Trustworthy AI Autonomy

M-0: Overview

Ding Zhao

Assistant Professor

Carnegie Mellon University

Carnegie Mellon University

Safe AI Lab @CMU

Good reasons to drop the course (to save ur time)

- I am not into maths/statistics. I mainly want to learn how to **implement** AI
  - This course aims to learn the fundamental and the advanced. Both have has a lot of boring maths.
- I want to take a well-polished mature course
  - I created this course last year and fully redesigned it now. Not mature at all. Neither the field.
- I want to have a slow-paced course. I want you to derive equations on the board.
  - This is an extremely fast-paced field/industry. So is this course.
- I do not have machine learning background
  - I assume you have it (details on the syllabus). If not, you may suffer, get lost, or lose interests to learn more.
- I have not used object-oriented programing language before. Can I use Matlab?
  - Projects are written in Python, starting on day one.
- I want an easy course. I want an in-person course …

Reasons to stay (short version)

- I am a PhD in the direction of Trustworthy AI
- I want to be a PhD
- I want to take a research positions in the industry
- I want to have a publication
- I have a lot of time to work on this course (< no more than two more heavy courses/research tasks)
- I want to be a leader in safety-critical applications e.g. **self-driving cars**, healthcare, assistant robotics, etc.

Plan for today

- Logistics
- Overview
  - What is Trustworthy AI AuTonomy (TAIAT)?
  - Why now?
  - Who cares?
  - How to learn Trustworthy AI?
  - Summary/reading materials

Online live lectures

- Live lectures on zoom, as interactive as possible
  - Ask questions!
    - By raising your hand
    - By entering the question in chat
- Participating the **live lectures** is critical as we have many in-class discussions/ presentations. Please slack me if you may have difficulty to do so.
- Join on time: I wrote a script to check your Zoom participation
- Open the camera if possible: Especially if you would like a recommendation letter from me

Logistics

- Everything is on Canvas/Syllabus. Read it carefully.

Assessment

- 55%: projects
  - 10%: Project 1 - robustness of deep learning
  - 10%: Project 2 - model-based reinforcement learning
  - 35%: Challenge
    - 5%: Proposal and proposal presentation
    - 5%: Milestone review
    - 10%: Presentation/Expo
    - 15%: Final report
- 35%: Paper review
  - 5%: Long review 1
  - 5%: Long review 2
  - 5%: Long review 3
  - 20%: Long review 1 presentation
- Zoom participation: 10%

- Emphasize on education rather than selection.

Supports

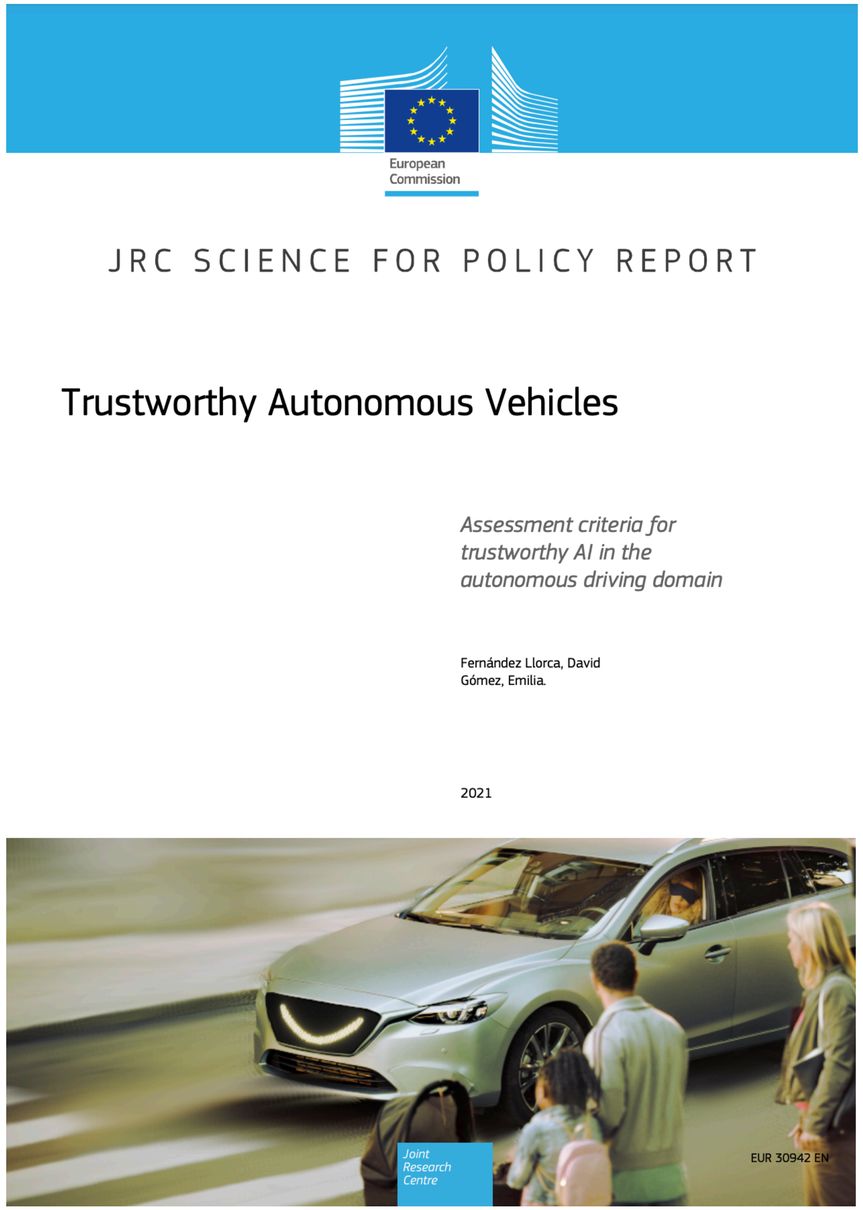
- Office hours every week day
- Campuswire hotline: get answers in 48 hours
- Project/challenge support camp (every Friday)

What is trustworthy AI?

An example with autonomous vehicles

What is trustworthy AI?

- The meaning of Trustworthy AI can be derived from the report of the AI HLEG on Ethical Guidelines for Trustworthy AI. Trustworthy should not be interpreted here in its literal sense, but rather as a comprehensive framework that includes multiple principles, requirements and criteria. Trustworthy AI systems are human-centric and rest on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom. Trustworthiness is conceived as a mean to maximise the benefits of AI systems while at the same time preventing and minimising their risks (AI HLEG, 2019)



Fernández Llorca D, Gómez E. Trustworthy Autonomous Vehicles. Joint Research Centre (Seville site); 2021 Dec.

What is trustworthy AI?

- Thus, trustworthy AI systems are grounded on three main components. First, they have to be lawful, complying with all applicable laws and regulations. Second, they have to be ethical, ensuring adherence to fundamental principles and values. And third, they have to be **robust** (in the general sense), both from a **technical** and social perspective

Table 2: Key Requirements for a Trustworthy AI.

Code	Requirements
KR1	Human agency and oversight
KR2	Technical robustness and safety
KR3	Privacy and data governance
KR4	Transparency
KR5	Diversity, non-discrimination and fairness
KR6	Societal and environmental wellbeing
KR7	Accountability

Fernández Llorca D, Gómez E. Trustworthy Autonomous Vehicles. Joint Research Centre (Seville site); 2021 Dec.

Key requirements (KR)

Req.	Code	Criteria
KR1		**Human agency and autonomy**
	CR1.1	Affects humans or society.
	CR1.2	Confusion as to whether the interaction is with a human or an AI
	CR1.3	Overreliance
	CR1.4	Unintended and undesirable interference with end-user decision-making
	CR1.5	Simulation of social interaction
	CR1.6	Risk of attachment, addiction and user behaviour manipulation
		Human oversight
	CR1.7	Self-learning or autonomous / Human-in-the-Loop / Human-on-the-Loop / Human-in-Command
	CR1.8	Training on how to exercise oversight
	CR1.9	Detection and response mechanisms for undesirable adverse effects
	CR1.10	Stop button
	CR1.11	Oversight and control of the self-learning or autonomous nature of the AI system

Req.	Code	Criteria
KR3		**Privacy**
	CR3.1	Privacy, integrity and data protection
	CR3.2	Flagging privacy issues
		Data Governance
	CR3.3	Use of personal data
	CR3.4	General Data Protection Regulation (GDPR) measures
	CR3.5	Privacy and data protection implications for Non-personal data
	CR3.6	Alignment with standards and protocols for data management and governance

Req.	Code	Criteria
KR2		**Resilience to Attack and Security**
	CR2.1	Effects on human safety due to faults, defects, outages, attacks, misuse, inappropriate or malicious use.
	CR2.2	Confusion as to whether the interaction is with a human or an AI
	CR2.3	Cybersecurity certification and security standards
	CR2.4	Exposure to cyberattacks
	CR2.5	Integrity, robustness and security against attacks
	CR2.6	Red teaming / Penetration testing
	CR2.7	Security coverage and updates
		General Safety
	CR2.8	Risks, risk metrics and risk levels
	CR2.9	Design faults, technical faults, environmental threats
	CR2.10	Stable and reliable behaviour
	CR2.11	Fault tolerance
	CR2.12	Review of technical robustness and safety
		Accuracy
	CR2.13	Consequences of low level accuracy
	CR2.14	Quality of the training data (up-to-date, complete and representative)
	CR2.15	Monitoring and documentation of system accuracy
	CR2.16	Invalid training data and assumptions
	CR2.17	Communication of expected level of accuracy
		Reliability, Fall-back plans and Reproducibility
	CR2.18	Consequences of low reliability and reproducibility
	CR2.19	Verification and validation of reliability and reproducibility
	CR2.20	Tested failsafe fallback plans
	CR2.21	Handling low confidence scores
	CR2.22	Online continuous learning

Fernández Llorca D, Gómez E. Trustworthy Autonomous Vehicles. Joint Research Centre (Seville site); 2021 Dec.

Key requirements (KR)

Req.	Code	Criteria
KR4		**Traceability**
	CR4.1	Measures to address traceability
		Explainability
	CR4.2	Explaining decisions to the users
	CR4.3	Continuous survey on users' understanding of decisions
		Communication
	CR4.4	Communicating the users that they are interacting with an AI system
	CR4.5	Inform users about purpose, criteria and limitations

Req.	Code	Criteria
KR5		**Avoidance of Unfair Bias**
	CR5.1	Unfair bias on data or algorithm design
	CR5.2	Diversity of end-users in the data
	CR5.3	Educational and awareness initiatives to avoid injecting bias
	CR5.4	Flagging bias, discrimination or poor performance issues
	CR5.5	Appropriate fairness definition
		Accessibility and Universal Design
	CR5.6	Correspondence to variety of preferences and abilities in society
	CR5.7	Usability of the user interface by people with special needs
	CR5.8	Universal design principles
	CR5.9	Impact on end-users
		Stakeholder Participation
	CR5.10	Stakeholder participation in the design and development

Req.	Code	Criteria
KR6		**Environmental Well-being**
	CR6.1	Negative impacts on the environment
	CR6.2	Environmental impact evaluation (development, deployment and use)
		Impact on Work and Skills
	CR6.3	Impacts on human work
	CR6.4	Consideration of impacted workers and their representatives
	CR6.5	Measures to ensure understanding of the impact on human work
	CR6.6	Risk of de-skilling
	CR6.7	Need of new (digital) skills
		Impact on Society at large or Democracy
	CR6.8	Negative impact on society at large or democracy

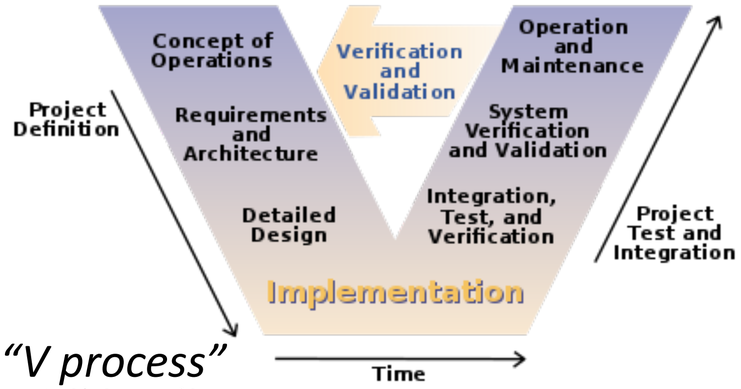
Req.	Code	Criteria
KR7		**Auditability**
	CR7.1	Auditability mechanisms
	CR7.2	Third parties audit
		Risk Management
	CR7.3	External guidance to oversee ethical concerns and accountability measures
	CR7.4	Risk training and applicable legal framework
	CR7.5	Ethics review board
	CR7.6	Adherence to the ALTAI ([1])
	CR7.7	Third party process to report vulnerabilities, risks or biases
	CR7.8	Redress by design

Fernández Llorca D, Gómez E. Trustworthy Autonomous Vehicles. Joint Research Centre (Seville site); 2021 Dec.

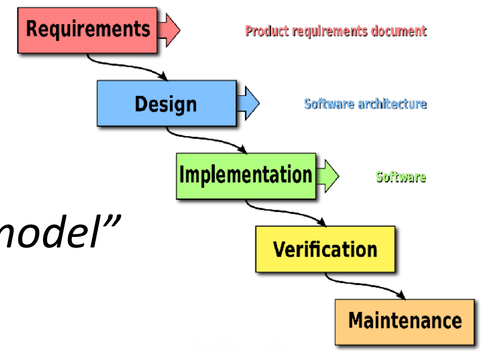
Why now?

And why I care about trustworthy AI autonomy

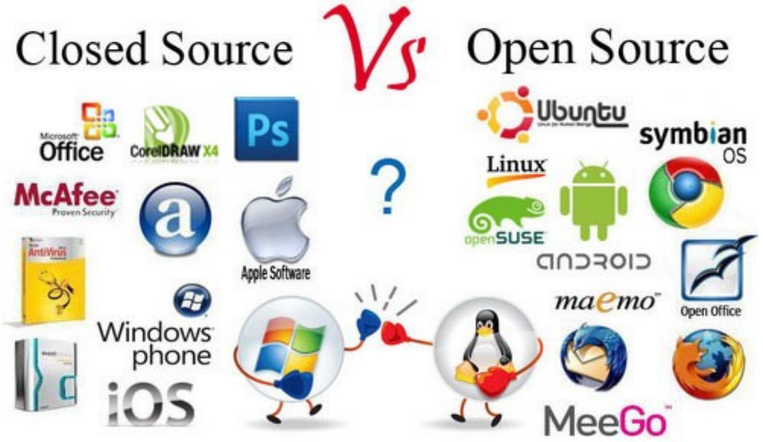
We are on the cusp to revolute the way to make machines



"V process" Wikipedia

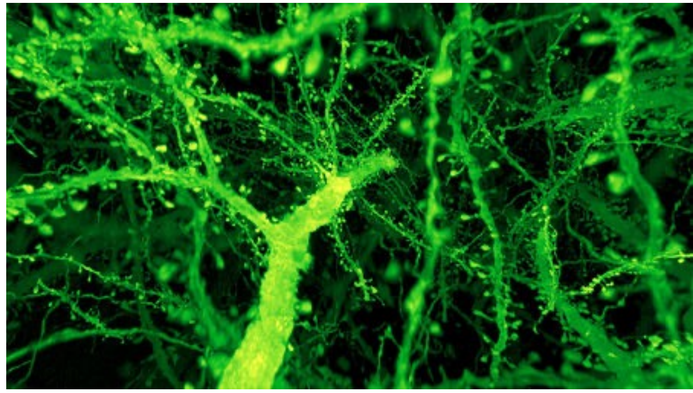


"Waterfall model"



Connected

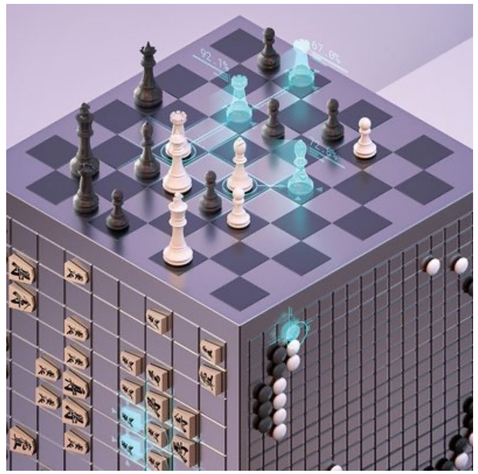
By complex structures



Neural Network
[Science, 2019]

Evolving

In a self-supervised way



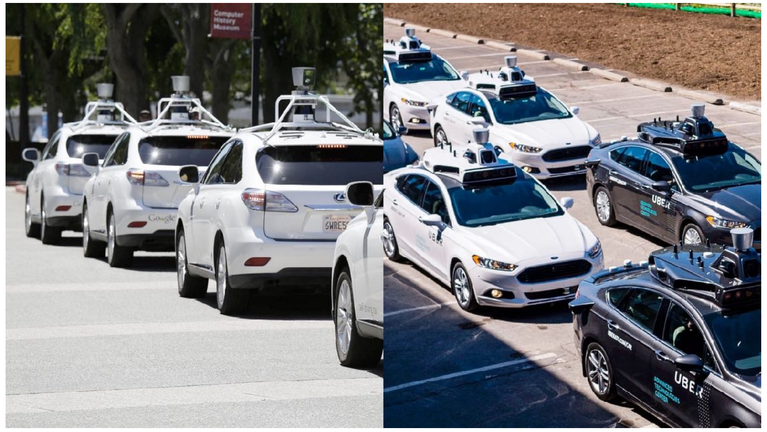
Reinforcement Learning
[Science, 2018]

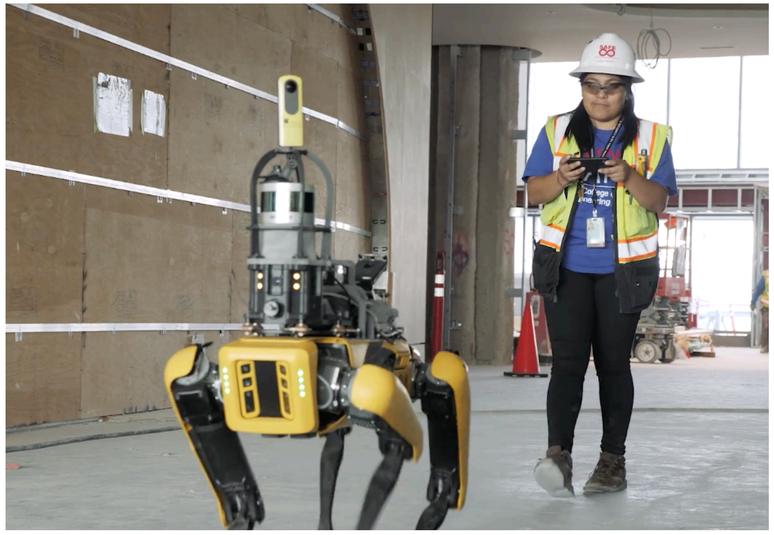
Sharing

With blackboxes and uncertainty

Open Code/data
[Science, 2017]

AI autonomy





Tesla
Fatal Crash, May, 2016

Waymo
Accident, May, 2018

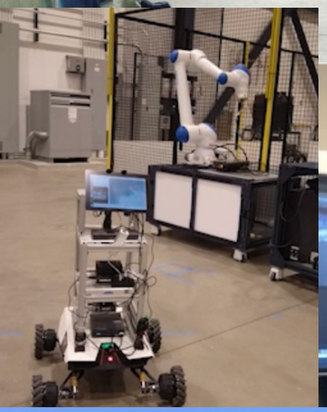
Uber
Kill a pedestrian, March, 2018

Things can go very wrong
... even for the best players

How to design and assess trustworthy AI autonomy?

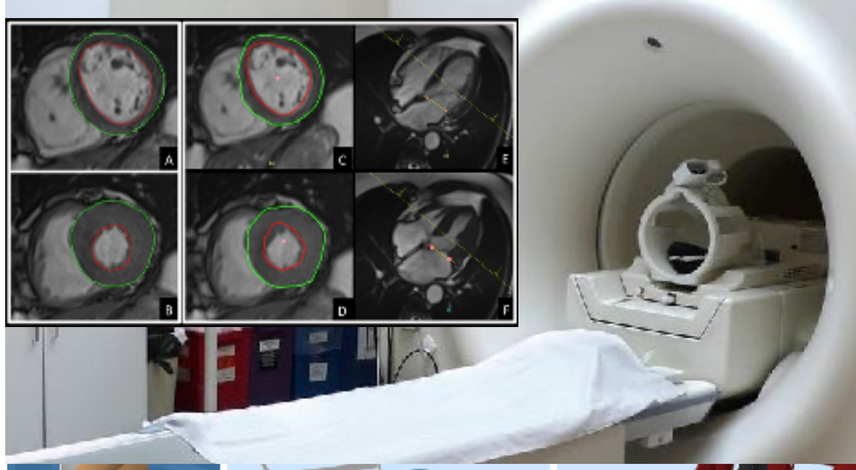
“ To develop **trustworthy** (robust, safe, generalizable, explainable, certifiable, and human centric) **AI** in the face of the uncertain, dynamic, multi-agent, and human-involved environment by bridging rigorous theories and practical technologies.

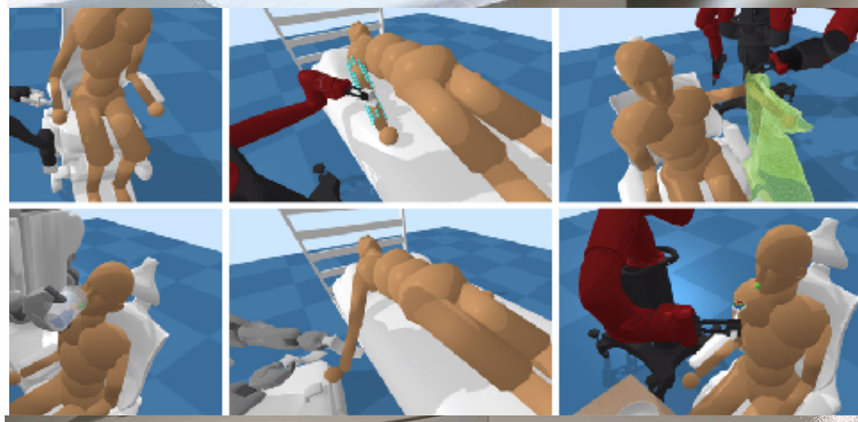
- Mission of *SafeAI Lab* @ CMU

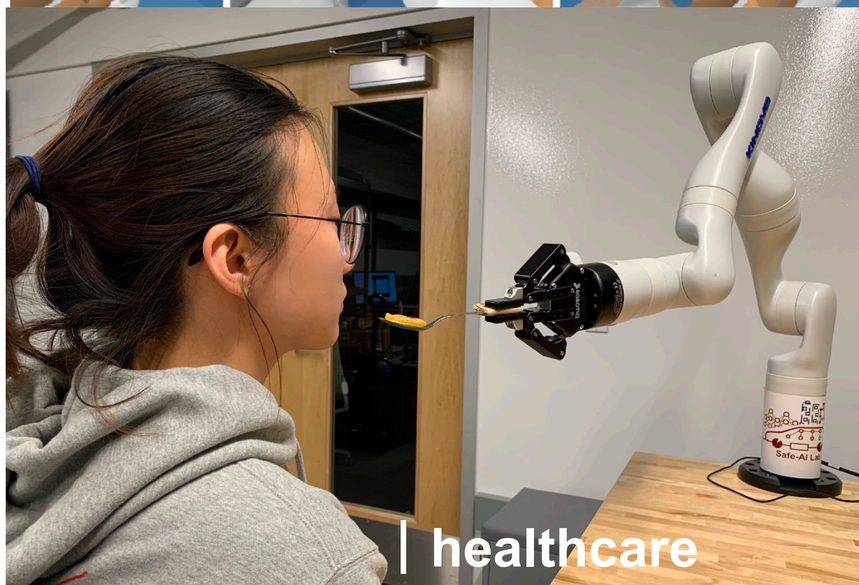


| Mobility

Safe AI Lab @CMU



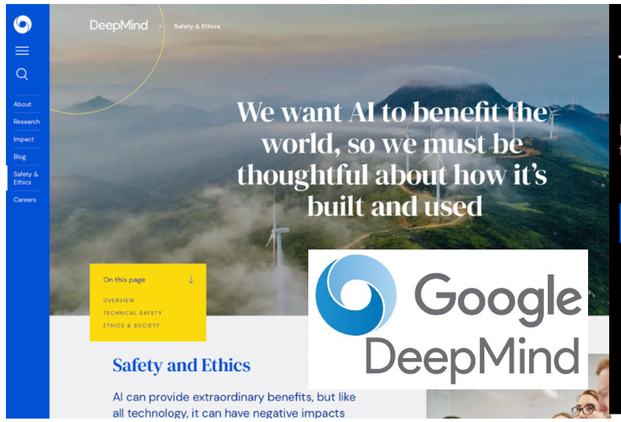


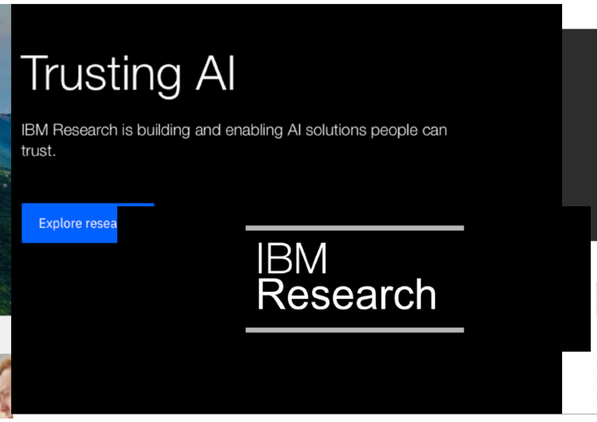


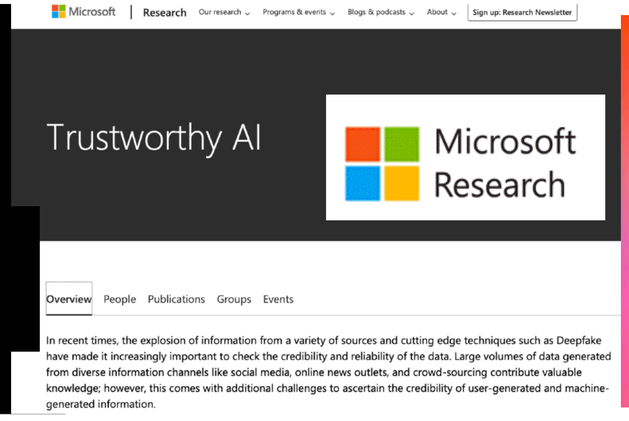
| healthcare

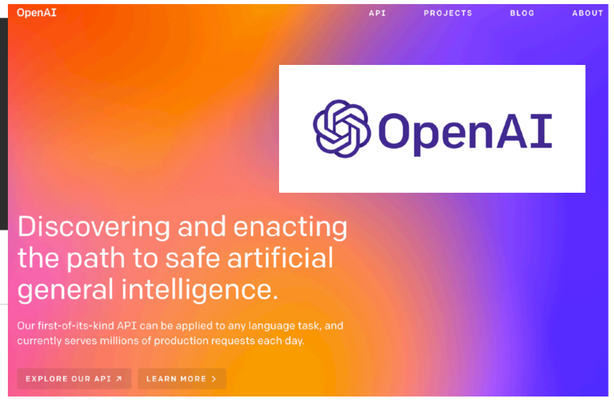
Who else cares?

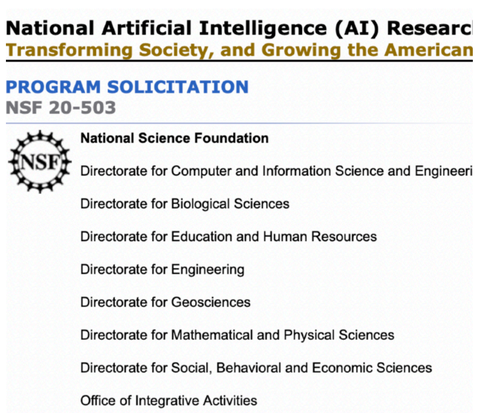
Who else cares? Like everybody?



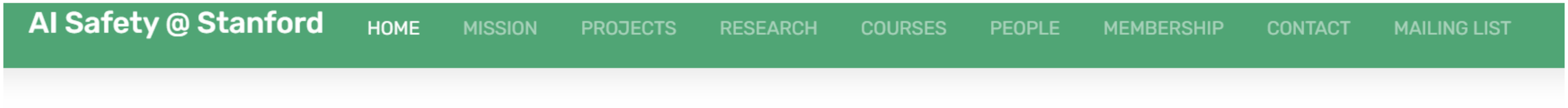


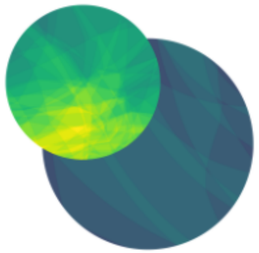






Why TAIAT? - Academia





Stanford University

The mission of the Stanford Center for AI Safety is to develop rigorous techniques for building safe and trustworthy AI systems and establishing confidence in their behavior and robustness, thereby facilitating their successful adoption in society.

Read more in the white paper

Why TAIAT? - Academia

Stanford Center for AI Safety

Clark Barrett, David L. Dill,
Mykel J. Kochenderfer, Dorsa Sadigh

1 Introduction

Software-based systems play important roles in many areas of modern life, including manufacturing, transportation, aerospace, and healthcare. However, developing these complex systems, which are expected to be smart and reliable, is difficult, expensive, and error-prone. A key reason for this difficulty is that the sheer complexity of many systems keeps growing, making it increasingly difficult for human minds to form a comprehensive picture of all relevant elements and behaviors of the system and its environment.

To mitigate this difficulty, research in the field of artificial intelligence (AI) has been promoting a different approach to programming. Instead of having a human engineer provide program logic for handling all possible inputs, al-

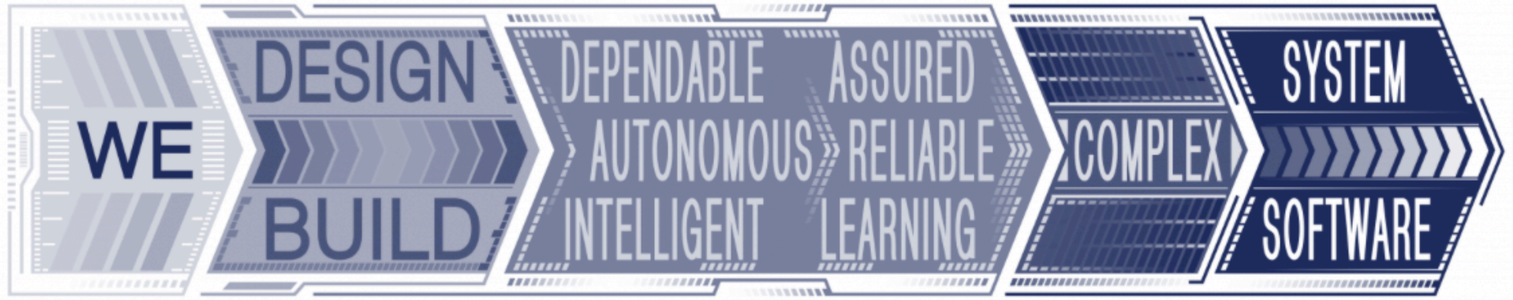
Stanford University

Why TAIAT? - Academia

PRECISE About Center for Safe AI Research Industry Education Events People

About

PRECISE Center for Safe AI



We are in the midst of a foundational shift. From self-driving cars and voice assistants to smart thermostats and recommendation engines, Artificial Intelligence (AI) and Machine Learning (ML) are becoming an integral part of our daily lives. The emergence of these technologies opens up countless opportunities to transform any industry and to revolutionize traditional ways of thinking, operating, and solving problems. But...

How do you know if you can trust AI?

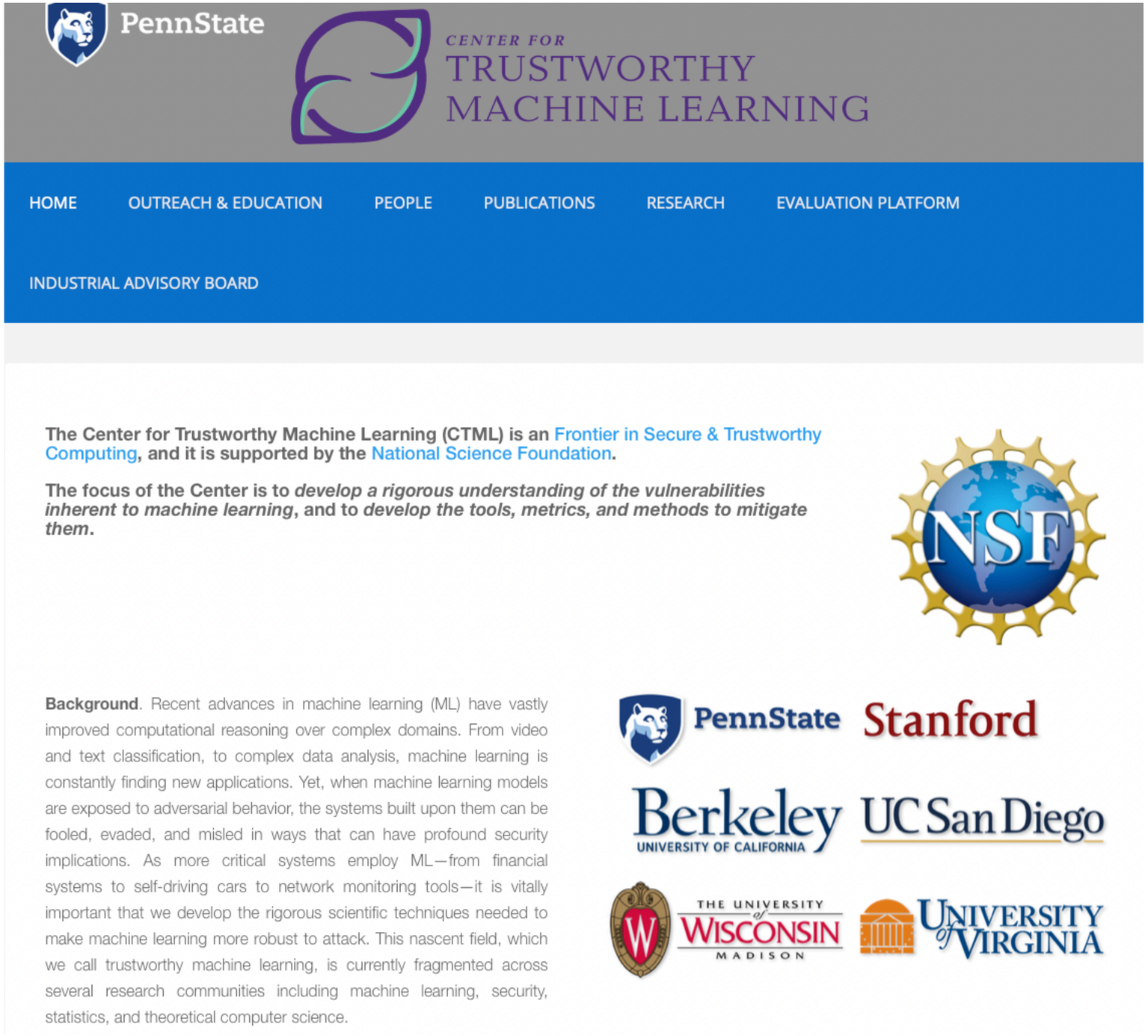
While AI+ML-based technologies have undoubtedly enhanced our daily lives, given that they require and rely on massive amounts of historical data to work effectively, they are by no means perfect. For many of these modern uses of AI+ML, 99% reliability may be sufficient. However, for mission-critical or life-dependent applications, 99% is not good enough. What about that 1% of uncertainty? Hospitals and doctors will remember the time that an autonomous system failed and someone died. Everyone remembers when a self-driving car or airplane crashes. It may not be often, but it does happen. That 1% is magnified when it is the difference between life and death.

Addressing this 1% of uncertainty is where PRECISE's Center for Safe AI is making significant inroads. We have a team of multidisciplinary experts developing highly-scalable tools and technologies that help companies and organizations verify safety in the edge cases of autonomous systems where failure is unacceptable. **The PRECISE Center for Safe AI is focused on working with existing AI designs and systems, making them safer, and providing formal verification of their safety.** We are building run-time monitoring for anomaly detection and taking a real-time systems approach to autonomy across multiple domains (e.g., healthcare, transportation, buildings, infrastructure). Through advanced automated techniques (using machine programming), we hope to accelerate the programmer productivity and software correctness, performance, and security.

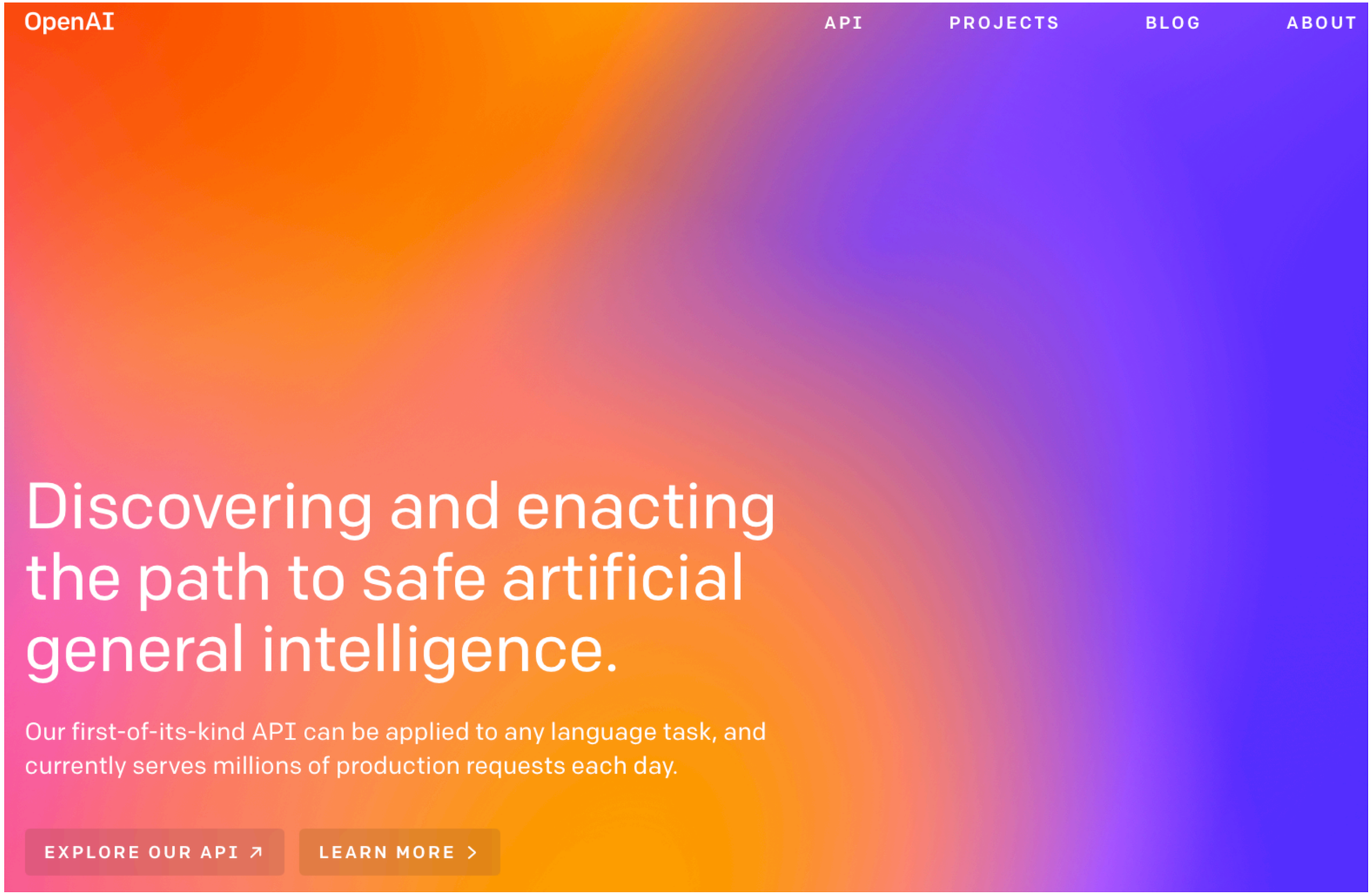
The Center's tools, technologies, and expertise are helping industries to answer the hard questions and giving them the full confidence in the safety of their autonomous systems in areas such as:

- Formally-verified ML models
- Robustness to adversarial attacks (via data and systems)
- Importance of simulation and its reliability

Why TAIAT? - Academia



Why TAIAT? - Industry



OpenAI

Why TAIAT? - Industry

Concrete problems in AI safety
D Amodei, C Olah, J Steinhardt, P Christiano… - arXiv preprint arXiv …, 2016 - arxiv.org
Rapid progress in machine learning and artificial intelligence (AI) has brought increasing attention to the potential impacts of AI technologies on society. In this paper we discuss one such potential impact: the problem of accidents in machine learning systems, defined as unintended and harmful behavior that may emerge from poor design of real-world AI systems. We present a list of five practical research problems related to accident risk, categorized according to whether the problem originates from having the wrong objective …
☆ Cited by 931 Related articles All 8 versions

Concrete Problems in AI Safety

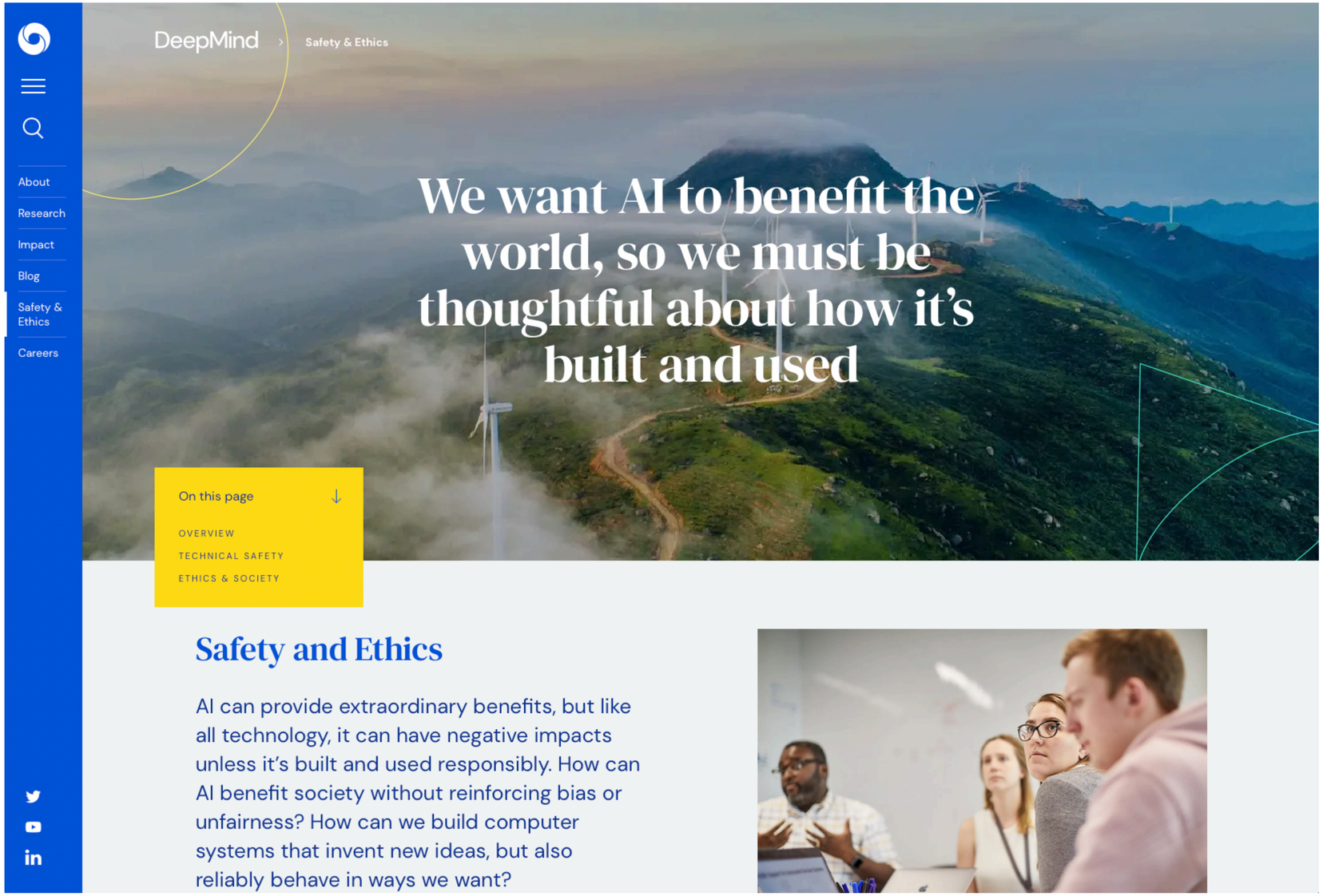
Dario Amodei* Google Brain
Chris Olah* Google Brain
Jacob Steinhardt Stanford University
Paul Christiano UC Berkeley
John Schulman OpenAI
Dan Mané Google Brain

Abstract

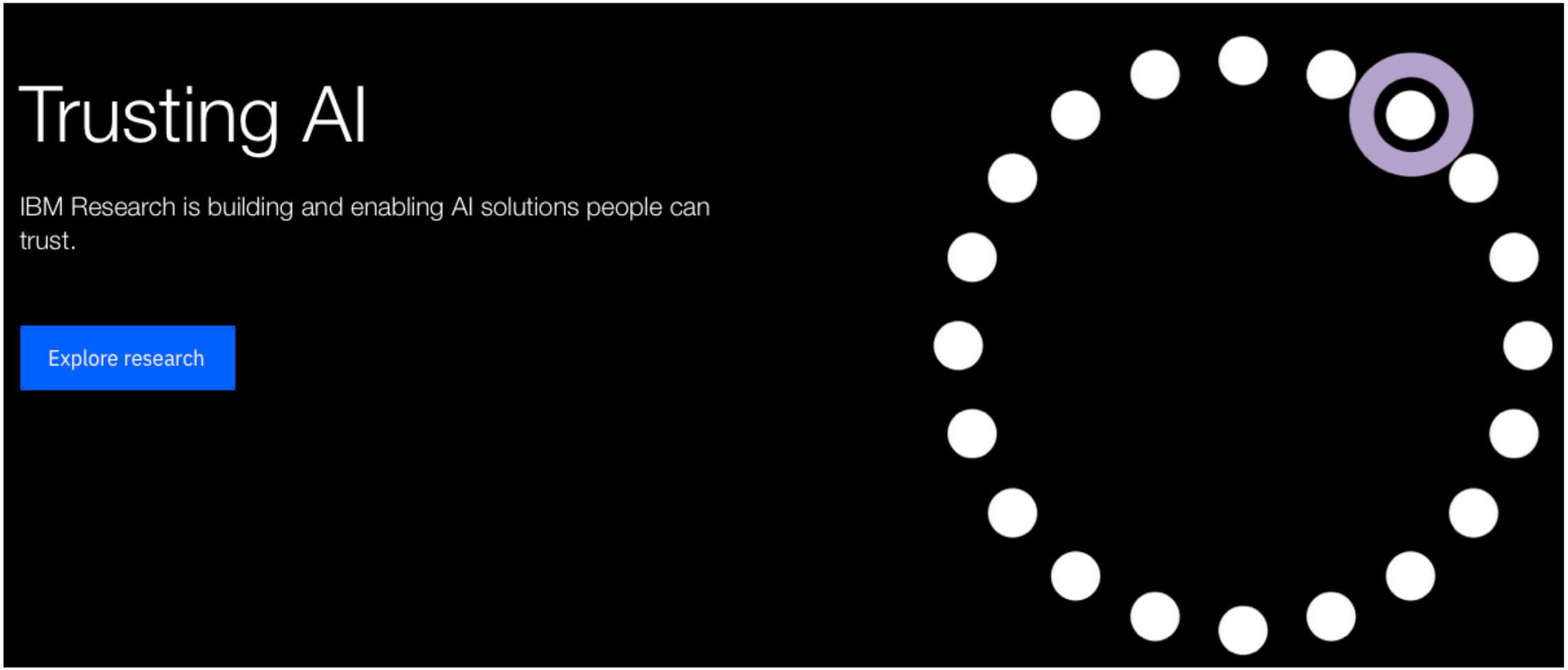
Rapid progress in machine learning and artificial intelligence (AI) has brought increasing attention to the potential impacts of AI technologies on society. In this paper we discuss one such potential impact: the problem of *accidents* in machine learning systems, defined as unintended and harmful behavior that may emerge from poor design of real-world AI systems. We present a list of five practical research problems related to accident risk, categorized according to whether the problem originates from having the wrong objective function ("avoiding side effects" and "avoiding reward hacking"), an objective function that is too expensive to evaluate frequently ("scalable supervision"), or undesirable behavior during the learning process ("safe exploration" and "distributional shift"). We review previous work in these areas as well as suggesting research directions with a focus on relevance to cutting-edge AI systems. Finally, we consider the high-level question of how to think most productively about the safety of forward-looking applications of AI.

OpenAI

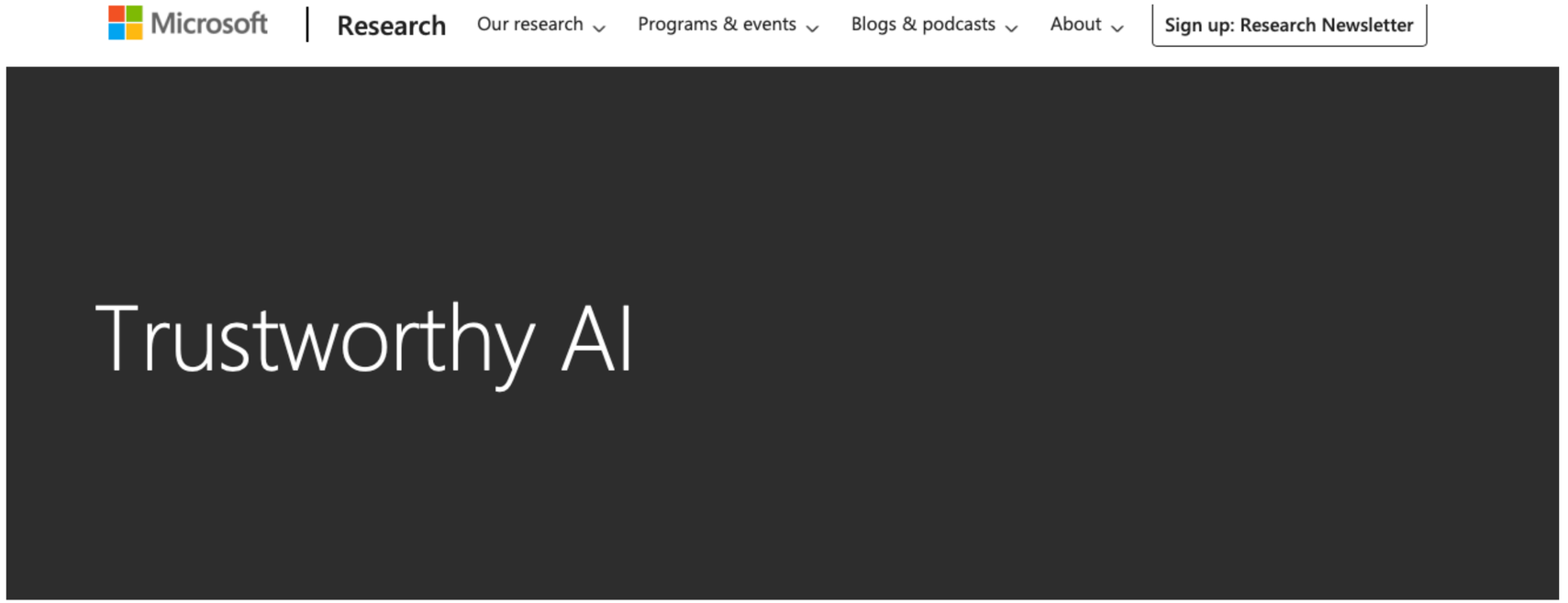
Why TAIAT? - Industry



Why TAIAT? - Industry



Why TAIAT? - Industry



Why TAIAT? - Industry

Home Who We are Tax Forms What we do Existential Risk Contact Donate

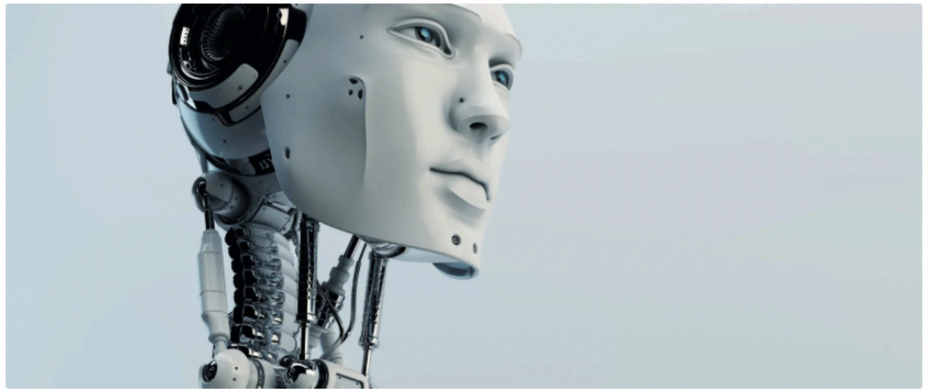
future of life INSTITUTE

Technology is giving life the potential to flourish like never before...

...or to self-destruct. Let's make a difference!

AI Biotech Nuclear Climate Podcasts

AI SAFETY RESEARCH



"What we really need to do is make sure that life continues into the future. [...] It's best to try to prevent a negative circumstance from occurring than to wait for it to occur and then be reactive."

-Elon Musk on keeping AI safe and beneficial

In spring of 2018, FLI launched our second AI Safety Research program, this time focusing on Artificial General Intelligence (AGI) and how to keep it safe and beneficial. By the summer, 10 researchers were awarded over $2 million to tackle the technical and strategic questions related to preparing for AGI, funded by generous donations from Elon Musk and the Berkeley Existential Risk Institute. You can read about their projects in the table below.

This research program comes as a sequel to our AI Safety grants competition in 2015, where generous donations from Elon Musk and the Open Philanthropy Project funded 37 researchers to begin various projects to help ensure that artificial intelligence remains safe and beneficial. Now, three years later, our grant winners have produced over 45 scientific publications and a host of conference events, which you can also read about below.

SUPPORT FLI

Donate directly using PayPal

Donate

Most benefits of civilization stem from intelligence, so how can we enhance these benefits with artificial intelligence without being replaced on the job market and perhaps altogether?

ABOUT ARTIFICIAL INTELLIGENCE

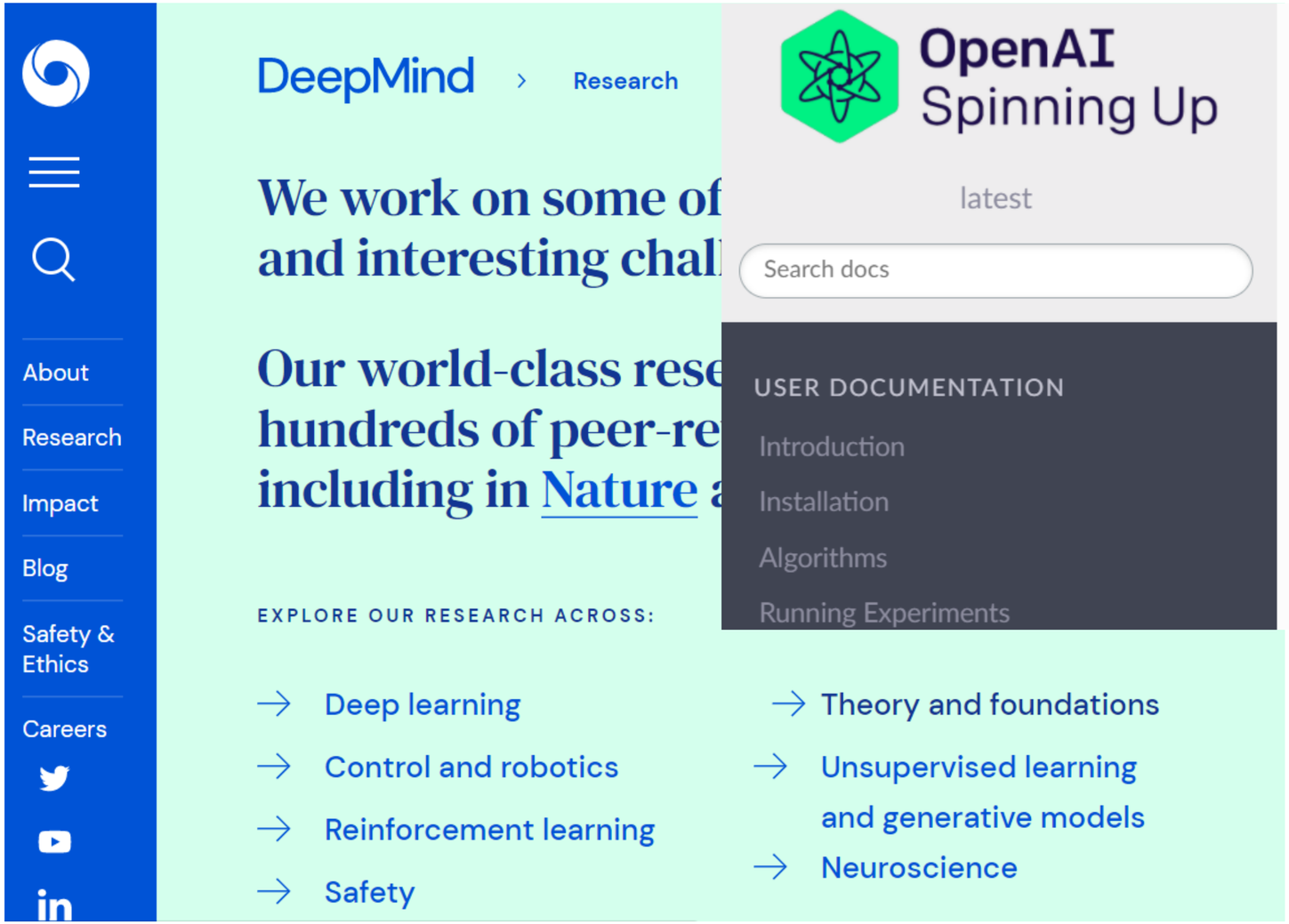
AI Safety Myths

Common myths about advanced AI distract from fascinating true controversies where even the experts disagree.

Benefits & Risks of Artificial Intelligence

From SIRI to self-driving cars, artificial

Why TAIAT? - Industry



- Key Papers in Deep RL
  - 1. Model-Free RL
  - 2. Exploration
  - 3. Transfer and Multitask RL
  - 4. Hierarchy
  - 5. Memory
  - 6. Model-Based RL
  - 7. Meta-RL
  - 8. Scaling RL
  - 9. RL in the Real World
  - 10. Safety
  - 11. Imitation Learning and Inverse Reinforcement Learning
  - 12. Reproducibility, Analysis, and Critique
  - 13. Bonus: Classic Papers in RL Theory or Review

Why TAIAT? - Government

National Artificial Intelligence (AI) Research Institutes Accelerating Research, Transforming Society, and Growing the American Workforce

PROGRAM SOLICITATION
NSF 20-503

National Science Foundation

Directorate for Computer and Information Science and Engineering

Directorate for Biological Sciences

Directorate for Education and Human Resources

Directorate for Engineering

Directorate for Geosciences

Directorate for Mathematical and Physical Sciences

Directorate for Social, Behavioral and Economic Sciences

Office of Integrative Activities

- Trustworthy AI;
- Foundations of Machine Learning;
- AI-Driven Innovation in Agriculture and the Food System;
- AI-Augmented Learning;
- AI for Accelerating Molecular Synthesis and Manufacturing; and
- AI for Discovery in Physics.

Challenges in TAIAT

- Safety critical applications
- Long tail problems/edge cases/imbalanced dataset/rare events
- Multi-modes/high-dimensional data modeling
- Exploration with cautiousness
- Make decision in evolving/new environment

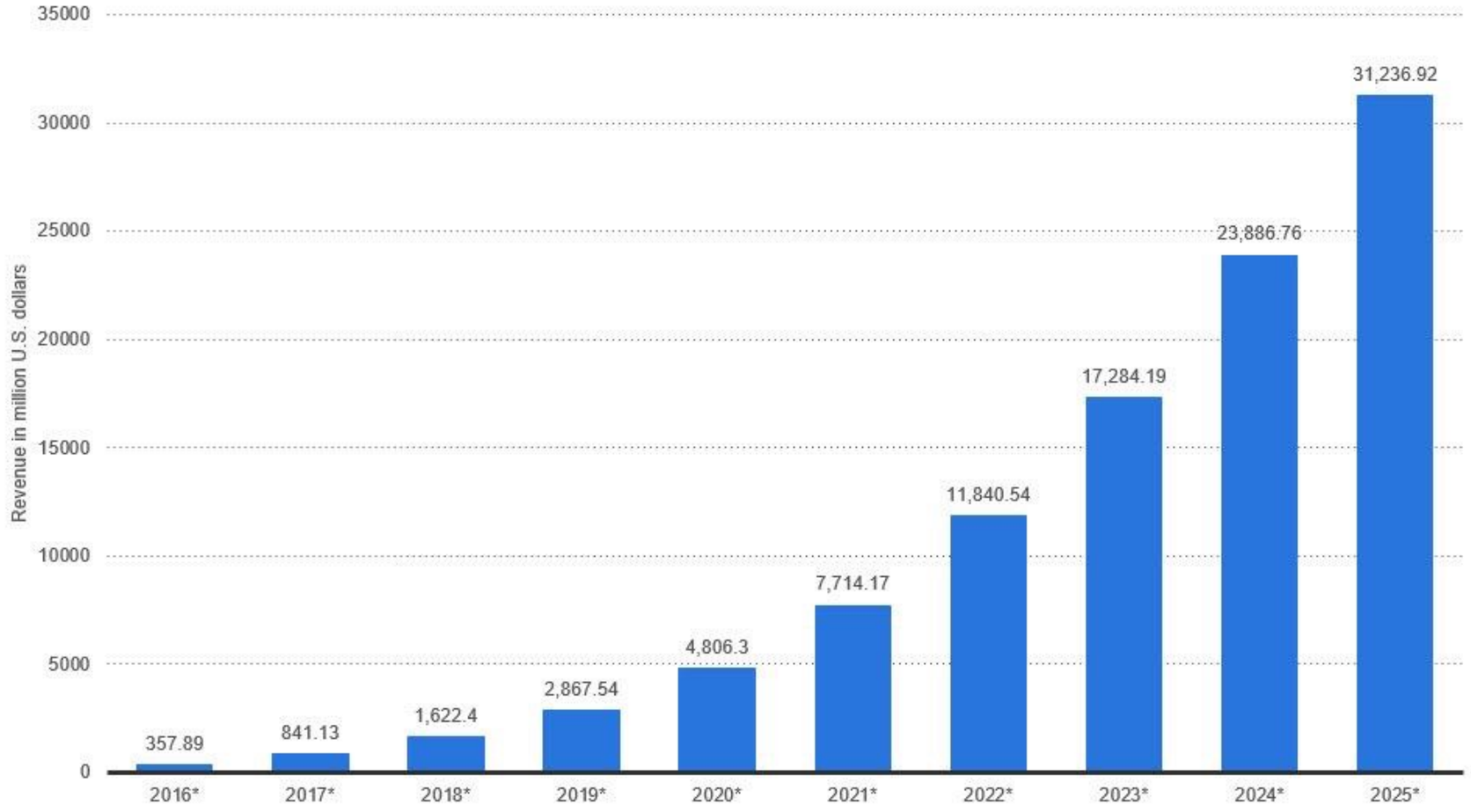
The biggest enemy of learning Trustworthy AI

- Lack of the big picture in an exciting and rapidly developing field of study

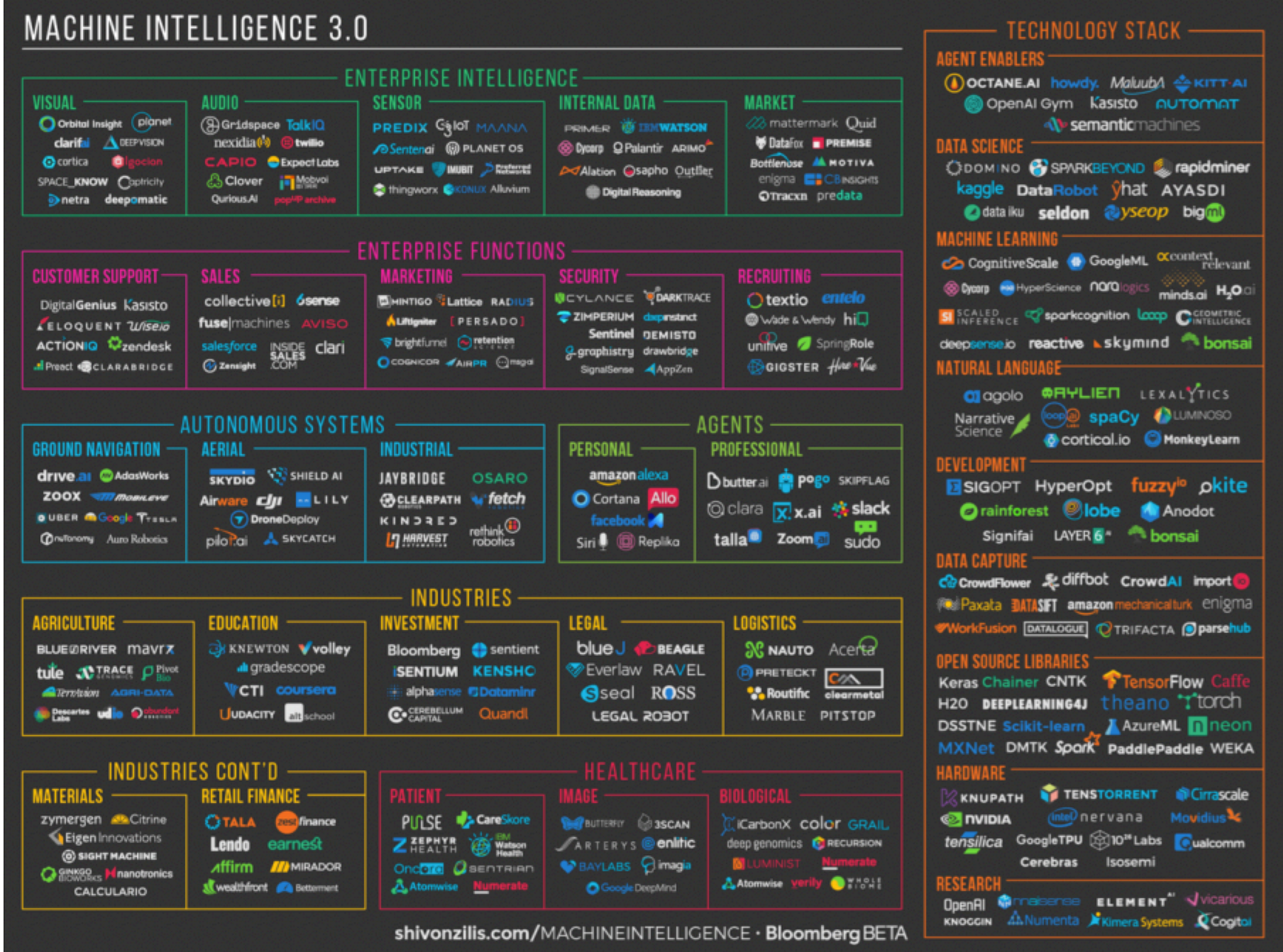
Money

Enterprise artificial intelligence market revenue worldwide 2016-2025

Revenues from the artificial intelligence for enterprise applications market worldwide, from 2016 to 2025 (in million U.S. dollars)

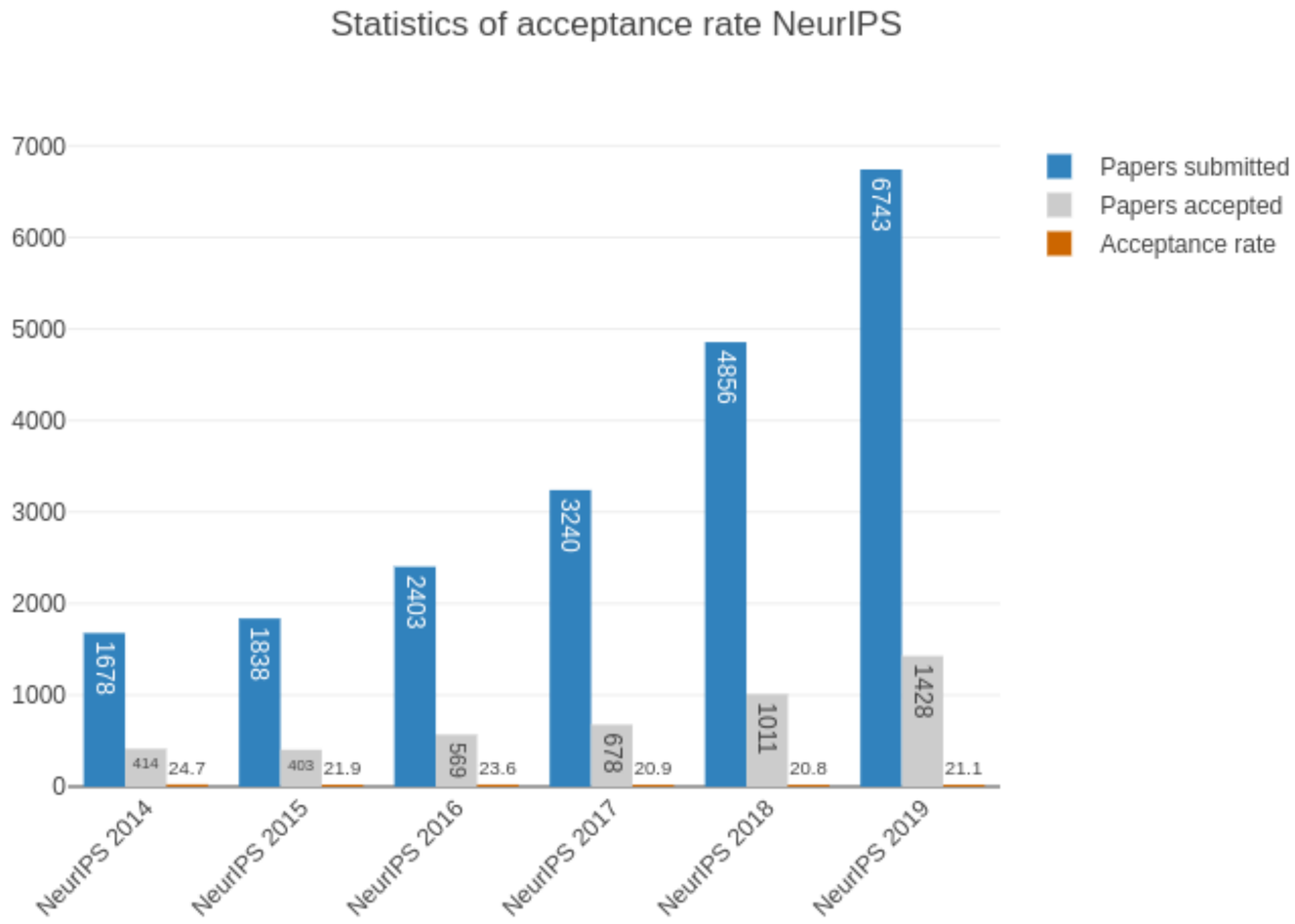


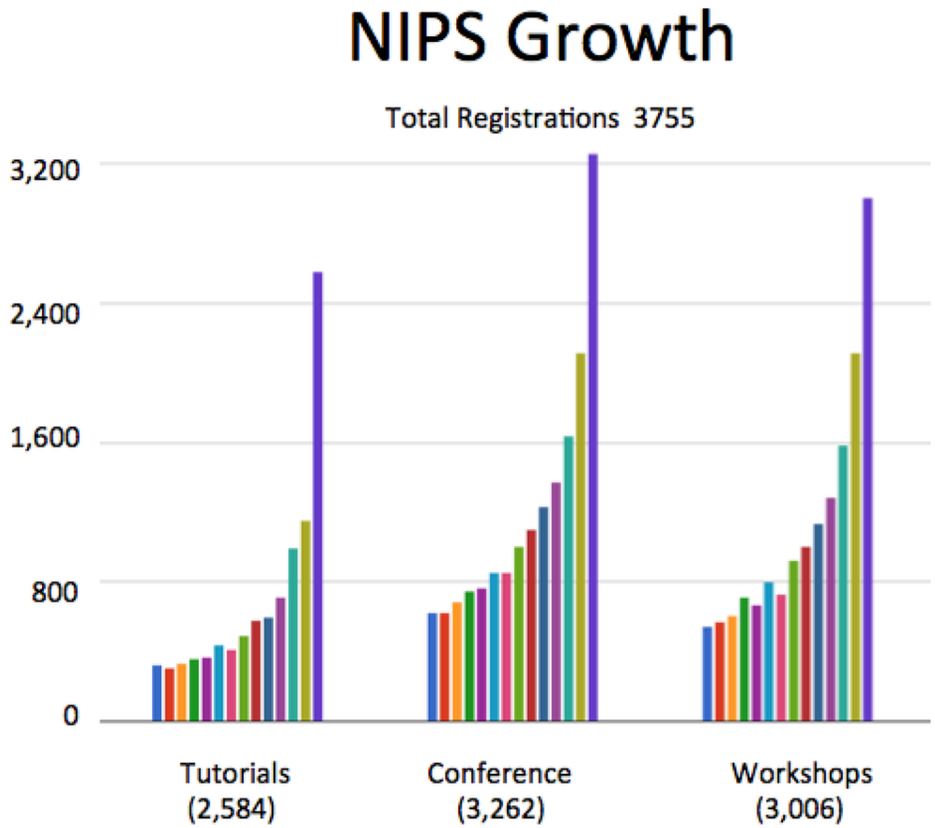
Job opportunities



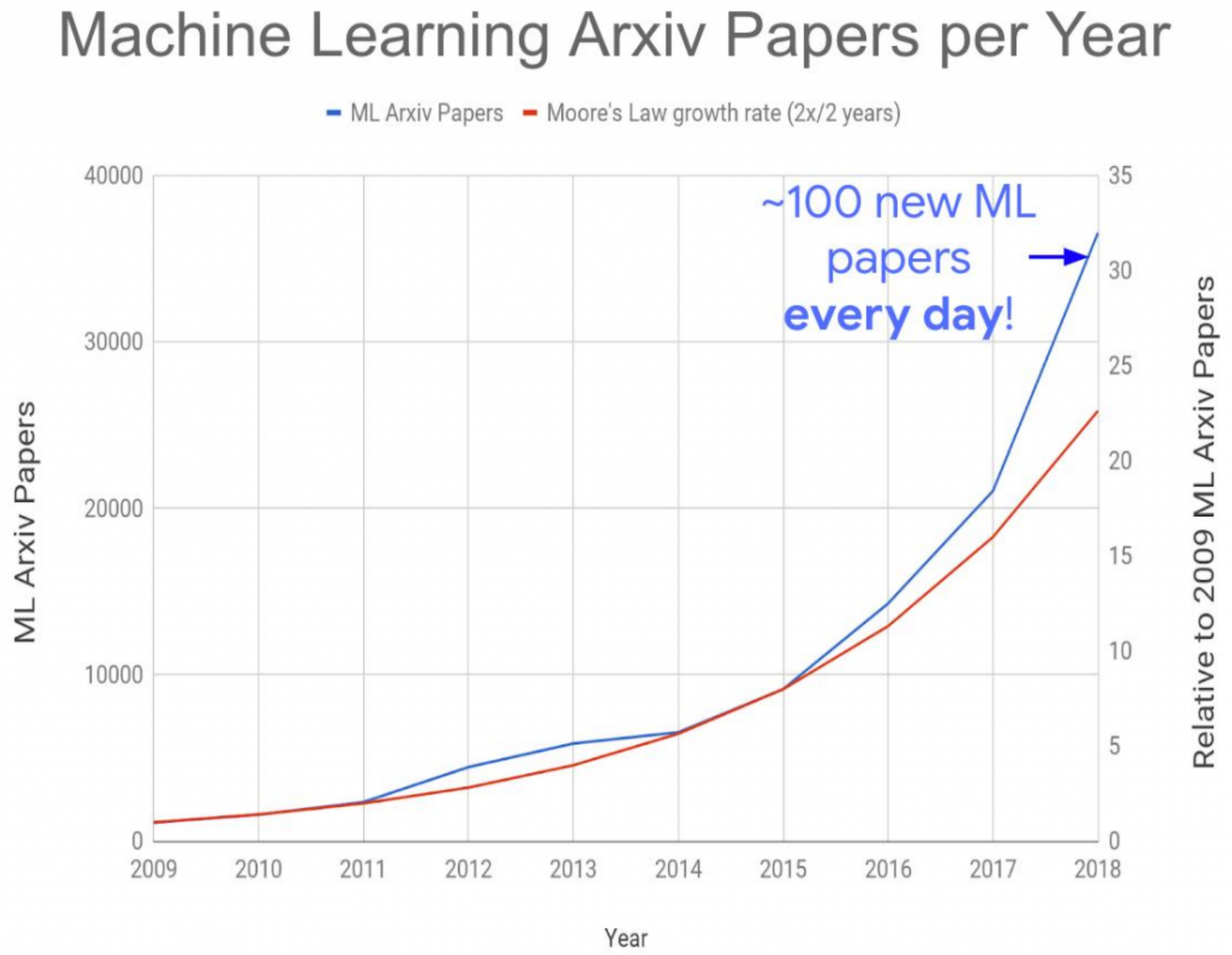
Ding Zhao | CMU

Exciting and new findings

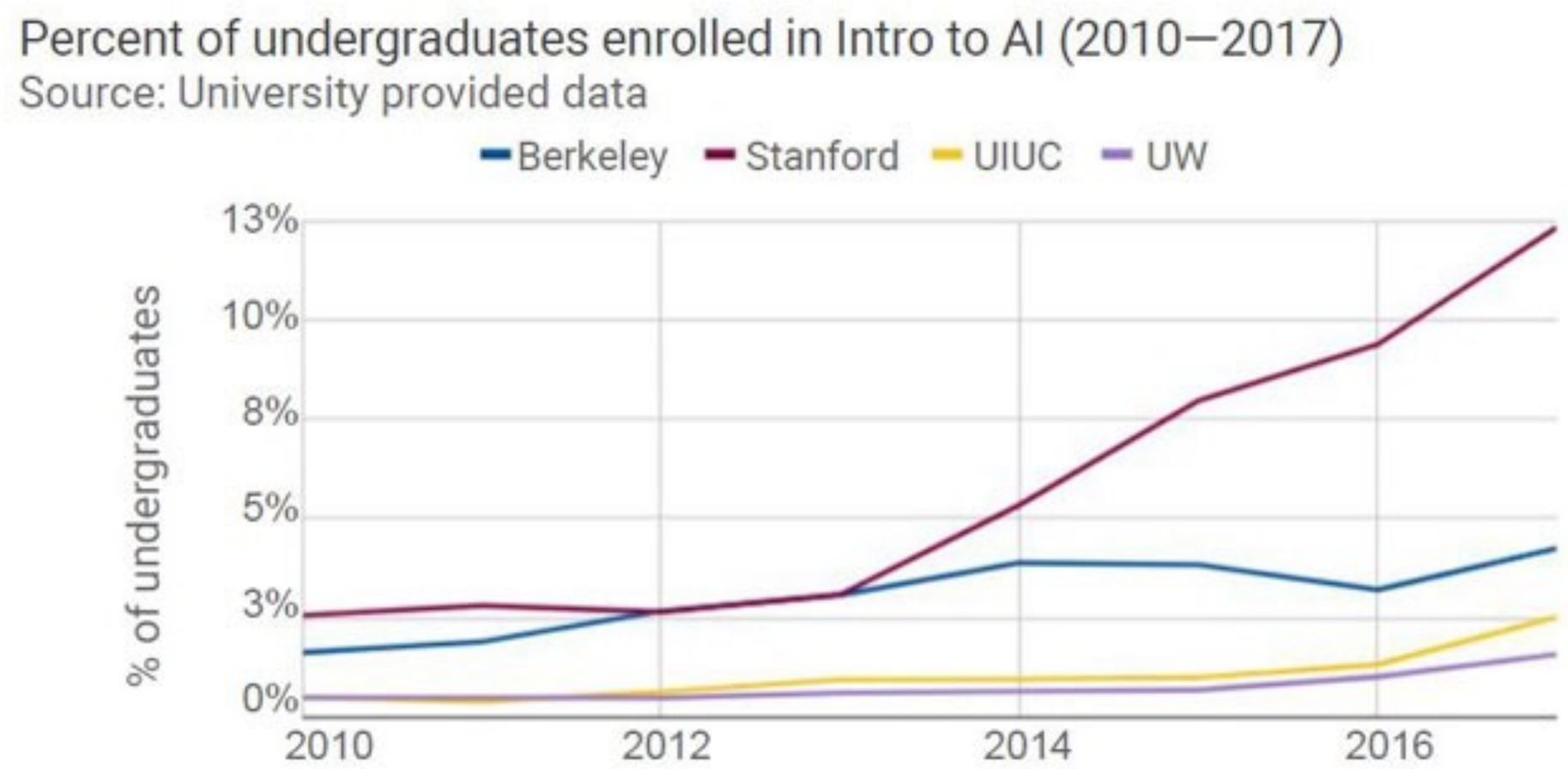




Many papers are on Arxiv



(Free) courses and online learning materials

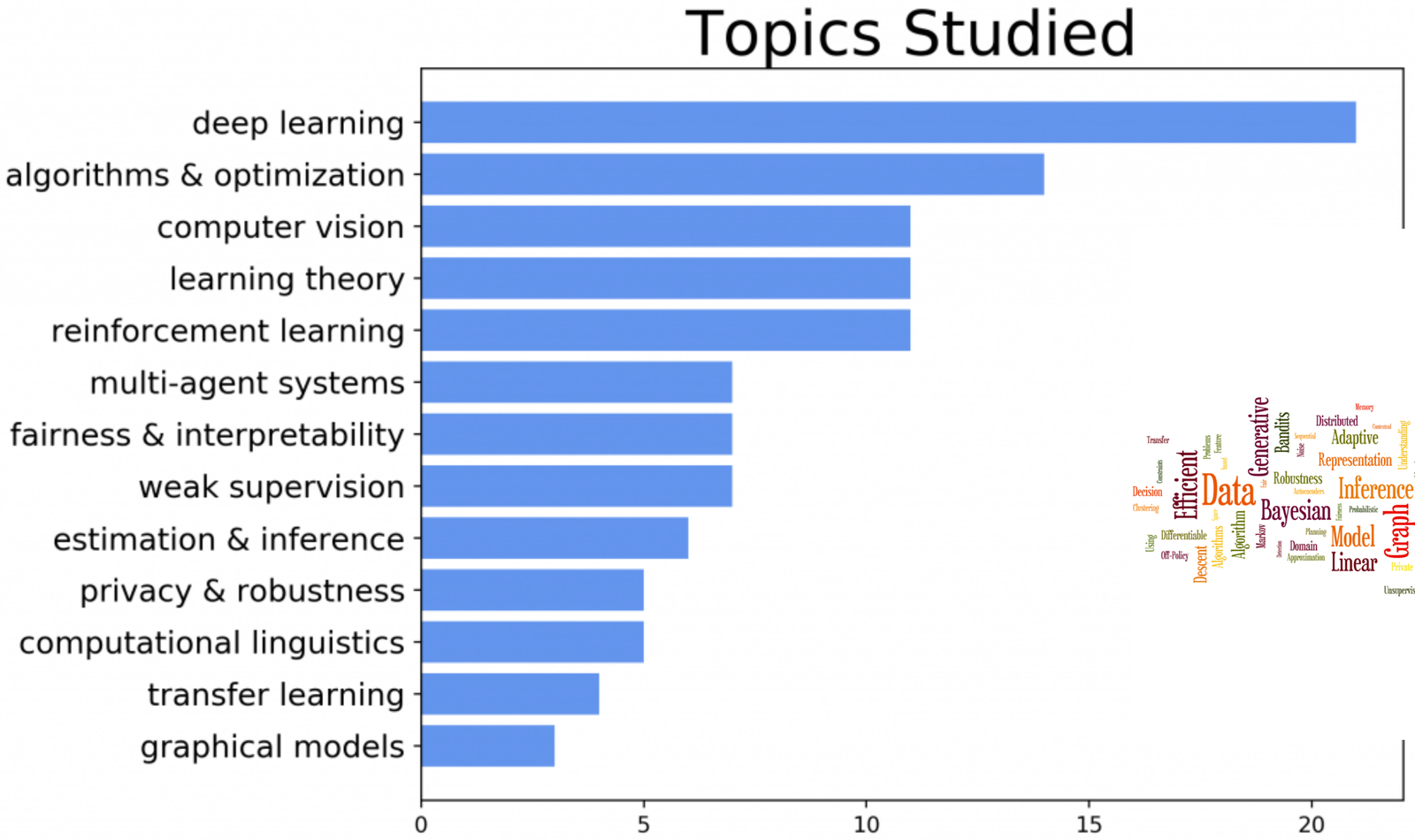


Word Cloud of Paper Titles at ICML 2020

Yet, a vast ocean of knowledge

https://chuvpilo.medium.com/whos-ahead-in-ai-research-in-2020-2009da5cd799

Moreover, topics are interconnected



https://blog.ml.cmu.edu/2020/12/04/carnegie-mellon-university-at-neurips-2020/

The biggest enemy of learning Trustworthy AI

- Lack of the big picture in an exciting and rapidly developing field of study

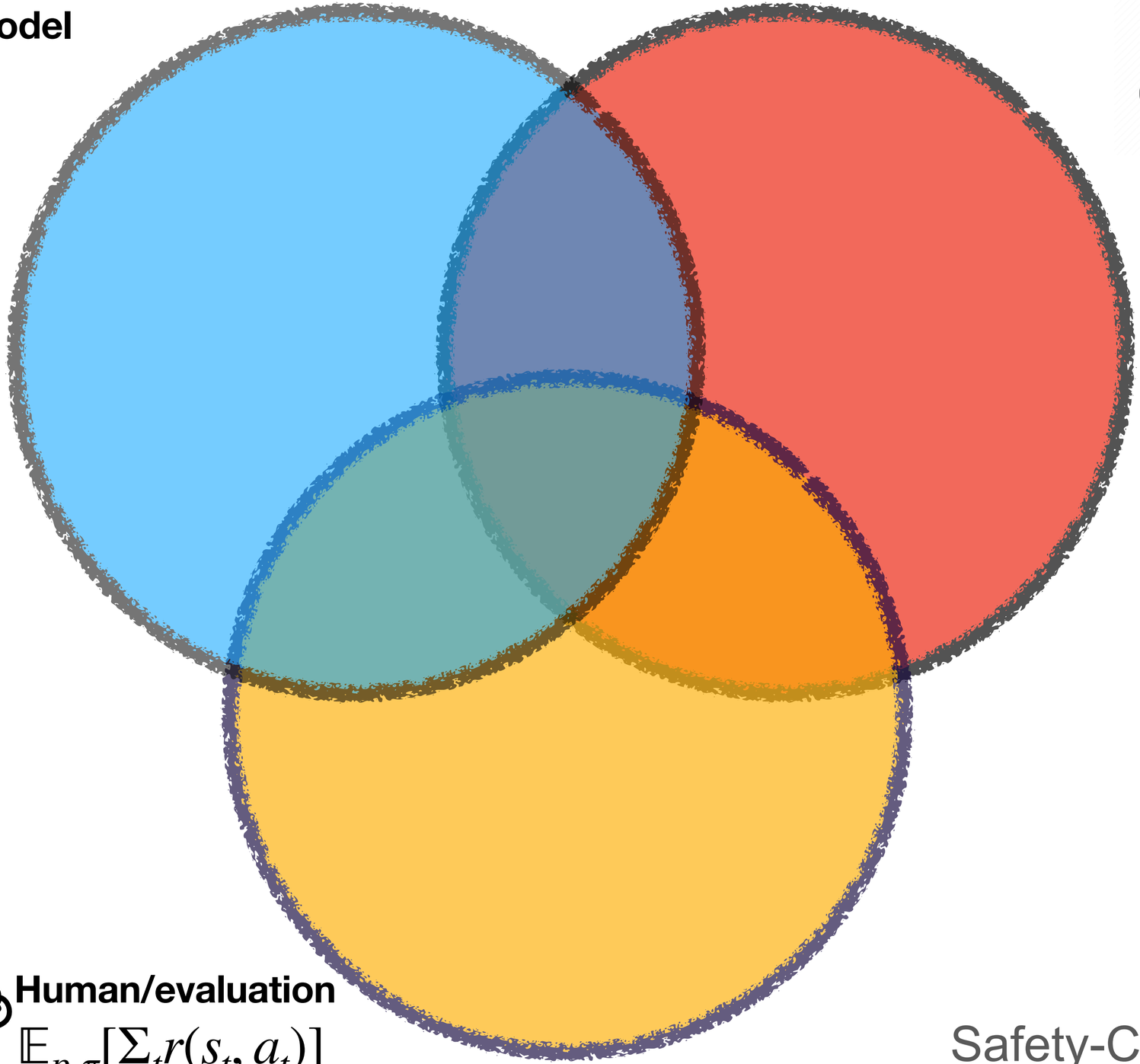
My way …

Environment/model

$p(s_{t+1} \mid s_t, a_t)$

Agent/decision

$\pi(s_t)$



Human/evaluation

$\mathbb{E}_{p,\pi}[\Sigma_t r(s_t, a_t)]$

Finished projects

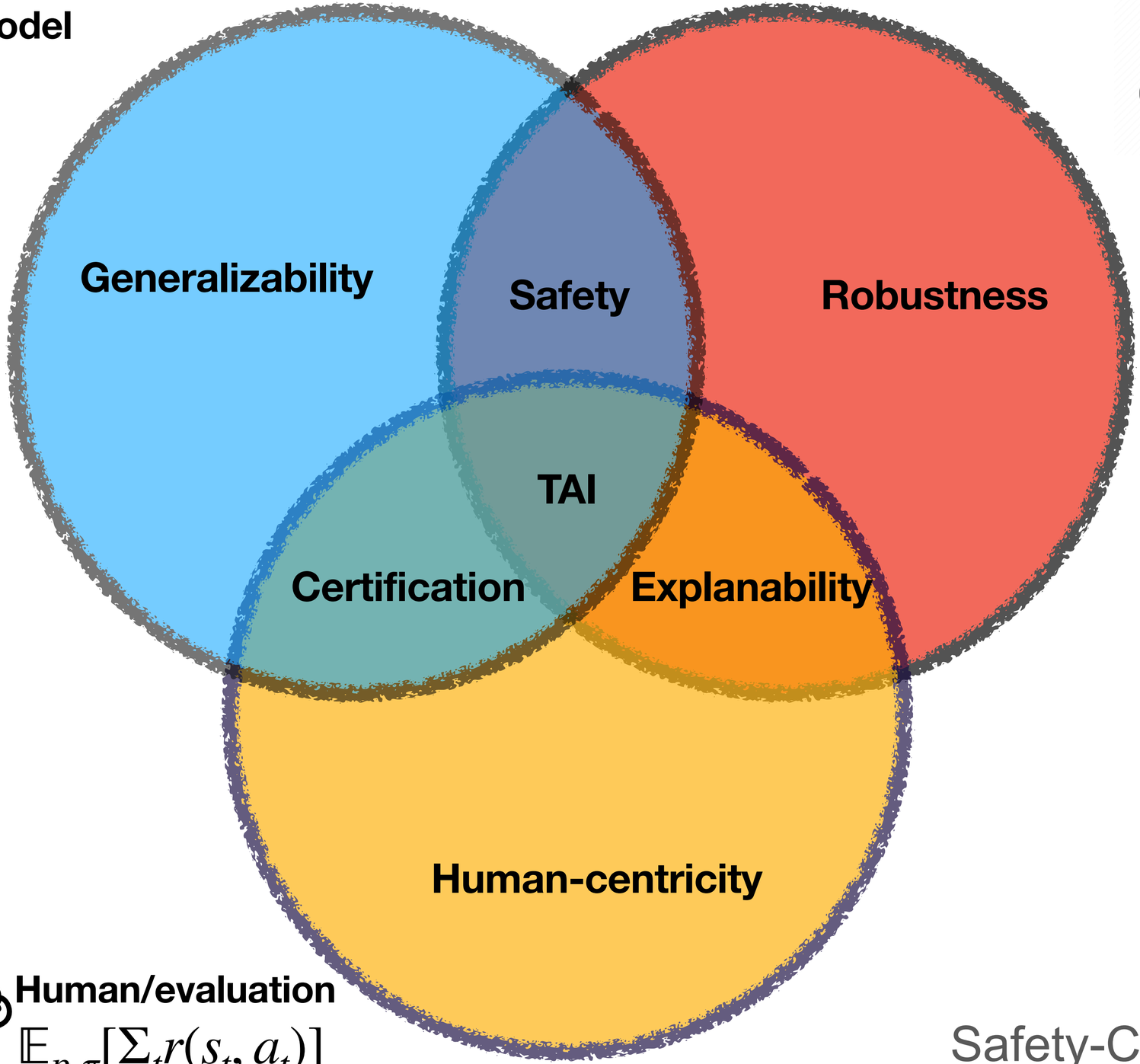
Trustworthy AI for
Safety-Critical Applications

Environment/model

$p(s_{t+1} \mid s_t, a_t)$

Agent/decision

$\pi(s_t)$



Human/evaluation

$\mathbb{E}_{p,\pi}[\Sigma_t r(s_t, a_t)]$

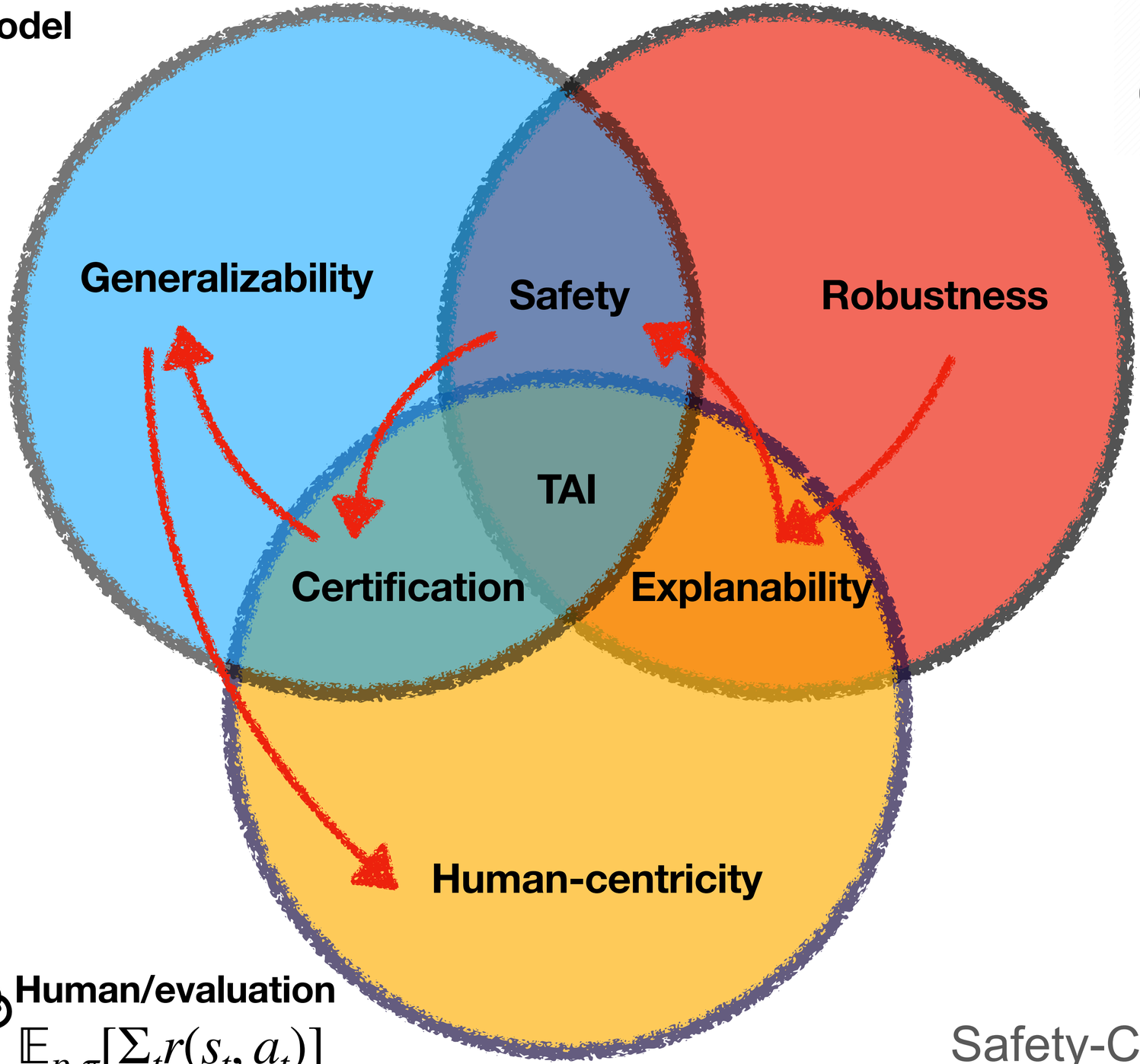
Trustworthy AI for
Safety-Critical Applications

Environment/model

$p(s_{t+1} \mid s_t, a_t)$

Agent/decision

$\pi(s_t)$



Human/evaluation

$\mathbb{E}_{p,\pi}[\Sigma_t r(s_t, a_t)]$

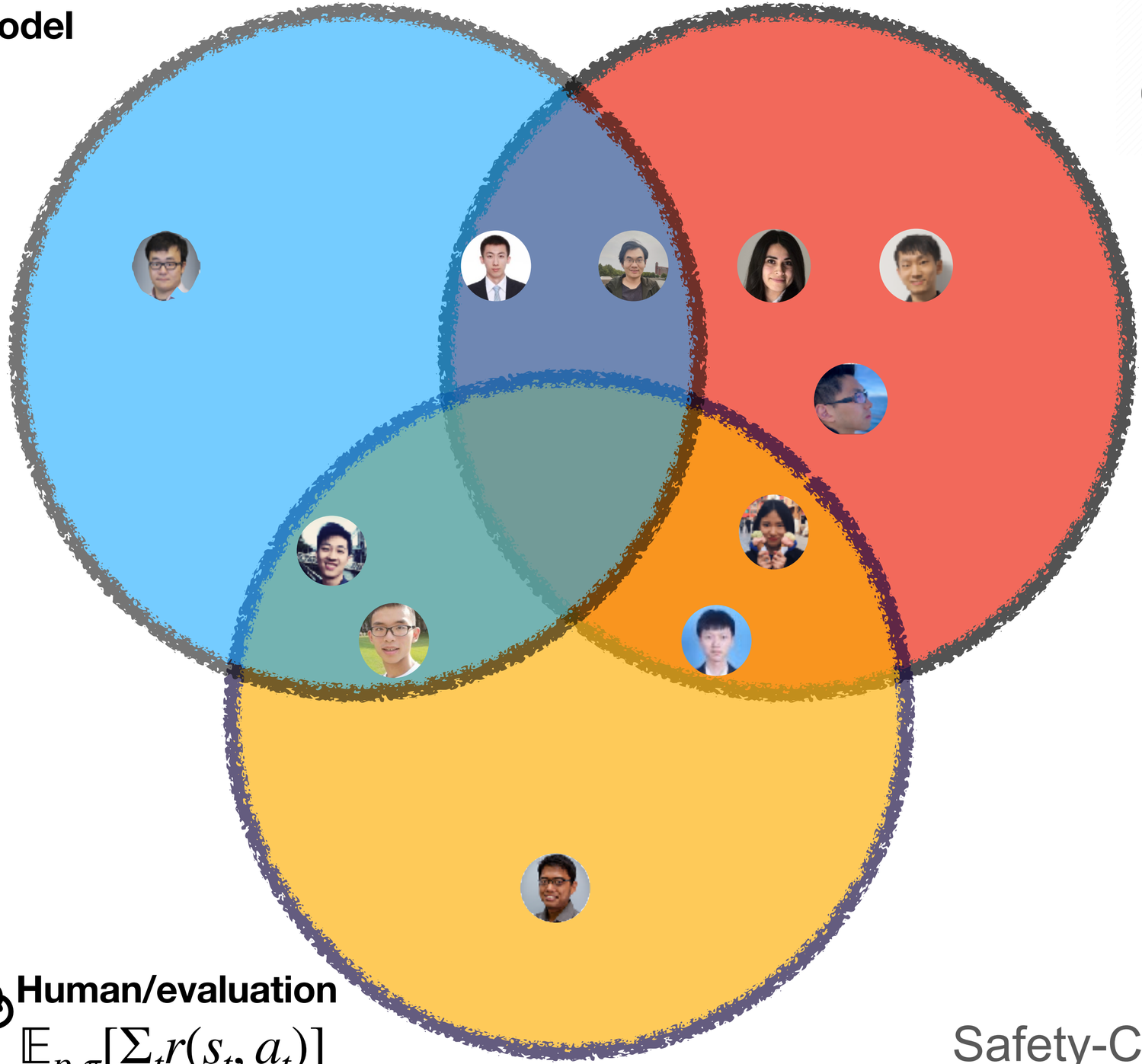
Trustworthy AI for
Safety-Critical Applications

Environment/model

$p(s_{t+1} \mid s_t, a_t)$

Agent/decision

$\pi(s_t)$



Human/evaluation

$\mathbb{E}_{p,\pi}[\Sigma_t r(s_t, a_t)]$

Finished projects

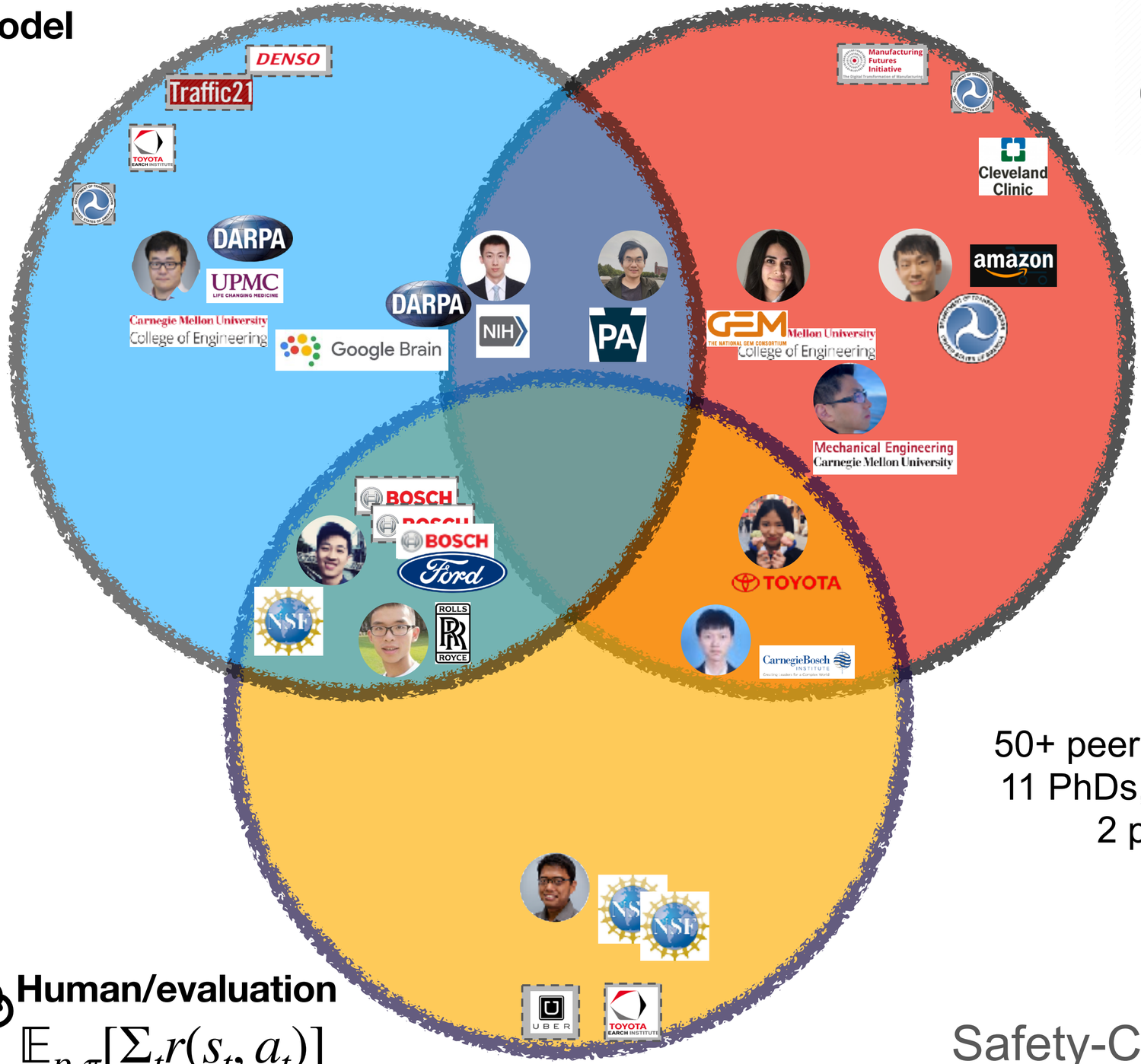
Trustworthy AI for
Safety-Critical Applications

Environment/model

$p(s_{t+1} \mid s_t, a_t)$

Agent/decision

$\pi(s_t)$



50+ peer reviewed publications
11 PhDs, 25 MS, 2 undergrads
2 postdoc, 2 visiting PhD
~30 projects

Human/evaluation

$\mathbb{E}_{p,\pi}[\Sigma_t r(s_t, a_t)]$

Finished projects

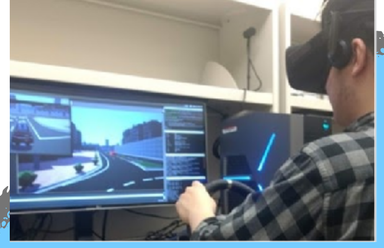
Trustworthy AI for
Safety-Critical Applications

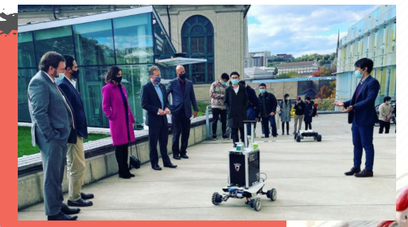
Environment/model

$p(s_{t+1} \mid s_t, a_t)$

Agent/decision

$\pi(s_t)$





Traffic Primitives

United database

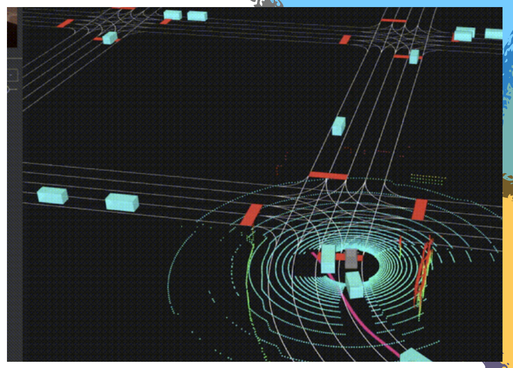
Rubst functional modular system

FOT: Functional Optimal transport

Hierarchical AI

Robust RL Coalitional game Delay aware RL

Critical scene generation



Train RL with IS

Deep IS

Certifiable Policy Evaluation

Human/evaluation

$\mathbb{E}_{p,\pi}[\Sigma_t r(s_t, a_t)]$

Finished projects

Trustworthy AI for Safety-Critical Applications

Structure of this course

- More details on the syllabus

Week	Day	Date	Lec #	Lecture
1	Tue	Jan. 18	1	**Overview, autonomy framework, trustworthy autonomy**
1	Thu	Jan. 20	**2**	Deep learning basics, vision models
1	Fri	Jan. 21	C1	**P1 camp**
2	Tue	Jan. 25	3	Latent space visualization, explanability
2	Thu	Jan. 27	4	Security attacks: poisoning, evasion, FGSM, robust physical attack
2	Fri	Jan. 28	C2	**Reading group presentation tutorial + P1 camp**
3	Tue	Feb. 1	**5**	Robustness-Adversarial and defensive ML: randomization, robust AI, certification
3	Thu	Feb. 3	6	Model-free decision making: imitation learning, reinforcement learning, Q learning
3	Fri	Feb. 4	C3	**P1 camp**
4	Tue	Feb. 8	7	R1: Adversarial AI
4	Thu	Feb. 10	8	Model-free Deep RL: REINFORCE, Actor-Critic
4	Fri	Feb. 11	C4	**P2 camp**
5	Tue	Feb. 15	9	Model-based Deep RL: MPC
5	Thu	Feb. 17	**10**	Gaussian processes: GP
5	Fri	Feb. 18	C5	**P2 camp**
6	Tue	Feb. 22	11	R2: RL for real world autonomy (and model based RL)
6	Thu	Feb. 24	12	Safety: CMDP, Lagrangian-based Method (TRPO-lag, PPO-lag), Constrained Optimization
6	Fri	Feb. 25	C6	**P2 camp**
7	Tue	Mar. 1	**13**	Safety: CMDP, Lagrangian-based Method (TRPO-lag, PPO-lag), Constrained Optimization
7	Thu	Mar. 3	14	Safety: reachability, Control Lyapunov, barrier function
7	Fri	Mar. 4	C7	**C camp**
8	Tue	Mar. 8		
8	Thu	Mar. 10		
8	Fri	Mar. 11		**spring break**
9	Tue	Mar. 15	15	Certification: overview, digital twin simulation, safety critical scenario generation
9	Thu	Mar. 17	**16**	R3: Safe RL
9	Fri	Mar. 18	C8	**Meetings with Prof Zhao to discuss the challenge 1**
10	Tue	Mar. 22	**17**	Digital twin - data-driven: VAE, GAN, and Flow
10	Thu	Mar. 24	**18**	Digital twin - adversarial: worst-case, IS, splitting
10	Fri	Mar. 25	**C9**	**C camp**
11	Tue	Mar. 29	**19**	R4: scenario generation, evaluation and certification
11	Thu	Mar. 31	**20**	Generalization: Working with real world robots, domain randominzation, DDPG, SAC
11	Fri	Apr. 1	C10	**Meetings with Prof Zhao to discuss the challenge 2**
12	Tue	Apr. 5	21	Generalization: Nonstationary environment: delay, RARL, meta learning, NP
12	Thu	Apr. 7		No Class
12	Fri	Apr. 8		Carnival
13	Tue	Apr. 12	22	**Midterm check**
13	Thu	Apr. 14	23	Generalization: hierachical AI, life long learning, DPGP
13	Fri	Apr. 15	C11	**C camp**
14	Tue	Apr. 19	24	R5: generalization
14	Thu	Apr. 21	**25**	**Human centricity: Privacy, fairness**
14	Fri	Apr. 22	S12	**C camp**
15	Tue	Apr. 26	26	**Alumni session: presentation**
15	Thu	Apr. 28	27	**Invited speaker: Bo Li**
15	Fri	Apr. 29	S14	**Meetings with Prof Zhao to discuss your and future career**

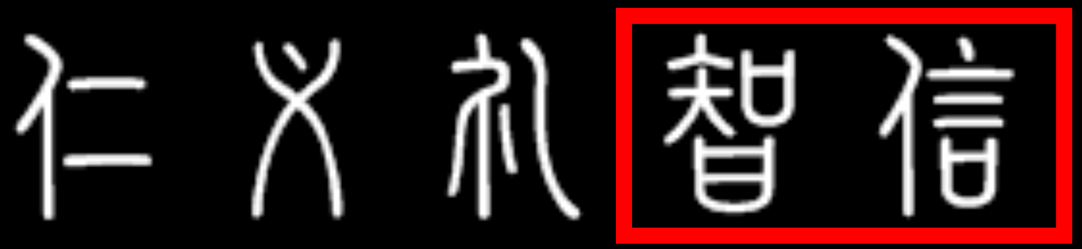
Structure of this course

- M1: Deep learning basics - Reading 1
  - **Explainability**
  - **Robustness**: adversarial ML/security attack/defense
- M2: Reinforcement learning - R2
  - Model-free
  - Model-based
- M3: **Safety**
  - Safe RL: Constrained MDP
  - Reachability, barrier function, Lyapunov stability
- M4: Digital twins/Metaverse
  - **Certification** methods
  - Critical scenario generation: data driven/adversarial robust/knowledge-based controllable generation
- M5: **Generalization**
  - Contextual MDP, domain randomization
  - Hierarchical AI, Life long learning
- M6: **Human centricity** and social good
  - Privacy, fairness

- NOT included (but also important)
  - Societal impact of AI/law/ethics

对人工智能的要求 (Requirement for AI)

赵鼎-2021



19xx-

道德
Ethics

公义
Solve real-world problems

2010-

规范
Standardized

自主解决复杂问题
Solve complex problems

2018-

可信赖的智能
Trustworthy AI

MIT Technology Review Innovators Under 35 China

Goal of this course

- Challenges in doing research on TAIAT?
  - A rapid growing field
  - A lot of interconnected background knowledge
- Goal of this course
  - Have a good sense of the TAIAT research landscape in academia and industry
  - Understand the basic concepts and where to find recourses
  - Quickly identify the key contributions/limitations of a new research
  - Propose and implement a novel idea with a team; know what is (im)possible; time management
  - Get better prepared for interviews of an AIML **research** position
  - A gateway to decide where to focus in your next step

Output of this course

- By the end of this course, in the area of TAIAT, you will be able to
  - Review fundamentals of AI
  - Understand the big picture
  - Quickly read papers
  - Identify a sub-area of interest for your master thesis/PhD research
  - Plan the whole publication cycle of your own research
  - Have a unique story to tell in your job interviews

Best resources

- Feifei Li, Stanford CS231n: Convolutional Neural Networks for Visual Recognition
- Sergey Levine, Berkerley CS 285: Deep Reinforcement Learning, Decision Making, and Control
- Pieter Abbeel, Berkeley CS287: Advanced Robotics
- Chelsea Finn, Stanford CS330: Multi-Task Learning & Transfer Learning Basics
- Dorsa Sadigh, Stanford CS333: Safe and Interactive Robotics
- Bo Li, UIUC CS 598: Adversarial Machine Learning
- Many other materials from Yee Whye Teh, Ian Goodfellow, J. Zico Kolter, etc

Worth reading

- Fernández Llorca D, Gómez E. Trustworthy Autonomous Vehicles. Joint Research Centre (Seville site), Dec, 2021. https://publications.jrc.ec.europa.eu/repository/handle/JRC127051

- Clark Barrett, David L. Dill, Mykel J. Kochenderfer, Dorsa Sadigh, Stanford Center for AI Safety White Paper http://aisafety.stanford.edu/whitepaper.pdf

- Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J. and Mané, D., 2016. Concrete problems in AI safety. https://arxiv.org/abs/1606.06565

- Standards for Trustworthy Autonomous Vehicles https://www.youtube.com/watch?v=rKj19umzYZM