

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 hÉigearthaigh

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



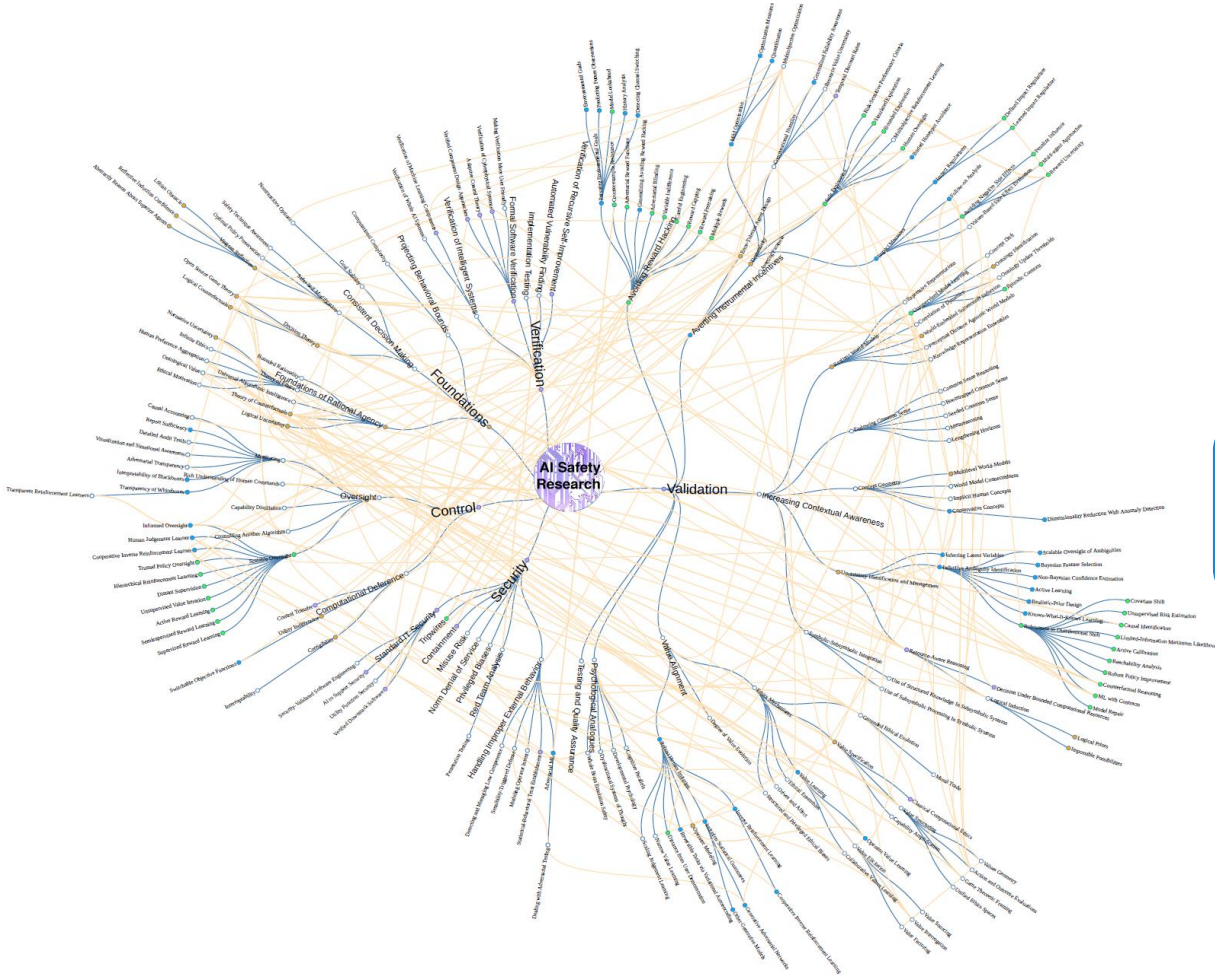
UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA



LEVERHULME CENTRE FOR THE
FUTURE OF INTELLIGENCE

SAFEAI2019 AAI 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

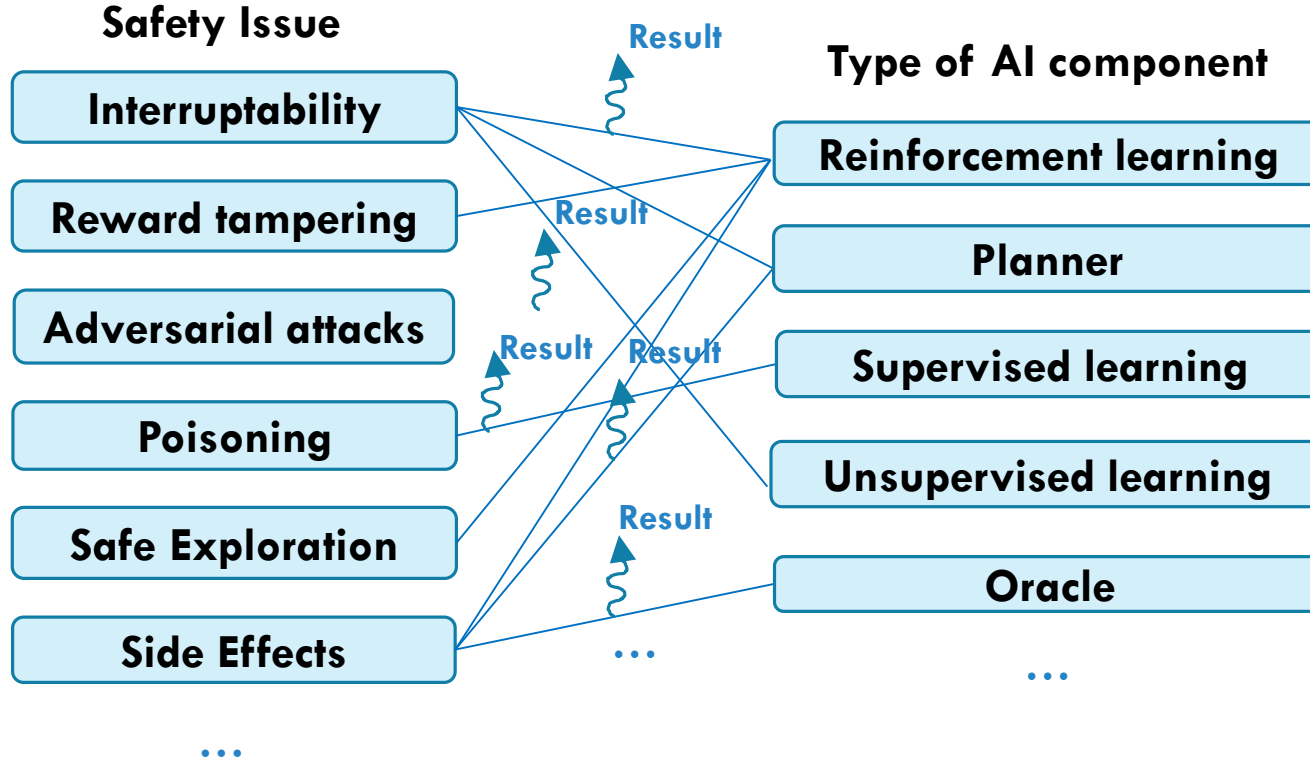
ARRANGING THE LANDSCAPE

- FLI / Russell et al. AAI 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



Specification (Define purpose of the system)	Robustness (Design system to withstand perturbations)	Assurance (Monitor and control system activity)
Design Bugs & inconsistencies Ambiguities Side-effects High-level specification languages Preference learning Design protocols	Prevention and Risk Risk sensitivity Uncertainty estimates Safety margins Safe exploration Cautious generalisation Verification Adversaries	Monitoring Interpretability Behavioural screening Activity traces Estimates of causal influence Machine theory of mind Tripwires & honeypots
Emergent Wireheading Delusions Metalearning and sub-agents Detecting emergent behaviour	Recovery and Stability Instability Error-correction Failsafe mechanisms Distributional shift Graceful degradation	Enforcement Interruptibility Boxing Authorisation system Encryption Human override
Theory (Modelling and understanding AI systems)		

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

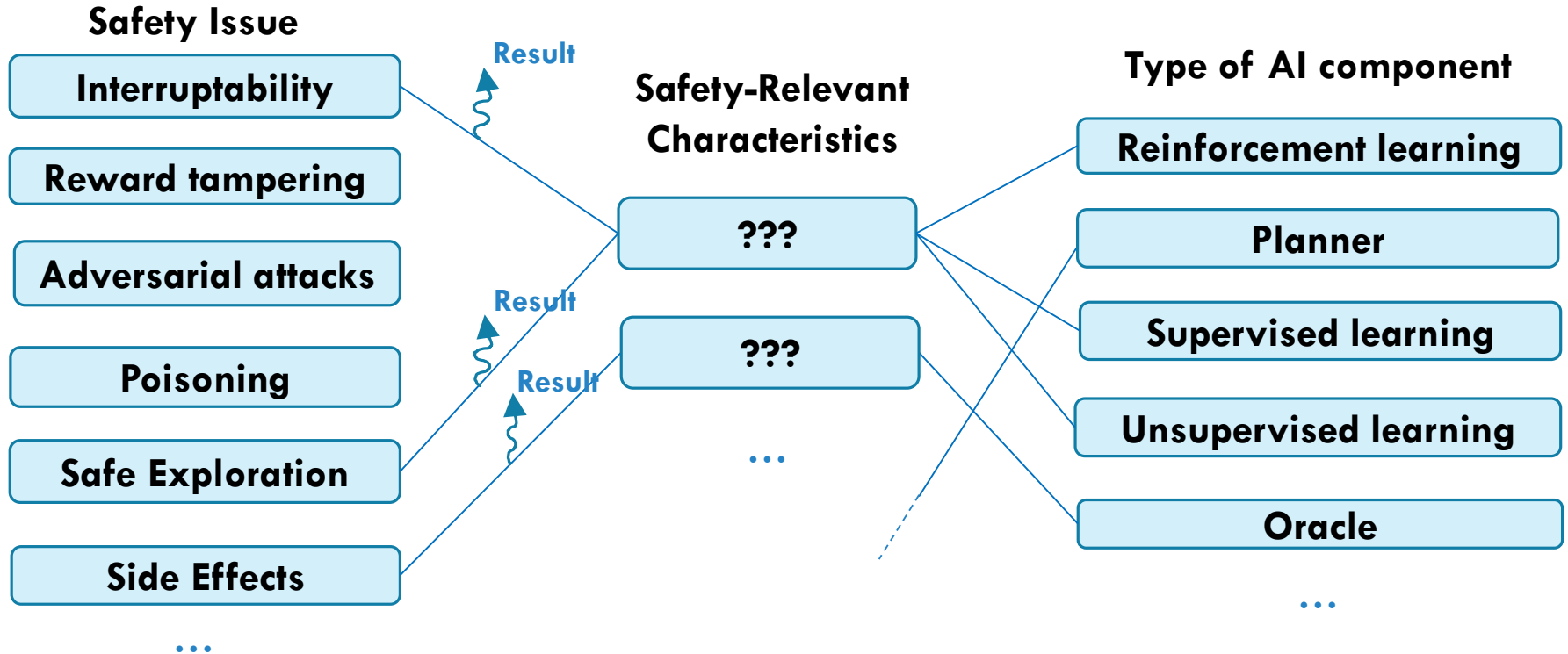
ARE WE EXPLORING ALL COMBINATIONS?

- 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

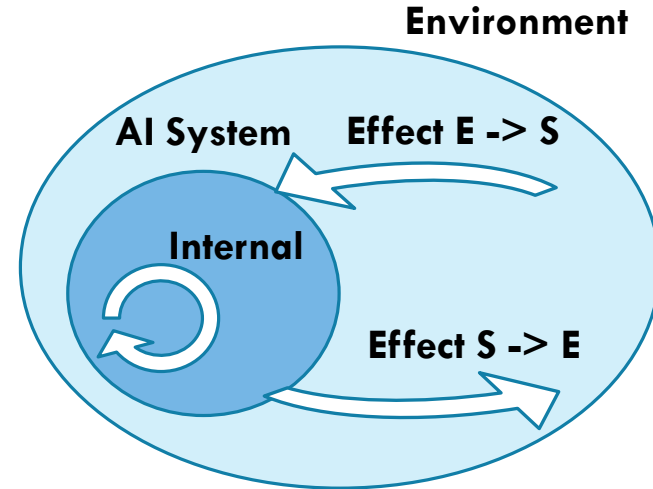
ISSUES, TYPES AND CHARACTERISTICS



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

POTENTIALLY SAFETY-RELEVANT CHARACTERISTICS (1/2)

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

POTENTIALLY SAFETY-RELEVANT CHARACTERISTICS (2/2)

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

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)