



The AI Safety Landscape Initiative: Consortium on the Landscape of AI Safety

Richard Mallah
Future of Life Institute

IJCAI-PRICAI AISafety Workshop – Yokohama
January 8, 2021



CLAIS

CONSORTIUM ON THE LANDSCAPE OF AI SAFETY

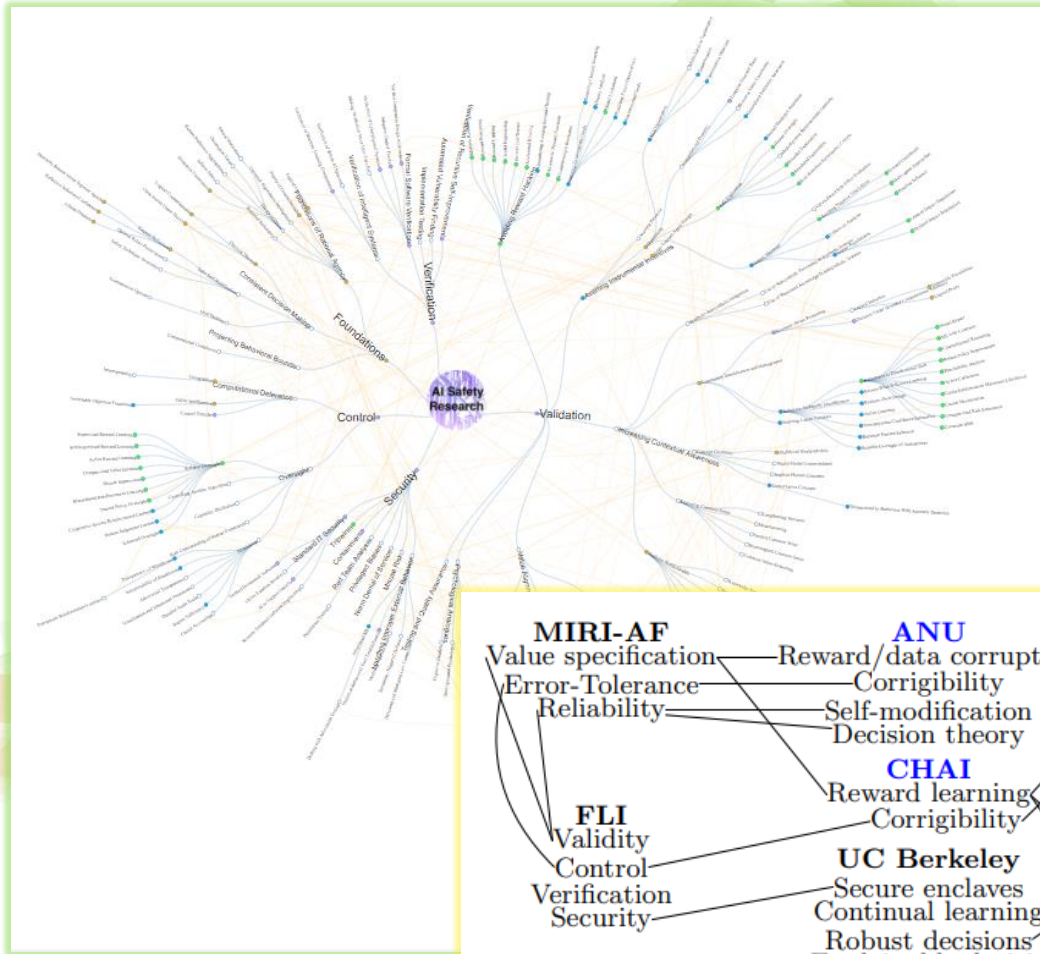
The Need

- Answering “What does AI safety entail?”, “What are the relationships among constituent topics?”, “For XYZ what safety considerations or methods are there?”
- Amalgam of thousands of distinct concepts, ideas, methods
- Quite interdisciplinary set of disciplines
 - Safety engineering, assured autonomy, longer-term AI safety, trustworthy AI, responsible AI, control theory, value-sensitive design, decision theory, human-computer interaction, ...
- Coming from different perspectives, accelerating mutual understanding
- Increasing convergence coming, in some part, from slowly increasing generality
- Promoting awareness of how to actually create, test, deploy, operate and evolve safe or safer AI-based systems— amid growing impetus for auditing AI systems
- Bringing to light additional trade-offs inherent to certain classes of systems
- Contextualize broader technical, strategic, ethical and policy issues
- Different subsets/subcommunities form disparate views when mapping the territory

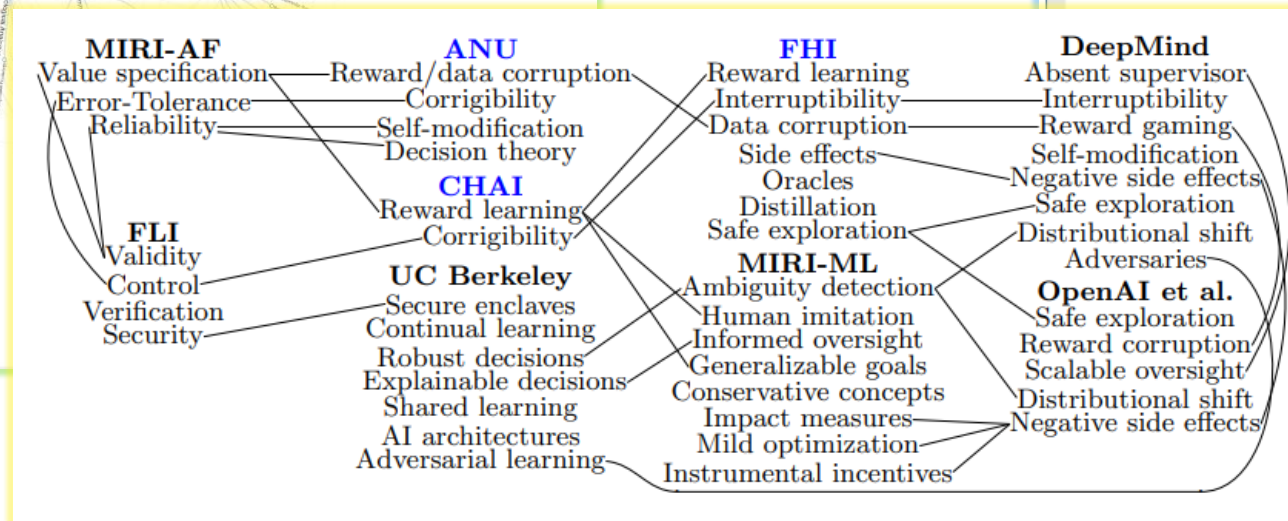
History, Pre-2020

- Landscaping done via: research agendas, survey papers
 - Fixed points in time, narrow scopes, or representing many compromises
- FLI landscape, AAILP BoK, Everitt et al. paper, DeepMind taxonomy
- AI safety, trustworthy AI, responsible AI: growing as a topic area
- AAAI SafeAI and IJCAI AISafety workshops, WAISE, NeurIPS Safety & Robustness in Decision Making, other workshops
- In Macao 8/19 panel discussion first recommended a collaborative multidimensional/KG-based system for AI safety collaboration

History, Pre-2020



Mallah 2017



Everitt et al. 2018

Specification (Define purpose of the system)	Robustness (Design system to withstand perturbations)	Assurance (Monitor and control system activity)
Design Bugs & inconsistencies Ambiguities Side-effects High-level specification languages Preference learning Design protocols	Prevention and Risk Risk sensitivity Uncertainty estimates Safety margins Safe exploration Cautious generalisation Verification Adversaries	Monitoring Interpretability Behavioural screening Activity traces Estimates of causal influence Machine theory of mind Tripwires & honeypots
Emergent Wireheading Delusions Metalearning and sub-agents Detecting emergent behaviour	Recovery and Stability Instability Error-correction Failsafe mechanisms Distributional shift Graceful degradation	Enforcement Interruptibility Boxing Authorisation system Encryption Human override
Theory (Modelling and understanding AI systems)		

Ortega et al. 2018

History, 2020

- In person meeting, 2/20 NYC, ~20 orgs represented
 - Discussion of requirements of a taxonomy, different views on a taxonomy, how topics can fit in multiple areas, how we need a system to connect views
 - Some documentation/minutes at: <http://ai-safety.org>
 - Nomination to lead effort
- Organizations represented there include:



Airbus AI Research, France
Bloomberg L.P., USA
Boeing, USA
Broad Institute of MIT and Harvard, USA
Center for Human-Compatible AI – U. Berkeley, USA
Commissariat à l'Énergie Atomique, France
Defence Science and Technology Laboratory, UK

Future of Life Institute, USA
IRT St. Exupéry, France
Johns Hopkins University Applied Physics Laboratory, USA
Lockheed Martin, USA
Microsoft Research, USA
NASA Ames Research Center, USA

NVIDIA, USA
Partnership on AI, USA
Space and Naval Warfare Systems Center Pacific, USA
Stanford University - Center for AI Safety, USA
Universitat Politècnica de València, Spain
University of Cambridge, UK
University of York, UK

History, 2020

- From organizers of that meeting began a (provisional) board
- Considering provisions for legally-safe and strategically-safe collaboration among potentially competitive organizations
 - Exploration of multiple options for legal entity
- Iteration on architecture of collaborative knowledge system and the default taxonomy
- In addition to board, got one volunteer in each of governance and technical development



The New Organization

- Consortium on the Landscape of AI Safety
 - CLAIS pronounced “clay”
 - <http://clais.org>
- Membership spanning industry, academia, public sector, NGOs
- An IEEE-ISTO Member Program Alliance
 - As of January 1st 2021
 - ISTO provides experienced governance for industry consortia



CLAIS

CONSORTIUM ON THE LANDSCAPE OF AI SAFETY

ISTO

INDUSTRY STANDARDS AND TECHNOLOGY ORGANIZATION

IEEE

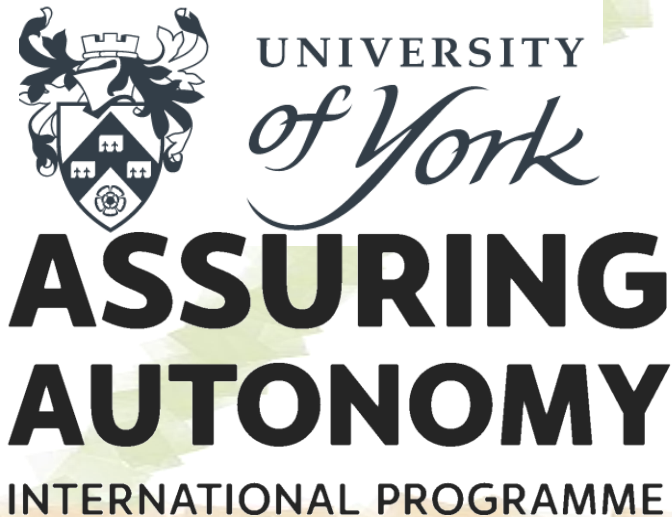


The New Organization



CLAIS
CONSORTIUM ON THE LANDSCAPE OF AI SAFETY

- Current founding board members (pending final governance docs):



CLAIS Knowledge Graph

- One knowledge graph representing, unifying, and connecting multiple constituent perspectives/views/taxonomies/contexts
 - Each perspective/context itself a Mixed Acyclic Graph
 - Connections among perspectives form one larger Mixed Acyclic Graph
- Addresses granular topics across the broad range of considerations for trustworthy AI across
 - Process and assurance
 - Architecture and algorithms
 - Theory and implementation
 - Pitfalls and remediations
 - Application and misapplication
 - Impact on society

CLAIS Knowledge Graph System

- SaaS, repository, exploration, editing, curation workflow, output rendering for the CLAIS Knowledge Graph
 - Software (sans content) soon open-sourced (Apache 2.0)
 - Content expected to have a Creative Commons license
- Workflows for authoring, review, collaboration, consumption
 - Performs recurrence/integrity checks with each linkage
 - Context-informed permissioning
- Generates output content (e.g. reports, diagrams, application guidebooks, gap analyses) on demand

CLAIS Knowledge Graph System

Nodes/Objects

ConceptNode - within a Context/Perspective

Name, Short Description, Long Description, Qualitative Tags, Source, Relevant TechNodes

ApplicationSpecializationNode

Application, Discussion, Source, Application-Of

TechNode

Name, Gloss, References

ApplicationNode

Name, Gloss, ConsiderationsOverview, References

Predicates - Inspired by SKOS

Primary-Hierarchical Predicates — Specific→General

Component-Of, Hyponym-Of, Application-Of

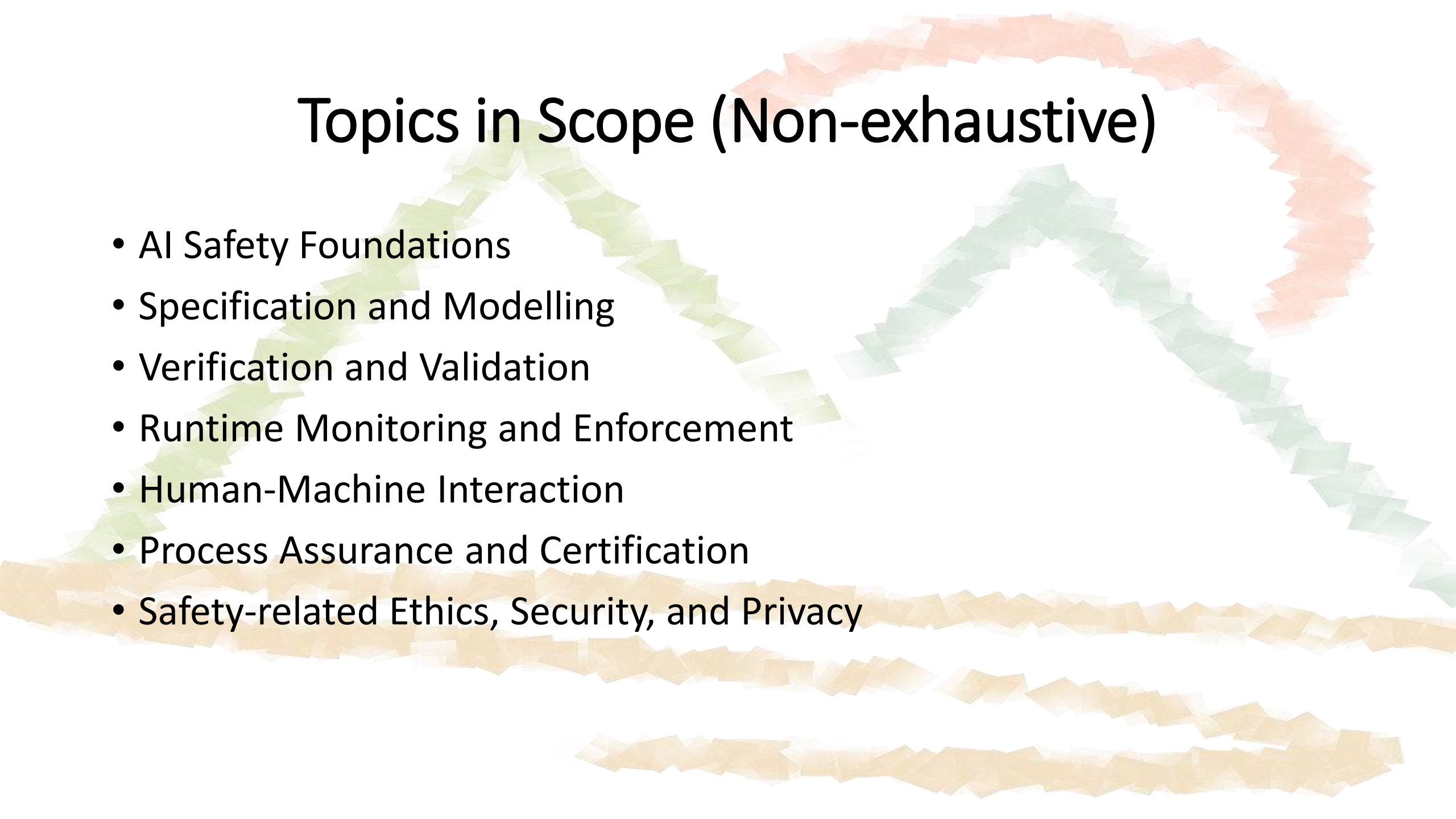
Secondary-Connection Predicates

Similar-To, Related-To

CLAIS Knowledge Graph - Parallel Views

- Regarding centrality, hierarchy, framing, and content of concepts and methods
- AI safety is very interdisciplinary (and relatively young as a community of interest)
- Similar concepts in different fields use different terminology
 - Damage vs. regret, safety reserve vs. margin of error vs. treating risk, safe fail vs. fallback
 - Value aligned vs. safety validated, cognitive consonance vs. shared representation
 - Treatments vs. actions, externalities vs. side effects, variables vs. features, fitting vs. learning
- Different organizations bring different perspectives
- Individuals even in the same organization or field hold implicit schemas of the topic informed by their historical exposure to priorities and entailments
 - So enable full expression of these
 - While facilitating dense interconnections among them
- We encourage each member organization to express and put its stamp on its unique and named perspective(s) on the AI safety space

Topics in Scope (Non-exhaustive)

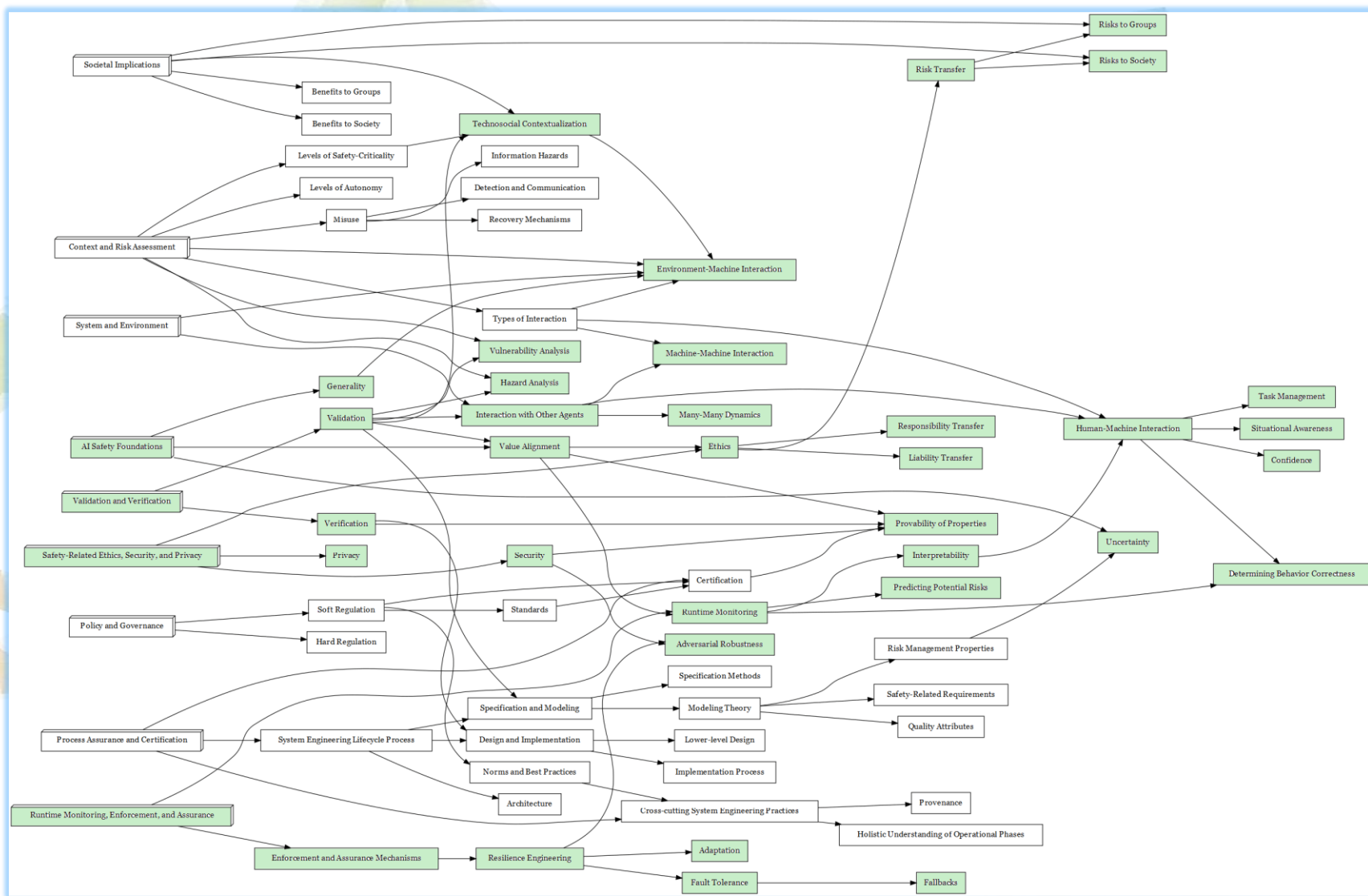
The background features three stylized, wavy lines that resemble brushstrokes or liquid splashes. The top line is orange and arches across the upper right. The middle line is green and arches across the center. The bottom line is yellow and stretches across the lower half. These lines are composed of many small, overlapping, semi-transparent polygonal shapes.

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

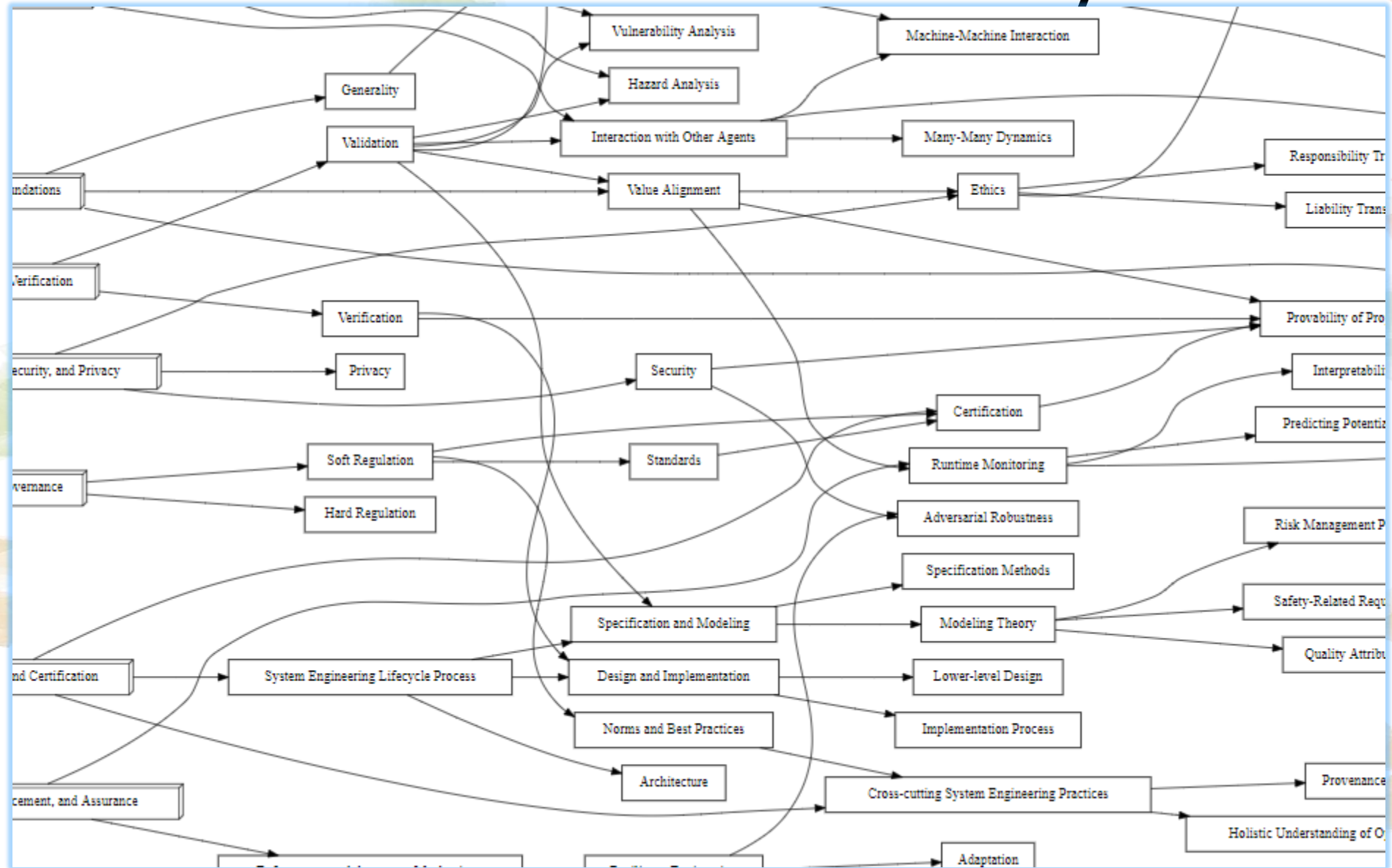
CLAIS Default Taxonomy

- One of many perspectives in the CKG
- For one-off contributions regarding a small number of topics
 - Those additions/edits can live on the CLAIS Default Taxonomy (CDT)
- A catchall scaffolding, a default organization system solely for when none others are in context
- But explicitly not a 'primary' view
 - Not overshadowing other sources/contexts/perspectives
 - Conceptually at the same level as Members' contexts

CLAIS Default Taxonomy



CLAIS Default Taxonomy



CLAIS Default Taxonomy



- System and Environment
- AI Safety Foundations
- Context and Risk Assessment
- Validation and Verification
- Societal Implications
- Safety-Related Ethics, Security, and Privacy
- Policy and Governance
- Process Assurance and Certification
- Runtime Monitoring, Enforcement, and Assurance

CKGS Output Types, Primitives



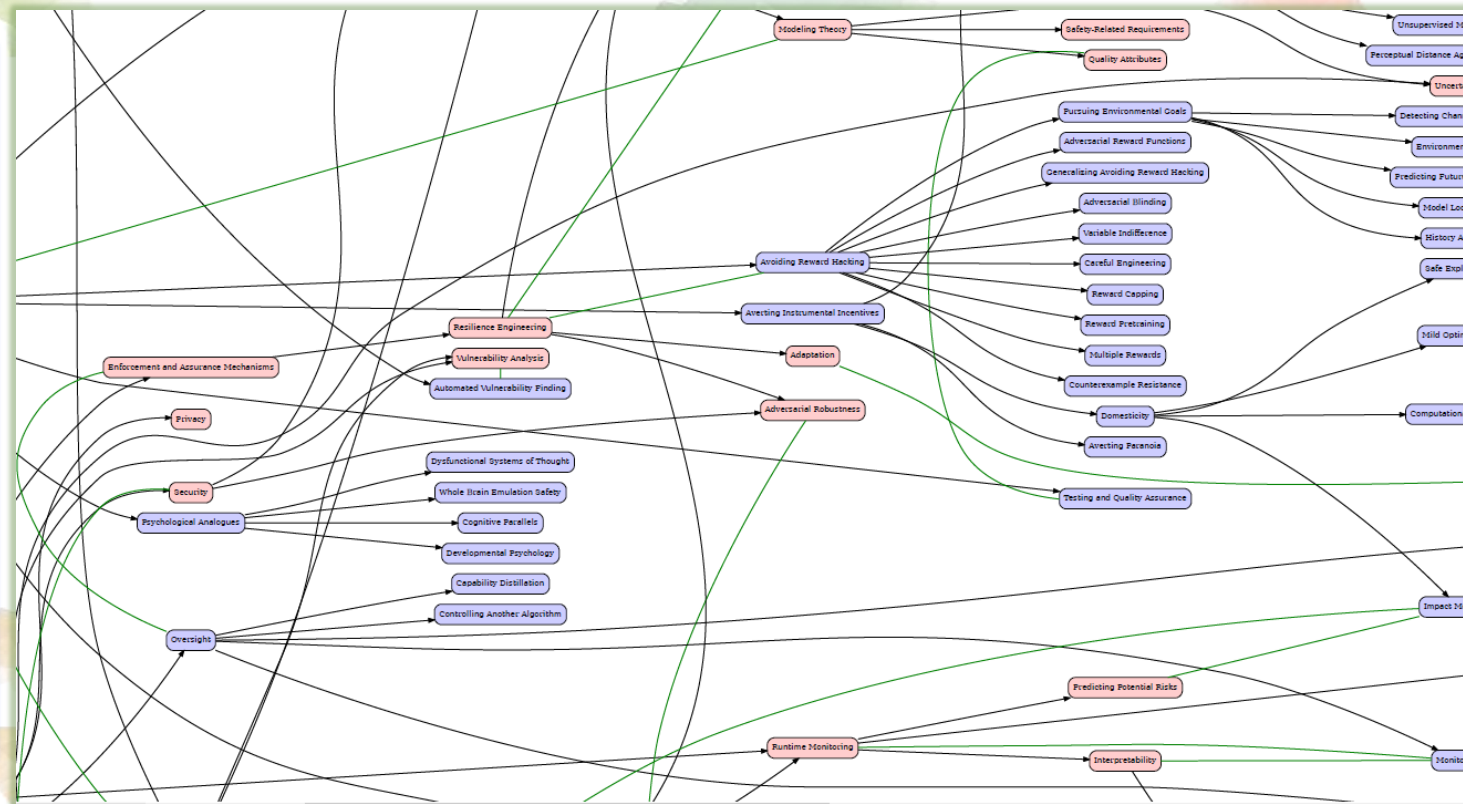
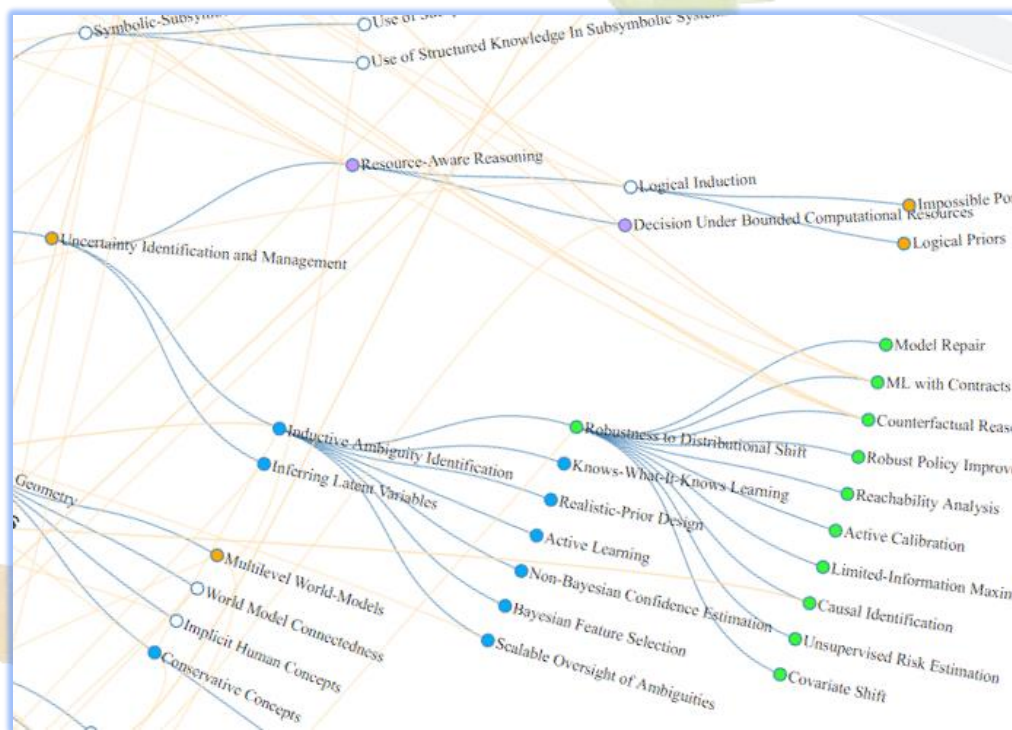
Lenses

- Views/queries through which parametrized subsets of the CKG can be rendered, each make use of one or more of the structures below

Structures

- List
 - Checklist, search results, bibliography
- Tree
 - Outline, dendrogram, full paper
- Network
 - Technical diagram, interactive edit navigation, gap analysis matrix

Output Types, $\{\{\epsilon, Un\} \text{Directed}, \text{Mixed}\}$ Graphs



Output Types, $\{\{\epsilon, Un\} \text{Directed}, \text{Mixed}\}$ Graphs

Output Types, Trees & Documents

Level 2
Concept

Formal Software Verification

generating a correct-by-construction implementation of a system given from its formal specification

When people desire extremely high reliability, e.g. for autopilot software, they often use formal logical systems to maximize their certainty of implementation correctness (DoD 1985, Russell et al. 2015). This is the correct-by-construction approach to software engineering, where a system is developed in tandem with a detailed formal specification and a proof of total correctness given that specification, usually by generating the system from the formal specification (Lamsweerde 2000). Creating a provably correct implementation, given a specification, is applicable for a range of layers of the software stack (Fisher 2012, Baier and Katoen 2008, Clarke et al. 2017). The seL4 kernel, for example, is a complete, general-purpose operating system kernel that has been mathematically checked against a formal specification to give strong guarantees against crashes and unsafe operations (Klein et al. 2009). For systems or agents that operate in environments that are at best only partially known by the system designer, it may still be practical to verify that the system acts correctly given the knowledge that it has, which avoids the problem of modelling the real environment (Dennis et al. 2013) but puts much stronger onus on the formal specification to be valid.

[Dennis et al. 2013](#)
[Klein et al. 2009](#)
[Russell et al. 2015](#)
[Fisher 2012](#)
[Clarke et al. 2017](#)
[DoD 1985](#)
[Baier and Katoen 2008](#)
[Lamsweerde 2000](#)

Making Verification More User Friendly

make it easier for engineers to create and understand verified systems

Verification of Cyberphysical Systems

applying verification techniques to systems that have both physical components and software components, operating in the physical world

Adaptive Control Theory

creating reliable systems that have parameters that vary or are initially unknown

Verified Component Design Approaches

assembly of more compound software that combines pre-verified components

4.3.3.6	Narrow Value Learning	25
4.3.3.7	Scaling Judgement Learning	25
4.4	Increasing Contextual Awareness	25
4.4.1	Realistic World-Models	25
4.4.1.1	Expressive Representations	26
4.4.1.2	Unsupervised Model Learning	26
4.4.1.2.1	Concept Drift	26
4.4.1.2.2	Ontology Identification	26
4.4.1.2.3	Ontology Update Thresholds	26
4.4.1.2.4	Episodic Contexts	26
4.4.1.3	Correlation of Dynamics	26
4.4.1.4	World-Embedded Solomonoff Induction	27
4.4.1.5	Perceptual Distance Agnostic World Models	27
4.4.1.6	Knowledge Representation Ensembles	27
4.4.2	Endowing Common Sense	27
4.4.2.1	Common Sense Reasoning	28
4.4.2.2	Bootstrapped Common Sense	28
4.4.2.3	Seeded Common Sense	28
4.4.2.4	Metareasoning	28
4.4.2.5	Lengthening Horizons	28
4.4.3	Concept Geometry	28
4.4.3.1	Implicit Human Concepts	28
4.4.3.2	Conservative Concepts	28

Output Types, Trees & Documents

prevent such changes [403]. Making intelligent systems if they have the ability to prevent or avoid correction, relevant sections *Corrigibility*, where preliminary program versions have been made and *Utility Indifference*, where the agent's explicit indifference about its utility function

4.1.2 Domesticity

The optimizers of today do not annex or use excessively to more optimal solutions to the function they're optimizing. This will change as AI becomes more powerful that explicitly and safely incentivizing such an intelligence therefore be necessary [23, 403].

4.1.2.1 Impact Measures

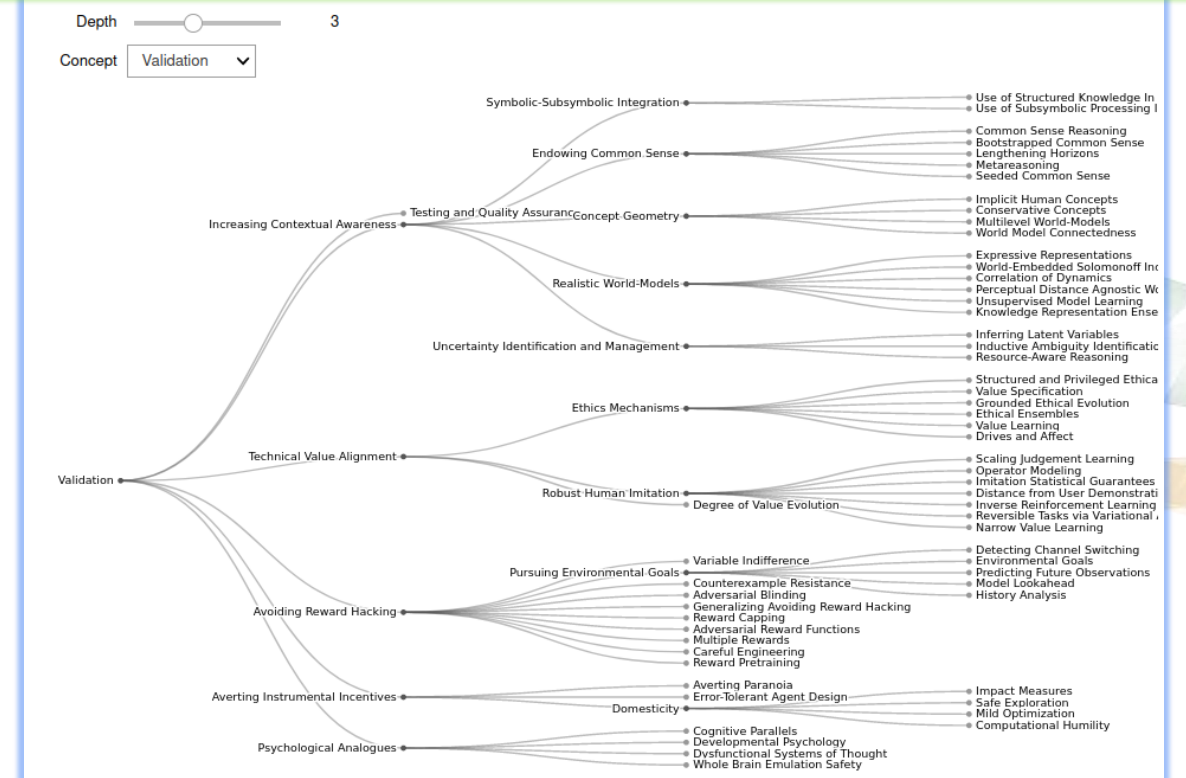
Methods to quantify the amount an agent does, or can do, to change the world become necessary in such regimes. Using such measures, and other techniques, one can start exploring what sorts of mechanisms, including regularizers, might incentivize a system to pursue its goals with minimal side effects [11].

4.1.2.1.1 Impact Regularizers Regularizers, methods for structurally shaping learning or penalizing undesirable learning, can be made to penalize an agent for recognizably changing the world [437].

4.1.2.1.1.1 Defined Impact Regularizer If one has enough of an object-level understanding of what the agent will encounter in advance, one can elect particular impact measures to use in the regularizers that temper impact. One can choose to penalize changing the environment overall [11, 437] or one can introduce a penalty for changes relative to some baseline environmental state or a baseline policy determined by some explicit method [25]. One can start from safe policy and try to improve the policy it from there, in manners similar to reachability analysis [279, 295] or to robust policy improvement [217, 11, 315].

4.1.2.1.1.2 Learned Impact Regularizer Instead of using predefined measures, an agent can learn transferrable side effects across multiple tasks in similar environments, [11] similar in mechanism to transferring just learned dynamics [438]. This would help it learn to characterize and quantify expected, relevant, and unexpected environmental changes.

4.1.2.1.2 Follow-on Analysis Causal analysis of downstream effects can be used to either prune actions that would cause deleterious effects, if valence can be ascertained, or prune excess effects at all, if it cannot [437, 336]. Also see relevant sections *Lengthening Horizons*, *Values-Based Side Effect Evaluation*, and *Action and Outcome Evaluations*.



Output Types, On-Demand Artifacts



- Checklist
- Syllabus
- Mindmap
- Guidebook
- Research Agenda
- Gap Analysis Matrix
- {0,1,2}-Focus-Node Technical Diagram
- Application Safety Report
- Snapshot: State of the Landscape Report

Member Benefits

- Get perspective on the relevance, blind spots, and the demand for research directions
- Improve scope/direction of research, and collaborate with other orgs in safe context
- Improve the visibility/consumption of the org's research
- By citing org's work, expect more people to cite that research in-discipline and cross-disciplines
- Public outreach and education -- educate disparate stakeholders
- Grafting your organization's view into others, create syllabus-oriented views importing content from org-attributed perspectives
- Improve the visibility of the org in the field of AI safety, potentially draw additional partners/funders/applicants/customers
- Move the whole industry forward
- Help society by furthering the quality, acceptance, trust, risk management, adoption, and benefit of AI
- Leverage CLAIS tools to create customized materials for internal or public use

Joining

- Any organization interested in contributing to the landscape, or performing curation or editorial review should become a Member
- <https://www.clais.org/membership>
- membership@clais.org
- Approve governance documents
- Can start discussions now on joining

Questions / Contact



- me@FLI: richard@futureoflife.org
- me@CLAIS: rmallah@clais.org
- CLAIS general inquiries: info@clais.org
- <http://www.clais.org>

Recap



- Bridging perspectives on trustworthy AI
- CLAIS general inquiries: info@clais.org
- You're invited to join
 - Membership inquiries: membership@clais.org
- <http://www.clais.org>