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

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?
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.
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.
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

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

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

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

2010-2018

Articulated as a complex network of techniques and artefacts, the diagram illustrates the evolution of safety issues in AI paradigms from 2010 to 2018. The circle represents the overarching field, with branches extending to various techniques and artefacts, each linked by arrows indicating the flow of ideas and developments.

Techniques:
- NLP
- Mult. Agents & Game Theory
- Interpretation
- Privacy & Integrity
- Reliability & Robustness

Arttefacts:
- Estimator
- Creator
- Dialogues
- Problem

Safety Issues:
- Confinement
- Goal alignment
- AI weapons
- Side effects

This visual representation highlights the interconnectedness of research and development in AI paradigms, emphasizing the critical role of safety issues in shaping the future of artificial intelligence.
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/)