Progress Summary
AI Safety Landscape

Feb 7, 2020
NYC, USA

Huáscar Espinoza, Commissariat a'l'Energie Atomique, France
Seán Ó hÉigeartaigh, University of Cambridge, UK
Xiaowei Huang, University of Liverpool, UK
José Hernández-Orallo, Universitat Politècnica de València, Spain
Mauricio Castillo-Effen, Lockheed Martin, USA
Xin Cynthia Chen, University of Hong Kong, China
Richard Mallah, Future of Life Institute, USA
John McDermid, University of York, UK

Why do we need a AI Safety Landscape?

AI Safety Landscape

- AI Safety has been recently recognized as a legitimate domain that is stretching the limits of the broader and more traditional discipline of safety engineering.
- More consensus in terminology and meaning is key towards aligning the understanding of engineering and socio-technical concepts, existing/available theory and technical solutions and gaps in the diversity of AI safety
- Focus on generally accepted knowledge so that the knowledge described is applicable to most AI Safety problems, by still expecting that some considerations will be more relevant to certain applications or algorithms.

What concrete aspects do we target?

AI Safety Landscape

- Bring together the most relevant initiatives and leaders interested on developing a map of AI Safety knowledge to seek consensus in structuring and outlining a generally acceptable landscape for AI Safety.
- Expected outcome is a series of workshop reports summarizing discussions about a landscape of AI Safety, the set of subfields that must be knowledgeable, including an outline of needs, challenges, practices and gaps.
- Align and synchronize the proposed activities and outcomes with other Related initiatives. Together with them, we expect to potentially evolve this landscape towards a more formal form, such as a body of knowledge.

Related Initiatives

AI Safety Landscape

- FLI's AI Safety Research Landscape

future of life INSTITUTE

- Assuring Autonomy International Programme (AAIP): Body of Knowledge

ASSURING AUTONOMY INTERNATIONAL PROGRAMME

- DeepMind's Specification, Robustness and Assurance Aspects in AI Safety

DeepMind

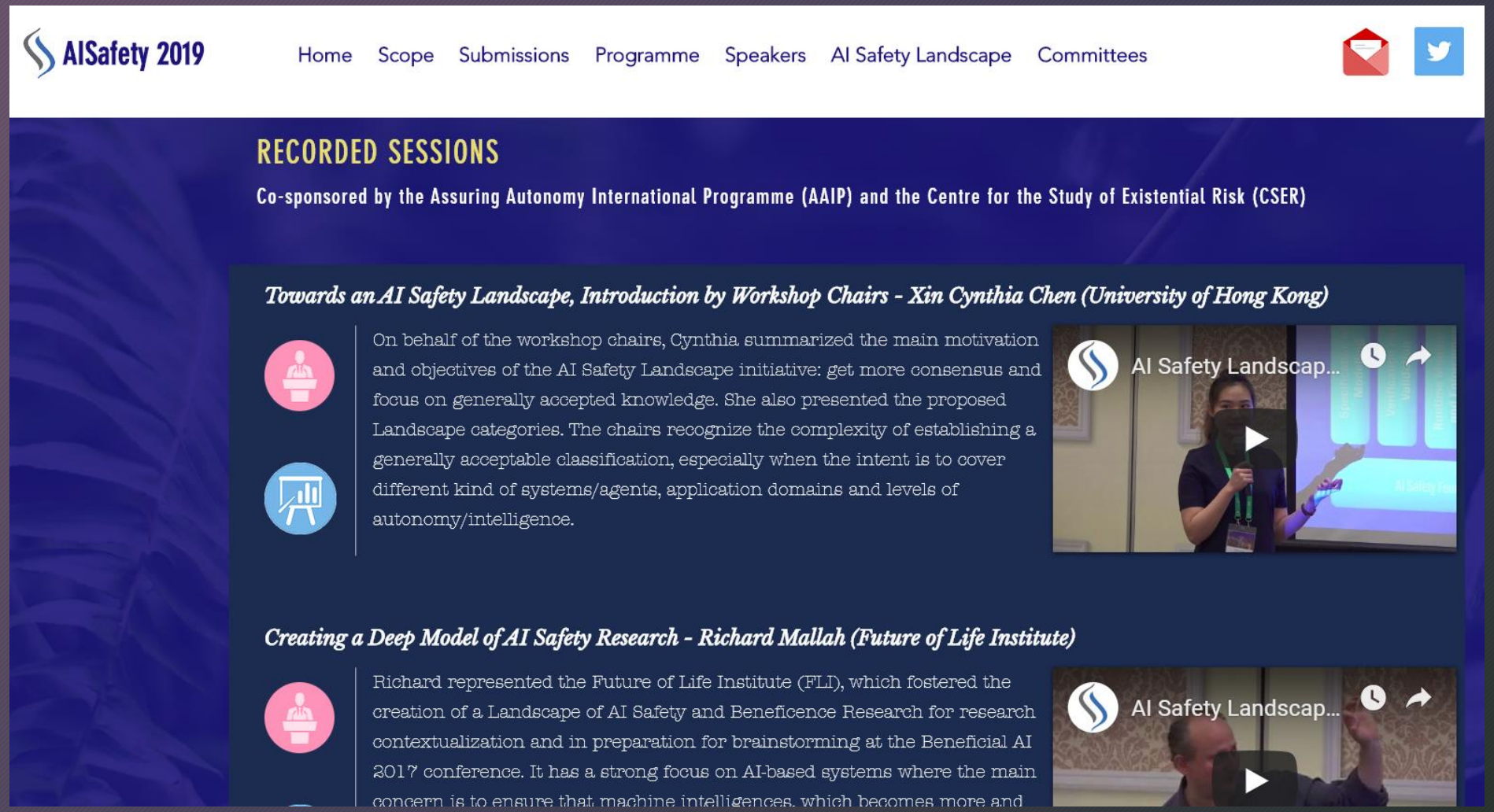
Way of Working

AI Safety Landscape

- The main interaction activities of this initiative are (open) face-to-face meetings that will take place together with the international workshops of AISafety (held at IJCAI) and SafeAI (held at AAAI).
- This first meeting focuses on getting preliminary agreement on the scope of the AI Safety field, outlining a straightforward and generally accepted high-level categorization of the AI Safety field, and planning follow-up actions to ensure effectiveness and coordination with other relevant initiatives.
- While it is clear that a first meeting is not enough to discuss much details of each category, we expect that the different talks and panels outline a preliminary view on its scientific and technical challenges, industrial and academic opportunities, as well as gaps and pitfalls.

1st Landscape Workshop @IJCAI-19: Results

AI Safety Landscape



2nd Workshop: Feb 6, 2020, Bloomberg Offices

AI Safety Landscape

- We looked for consensus on **expected landscape results**
  - Towards an online, interactive knowledge base for AI Safety. Basic governance: working group, moderation, incentives for active participation.
- We discussed key background topics:
  - **The Need of Paradigm Change for AI Safety**. What makes AI systems different? What are common and unique challenges?
  - **AI Safety Landscape Categories**. Creating the basis for AI Safety Foundations and high-level Landscape Categories. Getting proper and sufficient multidisciplinary vision. Can we build upon existing taxonomies, or do we need to start from scratch?
- Some **follow-up actions** were planned.

2nd Workshop: Expected Landscape Results

AI Safety Landscape

- What's for?
  - Basic principles derived to find steps and methodology, and tools to support…
  - Communities of potential users identified: AI / safety researchers (all levels), AI / safety engineers (in industry), regulators and policy-makers, other stakeholders (media, laypeople).
  - We must determine the value of the landscape for each of them.
- What is the expected "product" or artefact
  - Interactive tool that reconfigures based on the scenario (not clear if a wiki)
  - Think of the artefact as a guide (questions-centered?). Precursor of standardization efforts.
  - The artefact should have different parametrised lenses and entry points for different stakeholders
  - Focus on ontology and terminology mappings
- How to get there?
  - Determine the boundaries and evolution (timelines and intermediate deliverables, think big, start small)

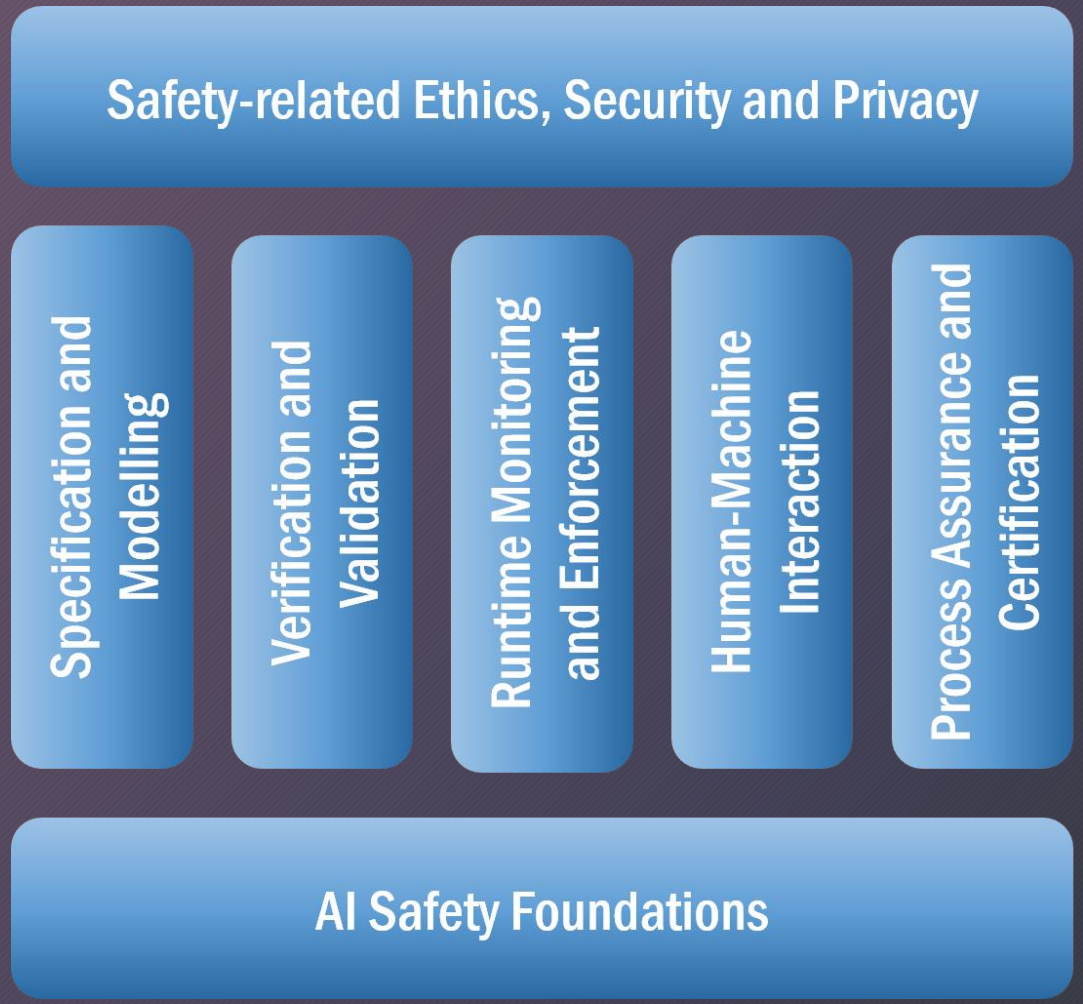
2nd Workshop: The Need for Paradigm Change

AI Safety Landscape

- Should we add AGI to the landscape?
  - A range of AGI safety issues have analogs in present day systems (e.g. value alignment, explainability) that link them in important ways.
  - But AGI is still hypothetical. Some AGI safety issues are different from present-day system issues, and will not manifest practically until more powerful future systems are developed, so they may be less relevant to AI safety engineers and regulators in the near term.
  - What are the overlaps between subfields?
- We should consider the way we validate, assure and approve systems with a lot of uncertainty (unknown unknowns)
  - Need to see the risk assessment from a different perspective (not as humans or usual approaches in safety engineering, but with a different paradigm: machine behaviour?)
- Social sciences must play an important role in the Landscape.
  - Intelligence safety, trust,…
  - Complex systems interaction, psychology can help in complex interactions
  - Economics

2nd Workshop: AI Safety Landscape Categories

AI Safety Landscape



- The top level category should be generalized for any other related trustworthiness property, or split in more categories (Ethics a separate cathegory?).
- Missing Policies and Governance.
- Design and Implementation?
- Architecture aspects (runtime monitoring a subgroup?)
- System engineering practices as part of process aspects?
- Machine-machine interaction?
- Split Foundations in Theoretical aspects and Context + Risk assessment aspects?

→ We will work in proposing a second version by listing bullets (what it convers) in each category

2nd Workshop: Follow-up Actions

AI Safety Landscape

- We collected inputs from participants
  - Are you willing to contribute and in which way? We have commitments for contributions of effort from individuals representing over a dozen organizations
  - How do you expect the Landscape to be useful for you? (value)
  - Likely pillar contributors include JHUAPL, DEEL, and AAIP
- We planned an specification document to be released in a couple of months:
  - Basic governance, and sustainability proposal
  - Concrete form of the Landscape
  - Categories revisited
- Next workshop in IJCAI 2020, Japan.