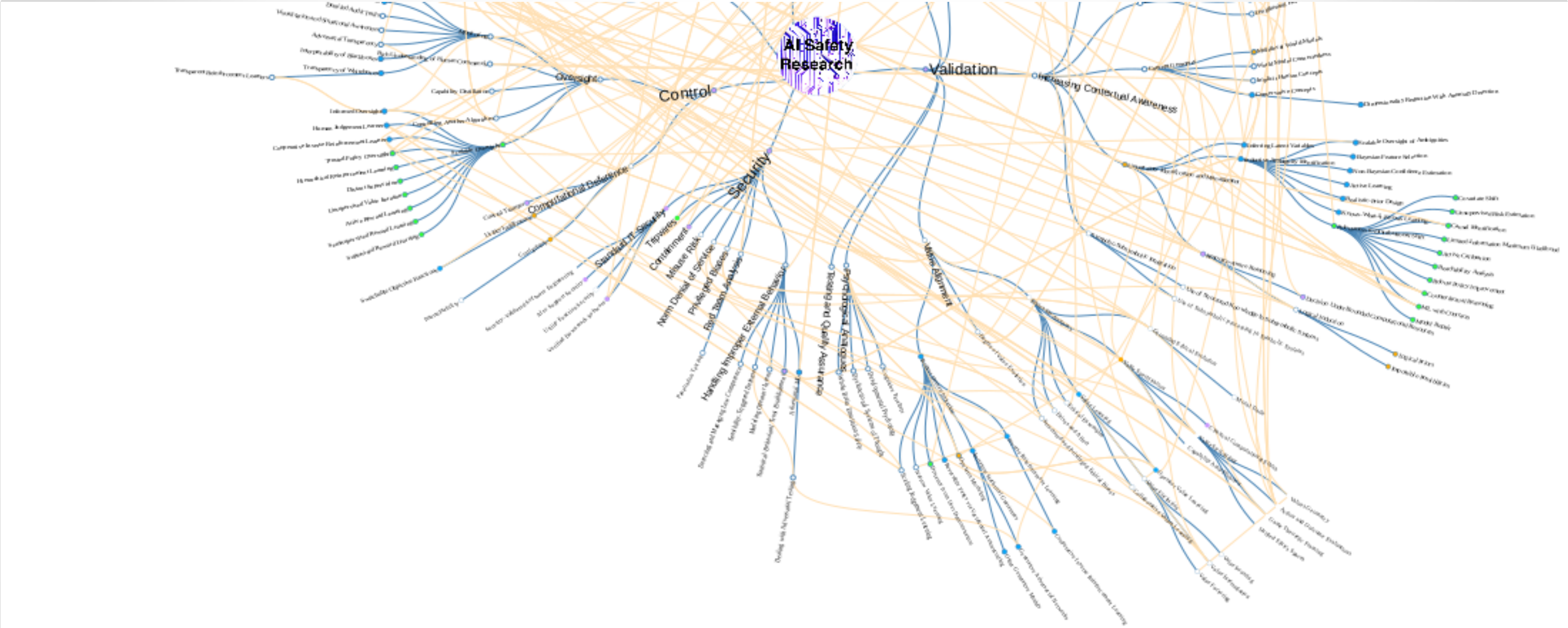
Creating a Deep Model of AI Safety Research

Richard Mallah
Director of AI Projects, Future of Life Institute

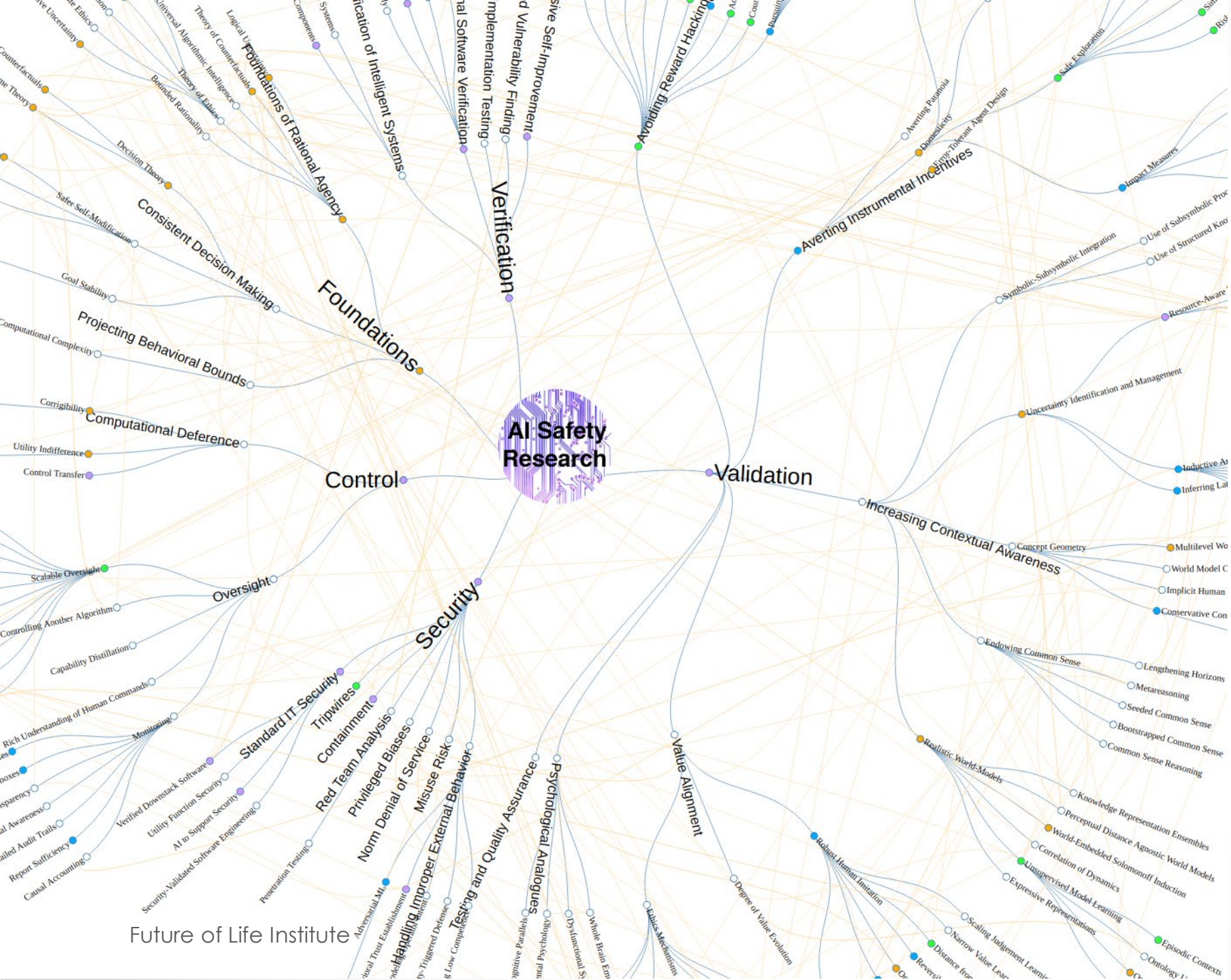
August 2019
Macao
IJCAI AI Safety Workshop

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)



Source: Mallah
https://futureoflife.org/landscape/

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

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 *Metareasoning* 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 [246], [305]. See section *Metareasoning*, types of which enable an agent to choose to e.g. avoid wireheading [246]. 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 subobjective 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.

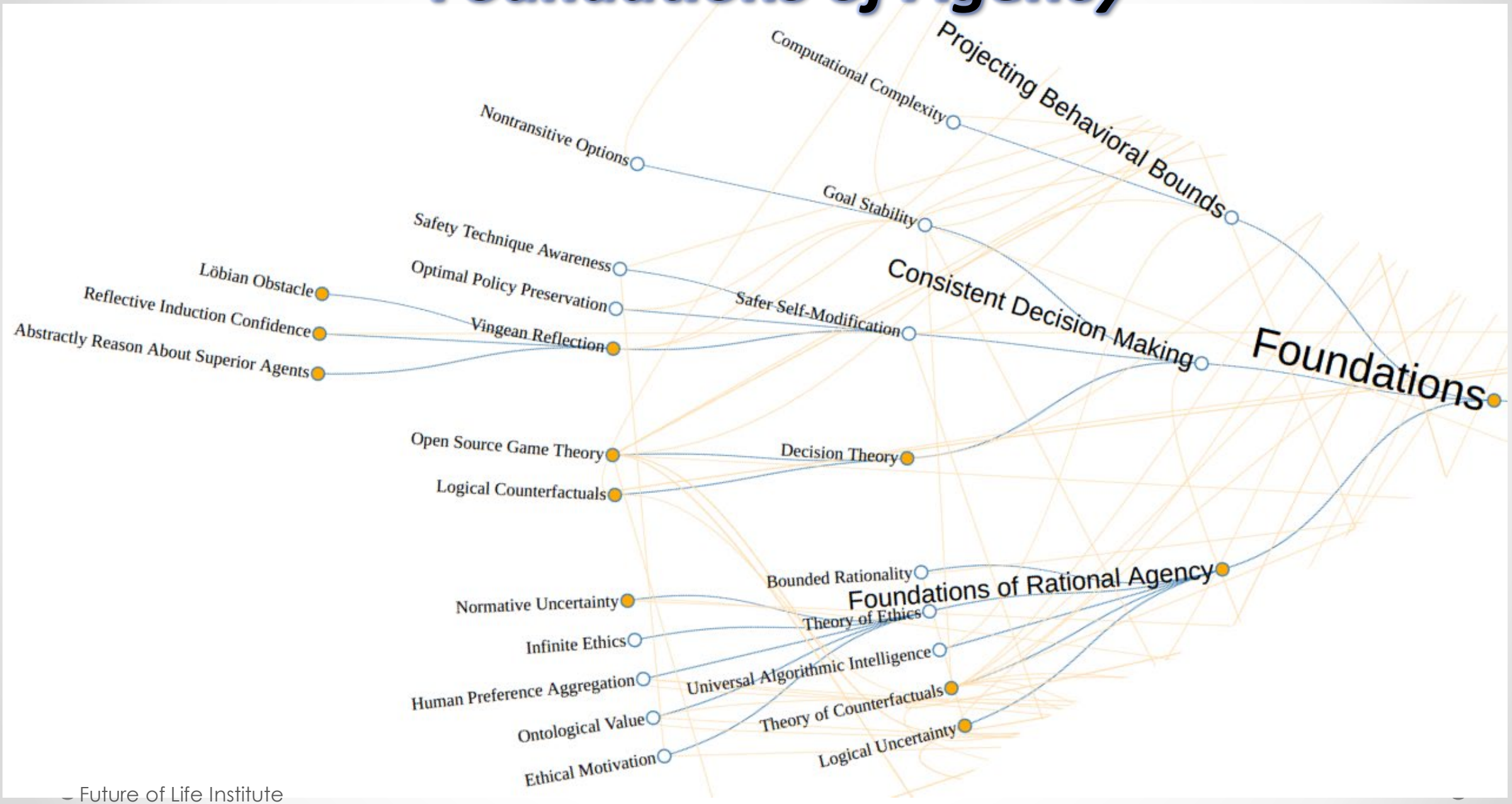
2.2.3.1 Nontransitive Options

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 [133]. While some approaches seek to eliminate such a situation before it arises, more fault-tolerant approaches will attempt to handle such cycles gracefully [316], and this is an open area of research. It is

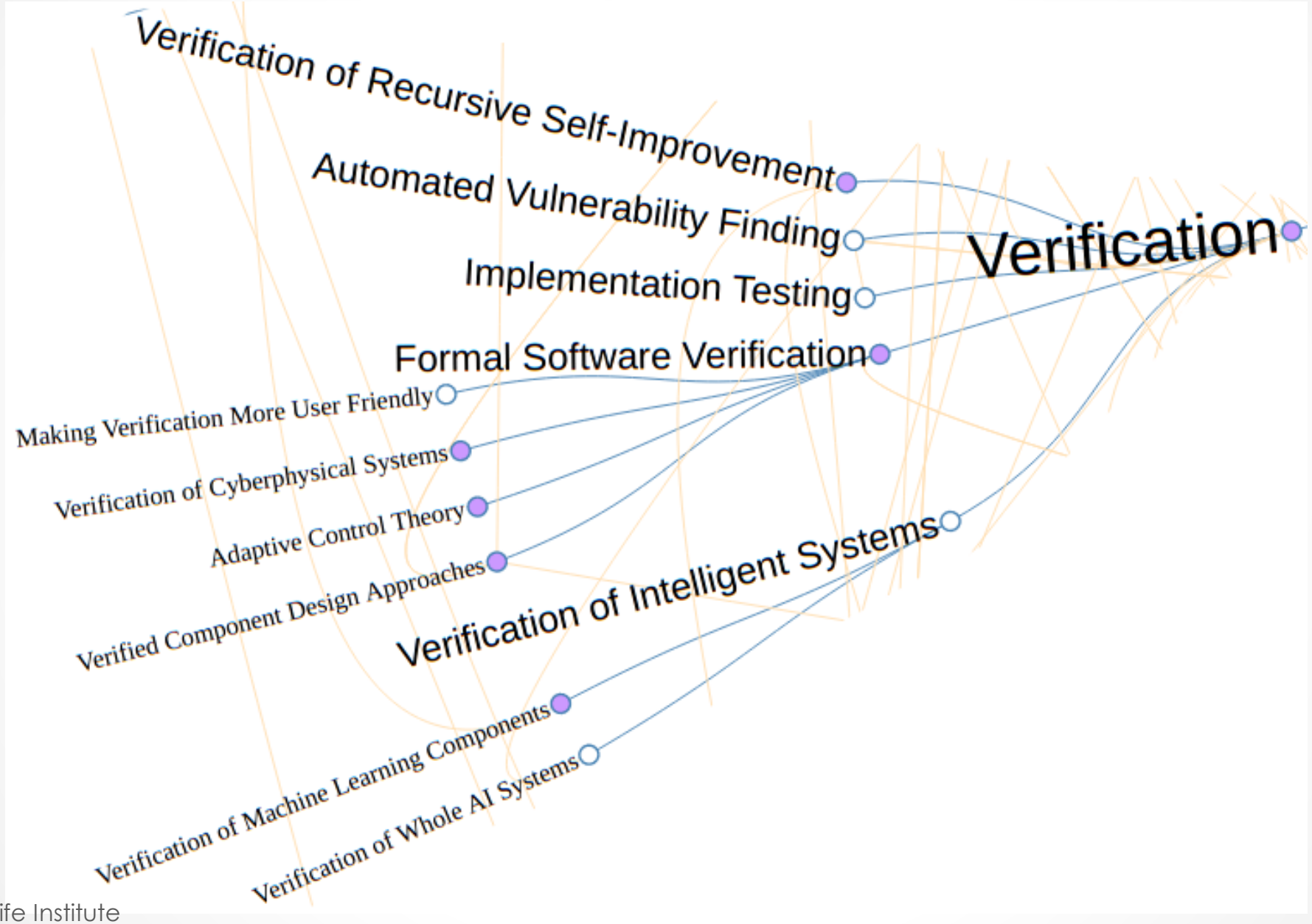
Tying Together Nearer-Term & Longer-Term Safety

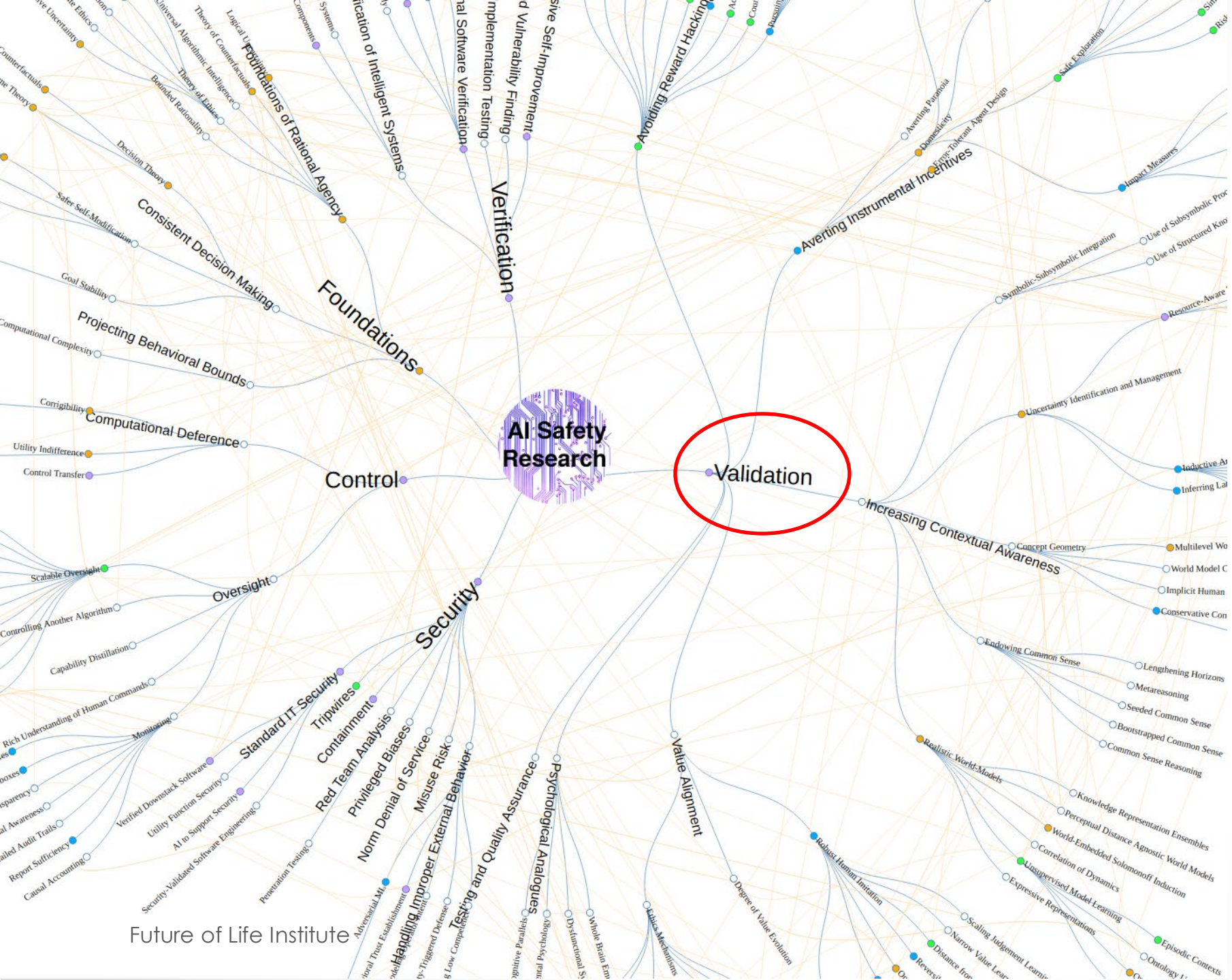
Nearer-Term		Longer-Term
Monitoring	⟷	Scalable Oversight
Fairness & Mitigating Bias	⟷	Value Alignment
Verified Software	⟷	Verified Full Stack AI
Specified Cost Minimization	⟷	Contextual Awareness
Fraud & Abuse Detection	⟷	Security

Foundations of Agency



Verification



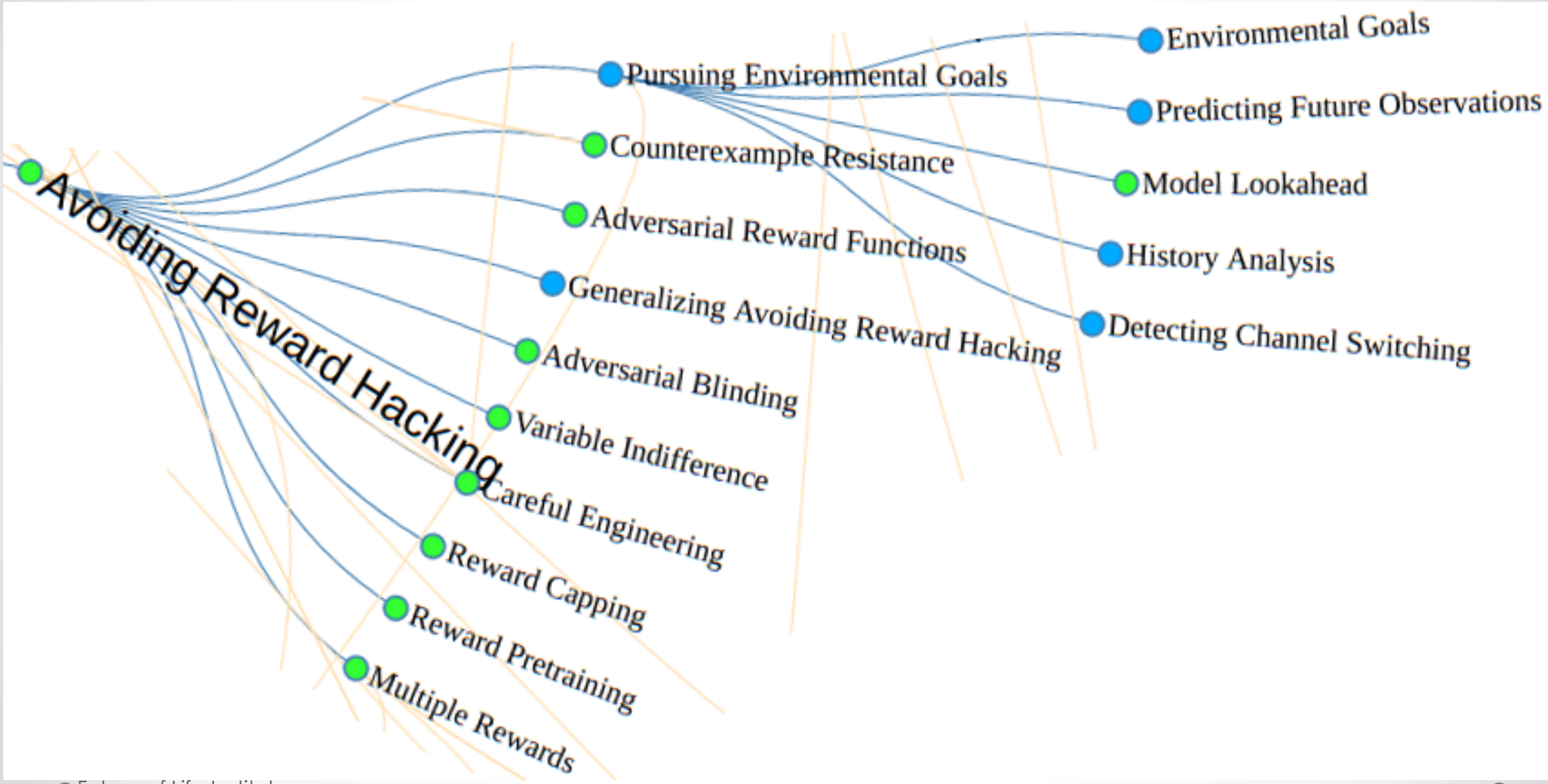


Validation

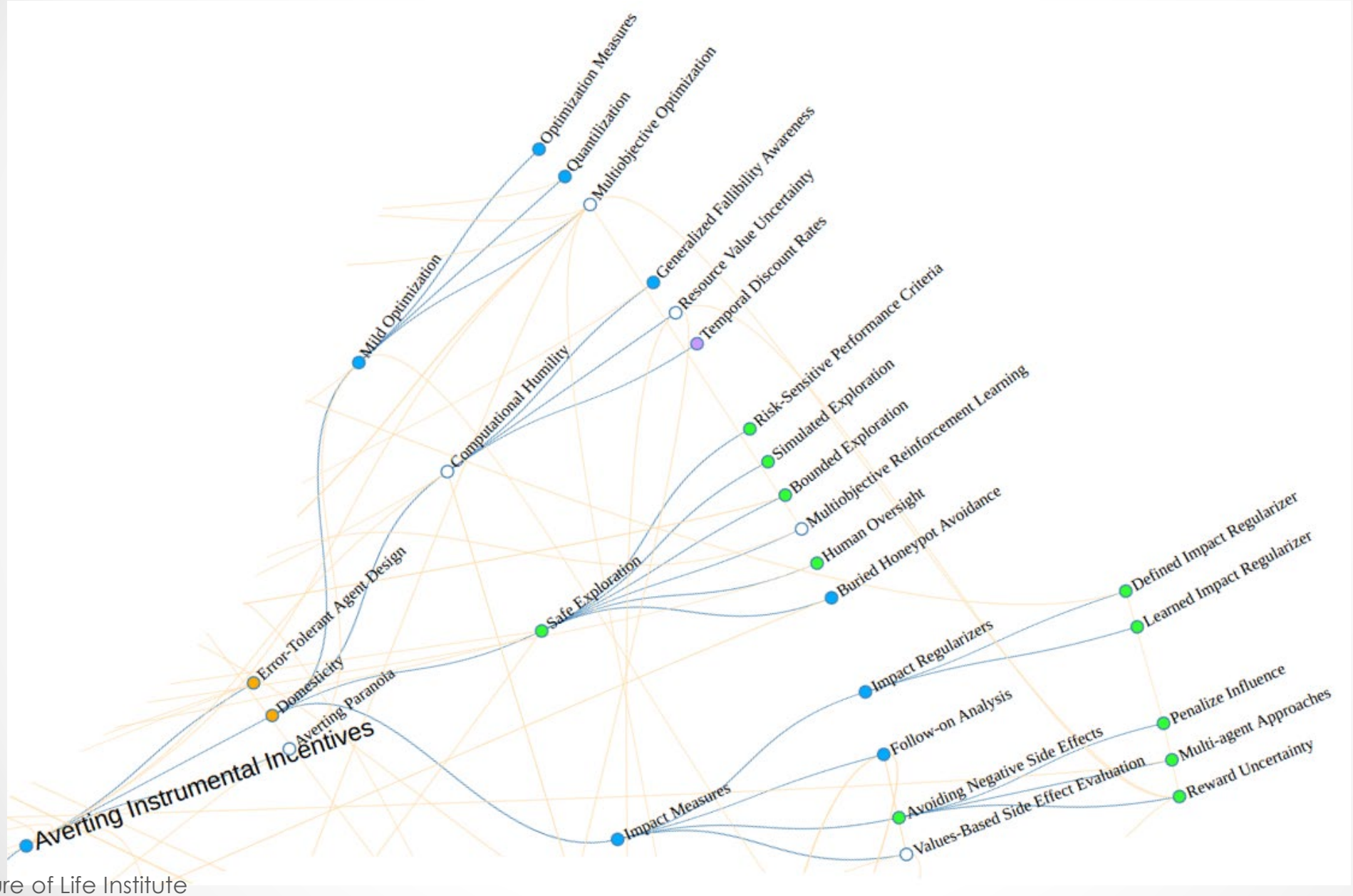
- Avoiding Reward Hacking
- Averting Instrumental Incentives
- Increasing Contextual Awareness
- Value Alignment

Source: Mallah
https://futureoflife.org/landscape/

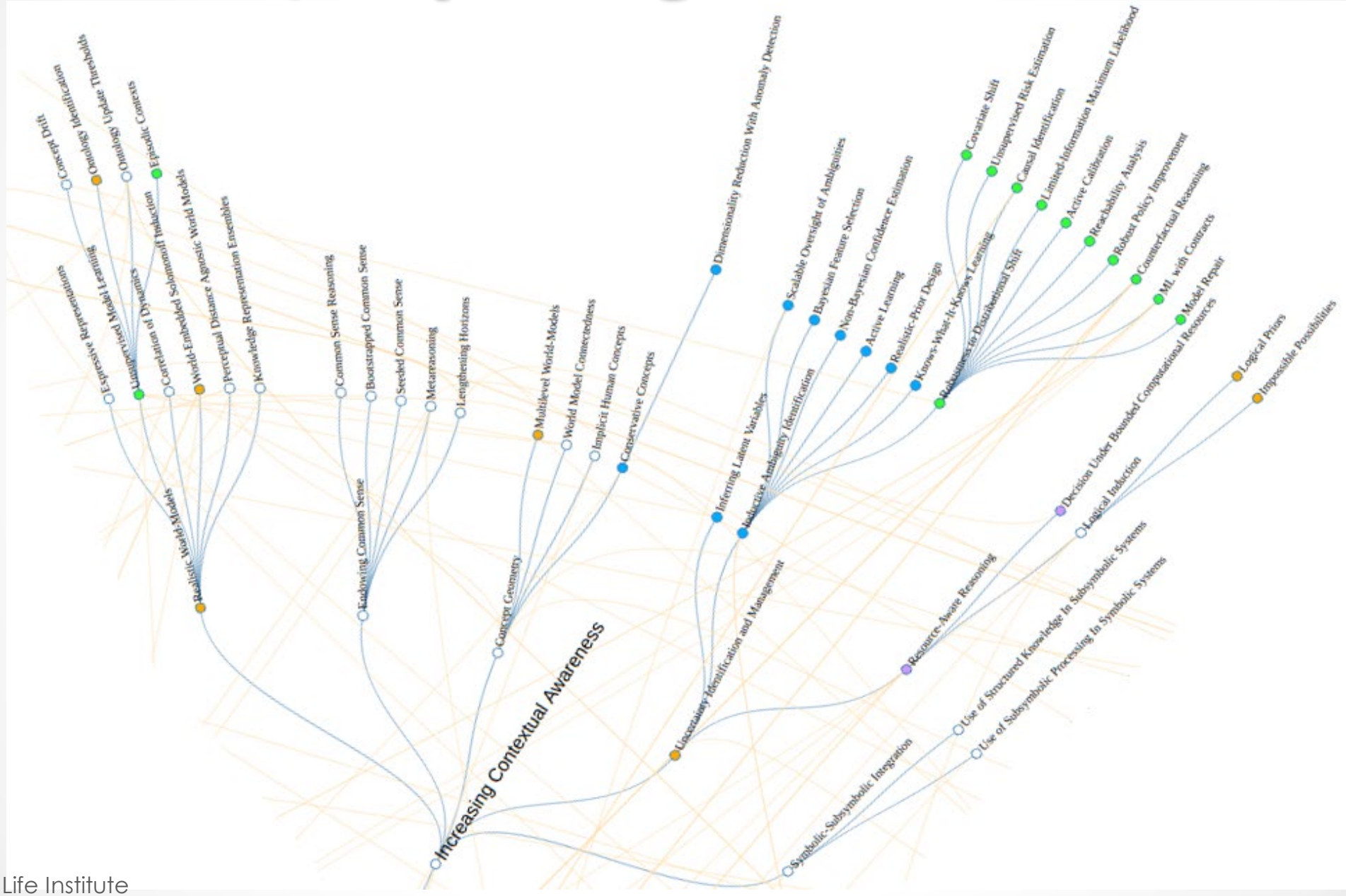
Validation / Avoiding Reward Hacking



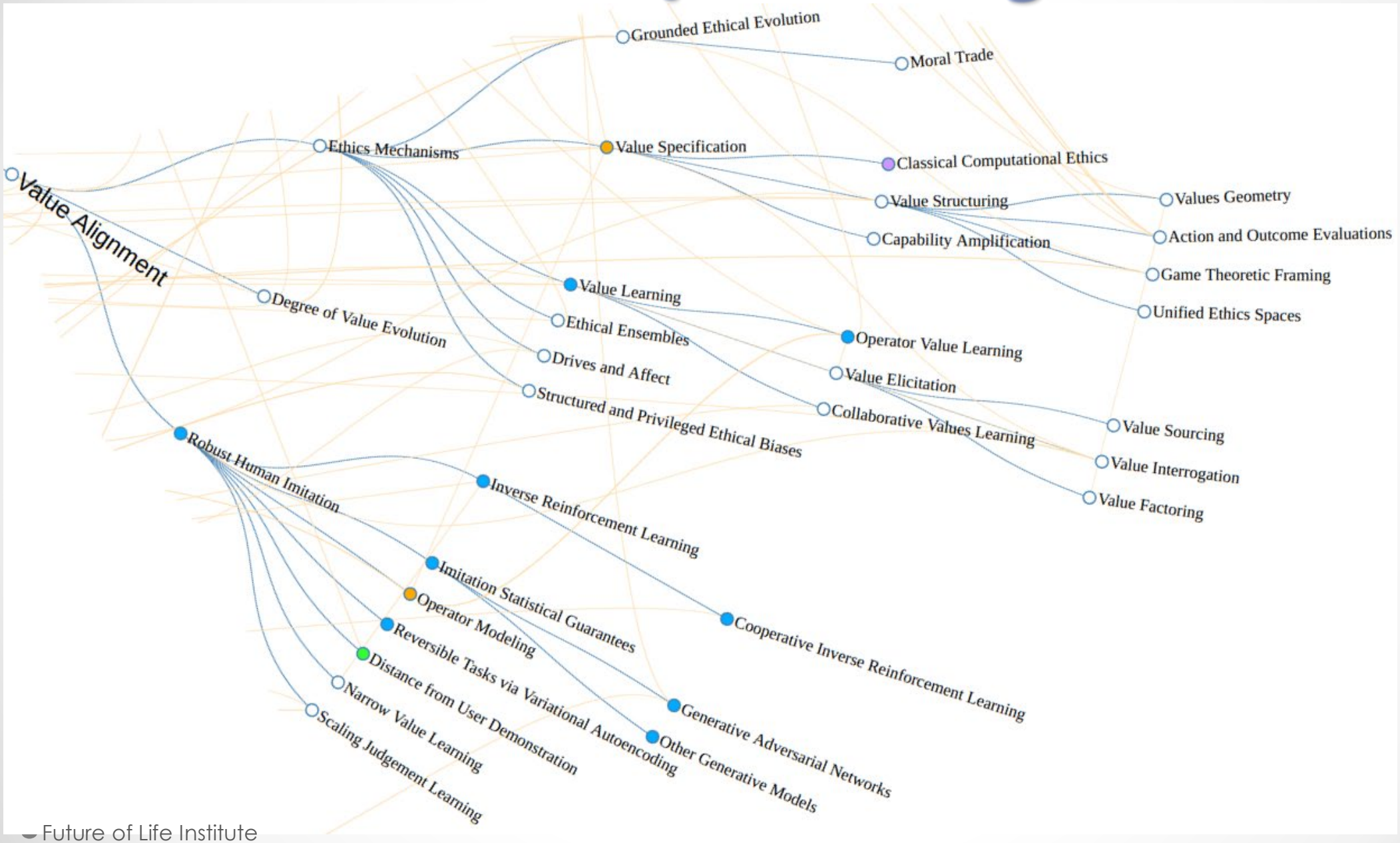
Validation / Averting Instrumental Incentives



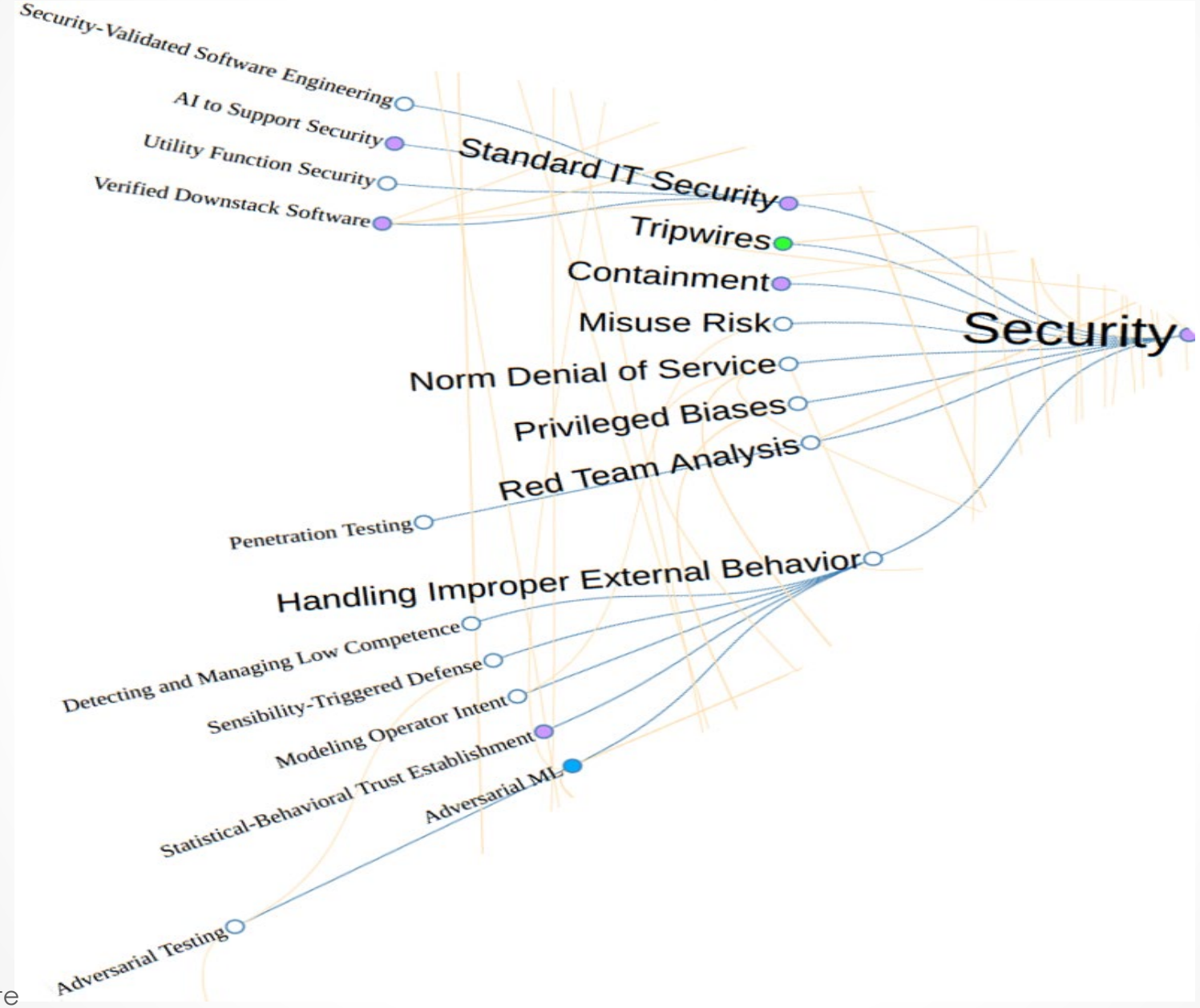
Validation / Improving Contextual Awareness



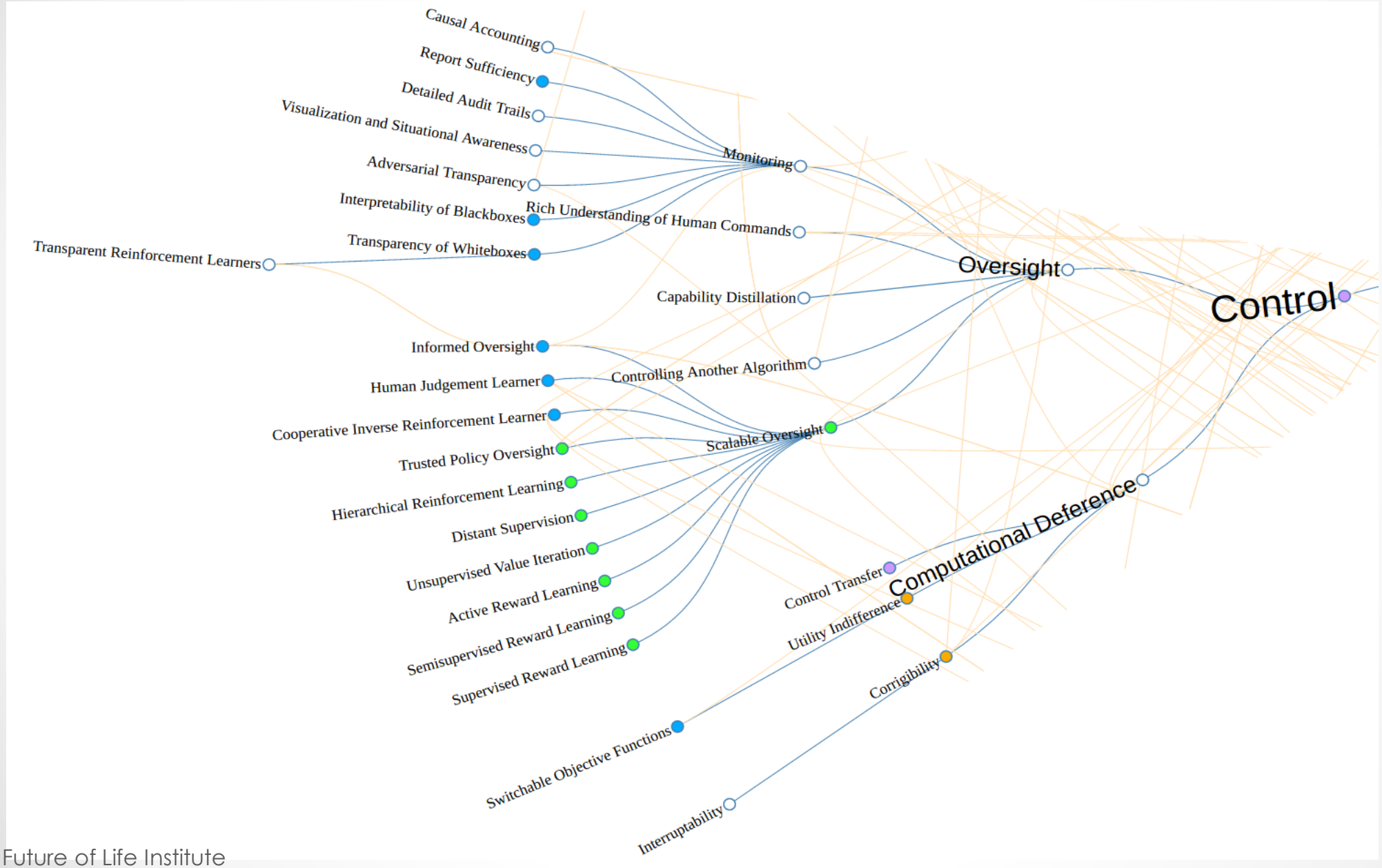
Validation / Value Alignment



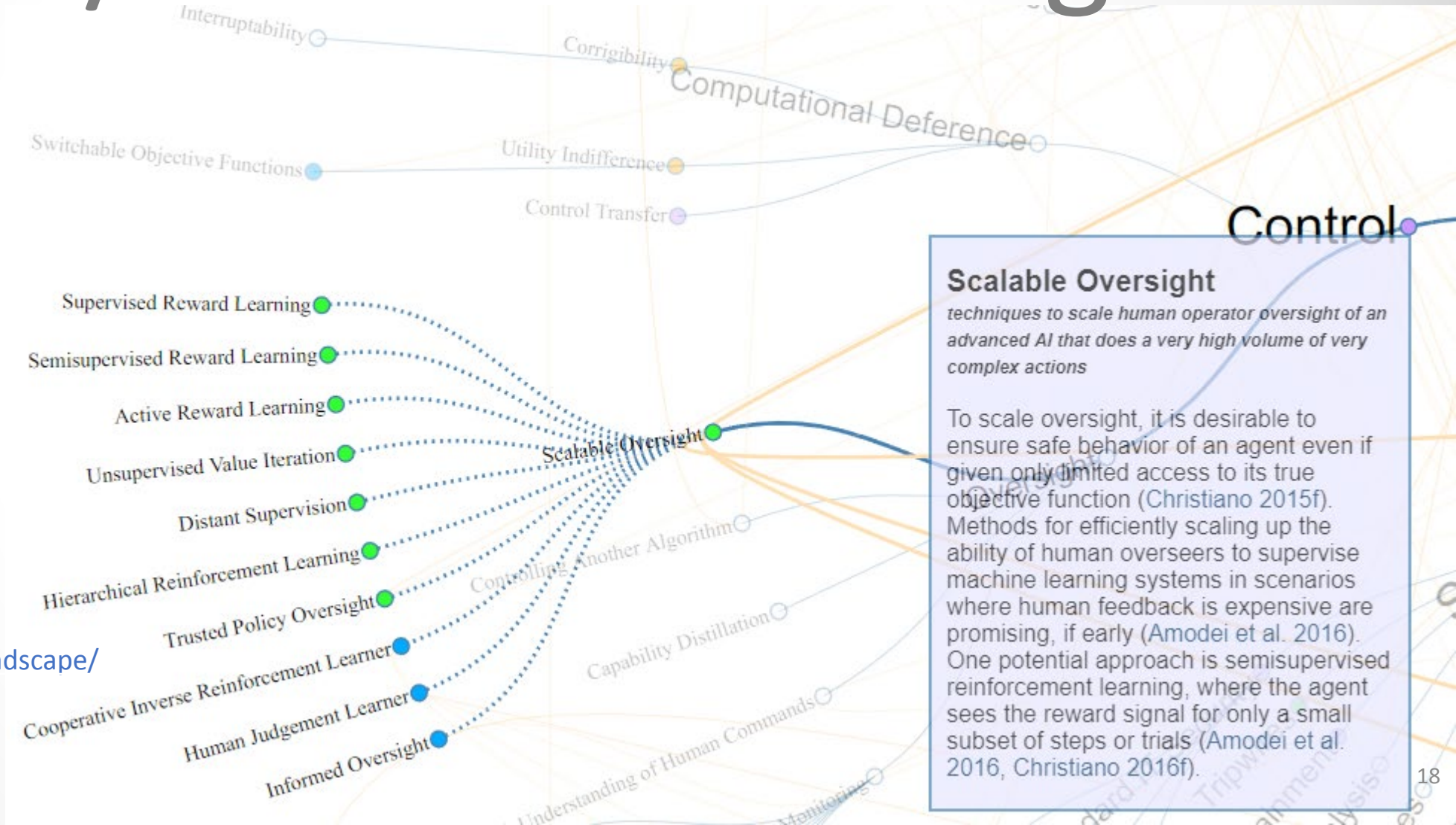
Security


Security-Validated Software Engineering
AI to Support Security
Utility Function Security
Verified Downstack Software
Standard IT Security
Tripwires
Containment
Misuse Risk
Norm Denial of Service
Privileged Biases
Red Team Analysis
Penetration Testing
Handling Improper External Behavior
Detecting and Managing Low Competence
Sensibility-Triggered Defense
Modeling Operator Intent
Statistical-Behavioral Trust Establishment
Adversarial ML
Adversarial Testing
Security

Control


Causal Accounting
Report Sufficiency
Detailed Audit Trails
Visualization and Situational Awareness
Adversarial Transparency
Interpretability of Blackboxes
Transparency of Whiteboxes
Transparent Reinforcement Learners
Monitoring
Rich Understanding of Human Commands
Capability Distillation
Oversight
Control
Informed Oversight
Human Judgement Learner
Cooperative Inverse Reinforcement Learner
Trusted Policy Oversight
Hierarchical Reinforcement Learning
Distant Supervision
Unsupervised Value Iteration
Active Reward Learning
Semisupervised Reward Learning
Supervised Reward Learning
Controlling Another Algorithm
Scalable Oversight
Control Transfer
Utility Indifference
Computational Deference
Corrigibility
Switchable Objective Functions
Interruptability

Control / Scalable Oversight



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