Creating a Deep Model of AI Safety Research

Richard Mallah
Director of AI Projects, Future of Life Institute

August 2019
Macao
IJCAI AI Safety Workshop
Agenda of This Talk

• Introduction to FLI’s AI Safety Research Landscape
  • Motivations, Format, Structure, Content
  • Tying Together Near-Term and Longer-Term Safety
  • Drilldowns Into Tree Branches
• Working Toward Consensus in the Creation Process
  • Stages, Feedback, Disagreements, Compromises
• Benefits of the Landscape Creation Process
This document borrows heavily from MIRI’s agent foundations technical agenda [403], Amodei et al.’s concrete problems review [11], MIRI’s machine learning technical agenda [437], and FLI’s previous research priorities document [371]. This document was written by Richard Mallah and was last edited January 2, 2017. It reflects feedback from Abram Demski, Andrew Critch, Anthony Aguirre, Bart Selman, Jan Leike, Jessica Taylor, Max Tegmark, Meia Chita-Tegmark, Nate Soares, Paul Christiano, Roman Yampolskiy, Stuart Armstrong, Tsvi Benson-Tilsen, Vika Krakovna, and Vincent Conitzer. (ver. 0.41)
Overview of the Landscape

https://futureoflife.org/landscape/

Foundations – rational agency, decision theory
Verification – provable implementations of AI/ML
Validation – goal and specification alignment
Security – active-managed biases & permissions
Control – monitoring, oversight, and deference
3 Verification

3.1 Formal Software Verification

3.1.1 Verified Component Design Approaches
3.1.2 Adaptive Control Theory
3.1.3 Verification of Cyberphysical Systems
3.1.4 Making Verification More User Friendly

3.2 Automated Vulnerability Finding

3.3 Verification of Intelligent Systems

3.3.1 Verification of Whole AI Systems
3.3.2 Verification of Machine Learning Components

3.4 Verification of Recursive Self-Improvement

3.5 Implementation Testing

4 Validation

4.1 Averting Instrumental Incentives

4.1.1 Error-Tolerant Agent Design
4.1.2 Domesticity

4.1.2.1 Impact Measures

4.1.2.1.1 Impact Regularizers

4.1.2.2 Defined Impact Regularizers

Source: Mallah

https://futureoflife.org/landscape/
2.2.2.3 Safety Technique Awareness

As a system gains skill in software development and algorithms research, explicit modeling and usage of the variety of safety techniques listed in this document can help it to improve its overall robustness and goal stability. See section **Ethical Motivation** as it will need to not only understand these techniques but want to apply them. See section **Meta-reasoning** as reasoning about the agent’s internal processes with cognizance of these techniques may aid long-term robustness.

2.2.3 Goal Stability

The maintenance of stability in the objectives of an advanced agent is challenging. See section **Meta-reasoning** types of which enable an agent to choose to e.g. avoid wireheading. Methods for safer self-modification can be used to help maintain goal stability in systems that self-improve. See section **Vingean Reflection** which is key to this. See section **Avoiding Reward Hacking** which is a crucial prerequisite to goal stability. See section **Corrigibility** for which goal stability is a prerequisite. When a static objective is acceptable, e.g. for some for short-to-medium-lived agents, one would want mechanisms to maintain the stability of the overall goals while providing flexibility as to situationally appropriate sub-objective priority. See section **Multiobjective Optimization** for this purpose, as multiobjectively optimized ensembles of subobjectives containing specifically formulated goal stability subobjectives might prevent reward function modification. See section **Degree of Value Evolution** which considers why one might want evolution of goals and how that might be managed.

2.2.3.1 Nontransitive Options

Whether individual operators, an aggregation of values of a large number of humans, or potentially the agent itself can have nontransitive preferences, where cycles or loops of comparative preferences occur. While some approaches seek to eliminate such a situation before it arises, more fault-tolerant approaches will attempt to handle such cycles gracefully, and this is an open area of research. It is
Tying Together Nearer-Term & Longer-Term Safety

**Nearer-Term**
- Monitoring
- Fairness & Mitigating Bias
- Verified Software
- Specified Cost Minimization
- Fraud & Abuse Detection

**Longer-Term**
- Scalable Oversight
- Value Alignment
- Verified Full Stack AI
- Contextual Awareness
- Security
Validation

- Avoiding Reward Hacking
- Averting Instrumental Incentives
- Increasing Contextual Awareness
- Value Alignment

Source: Mallah
https://futureoflife.org/landscape/
Validation / Avoiding Reward Hacking
Validation / Averting Instrumental Incentives
Control / Scalable Oversight

Scalable Oversight

Techniques to scale human operator oversight of an advanced AI that does a very high volume of very complex actions.

To scale oversight, it is desirable to ensure safe behavior of an agent even if given only limited access to its true objective function (Christiano 2015f). Methods for efficiently scaling up the ability of human overseers to supervise machine learning systems in scenarios where human feedback is expensive are promising, if early (Amodei et al. 2016). One potential approach is semisupervised reinforcement learning, where the agent sees the reward signal for only a small subset of steps or trials (Amodei et al. 2016, Christiano 2016).

Source: https://futureoflife.org/landscape/
Disagreements on Where Things Should Go
Some Received Suggestions – Accepted and Rejected

• We really need a top-level Foundations branch
• Monitoring and Informed Oversight should be the same
• Logical Uncertainty should fall under Decision Under Bounded Computational Resources
• Psychological Analogs and Increasing Contextual Awareness should go under Foundations of Rational Agency
• Consistent Decision Making, Foundations of Rational Agency, Avoiding Reward Hacking, and Averting Instrumental Incentives should all fall under Control
• Much disagreement on the relative hierarchy levels of Corrigibility, Computational Deference, Computational Humility, and Utility Indifference
Benefits of the Landscape Creation Process

- Prompts a more comprehensive way of thinking about the space
  - By its creators and its consumers
- Connects previous landscapes, agendas, and surveys together
  - And connects in additional relevant research
- Is a good process for explorations of unknown unknowns
  - Uncovers one’s own blind spots and those of the community
  - Finds divergent perspectives, identifying needed research
- Organizes the space well for particular constituencies
- Each landscape and each iteration brings the widening community closer to consensus, even if it can never be fully reached
richard@futureoflife.org