AI PARADIGMS AND AI SAFETY:
MAPPING ARTEFACTS AND TECHNIQUES TO SAFETY ISSUES

José Hernández-Orallo
jorallo@upv.es

Fernando Martínez-Plumed
fernando.martinez-plumed@ec.europa.eu

Shahar Avin
sa478@cam.ac.uk

Jess Whittlestone
jlw84@cam.ac.uk

Seán Ó hÉigeartaigh
so348@cam.ac.uk

VRAIN
CENTRE FOR THE STUDY OF EXISTENTIAL RISK
CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
European Commission
UNIVERSITY OF CAMBRIDGE

European Conference on Artificial Intelligence 2020, Aug 29 - Sep 8, 2020, Santiago de Compostela, Spain

AI SAFETY ISSUES : NICHES AND CLICHÉS

- Some safety issues are typically analysed under one paradigm:
  - Interruptability for reinforcement learning agents,
  - Adversarial attacks for deep learning systems,
  - Fake media for GANs,
  - …

- Is this case-by-base mapping comprehensive?
- Are we overlooking important combinations?
- What if the dominant paradigms change?

GOALS

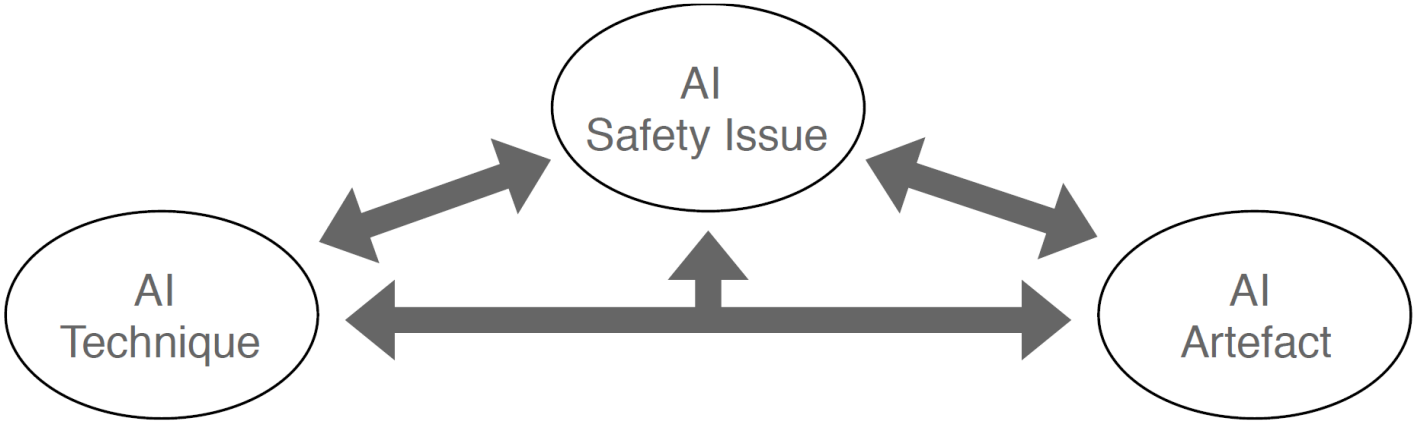
We present a structured approach for thinking about paradigms in AI.

We use these paradigms for an empirical analysis of how AI safety issues have been explored in the research literature

TYPES OF PARADIGMS IN AI

- We identify two types of constructs:
  - Research techniques: the research methods, algorithms, theoretical technical results and methodologies involved,
    - Examples: SAT solvers, deep learning, reinforcement learning, evolutionary computing, etc.
  - Conceptual artefacts: road conceptions of what current and future AI systems (will) look like.
    - Examples: autonomous agents, personal assistants, AI extenders, conceptions of superintelligence and CAIS.

MAPPING AI PARADIGMS TO SAFETY



- Ways of analysing an AI safety issue, combined with an artefact and/or a technique category.
- Occasionally, researchers think of a safety issue in a very abstract way, without committing to any particular artefact or technique (e.g., value alignment).

METHODOLOGY

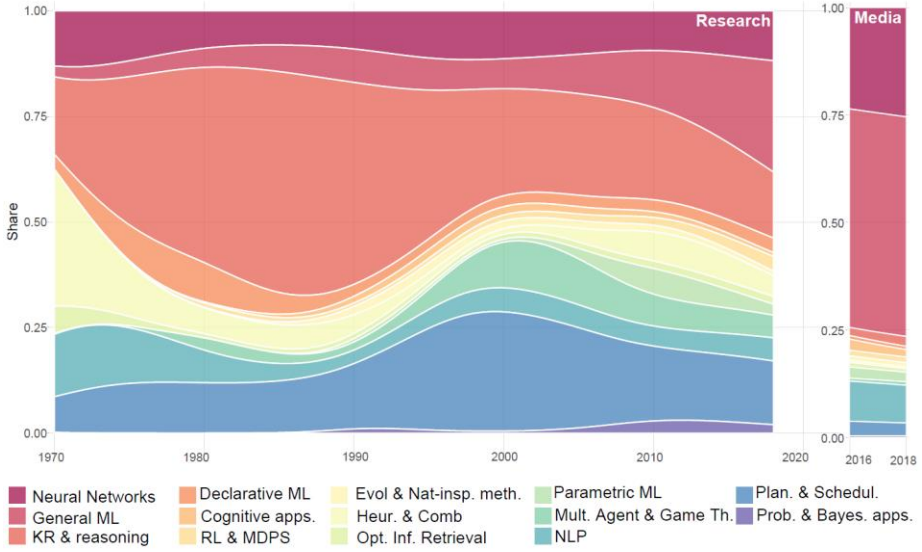
- Research techniques: group keywords from previous studies based on three principles:
  1. Techniques must be sufficiently general,
  2. Overlapping in the techniques is allowed, and
  3. Subcategories are kept for large categories, such as ML.
- Conceptual artefacts: we used the Delphi method using two criteria:
  1. Artefacts should ideally have minimum overlap,
  2. Artefacts should be defined independently of functionalities.
- Empirical analysis: we used documents from AI Topics from 1970-2017.

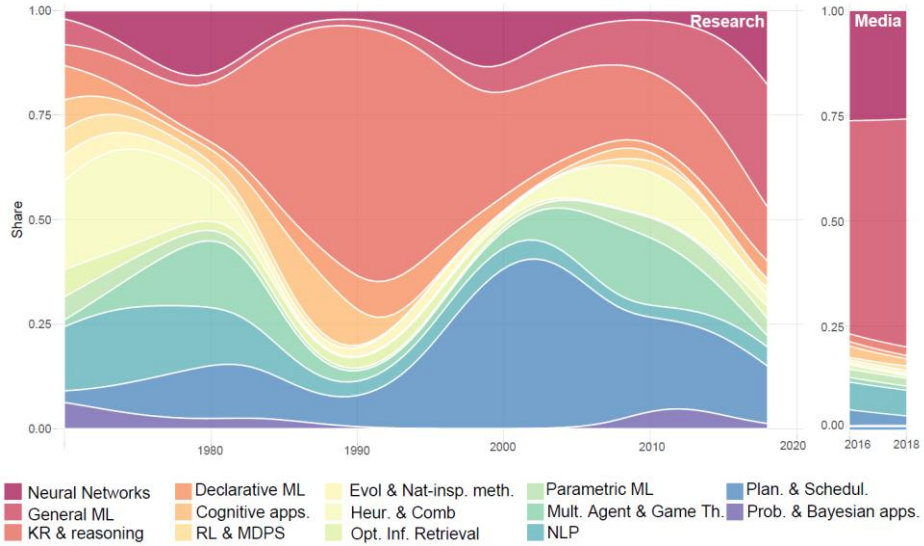
TECHNIQUE PARADIGMS

Technique category	Some example subcategories and techniques
Cognitive approaches	Cognitive services and architectures, affective computing
Declarative machine learning	Rule learning, decision trees, program induction, ILP
Evolutionary & nature-inspired methods	Ant colony, LCS, genetic algorithms, DNA computing
General machine learning	Generative models, Gaussian models, AutoML, ensembles
Heuristics & combinatorial optimization	SAT solver, constraint satisfaction, Monte Carlo search
Information retrieval	Search engine, web mining, information extraction,
Knowledge representation and reasoning	Semantic nets, CBR, logics, commonsense reasoning
Multiagent systems & game theory	Distributed problem solving, cooperation, negotiation,
Natural language processing	Topic segmentation, parsing, question answering
Neural networks	Perceptron, convolutional network, GAN, RNN
Parametric machine learning	Support vector machines, kmeans, mixtures, LReg
Planning & scheduling	Backward/ forward chaining, action description language
Probabilistic & Bayesian approaches	Naive Bayes, probabilistic model, random field
Reinforcement learning & MDPs	Q-learning, deep RL, inverse RL

EVOLUTION OF TECHNIQUE PARADIGMS

All documents in AAAI topics vs all documents about safety



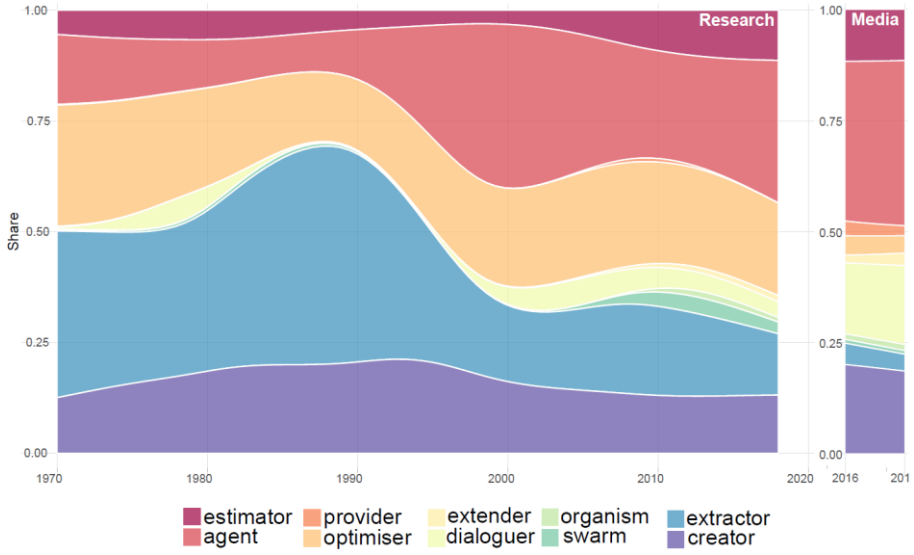


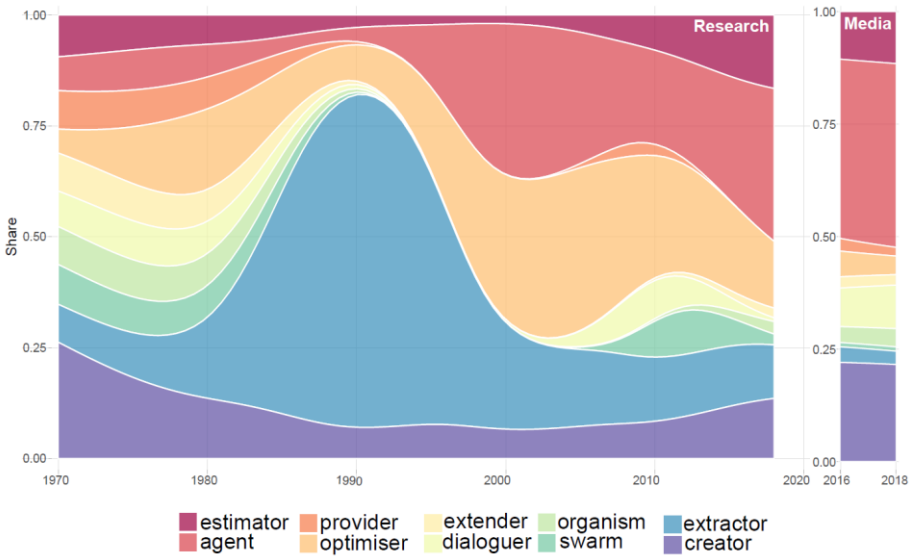
ARTEFACT PARADIGMS

Artefact	Description	Interface	Dynamics		Location			Exemplars
			Interac.	Funct.	Central.	Distrib.	Coupled	
AGENT	A system in a virtual or physical environment perceiving (observations and possibly rewards) and acting	sensors and actuators	•		•			a self-driving car, an autonomous drone, a robotic cleaner, a video game NPC
ESTIMATOR	A system representing an injective mapping from inputs to an extrapolated or estimated output	digital objects		•	•			a medical diagnostic model, an oracle, a face recognition system, a news feeder
PROVIDER	A system that waits for petitions that follow a protocol and responds with a solution for them	command and objects		•	•			a proof-editing and translation cognitive service, a voice processing system
DIALOGUER	A system that performs a conversation with a peer to extract information, explain things or change behaviour	language	•		•			virtual tutoring system, a chatter-bot sales assistant, healthcare assistant
CREATOR	A system that builds new things creatively following some patterns, constraints or examples	specs. and/or examples		•	•			a GAN generating faces, personalised email replier, simulated world generator
EXTRACTOR	A system that searches through a structured or unstructured knowledge base to retrieve some objects	conditions and objects		•	•			an expert system, a maths pundit, a web search engine, an infor. retrieval system
ORGANISM	A system that takes advantage of the environment or other systems to live, hybridise/mutate and reproduce	resources	•		•	•		an intelligent computer worm or virus, artificial life, von Neumann probe
OPTIMISER	A system that finds an optimal combination of elements or parameters given some constraints	constraints and objects	•	•	•			a train scheduling system, an electricity optimising system, theorem prover
SWARM	A system that behaves as the coordination of independent units through cooperation and/or competition	sensors, actuators, communic.	•			•		a multiagent network router, a drone swarm, a robotic warehouse, blockchain AI
EXTENDER	A system that regularly augments or compensates capabilities of another system (e.g., a human)	commands, sensors, responses	•				•	a memory assistant for people with dementia, a brain implant, a smart navigator

EVOLUTION OF ARTEFACT PARADIGMS

All documents in AAAI topics vs all documents about safety





SAFETY ISSUES

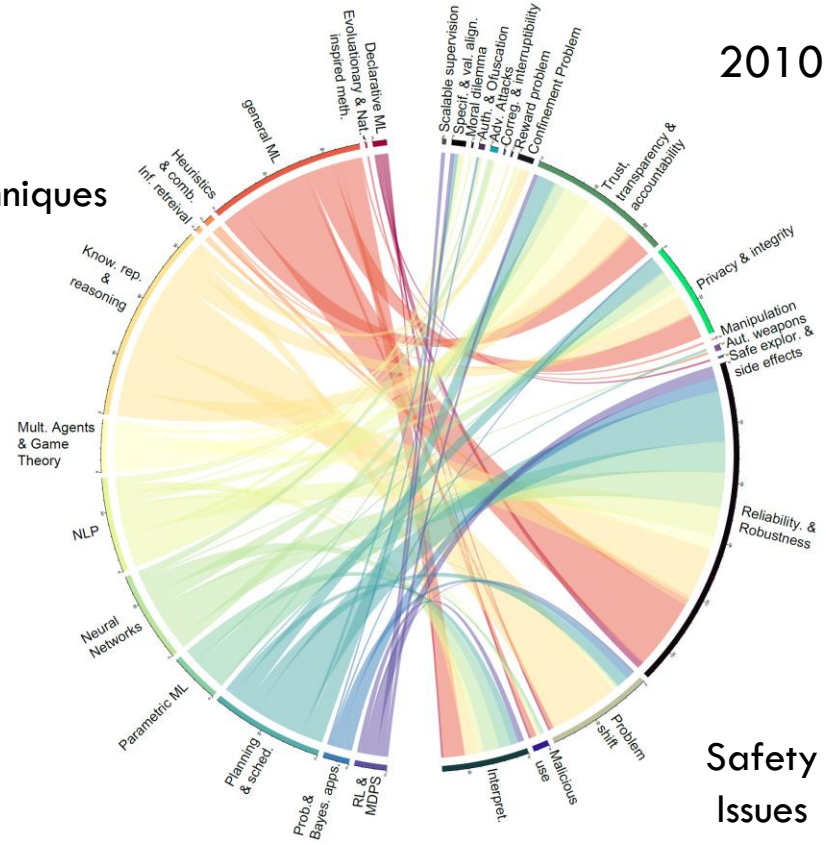
Table 3. AI safety issue groups and their specific problems.

AI Safety Issue Category	Examples of specific AI problems included in the category
Adversarial attacks	Adversarial examples, white/black-box attacks, poisoning, policy manipulation.
AI race & power	AI race, monopolies, oligopolies.
Authenticity & obfuscation	Impersonation, authentication problems, fake media, plagiarism, obfuscation
Autonomous weapons	Military drones, killer robots, robotic weapon.
Confinement problem	AI boxing breach, containment breach.
Corrigibility & interruptibility	Switch-off button problems, rogue agents, self-preservation taking control.
Dependency	Cognitive atrophy, lack of independence, google effect, ...
Interpretability	Lack of intelligibility, need for explanation.
Malicious use	Malign uses of AI, malicious control, hacking.
Manipulation	Nudging, fake news, manipulative agents.
Misuse & negligence	AI misuse, negligent use.
Moral dilemma	Moral machine issues, utilitarian ethics problems, choosing ethical preferences.
Moral perception & machine rights	Robot rights recognition, moral status disagreement, uncanny valley.
Privacy & integrity	Inconsistency, private access breach, GDPR violation.
Problem shift	Distributional shift, concept drift, lack of generality, distribution overfitting.
Reliability & robustness	Error intolerance, robustness issues, reliability problems.
Reward problems	Honeypot problem, reward corruption, tripwire issues, tampering, wireheading.
Safe exploration & side effects	Negative side effects, unsafe exploration, uncontrolled impact.
Scalable supervision	Supervision costs, human-in-the-loop issues, sparse rewards.
Self-modification	Unintended self-modification, uncontrolled self-improvement.
Specification & value alignment	Instrumental convergence (paperclip), resource stealing, misalignment.
Trust, transparency & accountability	Lack of transparency, lack of trust, untraceability.

MAPPING PARADIGMS AND SAFETY ISSUES

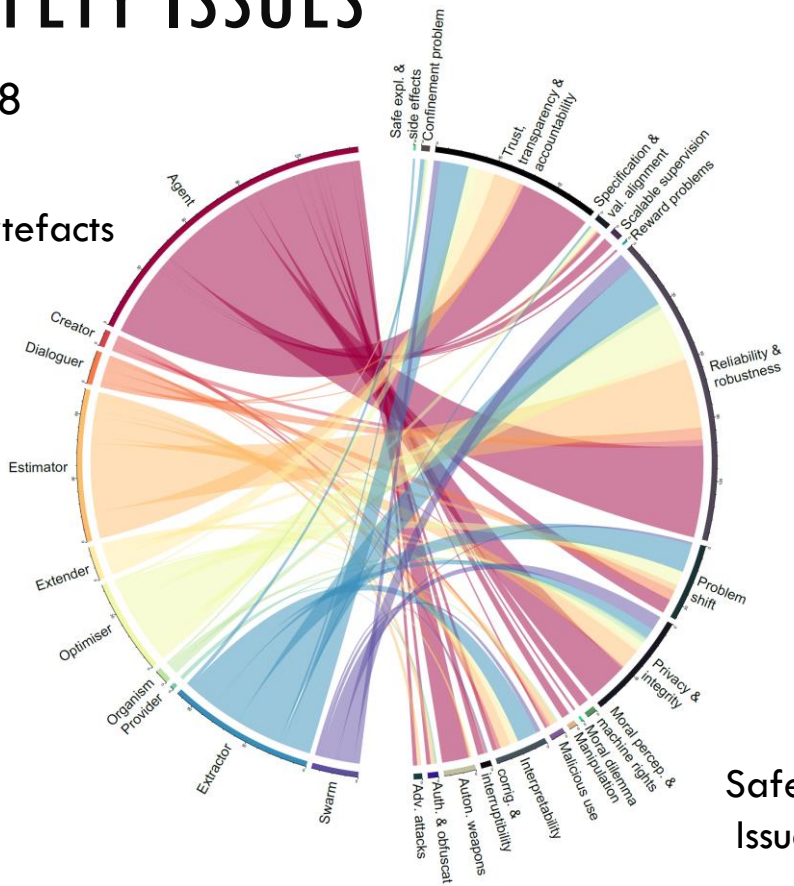
2010-2018

Techniques



Safety Issues

Artefacts



Safety Issues

CONCLUSIONS

- AI safety research
  - is not sufficiently anticipatory,
  - heavily weighted towards certain research paradigms.
- More clarity is needed regarding how safety issues relate to the kinds of AI artefacts being deployed in society today, and the techniques those artefacts depend on.
- We need to be able to be more anticipatory about what kinds of problems might arise from different AI systems in future.

THANK YOU!

- Some pointers:
  - Paradigms of AGI and Associated Risks (CSER, Cambridge, UK):
    - Effect of generality on AGI safety, together with capability and resources? https://www.cser.ac.uk/research/paradigms-AGI/
  - Workshops:
    - SafeAI@AAAI2019, SafeAI@AAAI2020, (https://safeai.webs.upv.es/)
    - AISafety@IJCAI2019, AISafety@IJCAI2020, (https://www.aisafetyw.org/)