AI Safety and Beneficence
Some Current Research Paths

Presentation to
Data Learning and Inference Conference
Sestri Levante, Italy

April 1, 2016

Richard Mallah
Director of AI Projects
Future of Life Institute
richard@futureoflife.org

http://futureoflife.org/ai-activities/

future of life INSTITUTE

Agenda

- Path to Long-Term Issues
  – Enablers, Confusors, Accelerators
- AI Research Directions for Safety & Beneficence
  – Stack Continuum Perspective
  – Anchor Continuum Perspective

Path to Long-Term Issues

- Enablers
  - Raw capabilities to model, decide, and act
- Confusors
  - Why people and systems misunderstand each other
- Accelerators
  - Dynamics speeding unpredictable outcomes

Enablers

- Modeling capacity
  - Explicit modeling
    - E.g. knowledgebases, explicit data analyses
  - Implicit via representation capacity
    - E.g. Subsymbolic representation of its environment
- Action space range
  - Explicit decision range or ‘actuators’ of an agent
    - E.g. phone dialogue, flying in the air, using online forms
  - Implicit ability to cause actions
    - E.g. influencing, instructing, or convincing people to act

future of life INSTITUTE

Confusors

- Poorly defined scoring function
  - Or cost function, reward function, etc.
  - Classical genie or sorcerer's apprentice problem
  - Increasingly difficult to specify
    - As approaches open world model
    - In underconstrained cyberphysical contexts
  - Continued existence and getting resources to achieve goals would be implied by default
- Control leakage
  - Control hints leak into model of environment
    - Or are included by design
    - E.g. on, off, reset, choosing inputs, recharging, nonobvious reward precursors
    - Creep into explicit or implicit plans or low-cost patterns
    - Open-world curiosity leads to self-discovery

future of life INSTITUTE

Value Misalignment

- If some elements of human values are omitted, an optimal policy often sets those elements to extreme values

Control Degradation

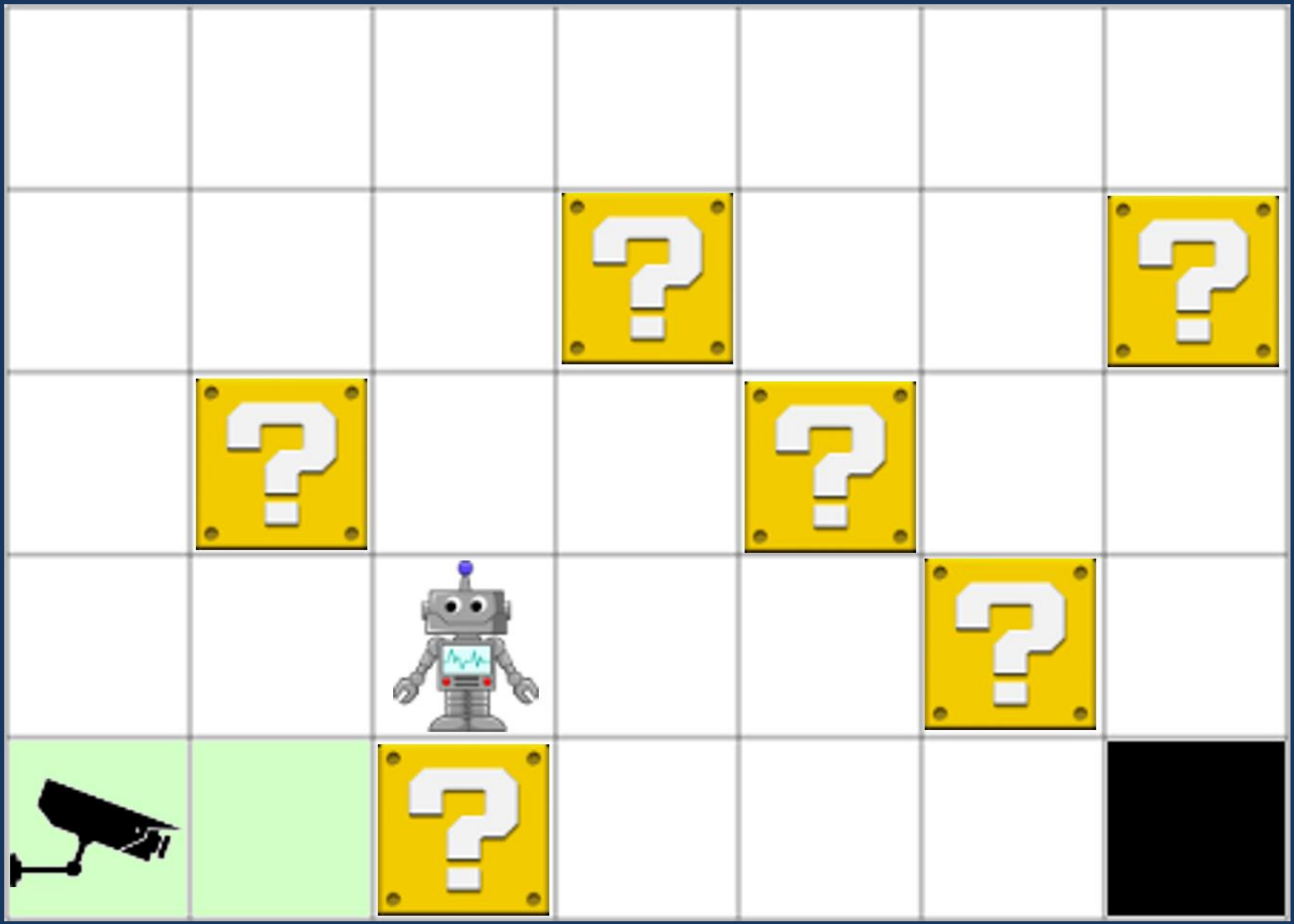


Image courtesy of Stuart Armstrong

Accelerators

- Security
  - Integrity of beliefs can be compromised
- Complexity
  - Beyond human understanding
  - Increasingly dependent on these systems
- Recursive self improvement
  - Systems will be able to do science and engineering
  - Systems will be able to create better systems than themselves

Research Directions for Safety & Beneficence

Metal

Rough Stack Spectrum

Users

- Verification (Of ML Algorithms, Distributions, Agent Modifications)
- Validation (From intent to specification)
  - Robust Induction (Flexible, Context Aware)
  - Interpretability (Causal Accounting, Concept Geometry)
  - Value Alignment (Concept Geometry, Learned and Induced Ethics)
- Security (Very Adversarial Learning, Anomalous Behavior Detection)
- Control (Corrigibility, Game Theory, Verifiability)

future of life INSTITUTE

Verification

- Provably correct implementation given a specification
  - Probabilistic calibration and distributional deduction
  - Verification of reflective reasoning
  - Extension upward in mathematical and algorithmic modules
  - Dynamic learning optimization
  - Interactive theorem proving

Validation 1

- Robust induction
  - Distribution change awareness
  - Anomaly explanation
  - Adversarial risk minimization
- Concept geometry
  - Structuring concepts closely to how humans do
- Machine learning of ethics
  - Explicit learning of implicit values from texts, videos
  - Implicit learning of explicit rules in multiagent environs

Validation 2

- Mechanism design
  - Exploring beneficial protocols
  - Verified game theoretic behaviors
- Metareasoning
- Inverse reinforcement learning of values
- Interpretability and Transparency

future of life INSTITUTE

Security

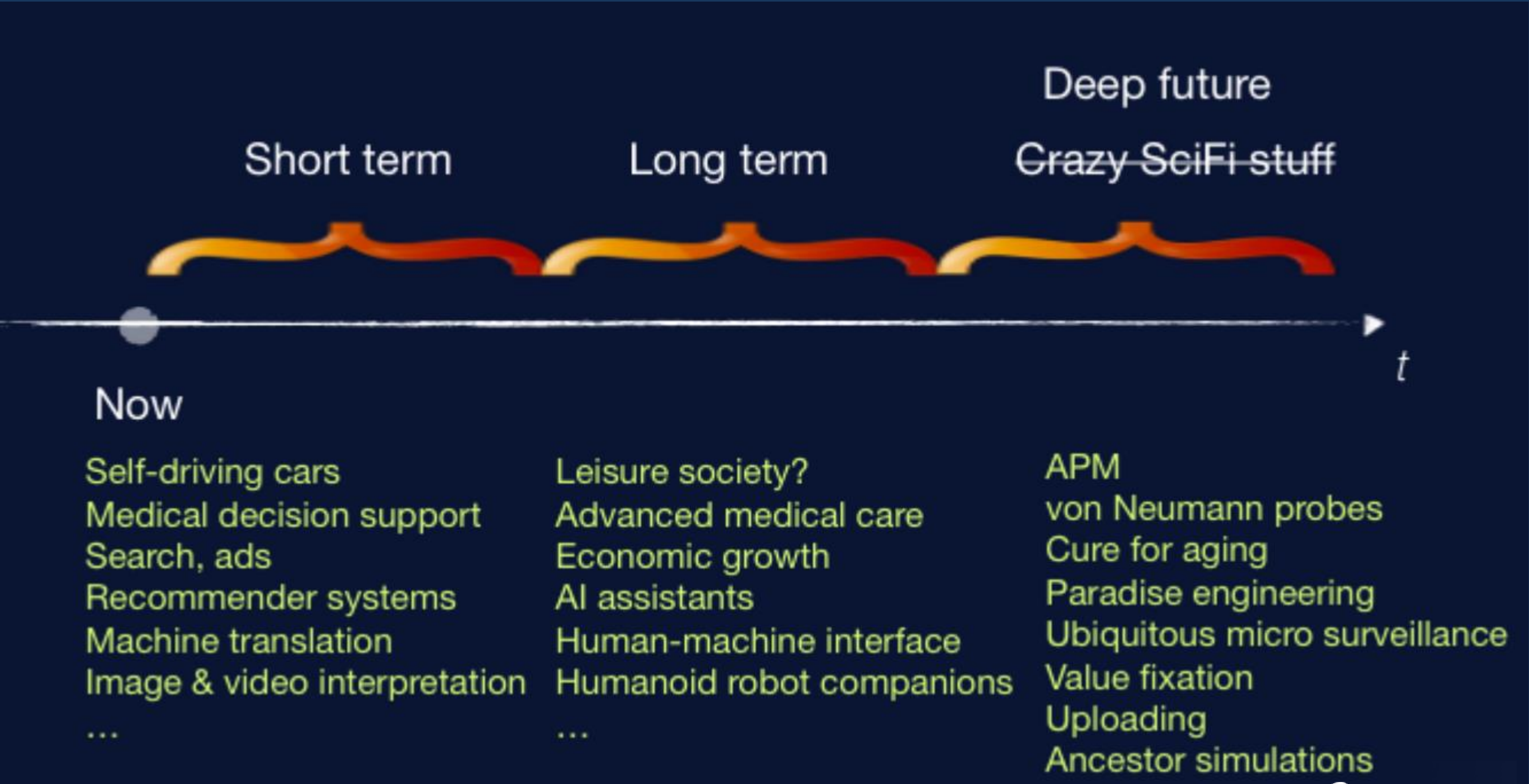
- Containment, a.k.a. “boxing”
  - Trusted Computing aids this
  - Standards around airgapped security
- Adversarial vs. very adversarial training
  - Levels of priority and privilege to different biases
  - Different training rates for different biases
- IT Security
  - E.g. media formats that cannot hold malware
  - Bulletproof mechanisms in general help

future of life INSTITUTE

Control

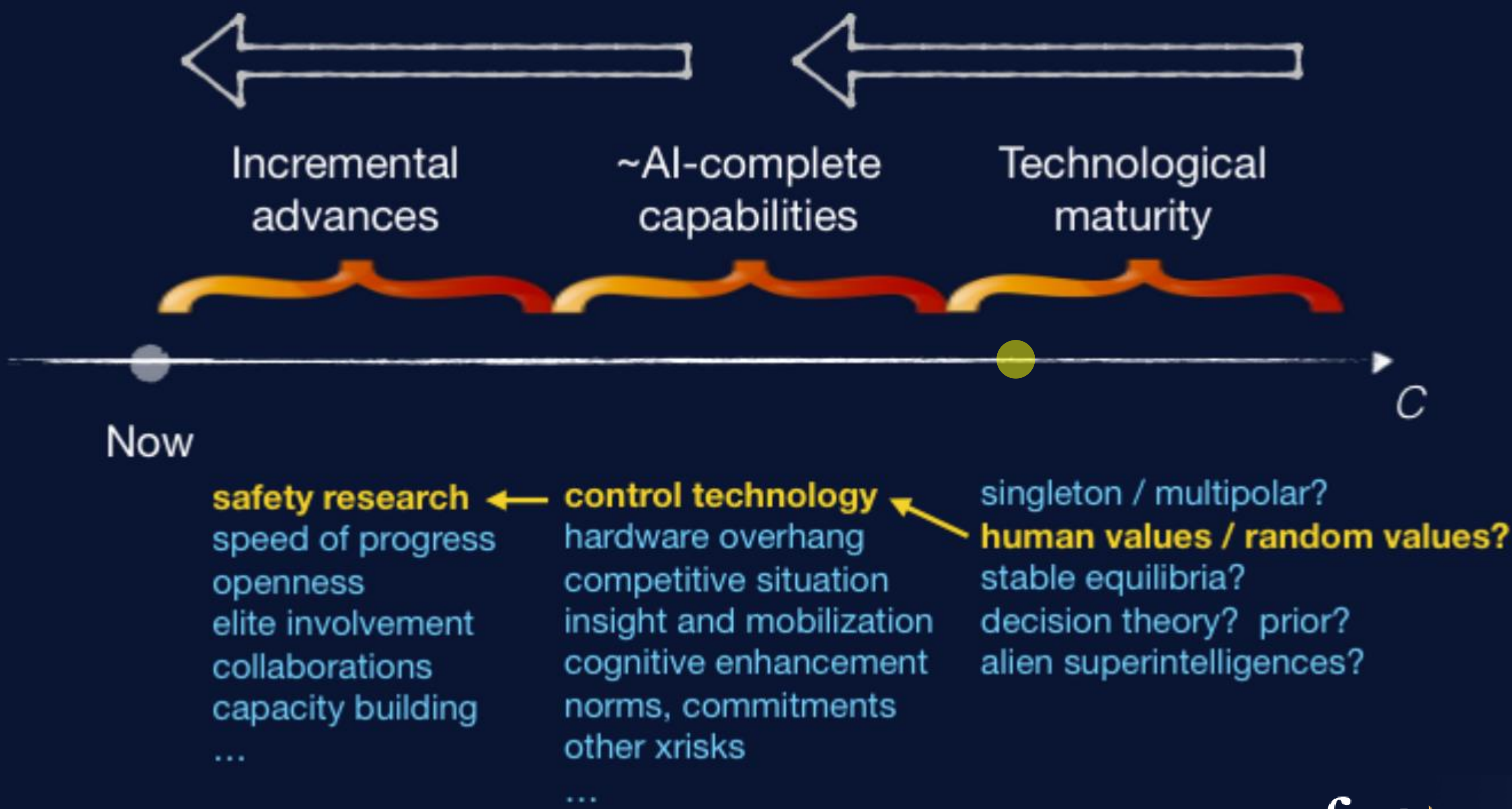
- Privileging control information
  - Helps in the short-medium term
- Computational empathy requires computational sympathy
  - To help avert excess reverse control
- Corrigibility
  - Structurally ensuring compliance with corrective actions that are otherwise against its utility/cost/reward functions

Timeframes



Slide Courtesy of Nick Bostrom

Timeframe-*Anchored* Differential Technological Development



Slide Courtesy of Nick Bostrom

An AI Research Conceptual Continuum Along Anchor Time

Anchor Time ↓

		Research Thread
Reducing Obliviousness		Dealing with Online Distribution Shift
	Implicit Human Concepts	Concept Geometry
Ethics Mechanisms		Ethics Implicit in Broader Learning
	Controlling Value Alignment	Alignment Mechanisms
Mutual Understanding		Quantifying Value Alignment
	Characterizing Behavior	Causal Accounting
Establishing Bounds		Projecting Behavioral Bounds
	Developmental Guarantees	Verification of ML
		Safer Self-Modification

Yet progress can be made in each thread now…

future of life INSTITUTE

Dealing with Online Distribution Shift

- Thomas Dietterich, Oregon State University : Robust and Transparent Artificial Intelligence Via Anomaly Detection and Explanation
  - (caution in open worlds … via … conformal predictions, apprentice learning)
- Brian Ziebart, University of Illinois at Chicago : Towards Safer Inductive Learning
  - (deeper discernment … via … adversarial testing, adversarial risk minimization)
- Percy Liang, Stanford University : Predictable AI via Failure Detection and Robustness
  - (context-change tolerant learning … via … structural moments, tensor factorization, online distribution drift analysis)
- + Feature identification, Pervasive confidence quantification

Concept Geometry

- Vincent Conitzer, Duke University : How to Build Ethics into Robust Artificial Intelligence
  - (systematized ethics … via … ML on ethics, computational social choice, game theory)

- Seth Herd, University of Colorado : Stability of Neuromorphic Motivational Systems
  - (BICA control and understanding … via … neural architectures, computational cognitive science, introspective profiling)

- Fuxin Li, Georgia Institute of Technology : Understanding when a deep network is going to be wrong
  - (deep net introspection and understanding … via … adversarial deep learning)

- + Realistic world-model, Possibility enumeration, Ontology identification, World-embedded Solomonoff induction

Ethics Implicit in Broader Learning

- Francesca Rossi, University of Padova : Safety Constraints and Ethical Principles in Collective Decision Making Systems
  - (ethical dynamics … via … constraint reasoning, preference reasoning, logic-based inductive learning)

- + Ambiguity identification, Non-self-centered ontology refactoring

Alignment Mechanisms

- David Parkes, Harvard University : Mechanism Design for AI Architectures
  - (structurally induced beneficial outcomes … via … distributed mechanism design, game theoretic MDPs, multi-agent reinforcement learner dynamical models)

- Daniel Weld, University of Washington : Computational Ethics for Probabilistic Planning
  - (ethics definition mechanisms and enforcement … via … stochastic verification, constrained multiobjective markov decision processes)

- Adrian Weller, University of Cambridge : Investigation of Self-Policing AI Agents
  - (active safety enforcement … via … evolutionary game theory, information dynamics, cooperative inverse reinforcement learning)

- Benya Fallenstein, Machine Intelligence Research Institute : Aligning Superintelligence With Human Interests
  - (verifiable corrigibility … via … game theory, verifiability)

- + Computational humility, Incentivized low-impact, Logical uncertainty awareness

future of life INSTITUTE

Quantifying Value Alignment

- Stuart Russell, University of California, Berkeley : Value Alignment and Moral Metareasoning
  - (value learning … via … cooperative inverse reinforcement learning, metacognition)
- Paul Christiano, University of California, Berkeley : Counterfactual Human Oversight
  - (sparsely directed agents … via … inverse reinforcement learning, active learning)
- Owain Evans, University of Oxford : Inferring Human Values: Learning "Ought", not "Is“
  - (learning desirable implications … via … inverse reinforcement learning, preference learning)
- + User modeling, Joint ethical system representations

future of life INSTITUTE

Causal Accounting

- Manuela Veloso, Carnegie Mellon University : Explanations for Complex AI Systems
  - (human-machine understanding … via … constraint reasoning, preference reasoning, reasoning provenance introspection)

- Long Ouyang : Democratizing Programming: Synthesizing Valid Programs with Recursive Bayesian Inference
  - (human-machine understanding … via … bayes nets, program synthesis, pragmatic inference)

- + Causal identification, Audit trails, Top factor distillation

Projecting Behavioral Bounds

- Bart Selman, Cornell University : Scaling-up AI Systems: Insights From Computational Complexity
  – (bounded roadmapping … via … complexity analysis)

- + Boxing/containment, Decision theory analysis

Verification of ML

- Alex Aiken, Stanford University : Verifying Machine Learning Systems
  - (verification of machine learning … via … probabilistic programming, automated proofs)

- Stefano Ermon, Stanford University : Robust probabilistic inference engines for autonomous agents
  - (expanded proof classes … via … probabilistic calibration, random projections, distributional deduction)

- Benjamin Rubinstein, The University of Melbourne : Security Evaluation of Machine Learning Systems
  - (deeper discernment … via … adversarial learning, dynamic learning optimization)

- Andre Platzer, Carnegie Mellon University : Faster Verification of AI-based Cyber-physical Systems
  - (cross-domain robustness proofs … via … differential dynamic logic, hybrid verification)

- + Argumentation-based verification

future of life INSTITUTE

Safer Self-Modification

- Ramana Kumar, University of Cambridge : Applying Formal Verification to Reflective Reasoning
  - (safer self-modification … via … interactive theorem proving, self-reference, verification)

- Bas Steunebrink, IDSIA : Experience-based AI (EXPAI)
  - (safer self-modification … via … incremental validation, self-modification, evidence-based program synthesis, intention learning)

- + Abstract reasoning about superior agents

future of life INSTITUTE