Towards an AI Safety Landscape

An Overview

https://www.ai-safety.org/

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

Table of Contents

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 sidelined 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).

Why do we need an AI Safety Landscape?

Despite the increasing number of researchers and practitioners worldwide working on AI Safety and the ubiquitous need of safe intelligent autonomous systems in our society, this field was only relatively recently recognized as a legitimate domain that is stretching the limits of the broader and more traditional discipline of safety engineering.

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, as a key step towards developing an AI Safety body of knowledge. Recognizing the need of an AI Safety Landscape, is pivotal because of the following reasons:

- **More consensus is crucial:** Achieving more consensus in terminology and meaning is a key step towards aligning the understanding of engineering and socio-technical concepts, existing/available theory and technical solutions and gaps in the diversity of AI safety. Increasing conceptual consensus has the power of accelerating the mutual understanding of the multiple disciplines working on how to actually create, test, deploy, operate and evolve safe AI-based systems, as well as ensuring awareness of broader strategic, ethical and policy issues. Also in any consensus there are many trade-offs and compromises we must make.
- **Focus on generally accepted knowledge:** "Generally accepted" means 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. We also expect to be somewhat forward-looking in the different interpretations by taking into consideration not only what is generally accepted today, but we expect will be generally accepted in a longer timeframe, with the dawn of systems whose cognitive capabilities approach those of humans.

What concrete aspects do we target?

The main goal of this initiative is to 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.

The core expected outcome is a single document identifying and describing a landscape of AI Safety, the set of subfields that must be knowledgeable not only in the engineering discipline but also in other socio-technical disciplines, including an outline of ***needs, challenges, practices and gaps***. The goal of this initiative is not to inventory everything related to AI Safety, but the core knowledge.

One important ambition of this initiative is to 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.

As a starting point for an efficient discussion, we propose **A preliminary set of landscape Categories**, which could be refined during the process of consensus and development of an AI Safety Landscape.

Related initiatives

There are a number of relevant initiatives aiming to define a map of knowledge in AI Safety or very close fields:

The **Future of Life Institute (FLI)** 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. This document borrows heavily from MIRI's agent foundations technical agenda [1], Amodei et al.'s concrete problems review [2], MIRI's machine learning technical agenda [3], FLI's previous research priorities document [4], and adds a large number of nodes from other literature as well. The document is available at:

https://futureoflife.org/landscape/ResearchLandscapeExtended.pdf

Another initiative is driven by the **Assuring Autonomy International Programme (AAIP)** to develop a Body of Knowledge (BoK) intended, in time, to become a reference source on assurance and regulation of Robotics and Autonomous Systems (RAS). This initiative is using the feedback from industry, regulators and researchers to outline the structure, specification and scope of this BoK. A draft document, which aims to constantly mature and evolve, can be found at:

https://www.york.ac.uk/assuring-autonomy/research/body-of-knowledge/

While these initiatives underlie the most systematic works, Ortega et al (**DeepMind**) also published an article proposing a straightforward structure of the technical AI safety field. The article is available at:

https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1

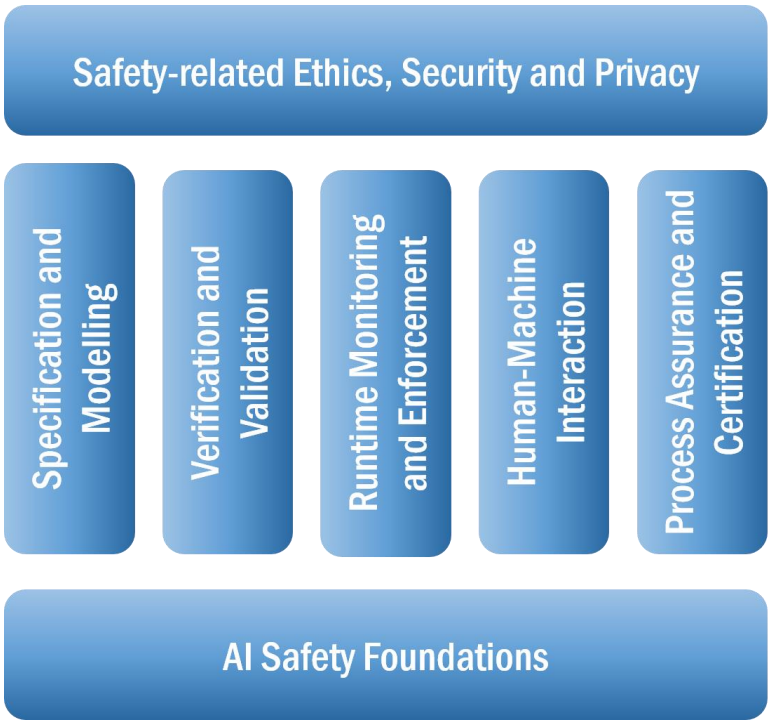
We intend to build on top of these efforts in the quest of consensus to build a common landscape for AI Safety.

A preliminary set of landscape Categories

The figure below proposes a draft schema of seven categories for classifying and discussing AI Safety knowledge. Each of these categories interacts and depends on the others. The purpose of this classification is to promote structured discussions towards a consistent view of AI Safety, as well as to understand the place of this field with respect to other disciplines in computer science, project management or social sciences. 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.



AI Safety Foundations

This category covers a number of foundational concepts, characteristics and problems related to AI safety that need special consideration from a theoretical perspective. This includes concepts such as uncertainty, generality or value alignment, as well as characteristics such autonomy levels, safety criticality, types of human-machine and environment-machine interaction. This group intends to collect any cross-category concerns in AI Safety.

Specification and Modelling

The main scope of this category is on how to describe needs, designs and actual operating safety-critical systems from different perspectives (technical concerns) and abstraction levels. This includes the specification and modelling of risk management properties (e.g., hazards, failures modes, mitigation measures), as well as safety-related requirements, training, behavior or quality attributes in AI-based systems.

Verification and Validation

This category concerns design-time approaches to ensure that an AI-based system meets its requirements (verification) and behaves as expected (validation). The range of techniques covers any formal/mathematical, model-based simulation or testing approach that provides evidence that an AI-based system satisfies its defined (safety) requirements and does not deviate from its intended behavior and causes unintended consequences.

Runtime Monitoring and Enforcement

The increasing autonomy and learning nature of AI-based systems is particularly challenging for their verification and validation (V&V), due to our inability to collect an epistemologically sufficient

quantity of evidence to ensure correctness. Runtime monitoring is useful to cover the gaps of design-time V&V by observing the internal states of a given system and its interactions with external entities, with the aim of determining system behavior correctness or predicting potential risks. Enforcement deals with runtime mechanisms to self-adapt, optimize or reconfigure system behavior with the aim of supporting fallback to a safe system state from the (anomalous) current state.

Human-Machine Interaction

As autonomy progressively substitute cognitive human tasks, some kind of human-machine interaction issues become more critical, such as the loss of situation awareness or overconfidence. Other issues include: collaborative missions that need unambiguous communication to manage self-initiative to start or transfer tasks; safety-critical situations in which earning and maintaining trust is essential at operational phases; or cooperative human-machine decision tasks where understanding machine decisions are crucial to validate safe autonomous actions.

Process Assurance and Certification

Process Assurance is the planned and systematic activities that assure system lifecycle processes conform to its requirements (including safety) and quality procedures. In our context, it covers the management of the different phases of AI-based systems, including training and operational phases, the traceability of data and artefacts, and people. Certification implies a (legal) recognition that a system or process complies with industry standards and regulations to ensure it delivers its intended functions safely. Certification is challenged by the inscrutability of AI-based systems and the inability to ensure functional safety under uncertain and exceptional situations prior to its operation.

Safety-related Ethics, Security and Privacy

While these are quite large fields, we are interested in their intersection and dependencies with safety issues. Ethics becomes increasingly important as autonomy (with learning and adaptive abilities) involves the transfer of safety risks, responsibility, and liability, among others. AI-specific security and privacy issues must be considered with regard to its impact on safety. For example, malicious adversarial attacks can be studied with focus on situations that compromise systems towards a dangerous situation.

Comparison with related initiatives

As a preliminary justification of the proposed landscape scheme, we outline a comparison matrix of the proposed categorization with the structure of other **Related initiatives**.

This AI Safety Landscape	FLI-coordinated AI Safety Landscape	Assured Autonomy (AAIP) Body of Knowledge	DeepMind's AI Safety Technical Areas
• AI Safety Foundations	• Foundations		
• Specification and Modelling	[Specification and Modeling is part of other categories]	• Defining required behavior (it also includes requirement validation)	• Specification: Design Specification: Emergent
• Verification and Validation	• Verification • Validation	• Implementation to provide the required behaviour	• Robustness: Prevention and Risks
• Runtime Monitoring and Enforcement	• Control	• Understanding and controlling deviations from required behaviour	• Assurance: Monitoring Assurance: Enforcement

● Human-Machine Interaction	[Human considerations are part of other categories]	[Human considerations are part of other categories]	[Human considerations are part of other categories]
● Process Assurance and Certification		● Gaining approval for operation	
● Safety-related Ethics, Security and Privacy	[Ethics is part of (most of) other categories] ● Security (including privacy)		[Ethics and Security and issues are part of other categories]

Please consider this comparison as a simplified comparison that needs further study.

The ***FLI-coordinated AI Safety Landscape*** has a strong focus on AI-based systems where the main concern is to ensure that machine intelligences, which becomes 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. This landscape does not consider some (near-term) organizational issues in AI safety, such as process assurance and certification for operation.

The ***AAIP's (in-progress) Body of Knowledge*** focuses on Robotics and Autonomous Systems (RAS), where AI and ML are specific (but still important) topics of the initiative. This map has a strong emphasis on safety engineering, including gaining approval for operation from regulatory/certification entities. The whole approach is organized around categories that define assurance (or regulatory) objectives, contextual description and approaches for demonstration. It does not consider (yet) any ethical issues or systems whose cognitive capabilities approach or overcome those of humans.

The ***DeepMind's technical AI safety*** categorization offers a strong approach, as a first (DeepMind) attempt to map the AI safety knowledge, but is light on detail. The categories do not explicitly mention any consideration for process assurance and certification.

None of the related initiatives considers human-in-the-loop autonomy aspects to be a separate category. In our landscape, we consider this area of special importance so that we treat it as a separate area.

Way of Working

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). During these workshops, there will be slots to both get the input of the general community and by-invitation talks and panels.

Plan for AISafety 2019

The first meeting, which will take place at AISafety 2019, focuses on the following specific objectives:

1. Get preliminary agreement on the scope of the AI Safety field.
2. Outlining a straightforward and generally accepted high-level categorization of the AI Safety field.
3. Understanding the content, boundaries and connections between the proposed categories.
4. Plan follow-up actions to ensure effectiveness and coordination with other relevant initiatives.

AISafety 2019 is planned as a two-days workshop with general AI Safety topics in the first day and AI Safety Landscape talks and panels during the second day. In additional to general topics during the

first day, the workshop will also include position paper presentations specifically focused on the proposed AI Safety Landscape, open to the community through the Call for Contributions.

The AI Safety Landscape sessions, during the second day, will be structured in by-invitation pitches for each of the landscape Categories and panels with structured discussions.

The by-invitation pitches shall focus on the above-mentioned specific objective 3, but shall also provide inputs to the other specific objectives. While it is clear that a first meeting is not enough to discuss much details of ***each category***, each pitch shall outline a preliminary view on its ***scientific and technical challenges, industrial and academic opportunities, as well as gaps and pitfalls***.

The panel sessions shall focus on the aforementioned specific objectives 1, 2 and 4. Like for the pitches, getting onsite consensus is an ambitious task, so the main aim is to get as much inputs as possible to follow-up any related action in a well-informed way.

References

[1] Nate Soares and Benja Fallenstein. Aligning Superintelligence with Human Interests: A Technical Research Agenda. Tech. rep. Forthcoming 2017 in ”The Technological Singularity: Managing the Journey” Jim Miller, Roman Yampolskiy, Stuart J. Armstrong, and Vic Callaghan, Eds. Berkeley,CA: Machine Intelligence Research Institute. Machine Intelligence Research Institute, 2014. url: http://intelligence.org/files/TechnicalAgenda.pdf.

[2] Dario Amodei et al. “Concrete Problems in AI Safety”. In: CoRR abs/1606.06565 (2016). url: http://arxiv.org/abs/1606.06565.

[3] Jessica Taylor et al. Alignment for Advanced Machine Learning Systems. Tech. rep. Machine Intelligence Research Institute, 2016. url: https://intelligence.org/files/AlignmentMachineLearning.pdf

[4] Stuart Russell, Daniel Dewey, and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence”. In: AI Magazine 36.4 (2015). url: http://futureoflife.org/data/documents/research_priorities.pdf.