Report on the IJCAI-19 Workshop on Artificial Intelligence Safety (AISafety 2019)

Huáscar Espinoza[1], Han Yu[2], Xiaowei Huang[3], Freddy Lecue[4], Cynthia Chen[5], José Hernández-Orallo[6], Seán Ó hÉigeartaigh[7], Richard Mallah[8], John McDermid[9]

[1]Commissariat à l´Energie Atomique, France
[2]Nanyang Technological University, Singapore
[3]University of Liverpool, UK
[4]Thales, Canada
[5]University of Hong Kong, China
[6]Universitat Politècnica de València, Spain
[7]University of Cambridge, UK
[8]Future of Life Institute, USA
[9]University of York, UK

The IJCAI-19 Workshop on Artificial Intelligence Safety (AISafety 2019), was held at the 28th International Joint Conference on Artificial Intelligence (IJCAI) on August 11-12, 2019 in Macao, China. The proceedings were published with CEUR (http://ceur-ws.org/Vol-2419/). The videos and slides of the second day are available at https://www.ai-safety.org/recorded-sessions.

Introduction

In the last decade, there has been a growing concern on risks of Artificial Intelligence (AI). Safety is becoming increasingly relevant as humans are progressively side-lined from the decision/control loop of intelligent and learning-enabled machines. In particular, the technical foundations and assumptions on which traditional safety engineering principles are based, are inadequate for systems in which AI algorithms, and in particular Machine Learning (ML) algorithms, are interacting with people and/or the environment at increasingly higher levels of autonomy. We must also consider the connection between the safety challenges posed by present-day AI systems, and more forward-looking research focused on more capable future AI systems, up to and including Artificial General Intelligence (AGI).

The IJCAI-19 Workshop on Artificial Intelligence Safety (AISafety 2019) seeks to explore new ideas on AI safety with particular focus on addressing the following questions:

- How can we engineer trustable AI software architectures?
- Do we need to specify and use bounded morality in system engineering to make AI-based systems more ethically aligned?
- What is the status of existing approaches in ensuring AI and ML safety and what are the gaps?
- What safety engineering considerations are required to develop safe human-machine interaction in automated decision-making systems?
- What AI safety considerations and experiences are relevant from industry?
- How can we characterise or evaluate AI systems according to their potential risks and vulnerabilities?
- How can we develop solid technical visions and paradigm shift articles about AI Safety?
- How do metrics of capability and generality affect the level of risk of a system and how trade-offs can be found with performance?
- How do AI system feature for example ethics, explainability, transparency, and accountability relate to, or contribute to, its safety?
- How to evaluate AI safety?

The main interest of AISafety 2019 is to look holistically at AI and safety engineering, jointly with the ethical and legal issues, to build trustable intelligent autonomous machines. The first edition of AISafety was held in August 11-12, 2019, in Macao (China), as part of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19). The AISafety workshop is organized as a "sister workshop" to other two workshops: WAISE (https://www.waise.org/) and to SafeAI (http://www.safeai2019.org).

As part of this IJCAI workshop, we also started the ***AI Safety Landscape*** initiative. This initiative aims at defining an AI safety landscape providing a "view" of the current needs, challenges and state of the art and the practice of this field. Further information about this initiative can be found at: https://www.ai-safety.org/ai-safety-landscape

Programme

The Programme Committee (PC) received 36 submissions, in the following categories:

- Short position papers – 9 submissions.
- Full scientific contributions – 23 submissions.
- Proposals of technical talks – 4 submissions.

Each of the papers was peer-reviewed by at least three PC members, by following a single-blind reviewing process. The committee decided to accept 13 papers (2 position papers and 11 scientific papers) and 2 talks, resulting in an overall acceptance rate of 42%. We additionally invited 1 talk, which was not submitted to the call, and accepted 7 submissions as short papers for poster presentation.

AISafety 2019 was planned as a two-days' workshop with general AI Safety topics in the first day and AI Safety Landscape talks and discussions during the second day. We summarise some of the presentations and discussions below.

First Workshop Day (Aug 11)

The day included four thematic sessions, one keynote and two invited talks. The thematic sessions followed a highly interactive for-mat. They were structured into short talks and a common panel slot to discuss both individual paper contributions and shared topic issues. Three specific roles were part of this format: session chairs, presenters and session discussants.

- *Session Chairs* introduced sessions and participants. The Chair moderated session and plenary discussions, took care of the time, and gave the word to speakers in the au-dience during discussions.
- *Presenters* gave a paper talk in 10 minutes and then par-ticipated in the debate slot.
- *Session Discussants* prepared the discussion of individual papers and the plenary debate. The discussant gave a crit-ical review of the session papers.

After the introduction to the workshop, the first keynote, Joel Lehman (Uber AI Labs, USA), gave a talk about "AI Safety for Evolutionary Computation, Evolutionary Computation for AI Safety". He built on the three broad pillars of AI Safety (from Leike and Chiappa's tutorial 2019): specification, robustness and assurance. He developed on the intersection of AI safety and evolutionary computation in two directions: evolutionary computation for AI safety (e.g., we can use biological evolution as example, and evolutionary algorithms may be able to help stress testing other ML algorithms) and AI safety for evolutionary computation (e.g., to control systems that self-improve).

The first session, on **Safe Learning**, included two presentations. The first one, “Learning Modular Safe Policies in the Bandit Setting with Application to Adaptive Clinical Trials” by Hossein Aboutalebi, Doina Precup and Tibor Schuster, presented adaptive approaches using one-armed bandit machine that go quicker than Randomized Controlled Trials (RCTs). The approach includes a modular regret definition – building on modular ideas from software engineering. The second paper, “Metric Learning for Value Alignment” by Andrea Loreggia, Nicholas Mattei, Francesca Rossi and Kristen Brent Venable works on the concept of ethically bounded AI, using CP-nets to model preferences and ethical priorities. CP-nets are used as a way of expressing preferences – the net expands into a partial order, which may work well for transitive preference but may be computationally costly.

The second session in the morning, dealt with **Reinforcement Learning Safety,** and included for presentations. The first one, “Penalizing side effects using stepwise relative reachability” by Victoria Krakovna, Laurent Orseau, Miljan Martic and Shane Legg, try to avoid undesirable side effects that are not in the reward function, for cases where the environment is more complex than imagined. They use the box environment, and other examples from recent DeepMind papers. The second paper, Conservative Agency by Alexander Turner, Dylan Had-field-Menell and Prasad Tadepalli present attainable utility preservation, a way of reward specification that penalises a decrease in the ability to do (as yet) unspecified actions. The third paper, Modeling AGI Safety Frameworks with Causal Influence Diagrams, by Tom Everitt, Ramana Kumar, Victoria Krakovna and Shane Legg introduce causal inference diagrams (CID) for the analysis of incentives in AI systems. A CID is a CG with decision nodes whose values are determined by a policy, and used to optimise the value of reward (utility) nodes. The fourth paper, Detecting Spiky Corruption in Markov Decision Processes by Alok Singh, Jason Mancuso, David Lindner and Tomasz Kisielewski, uses Lipschitz models and violations to model ‘bad’ states (novelty is spyky).

After lunch, we resumed with an invited talk by Shlomo Zilberstein (University of Massachusetts Amherst, USA), who gave a presentation about AI Safety Based on Competency Models. He distinguishes safety when assumptions (implicit or explicit) are satisfied, and when they are violated, and recognises the challenge of incomplete and inaccurate model of the environment. He presents a particular approach using separate decision models (POMDPs) on a pairwise basis and instantiates them at run time. Competency models are used to deal with difficult situations, and reason about when the system needs help from people.

The third session focused on **Safe Autonomous Vehicles**. The first presentation, On the Susceptibility of Deep Neural Networks to Natural Perturbations, by Mesut Ozdag, Sunny Raj, Steven L. Fernandes, Alvaro Velasquez, Laura Pullum and Sumit Kumar Jha, discusses experiments trying to ‘defeat’ image classifiers and research on this kind of perturbations is useful to help develop defences against fog and other natural perturbations. The second presentation, Managing Uncertainty of AI-based Perception for Autonomous Systems. Maximilian Henne, Adrian Schwaiger and Gereon Weiss, focuses on real cases such as vehicle perception, with a state space that is too large for formal verification with traditional methods. Predictions are often over-confident – the softmax value cannot be treated as a probability, and they suggest the use of Bayesian Deep Learning. The third presentation, A Framework for Safety Violation Identification and Assessment in Autonomous Driving by Lukas Heinzmann, Sina Shafaei, Mohd Hafeez Osman, Christoph Segler and Alois Knoll, presents a framework for mapping safety-critical situations based on safety measures on RL scenarios for driving in the CARLA simulator. Several safety critical situations about machine learning agents can be identified.

The last invited talk of the day was given by Yang Liu (WeBank, China) about User Privacy, Data Confidentiality and AI Safety in Collaborative Learning, they present scenarios where they distribute models to users – to allow training on user data, but they ask the question of whether the updates secure and privacy preserving. They use a form of federated learning, and use encryption to protect intermediate results. He also presented an open source version of the system, known as WeBank FATE and some pointers and indications about standards.

The fourth and last session dealt with AI **Value Alignment, Ethics and Bias.** The first presentation**,** The Glass Box Approach: Verifying Contextual Adherence to Values by Andrea Aler Tubella and Virginia Dignum emphasised that high-level values have different interpretations in different contexts and cultures. One approach is the 'design for values' perspective. Their glass box approach, formalised in multi-modal logic, separates out the values into norms and then requirements. The second presentation, Requisite Variety in Ethical Utility Functions for AI Value Alignment by Nadisha-Marie Aliman and Leon Kester, discusses on the variety in embodied ethical utility functions, seen as a security issue, and the focus is on value alignment. They summarize variety-relevant background knowledge from neuroscience and psychology and present the design of approximate ethical goal functions. The third presentation, Slam the Brakes: Perceptions of Moral Decisions in Driving Dilemmas by Holly Wilson, Andreas Theodorou and Joanna Bryson, elaborates on the well-known trolley problem, with a simulator that shows some issues of the moral machine, with and without explanation. The fourth presentation, Understanding Bias in Datasets using Topological Data Analysis by Ramya Srinivasan and Ajay Chander, examines various stages of the AI pipeline, focusing on software engineering, using topological data analysis.

The day was completed by spotlight presentations and discussions around the following posters:

- Computational Strategies for the Trustworthy Pursuit and the Safe Modeling of Probabilistic Maintenance Commitments. Qi Zhang, Edmund Durfee and Satinder Singh
- Categorizing Wireheading in Partially Embedded Agents. Arushi Majha, Sayan Sarkar and Davide Zagami
- Adversarial Exploitation of Policy Imitation. Vahid Behzadan and William Hsu.
- The Challenge of Imputation in Explainable Artificial Intelligence Models. Muhammad Ahmad, Carly Eckert and Ankur Teredesai
- On the importance of system testing for assuring safety of AI systems. Franz Wotawa
- Towards Empathic Deep Q-Learning. Bart Bussmann, Jacqueline Heinerman and Joel Lehman
- Watermarking of DRL Policies with Sequential Triggers. Vahid Behzadan and William Hsu.

Second workshop day : Landscape (Aug 12)

The second-day workshop (AI Safety Landscape) sessions on August 12 were organized into by-invitation talks and panels with structured discussions. The by-invitation talks focused on diverse topics contributing to understand the AI Safety Landscape, in terms of their scientific and technical challenges, industrial and academic opportunities, as well as gaps and pitfalls.

The day started with an introduction by Cynthia Chen (University of Hong Kong), one of the workshop chairs, of the motivation and goals of progressing *Towards an AI Safety Landscape*. On behalf of the workshop chairs, Cynthia summarized the main motivation and objectives of the AI Safety Landscape initiative: get more consensus and focus on generally accepted

knowledge. She also presented the proposed Landscape categories. The chairs 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.

The first invited presentation, Creating a Deep Model of AI Safety Research, by Richard Mallah (Future of Life Institute), represented the Future of Life Institute (FLI), which fostered the creation of a Landscape of AI Safety and Beneficence Research for research contextualization and in preparation for brainstorming at the Beneficial AI 2017 conference at Asilomar. It has a strong focus on AI-based systems where the main concern is to ensure that machine intelligences, which become more and more general and broad in their capability, remain beneficial for the humanity. In this sense, both "AI" and "safety" cover very broad problems, including AGI and superintelligent agents as well as ethics and security. FLI's landscape is a tree-based graphical structure (but that doesn't exclude some other connections) accompanied by a paper (https://futureoflife.org/landscape/), covering four main areas, foundations – rational agency, decision theory, verification – provable implementations of AI/ML, validation – goal and specification alignment, security – active-managed biases & permissions, and control – monitoring, oversight, and deference. Richard also distinguished issues that are usually near-term, such as monitoring, fairness and mitigating bias, verified software, specified cost minimization, and fraud & abuse detection, and those that are longer-term, such as scalable oversight, value alignment, verified full stack AI, contextual awareness and security.

The second presentation, Towards a Framework for Safety Assurance of Autonomous Systems by John McDermid (University of York, and Director of the Lloyd's Register Foundation funded Assuring Autonomy International Programme, AAIP). His talk addressed the challenges of safety assurance of autonomous systems and proposes a novel framework for safety assurance that, inter alia, uses machine learning to provide evidence for a system safety case and thus enables the safety case to be updated dynamically as system behaviour evolves. AAIP develops a Body of Knowledge (BoK) intended to become a reference source on assurance and regulation of Robotics and Autonomous Systems (RAS). The talk covered Hollnagel's distinction between Safety-I (eliminating failures) and Safety-II (reinforcing right behaviour) and Ashmore et al's Emphasis on the ML Life Cycle, extending Hollnagel's framework highlighting the discrepancies between real world, world as imagined and world as observed.

These two talks were followed by a panel on The Challenge of Achieving Consensus, chaired by Xiaowei Huang with the speakers as discussants: Richard Mallah and John McDermid. During this session, Richard and John discussed their experience in the initiatives they lead to look for consensus in a related field, the challenges in AI Safety for getting such consensus? They also discussed to what extend we could get consensus in AI Safety, what priorities should be considered to get consensus for an AI Safety Landscape in the AI Safety field. Finally, they mentioned the kind of mechanisms they deem essential for finding consensus in the safety-critical systems domain considering AI and autonomy aspects.

The next session started with the presentation, AI Safety and The Life Sciences, by Gopal Sarma (Broad Institute of MIT and Harvard). Gopal discussed the need to consider Life Sciences in engineering future safe AI-based systems. He anticipated the narratives surrounding biotechnology controversies to become intertwined with concerns related to AI and AI safety.

From a public policy and public relations standpoint, this will create many novel challenges in crafting a set of national priorities that address both the concerns of elite scientists (such as the AI safety community) as well as the many fears the general public will have about the interplay between artificial intelligence and synthetic biology.

The second talk of the session, Formal Methods in Certifying Learning-Enabled Systems, by Xiaowei Huang (University of Liverpool) discussed the risk of using DNNs in safety-critical systems and the use of formal methods to guarantee robustness and safety in those systems. He summarized safety risks as related to robustness, generalisation, understanding, and interaction. He considers that current verification effort, especially for DNNs, is focused on robustness. He thinks we need to look at the other areas too! Also, he mentioned the need to develop better run-time monitoring and enforcement approaches for operational-time errors.

The final talk of the session, AI Safety and Evolutionary Computation, by Joel Lehman (Uber AI Labs), described the broad aspirations of evolutionary computation (EC), and the intersection of AI safety and evolutionary computation. Some communities within EC focus not on optimization of a fixed objective, but on understanding the algorithmic nature of evolution's divergent creativity, i.e. algorithms that are capable of continually innovating in an open-ended way. These kinds of evolutionary algorithms (EAs) offer a bottom-up path towards general artificial intelligence or AGI, known as AI Generating Algorithms (AI-GA), one where AGI emerges as a by-product of a larger open-ended creative project, as occurred in biological evolution.

The session was closed by a panel on The Need for Paradigm Change, chaired by Seán Ó hÉigeartaigh and with discussants: Gopal Sarma, Xiaowei Huang, Joel Lehman, Nadisha-Marie Aliman and Fredrik Heintz. This panel discussed how AI/ML/DL are stretching the (technical and non-technical) limits of the traditional system engineering disciplines in present-day intelligent systems, and more capable future AI-based systems. Discussants provided a view of the challenges to include new paradigms in AI Safety. They also discussed the priorities in terms of research and development should be considered to include new paradigms in AI Safety, as well as to what extent regulatory frameworks should be changed to tackle the challenges of safe AI-based intelligent autonomous systems.

The next session started with, AI Safety for Humans, an invited presentation by Virginia Dignum (University of Umeå). Virginia emphasised the socio-technical perspective of AI Safety. She looked at ways to ensure that behavior by artificial systems is aligned with human values and ethical principles. Given that ethics are dependent on the socio-cultural context and are often only implicit in deliberation processes, methodologies are needed to elicit the values held by designers and stakeholders, and to make these explicit leading to better understanding and trust on artificial autonomous systems. She particularly focused on the ART principles for AI: Accountability, Responsibility and Transparency, and emphasised ideas such as fair trade in AI and building ethical systems by design.

The next talk of the session, Towards Trustworthy Autonomous and Intelligent Systems, by Raja Chatila (Sorbonne University), focused on how to make autonomous systems trustworthy to reliably deliver the expected correct service. This must be the case even when components are imperfect or fail (e.g., an aircraft when one of the engines explodes). As decisions usually devoted to humans are being more and more delegated to machines, sometimes running

computational algorithms based on learning techniques using data, operating in complex and evolving environments, new issues have to be considered, such as lack of context and semantics. Raja discussed new technical and non-technical measures to be considered in the design process and in the governance of these systems. He emphasises the IEEE and EU work in this area, such as the dependability attributes (Availability, Reliability, Safety, Confidentiality, Integrity, Maintainability, Security) and standardisation (IEEE, P7000 Standardization Projects for Ethically Aligned Design).

The third and final talk of the session, AI Principles and Ethics by Design, by Jeff Cao (Tencent Research Institute) referred to Tencent RI's work, which is involved in transportation and healthcare sectors, where ethics is important. Jeff mentioned that there are three levels of AI safety: technical, physical and social safety. They have both found problems with e.g. Tesla systems, and hacked them. Tencent people have produced a research report on 'tech ethics'. He talks about ARCC principles: available, reliable, comprehensible, controllable. He stresses the need of multi-level governance: laws and regulations; industry self-regulation and; education and awareness raising.

The panel ending the session, Towards More Human-Centered and Ethics-Aware Autonomous Systems, chaired by Richard Mallah had Discussants: Virginia Dignum, Raja Chatila and Jeff Cao. This panel discussed which aspects of ethics and human-centred disciplines are of high priority when dealing with safety-critical AI-based systems. Developers and operators, at minimum – should have principles and guidelines for such organisations. We should start by looking at existing legal mechanisms. What are the incentives for an organisation to follow ethical guidelines? Positive differentiation!! Customer trust is another key differentiator; and this will influence success in the marketplace. But making things trustworthy may make them more expensive, so maybe there also needs to be an overarching regulatory framework. There was some discussion –and disagreement-- about whether safety can be ensured with systems that can explicitly reason or optimise for utility functions with ethical constraints or principles, but not designed for ethical principles (e.g., through norms). There was an agreement that a critical aspect is that humans usually disagree on values and terminology and quantities to express those values.

The last talk of the day was Specification, Robustness and Assurance Problems in AI Safety by Victoria Krakovna (Google DeepMind). She presented the DeepMind's categories for AI Safety, as a first (DeepMind) attempt to map the AI safety knowledge. This includes near and long-term AI safety issues. She discussed the three areas of technical AI safety: specification, robustness, and assurance. Specification captures issues such as Goodhart's law, specification gaming (tinyurl.com/specification-gaming), side effects and reward tampering. Robustness covers safe exploration, distributional shift, etc., and assurance includes interpretability, privacy, off-switch, containment, etc. Particular focus has been put on specification (ideal, design and revealed specification) and some examples are illustrated with the safety grid worlds. DeepMind feels these three areas cover a sufficiently wide spectrum to provide a useful categorisation for ongoing and future research. DeepMind made progress in some of these areas but many open problems remain.

Finally, to close the day and the workshop, there was a longer panel on Building an AI Safety Landscape: Perspectives and Future Work, chaired by John McDermid with discussants: Richard Mallah, Seán Ó hÉigeartaigh, Xiaowei Huang and Andrea Aler Tubella. This panel

focused on the questions of gaining consensus, terminology, and connections for building the landscape – safety engineering, machine learning, legal and ethical expertise, cognitive science, etc. There was some discussion and disagreement about the relevance of issues such as AI boxing or off-switch (we may already be lost if we need this). The debate moved towards recognising what areas may be missing or should be better emphasised in the landscape. It was suggested that formal methods expertise is needed; as well as to understand human factors; and to consider long-term monitoring of systems and their unintended effects. There is a need to consider certification. DeepMind team's work – more in foundations & specification and modelling – and some work on verification needs scaling from grid worlds and theoretical scenarios to real-world cases. Legal issues, open vs closed worlds. Importance of system modelling (both agents and environment) – we need to include this aspect in the landscape.

Acknowledgements

We thank all those who submitted papers to AISafety 2019 and congratulate the authors whose papers and posters were selected for inclusion into the workshop program and proceedings.

We specially thank our distinguished PC members, for reviewing the submissions and providing useful feedback to the authors:
• Stuart Russell, UC Berkeley, USA
• Victoria Krakovna, Google DeepMind, UK
• Peter Eckersley, Partnership on AI, USA
• Riccardo Mariani, Intel, Italy
• Brent Harrison, University of Kentucky, USA
• Siddartha Khastgir, University of Warwick, UK
• Emmanuel Arbaretier, Apsys-Airbus, France
• Martin Vechev, ETH Zurich, Switzerland
• Sandhya Saisubramanian, University of Massachusetts Amherst, USA
• Alessio R. Lomuscio, Imperial College London, UK
• Mauricio Castillo-Effen, Lockheed Martin, USA
• Yi Zeng, Chinese Academy of Sciences, China
• Brian Tse, Affiliate at University of Oxford, China
• Sandeep Neema, DARPA, USA
• Michael Paulitsch, Intel, Germany
• Elizabeth Bondi, University of Southern California, USA
• Hélène Waeselynck, CNRS LAAS, France
• Rob Alexander, University of York, UK
• Vahid Behzadan, Kansas State University, USA
• Simon Fürst, BMW, Germany
• Chokri Mraidha, CEA LIST, France
• Fuxin Li, Oregon State University, USA
• Francesca Rossi, IBM and University of Padova, Italy
• Ian Goodfellow, Google Brain, USA
• Yang Liu, Webank, China
• Ramana Kumar, Google DeepMind, UK
• Javier Ibañez-Guzman, Renault, France

• Dragos Margineantu, Boeing, USA
• Joanna Bryson, University of Bath, UK
• Heather Roff, Johns Hopkins University, USA
• Raja Chatila, Sorbonne University, France
• Hang Su, Tsinghua University, China
• François Terrier, CEA LIST, France
• Guy Katz, Hebrew University of Jerusalem, Israel
• Alec Banks, Defence Science and Technology Laborato-ry, UK
• Gopal Sarma, Emory University, USA
• Lê Nguyên Hoang, EPFL, Switzerland
• Roman Nagy, BMW, Germany
• Nathalie Baracaldo, IBM Research, USA
• Toshihiro Nakae, DENSO Corporation, Japan
• Peter Flach, University of Bristol, UK
• Richard Cheng, California Institute of Technology, USA
• José M. Faria, Safe Perspective, UK
• Ramya Ramakrishnan, Massachusetts Institute of Tech-nology, USA
• Gereon Weiss, Fraunhofer ESK, Germany
• 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 Valèn-cia, Spain
• Seán Ó hÉigeartaigh, University of Cambridge, UK
• Richard Mallah, Future of Life Institute, USA
As well as the additional reviewers:
• George Amariucai, Kansas State University, USA
• Neale Ratzlaff, Oregon State University, USA

We thank Joel Lehman, Shlomo Zilberstein, Yang Liu, Richard Mallah, John McDermid, Gopal Sarma, Xiaowei Huang, Virginia Dignum, Raja Chatila, Jeff Cao and Victoria Krakovna for their interesting talks on the current challenges of AI safety.

We would like to specially thank our sponsors, which funded the Best Paper Award and the video-recording of the AI Safety Landscape sessions:

• Assuring Autonomy International Programme (AAIP).

• Partnership on AI.

• The Centre for the Study of Existential Risk (CSER).

Finally, yet importantly, we thank the IJCAI-19 organization for providing an excellent framework for AISafety 2019.