AISafety 2019

Towards an AI Safety Landscape

Aug 12, 2019
Macao, China

Huáscar Espinoza, Commissariat à l´Energie Atomique, France
Han Yu, Nanyang Technological University, Singapore
Xiaowei Huang, University of Liverpool, UK
Freddy Lecue, Thales, Canada
Cynthia Chen, University of Hong Kong, China
José Hernández-Orallo, Universitat Politècnica de València, Spain
Seán Ó hÉigeartaigh, University of Cambridge, UK
Richard Mallah, Future of Life Institute, USA

Why do we need an AI Safety Landscape?

AISafety 2019

- 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?

AISafety 2019

- 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

AISafety 2019

- FLI's AI Safety Research Landscape

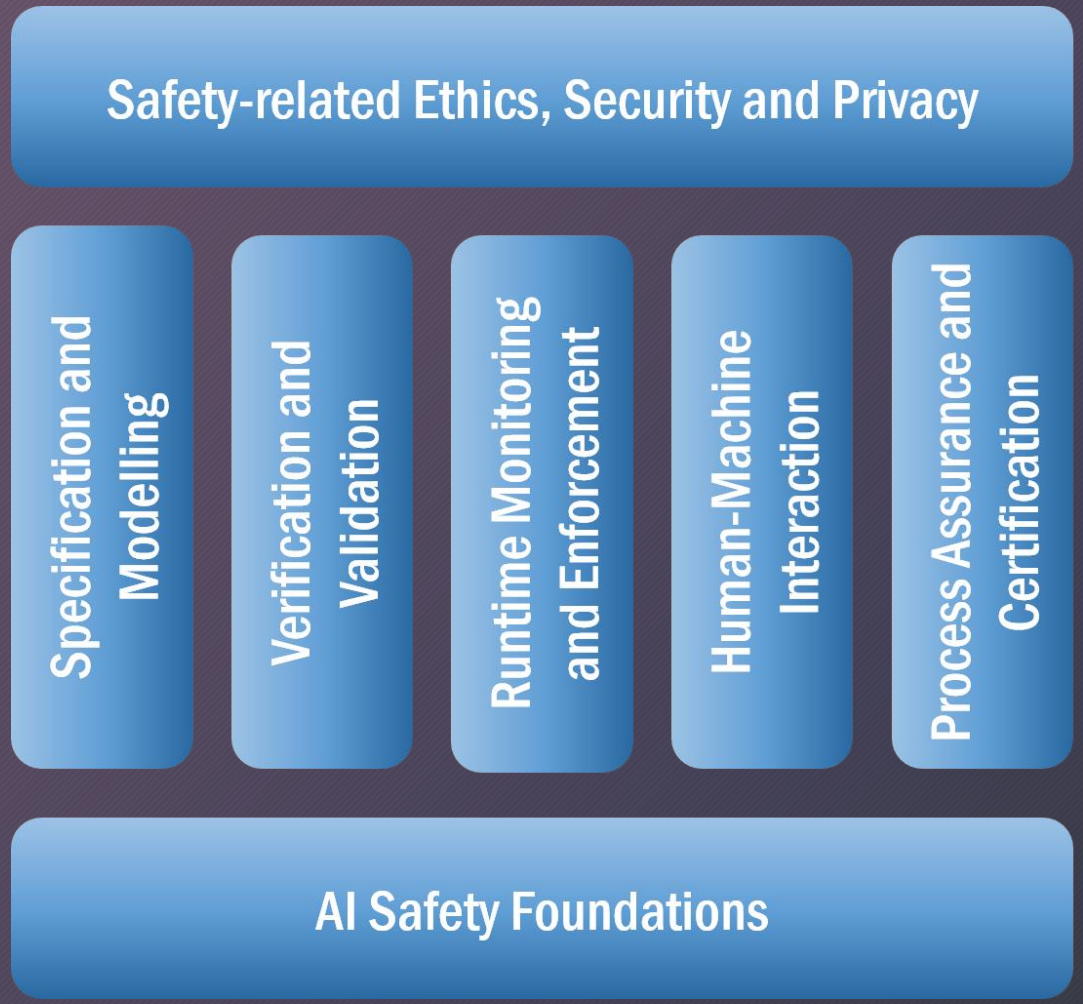
- Assuring Autonomy International Programme (AAIP): Body of Knowledge

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

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

Tentative Landscape Categories

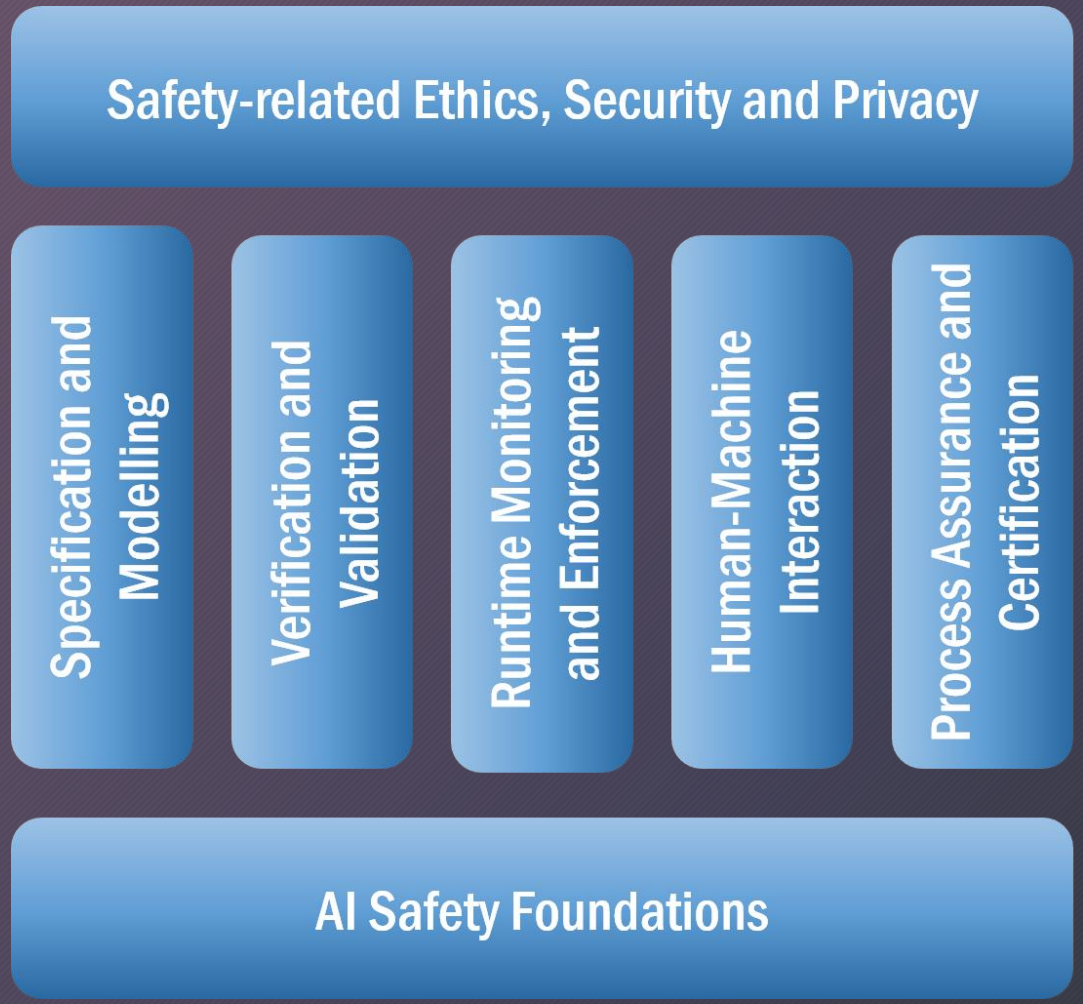
AISafety 2019



- AI Safety Foundations
- Specification and Monitoring
- Verification and Validation
- Runtime Monitoring and Enforcement
- Human-Machine Interaction
- Process Assurance and Certification
- Safety-related Ethics, Security and Privacy

Tentative Landscape Categories

AISafety 2019



- Its goal is to promote structured discussions towards a consistent view of AI Safety.
- This taxonomy is fully open to be amended during the workshop or future meetings.
- We recognize the complexity of establishing a generally acceptable classification, especially when the intent is to cover different kind of systems/agents, application domains and levels of autonomy/intelligence.
- This preliminary classification collects, in our view, the best aspects of Related initiatives at coarse-grained level, which shall be broken down in subcategories later in the process.

Way of Working

AISafety 2019

- 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.

Invited Speakers

AISafety 2019

Dr. Joel Lehman
Uber AI Labs (USA)

Prof. Shlomo Zilberstein
University of Massachusetts Amherst (USA)

Dr. Yang Liu
WeBank (China)

Dr. Victoria Krakovna
Google DeepMind (UK)

Richard Mallah
Future of Life Institute (USA)

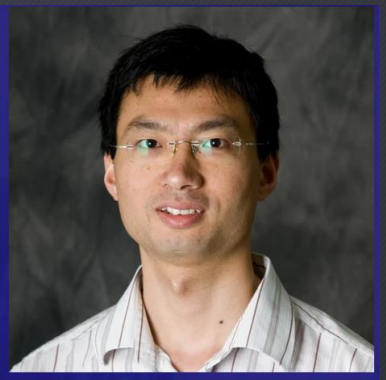
Prof. John McDermid
University of York (UK)

Jeff Cao
Tencent Research Institute (China)

Prof. Virginia Dignum
Umeå University (Sweden)

Prof. Raja Chatila
Sorbonne University (France)

Dr. Gopal Sarma
Broad Institute of MIT and Harvard (USA)



Dr. Xiaowei Huang
University of Liverpool (UK)

Program (Aug 12, Morning)

AISafety 2019

8:30-9:00	Towards an AI Safety Landscape, Introduction by Workshop Chairs
9:00-9:20	***Invited Talk: Richard Mallah (Future of Life Institute, USA): Creating a Deep Model of AI Safety Research***
9:20-9:40	***Invited Talk: John McDermid (University of York, UK): Towards a Framework for Safety Assurance of Autonomous Systems***
9:40-10:10	**Panel 1: The Challenge of Achieving Consensus - Chair: Xiaowei Huang** Session Discussants: Richard Mallah, John McDermid, Huascar Espinoza
10:10-10:40	Coffee Break
10:40-11:00	***Invited Talk: Gopal Sarma (Broad Institute of MIT and Harvard, USA): AI Safety and The Life Sciences***
11:00-11:20	***Invited Talk: Xiaowei Huang (University of Liverpool, UK): Formal Methods in Certifying Learning-Enabled Systems***
11:20-11:40	***Invited Talk: Joel Lehman (Uber AI Labs, USA): AI Safety and Evolutionary Computation***
11:40-12:30	**Panel 2: The Need for Paradigm Change - Chair: Seán Ó hÉigeartaigh** Session Discussants: Gopal Sarma, Xiaowei Huang, Joel Lehman
12:30-14:00	Lunch

Program (Aug 12, Afternoon)

AISafety 2019

14:00-14:20	***Invited Talk: Virginia Dignum (University of Umeå, Sweden): AI Safety for Humans***
14:20-14:40	***Invited Talk: Raja Chatila (Sorbonne University, France): Towards Trustworthy Autonomous and Intelligent Systems***
14:40-15:00	***Invited Talk: Jeff Cao (Tencent Research Institute, China): AI Principles and Ethics by Design***
15:00-15:30	**Panel 3: Towards More Human-Centered and Ethics-Aware Autonomous Systems - Chair: Richard Mallah** Session Discussants: Virginia Dignum, Raja Chatila, Jeff Cao
15:30-16:00	Coffee Break
16:00-16:20	***Invited Talk (via Skype): Victoria Krakovna (Google DeepMind, UK): Specification, Robustness and Assurance Problems in AI Safety***
16:20-17:20	**Panel 4: Building an AI Safety Landscape: Perspectives and Future Work - Chair: Huascar Espinoza** Session Discussants: Richard Mallah, Seán Ó hÉigeartaigh, Xiaowei Huang, Andrea Aller Tubella
17:20-17:30	Wrap-up

Sponsorship

AISafety 2019

ASSURING AUTONOMY
INTERNATIONAL PROGRAMME

The Assuring Autonomy International Programme is delighted to be supporting the AI Safety Workshop 2019. The Programme is advancing the safety of robotics and autonomous systems (RAS) across the globe. It is a £12million partnership between Lloyd's Register Foundation and the University of York that is working with an international community of developers, regulators, researchers and others to ensure the public can benefit from the safe, assured and regulated introduction and adoption of RAS. The Programme is addressing core technical issues underlying the assurance of RAS, supporting industrial demonstrator projects, delivering training and education, and creating an online Body of Knowledge that will reflect the evolving state-of-practice in assuring and regulating RAS.

PARTNERSHIP ON AI

Partnership on AI (full name Partnership on Artificial Intelligence to Benefit People and Society) is a technology industry consortium focused on establishing best practices for artificial intelligence systems and to educate the public about AI.

CENTRE FOR THE STUDY OF
EXISTENTIAL RISK

The Centre for the Study of Existential Risk (CSER) is an interdisciplinary research centre at the University of Cambridge dedicated to the study and mitigation of existential risks posed by present or future technology.