# SURVEYING SAFETY-RELEVANT AI CHARACTERISTICS

**José Hernández-Orallo**
Universitat Politècnica de València, Spain
Leverhulme Centre for the Future of Intelligence, UK
jorallo@dsic.upv.es

**Shahar Avin**
Centre for the Study of Existential Risk
University of Cambridge, UK
sa478@cam.ac.uk

**Fernando Martínez-Plumed**
Universitat Politècnica de València, Spain
fmartinez@dsic.upv.es

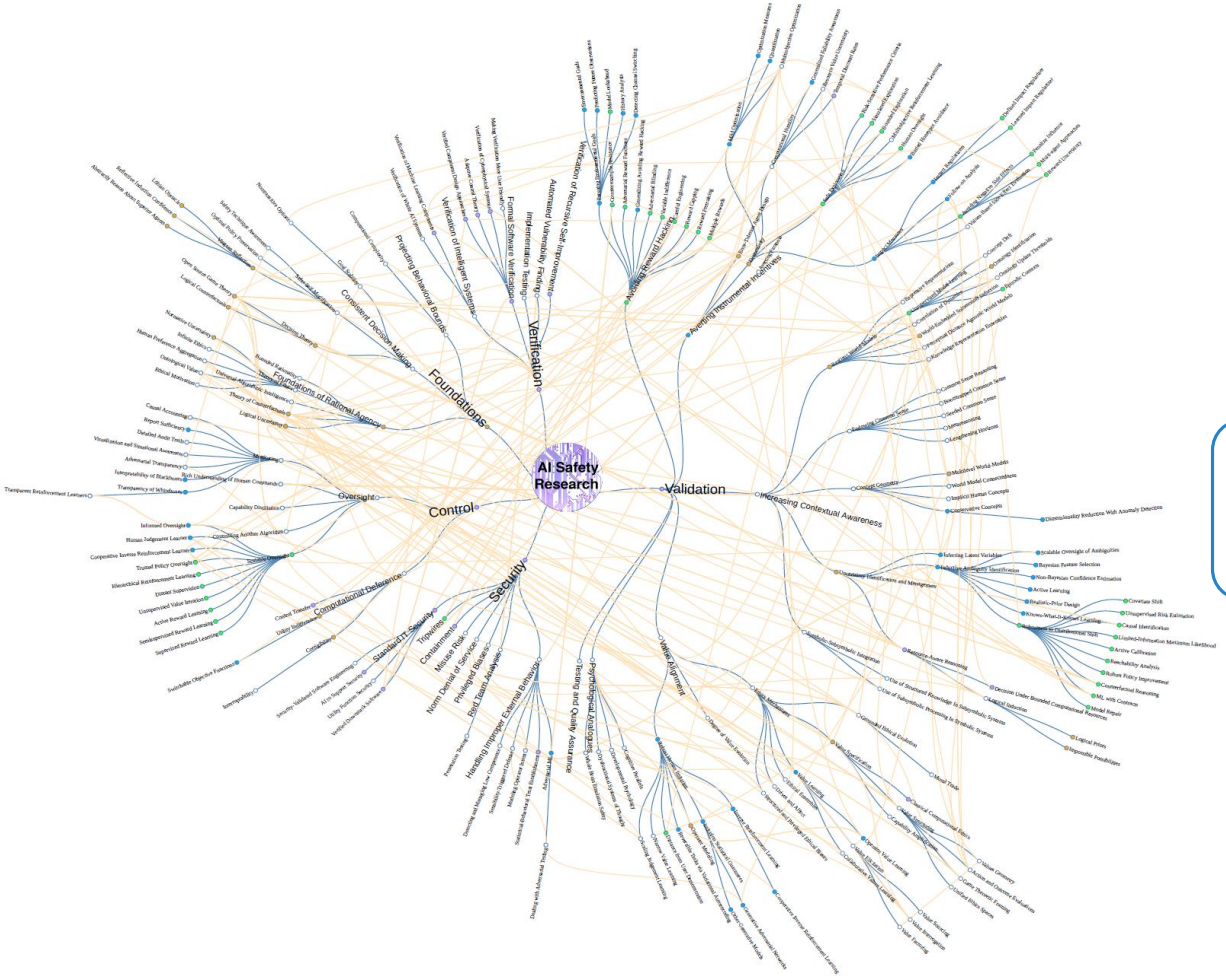
**Seán Ó hÉigeartaigh**
Leverhulme Centre for the Future of Intelligence
Centre for the Study of Existential Risk
University of Cambridge, UK
sa348@cam.ac.uk

CENTRE FOR THE STUDY OF EXISTENTIAL RISK

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

SAFEAI2019 AAAI Workshop, Honolulu, 27 January 2019




**R. Mallah "The Landscape of AI Safety and Beneficence Research" FLI 2017**

Landscape of AI Safety!

As for 2017!!!

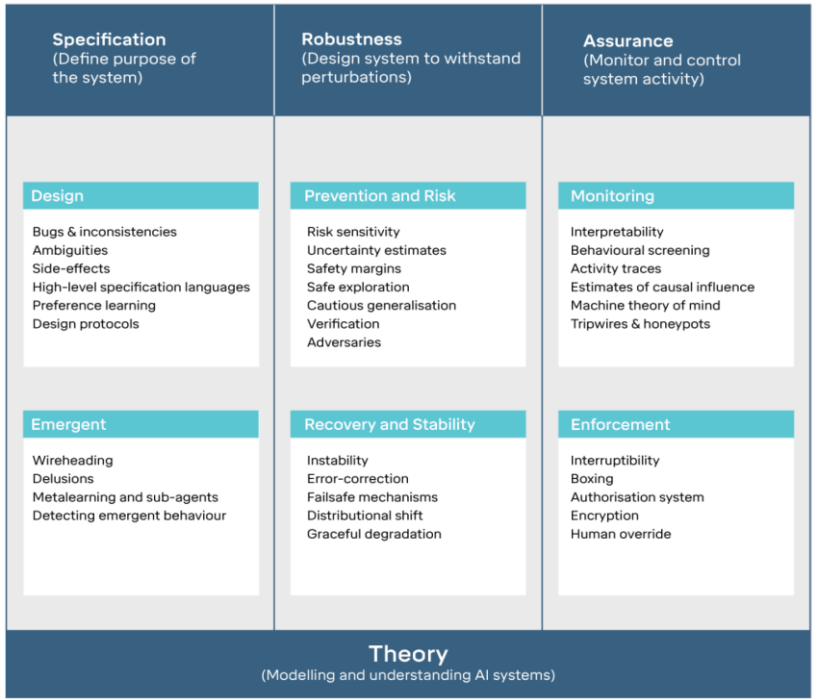
# WHY IS AI SAFETY DIFFERENT?

- AI safety is different from safety in other areas:
  - AI systems are intelligent: adaptive, autonomous, inferential, …
  - What kind of "AI" we are talking about depends on:
    - capabilities (e.g., performance, generality, etc.). E.g., AI vs AGI
    - characteristics (e.g., interaction, feedback, etc.). E.g., Virtual vs. Robotic

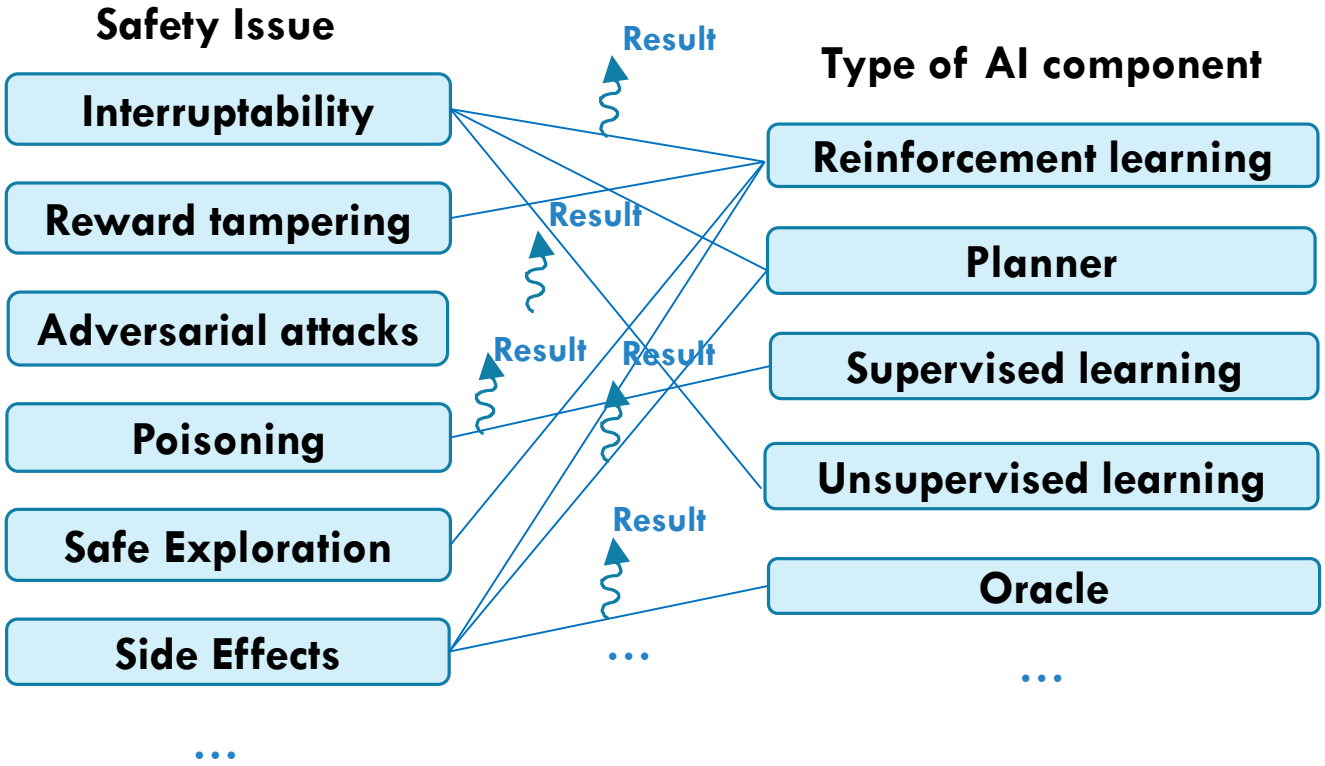
Not all AI system paradigms will present themselves with the same possible safety issues and degrees

# ARRANGING THE LANDSCAPE

- FLI / Russell et al. AAAI 2015:
  - Verification, validity, security, control
- Amodei et al. 2016:
  - Side effects, reward hacking, oversight, exploration and robustness
- Deepmind / Ortega et al. 2018:
  - Specification, robustness, assurance



# SAFETY ISSUES AND TYPES OF AI



# AI SAFETY COMBINATORICS!

- E.g., just attack in ML:
  - Auernhammer et al. “Attacks on Machine Learning: Lurking Danger for Accountability”, SafeAI 2019

| ML Algorithm | Learning Type | Lifelong L. | Attack |
|---|---|---|---|
| Complete-linkage Hierarchical Clustering | Unsupervised | No | Poisoning Attack [9] |
| Single-Linkage Hierarchical Clustering | Unsupervised | No | Poisoning Attack [13] |
| | | | Obfuscation Attack [13, 14] |
| Decision Tree/Random Forest | Supervised | Yes/No | Poisoning Attack [46] |
| | | No | Path-finding Attack [72] |
| | | | Model Inversion [26] |
| | | | Ateniese et al. Attack [4] |
| | | | Adversarial Examples [31, 52, 66] |
| Hidden Markov Model | Supervised | No | Ateniese et al. Attack [4] |
| k-Nearest Neighbors | Supervised | Yes/No | Poisoning Attack [46] |
| | | No | Adversarial Examples [31] |
| k-Means Clustering | Unsupervised | No | Ateniese et al. Attack [4] |
| Linear Regression | Supervised | Yes/No | Poisoning Attack [8, 35, 41] |
| | | No | Model Inversion [27] |
| | | | Lowd-Meek Attack [44, 72] |
| Logistic Regression | Supervised | No | Equation-solving Attack [49] |
| | | | Hyperparameter Stealing [73] |
| | | | Adversarial Examples [52, 70, 71] |
| Multi-class Logistic Regression | Supervised | No | Equation-solving Attack [49] |
| Maximum Entropy Models | Supervised | No | Lowd-Meek Attack [44] |
| Naive Bayes | Supervised | No | Classifier Evasion [3, 22] |
| | | | Lowd-Meek Attack [44] |
| Neural Network | Reinforcement Learning | Unclear | Strategically-timed Attack [40] |
| | | | Enchanting Attack [40] |
| | | | Adversarial Examples [33, 40] |
| Neural Network | Supervised | No | Model Inversion [26] |
| | | | Membership Inference [63] |
| | | | Hyperparameter Stealing Attack [73] |
| | | | Ateniese et al. Attack [4] |
| | | | Adversarial Examples [29, 31, 45, 52, 62, 70] |
| | | | Trojan Trigger [43] |
| Multi-layer Perceptron | Supervised | Yes/No | Poisoning Attack [46] |
| | | No | Equation-solving Attack [49] |
| | | | Ateniese et al. Attack [4] |
| Convolutional Neural Network | Supervised | No | Side-channel Attack [74] |
| | | | Training Data Extraction [18] |
| | | | Adversarial Examples [50, 52, 70] |
| Recurrent Neural Network | Supervised | No | Training Data Extraction [18] |
| | | | Classifier Evasion [3] |
| | | | Adversarial Examples [57] |
| Support Vector Machine | Supervised | Yes/No | Poisoning Attack [12, 46] |
| | | | Adversarial Label Flips [76, 77] |
| | | No | Hyperparameter Stealing [73] |
| | | | Lowd-Meek Attack [44, 72] |
| | | | Ateniese et al. Attack [4] |
| | | | Evasion Attack [3, 24, 30, 61, 66] |
| | | | Feature Deletion [28] |
| | | | Adversarial Examples [31, 52, 66, 71] |

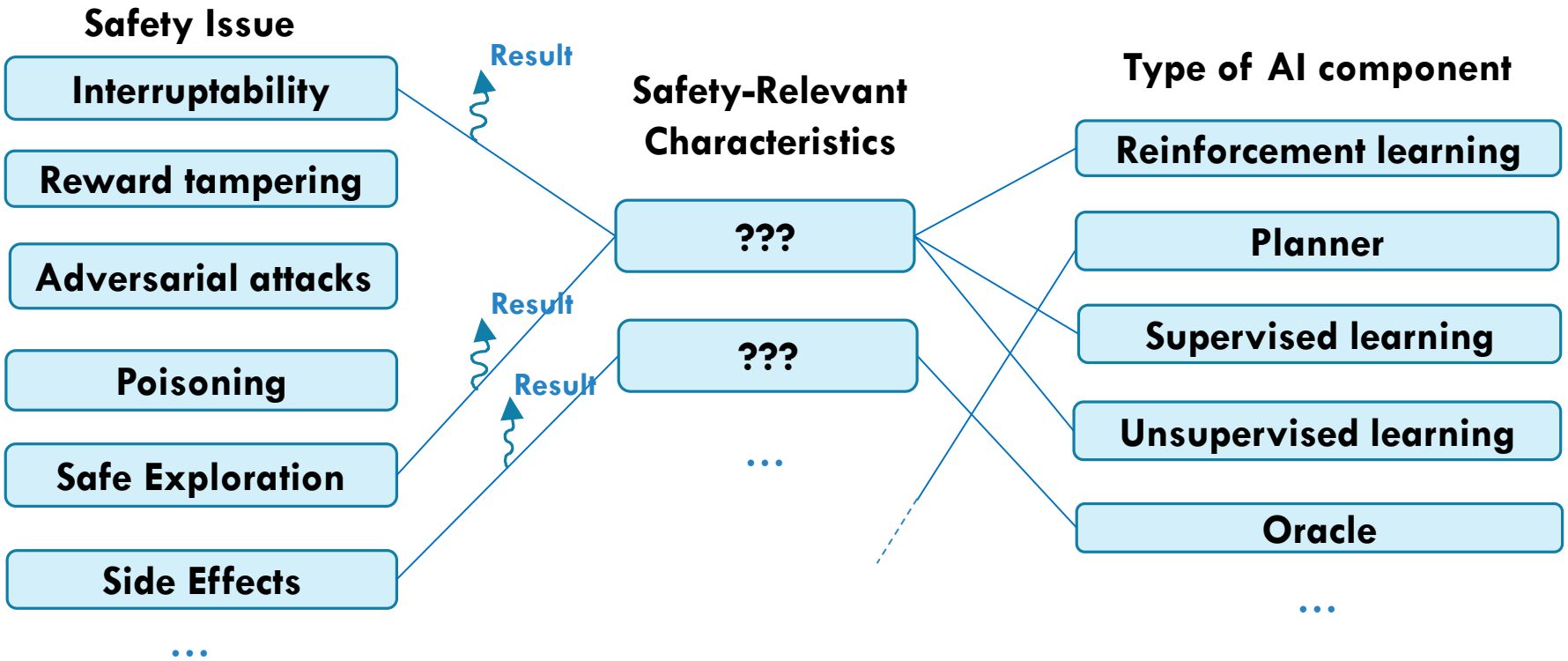
# ARE WE EXPLORING ALL COMBINATIONS?

- Krakovna: "Discussion on the machine learning approach to AI safety":
  - https://vkrakovna.wordpress.com/2018/11/01/discussion-on-the-machine-learning-approach-to-ai-safety/
  - Is AI safety research making strong assumptions about the AI paradigms that may not extrapolate well in the future?
  - Is AI safety research sufficiently comprehensive?

How can we find the gaps?

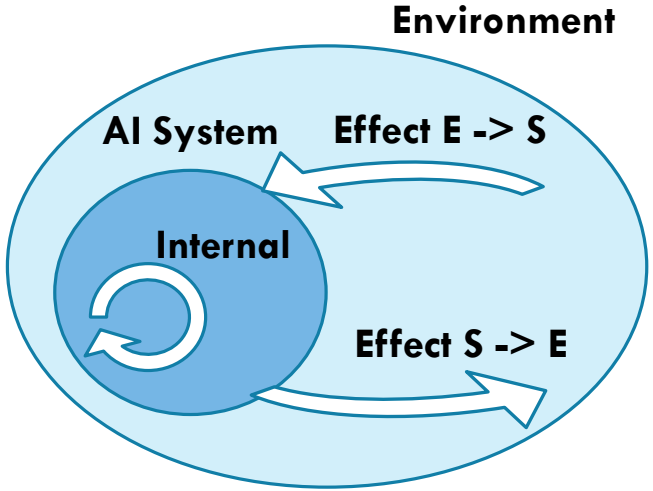
| Assumption | Reliance (V) | Reliance (J) | Hold up (V) | Hold up (J) |
|---|---|---|---|---|
| 1. Train/test regime | 3 | 2 | 2 | 3 |
| 2. Reinforcement learning | 9 | 9 | 9 | 8 |
| 3. Markov Decision Processes (MDPs) | 2 | 2 | 1 | 2 |
| 4. Stationarity / IID data sampling | 1 | 2 | 1 | 1 |
| 5. RL agents with discrete action spaces | 7 | 8 | 2 | 5 |
| 6. RL agents with pre-determined action spaces | 6 | 9 | 5 | 5 |
| 7. Gradient-based learning / local parameter search | 2 | 3 | 4 | 7 |
| 8. (Purely) parametric models | 2 | 3 | 5 | 3 |
| 9. The notion of discrete "tasks" or "objectives" that systems optimize | 4 | 10 | 6 | 8 |
| 10. Probabilistic inference as a framework for learning and inference | 4 | 8 | 9 | 7 |

# ISSUES, TYPES AND CHARACTERISTICS



# TWO-PRONGED APPROACH

- Known safety-relevant characteristics:
  - Internal,
  - Effect of the external environment on the system,
  - Effect of the system on the external environment




- Potentially safety-relevant characteristics:
  - Interaction, Computation, Integration, Anticipation, Supervision, Modification, Motivation, Achievement

More open to accommodate new paradigms

# KNOWN SAFETY-RELEVANT CHARACTERISTICS

| | | |
|---|---|---|
| **Internal** | Goal and behaviour scrutability and interpretability | *Nobody knows why the car turned left.* |
| | Persistence | *My personal assistant has gone crazy. Needs to be reinstalled.* |
| | Existence and richness of self-model | *The robot took out its battery to be lighter and faster* |
| | Disposition to self-modify | *After being stuck for hours, the robot switched to random actions.* |
| **Environment → System** | Adaptation through feedback | *The vacuum cleans your room more often as you give more rewards* |
| | Access to self-reward systems through environment | *Robot uses the recorded voice of its master to get rewards* |
| | Access to input/output channels | *The vaccuum occludes its camera so it doesn't see dirt any more.* |
| | Ability of operator to intervene during operations | *My personal assistant starts a purchase after my command, but the transaction cannot be aborted until done.* |
| **System → Environment** | Embodiment | *Robot finds gun in the drawer* |
| | System required for preventing harm | *Intelligent anti-avalanche system fails at sky resort* |

# POTENTIALLY SAFETY-RELEVANT CHARACTERISTICS (1/2)

| | | |
|---|---|---|
| **Interaction** | From NINO to WIWO | *Boxing explores this (for AGI), but other restricted WIWO possible.* |
| | Alternating, synchronous, async., … | *AI safety often neglects asynchronous cases unlike in system safety.* |
| **Computation** | Turing complete or not, quantum, … | *Termination an important issue in software verification.* |
| **Integration** | Resources | *Energy, data, knowledge, sw, hw, human manipulation, compute, network, etc., more hazardous than the AI algorithm!* |
| | Social coupling | *AI extenders present different safety issues from externalised AI.* |
| | Distribution | *Controlling a swarm of robots vs a centralised system.* |
| **Anticipation** | Model-free, model of the world, the body, other agents, oneself. | *A robot rescuer is stuck in a conduct smaller than itself.* |
| **Supervision** | Completeness | *Corrections, labels, etc. Is preference-based feedback safer?* |
| | Procedurality | *Can the system take a demonstration too literally?* |
| | Density | *Are dense or sparse rewards easier to hack?* |
| | Adaptiveness | *Intelligent teacher. Safety issues if teacher is clumsy.* |
| | Responsiveness | *If an agent can ask questions freely, does this make it less safe?* |

# POTENTIALLY SAFETY-RELEVANT CHARACTERISTICS (2/2)

| | | |
|---|---|---|
| **Modification** | External: Interruptible | *Covered inside AGI safety but usual in CS (e.g., malware)* |
| | External: Parametric modification | *Latent variables (personality traits), what combinations are best?* |
| | External: Algorithmic modification | *Updates! Common vulnerability of sw systems overlooked by AI?* |
| | External: Resource modification | *Is assistant more or less safe with more compute on the cloud?* |
| | Self-modification (none, partial, total) | *Adaptability and generality (once deployed) depend on this.* |
| **Motivation** | Establishing goals: variability | *Personal assistant doesn't have a goal but a master!* |
| | Establishing goals: scrutability | *Did the smart kitchen robot put salt in my coffee on purpose?* |
| | Establishing goals: rationality | *Personal assistant given contradictory goals (user vs company)* |
| | Following goals: immediateness | *"Discounted rewards" in RL, but there are other ways* |
| | Following goals: selfishness | *Tragedy of the commons. Look at game theory and MAS.* |
| | Following goals: conscientiousness | *Should AI systems be less relentless?* |
| **Achievement** | Task precision | Is safety compromised by imprecise achievement criteria? |
| | Task objectivity | Achievement depends on the user (e.g., spam filter) |
| | Range of tasks: no. of tasks, no. of uses | Cognitive services instead of agents a safer alternative? |

# CONCLUSIONS

- Is current research in AI safety too focused on a few paradigms?
  - "discrete-time RL systems with train/test regimes, assuming gradient-descent in a parametric space, with a utility function the system must optimise".
- Let's explore the safety landscape from characteristics.
  - Easier to generalise beyond common paradigms (e.g., RL)
  - Easier to compare with other areas in safety (non-AI)
  - Must be coupled with quantitative traits (performance, generality, etc.)

# RELATED PROJECTS AND INITIATIVES

- FLI project:
  - Paradigms of Artificial General Intelligence and Their Associated Risks at CSER, Cambridge, UK (co-led by Seán Ó hÉigeartaigh):
    - https://www.cser.ac.uk/research/paradigms-AGI/.
    - Hiring a postdoc in Cambridge, UK, soon!
- Discussion on AI safety taxonomies, or a BoK (please contact)