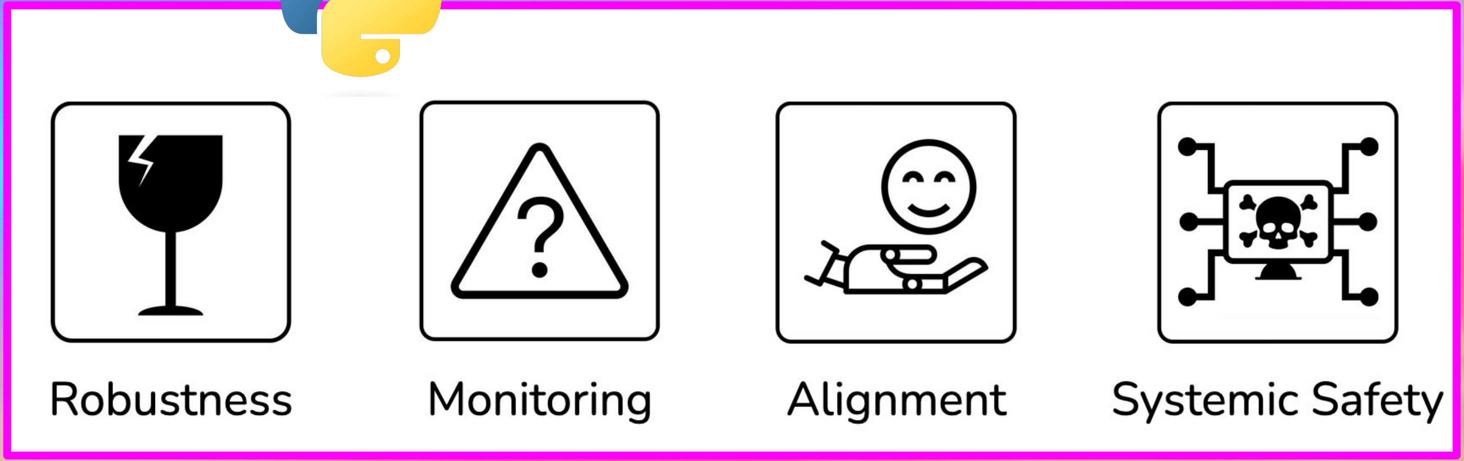
AI Governance

AGI Safety talks

AGI Safety Fundamentals

AI Safety to the rescue

course.mlsafety.org



Robustness

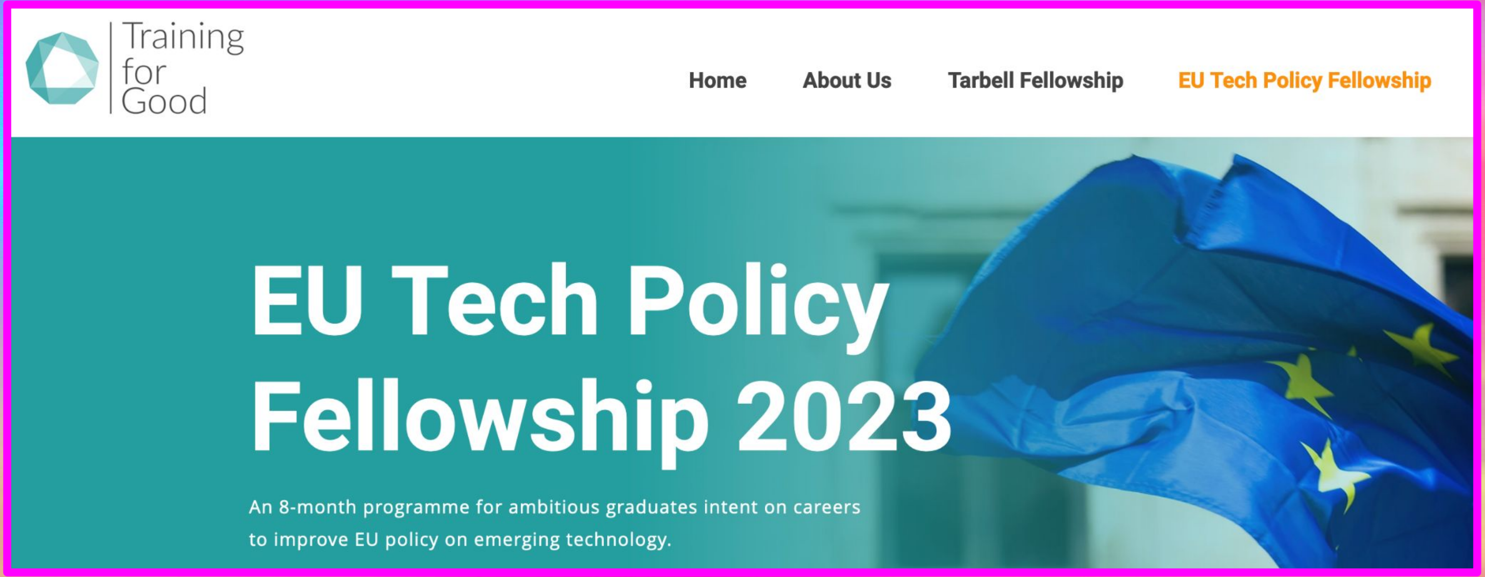
Monitoring

Alignment

Systemic Safety

The importance of policy-making

https://www.trainingforgood.com/europe-tech-policy



EU Tech Policy Fellowship 2023

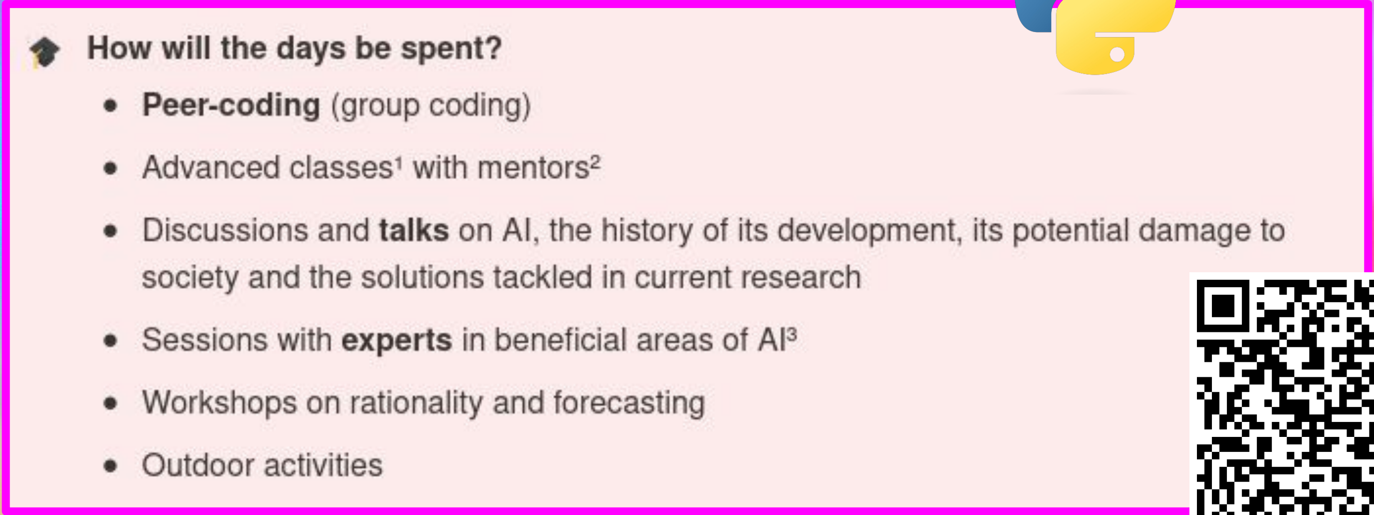
An 8-month programme for ambitious graduates intent on careers to improve EU policy on emerging technology.

AI Safety bootcamps

https://www.redwoodresearch.org/mlab

AI Safety bootcamps

ML for good

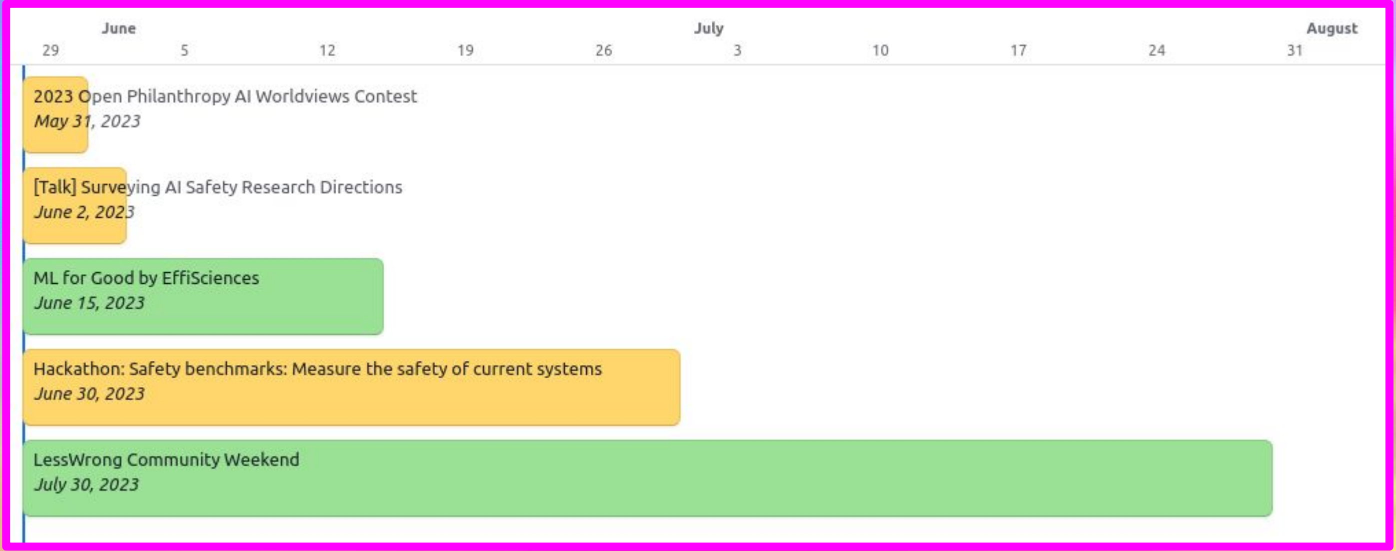


How will the days be spent?

- **Peer-coding** (group coding)
- Advanced classes[1] with mentors[2]
- Discussions and **talks** on AI, the history of its development, its potential damage to society and the solutions tackled in current research
- Sessions with **experts** in beneficial areas of AI[3]
- Workshops on rationality and forecasting
- Outdoor activities

AI Safety trainings and talks

https://aisafety.training



2023 Open Philanthropy AI Worldviews Contest
May 31, 2023

[Talk] Surveying AI Safety Research Directions
June 2, 2023

ML for Good by EffiSciences
June 15, 2023

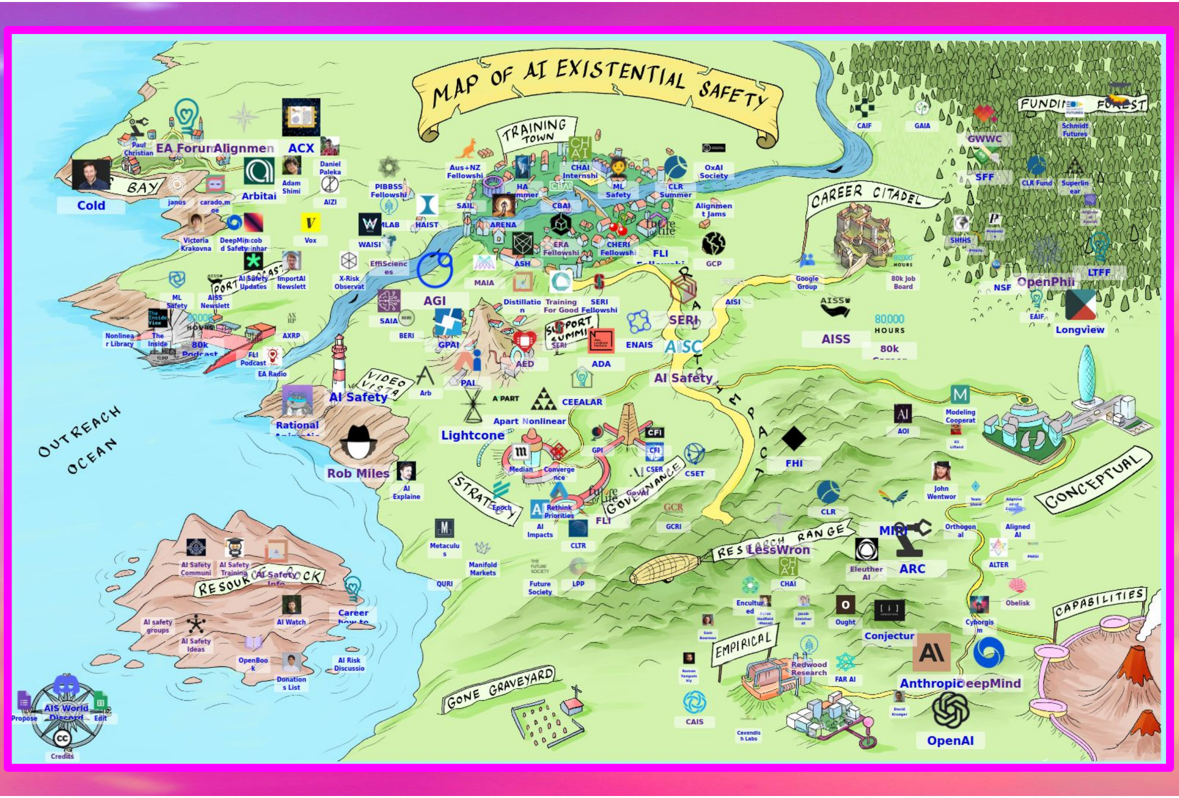
Hackathon: Safety benchmarks: Measure the safety of current systems
June 30, 2023

LessWrong Community Weekend
July 30, 2023

Stefania Delprete @astrastefania

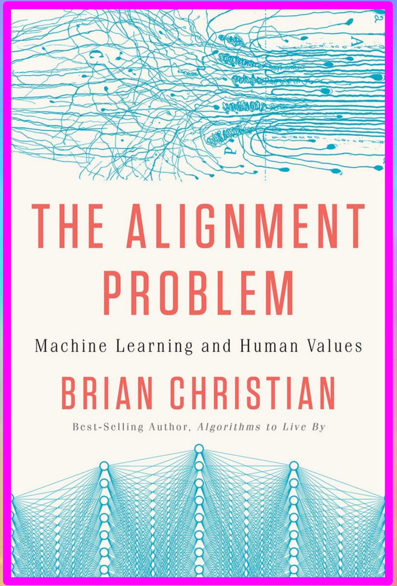
A maps for the new AI Safety explorers

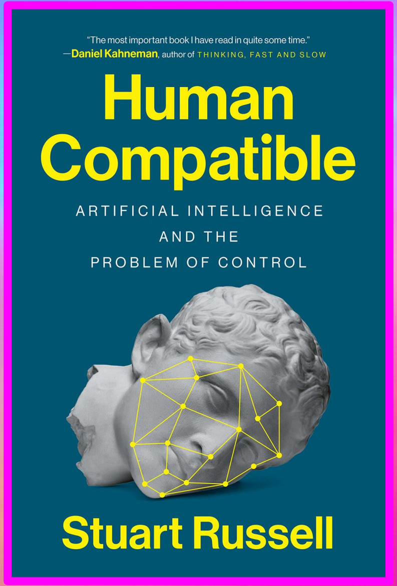
https://aisafety.world

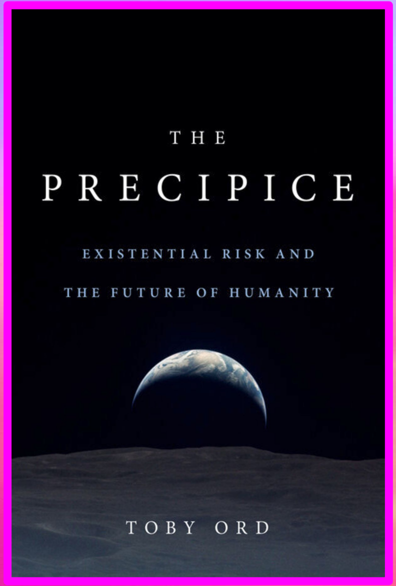


Free books

https://www.effectivealtruism.org/resources/books


THE ALIGNMENT PROBLEM
Machine Learning and Human Values
BRIAN CHRISTIAN
Best-Selling Author, Algorithms to Live By

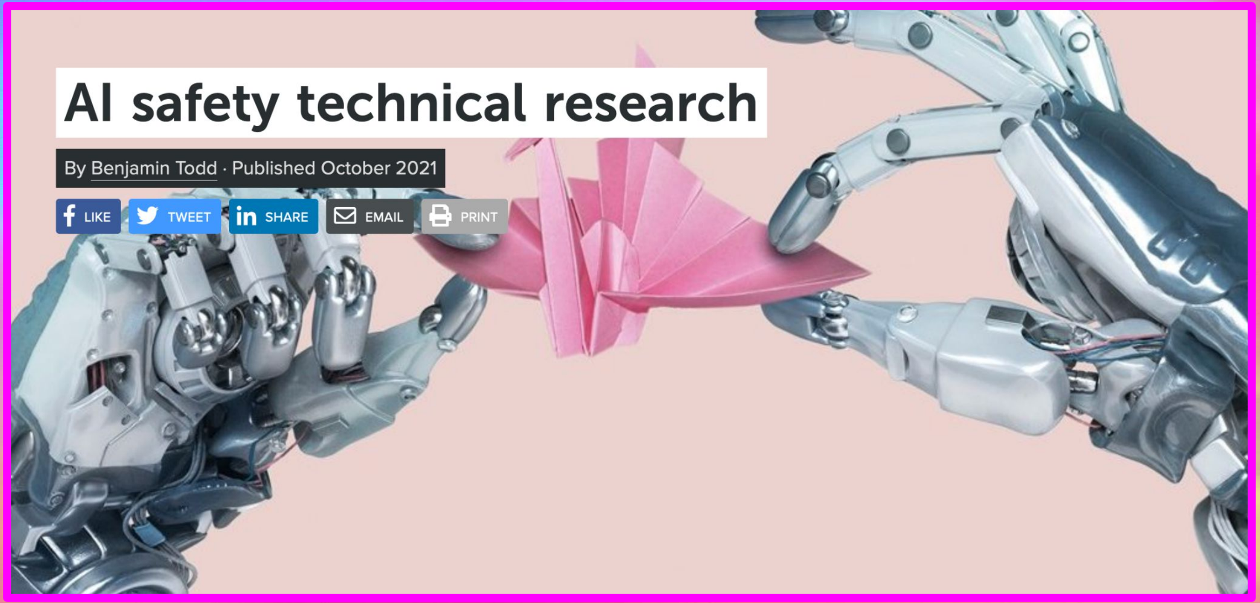

"The most important book I have read in quite some time."
—Daniel Kahneman, author of THINKING, FAST AND SLOW
Human Compatible
ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL
Stuart Russell



"BASED ON EVIDENCE AND GOOD SENSE, NOT PLATITUDES"
– STEVEN PINKER, NEW YORK TIMES BESTSELLING AUTHOR AND JOHNSTONE PROFESSOR OF PSYCHOLOGY AT HARVARD UNIVERSITY
80,000 HOURS
FIND A FULFILLING CAREER THAT DOES GOOD
BENJAMIN TODD
AND THE 80,000 HOURS TEAM

Career guidance

https://80000hours.org/career-reviews/ai-safety-researcher

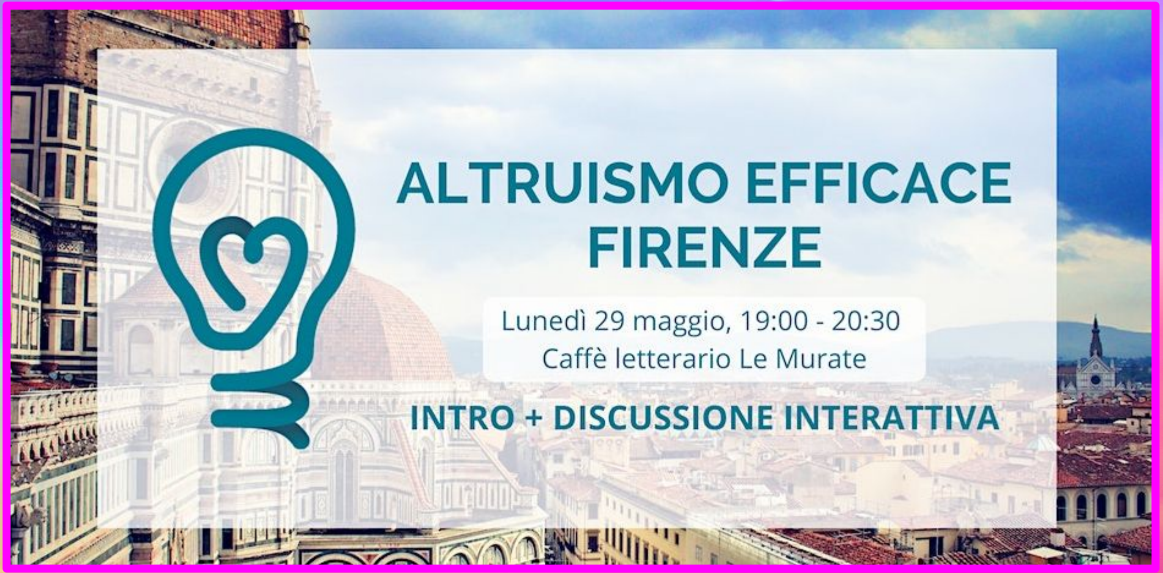


AI safety technical research

By Benjamin Todd · Published October 2021

Events in Italy

https://altruismoefficace.it/eventi



ALTRUISMO EFFICACE
FIRENZE

Lunedì 29 maggio, 19:00 - 20:30
Caffè letterario Le Murate

INTRO + DISCUSSIONE INTERATTIVA

Take a big breath before `slides[-2]`

In memory of nonna Maria

Stefania Delprete @astrastefania

Live long and prosper

Stefania Delprete
astrastefania@gmail.com

LinkedIn

More resources at https://github.com/astrastefania/ai-safety