AGI Safety Literature Review

Tom Everitt, Gary Lea, Marcus Hutter
Australian National University

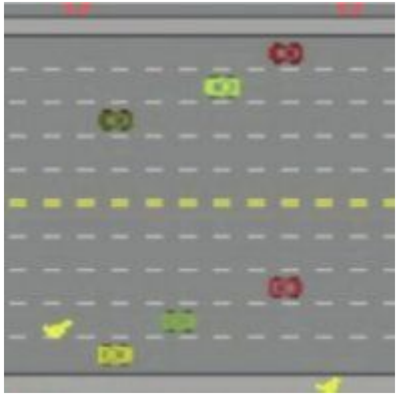
- Understanding AGI
  - Definition & Formalization
  - Orthogonality
  - Instrumental Convergence
  - Alternative views
- Predicting AGI
  - Surveys
  - Singularity (sooner problems)
- Problems with AGI
  - Big figure -- will it fit on a slide?

- Proposed solutions
  - Value specification
    - Reward learning
    - Reward corruption
    - IDA
  - Corrigibility
    - Uncertainty
    - Indifference
  - Intelligibility?
  - Oracles
- Public Policy?

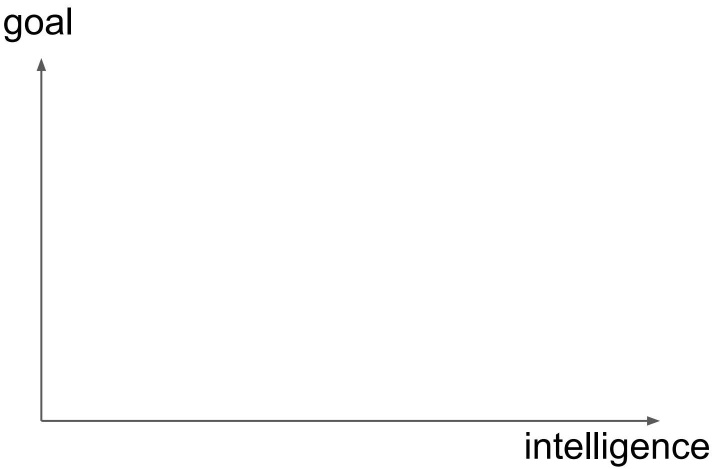
Understanding AGI

Defining intelligence

“Intelligence is the ability to achieve a wide range of goals in a wide range of environments” (Legg & Hutter, 2007)



Orthogonality & Convergence



Increasing intelligence won't make the goal more "intelligent"
(Bostrom 2012, 2014)

Humans value very specific things
(Yudkowsky, 2009)

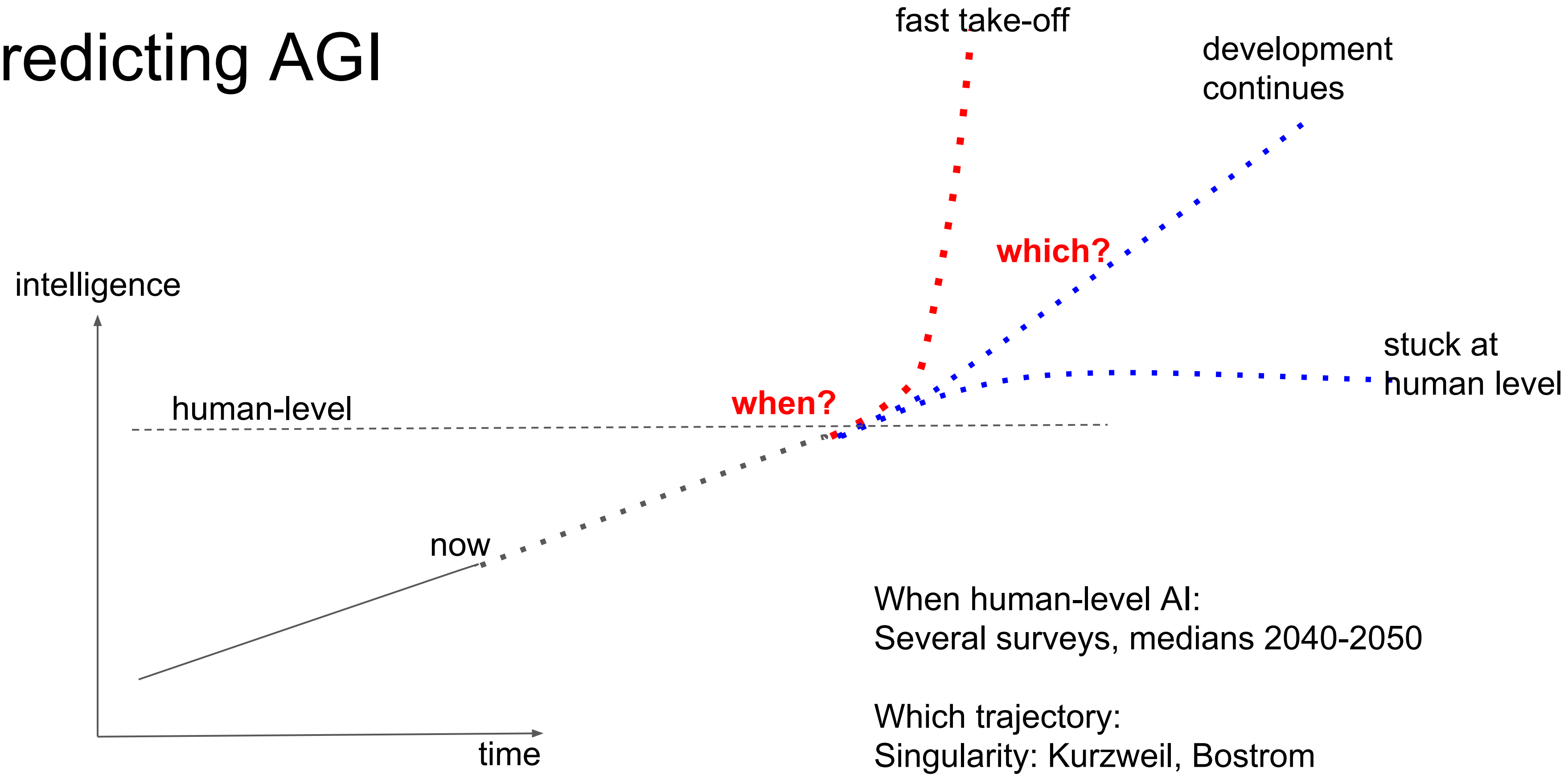
For achieving almost any goal, it is helpful to first:

- Acquire lots of resources
- Self-improve
- Protect one's utility function

(Omohundro, 2008)

Predicting AGI

Predicting AGI

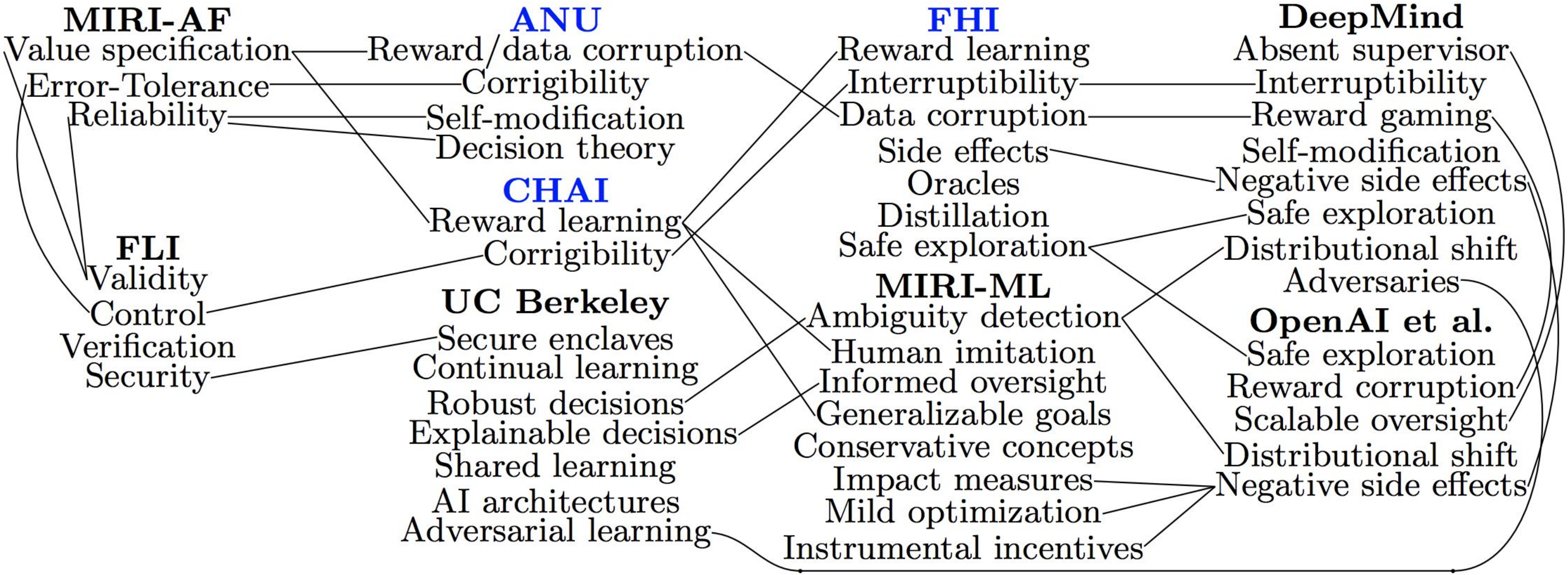


When human-level AI:
Several surveys, medians 2040-2050

Which trajectory:
Singularity: Kurzweil, Bostrom
Continuity: Hanson
Development stalls: Modis

AGI Safety Research Problems

Problems AGI



MIRI-AF
Value specification
Error-Tolerance
Reliability

FLI
Validity
Control
Verification
Security

ANU
Reward/data corruption
Corrigibility
Self-modification
Decision theory

CHAI
Reward learning
Corrigibility

UC Berkeley
Secure enclaves
Continual learning
Robust decisions
Explainable decisions
Shared learning
AI architectures
Adversarial learning

FHI
Reward learning
Interruptibility
Data corruption
Side effects
Oracles
Distillation
Safe exploration

MIRI-ML
Ambiguity detection
Human imitation
Informed oversight
Generalizable goals
Conservative concepts
Impact measures
Mild optimization
Instrumental incentives

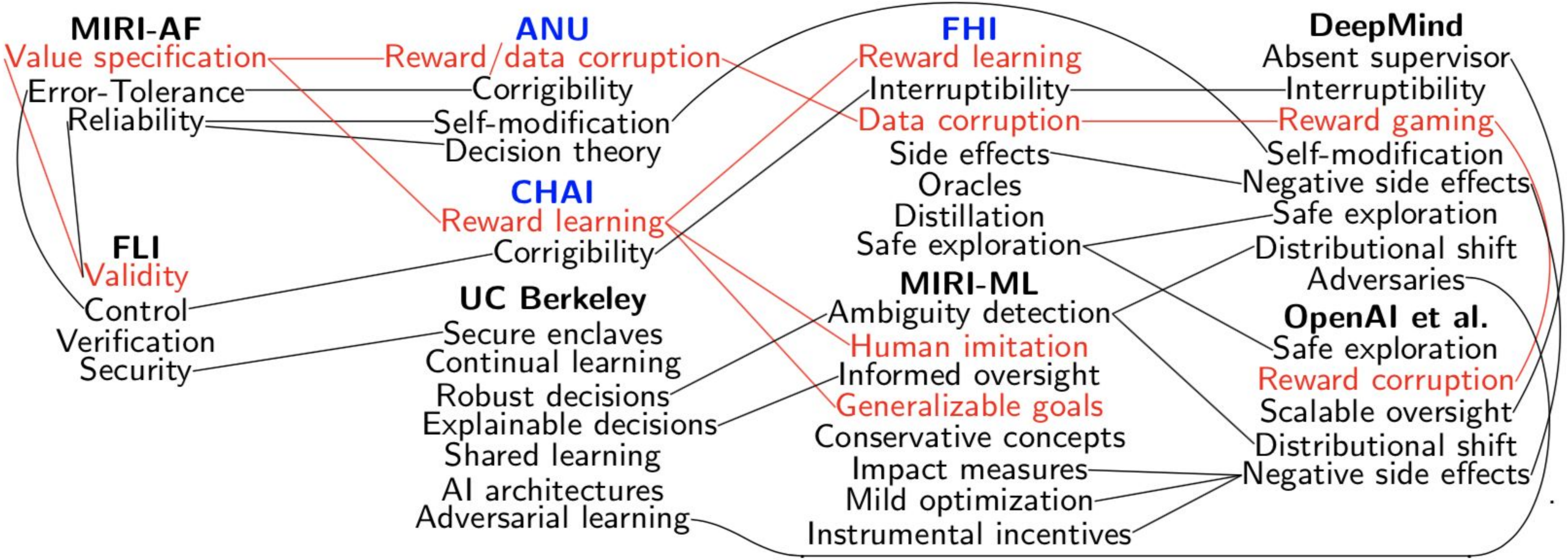
DeepMind
Absent supervisor
Interruptibility
Reward gaming
Self-modification
Negative side effects
Safe exploration
Distributional shift
Adversaries

OpenAI et al.
Safe exploration
Reward corruption
Scalable oversight
Distributional shift
Negative side effects

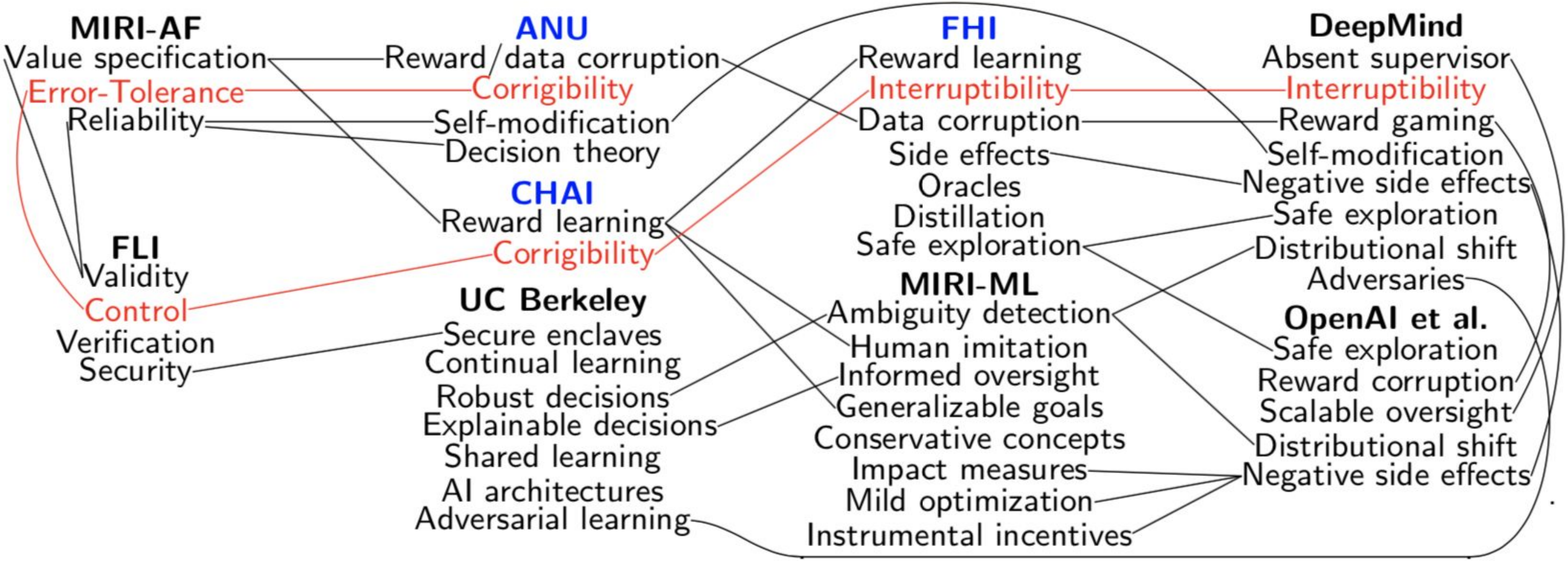
Clusters:

- Value specification
- Reliability
- Corrigibility
- Security
- Safe learning
- Intelligibility
- Social consequences

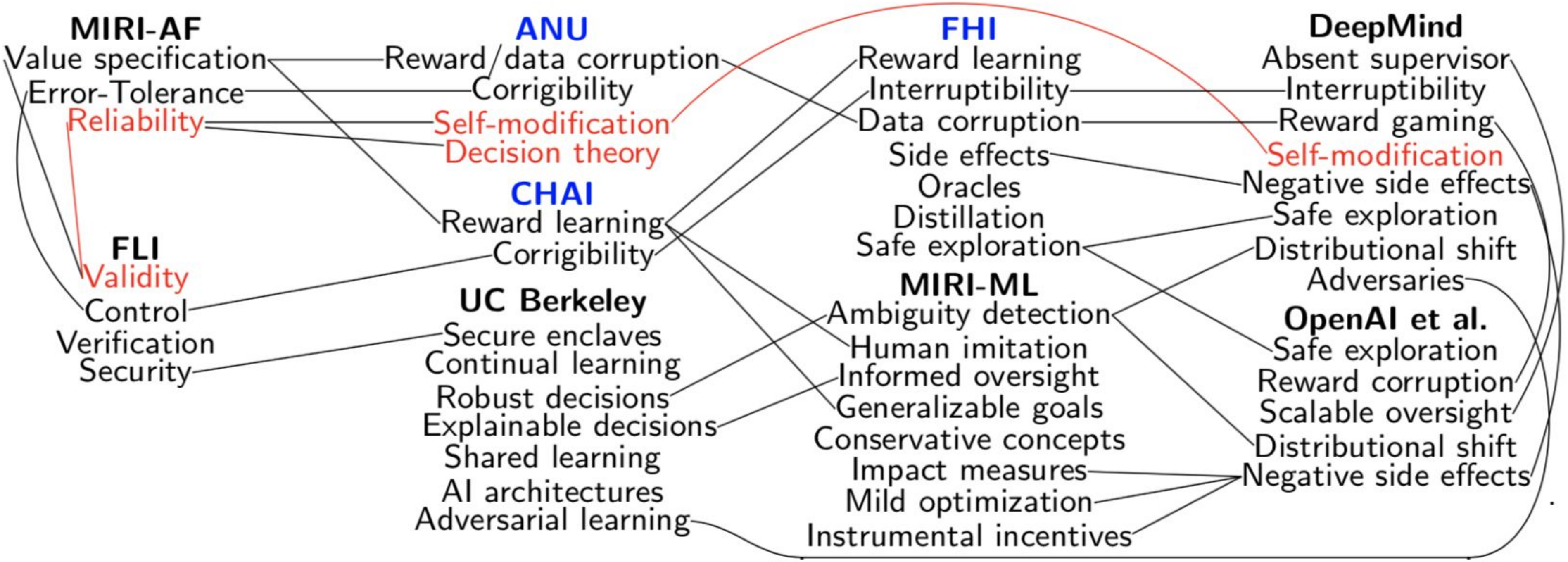
Problems AGI - Value Specification



Problems AGI - Corrigibility



Problems AGI - Reliability



MIRI-AF
Value specification
Error-Tolerance
Reliability

FLI
Validity
Control
Verification
Security

ANU
Reward/data corruption
Corrigibility
Self-modification
Decision theory

CHAI
Reward learning
Corrigibility

UC Berkeley
Secure enclaves
Continual learning
Robust decisions
Explainable decisions
Shared learning
AI architectures
Adversarial learning

FHI
Reward learning
Interruptibility
Data corruption
Side effects
Oracles
Distillation
Safe exploration

MIRI-ML
Ambiguity detection
Human imitation
Informed oversight
Generalizable goals
Conservative concepts
Impact measures
Mild optimization
Instrumental incentives

DeepMind
Absent supervisor
Interruptibility
Reward gaming
Self-modification
Negative side effects
Safe exploration
Distributional shift
Adversaries

OpenAI et al.
Safe exploration
Reward corruption
Scalable oversight
Distributional shift
Negative side effects

Value specification

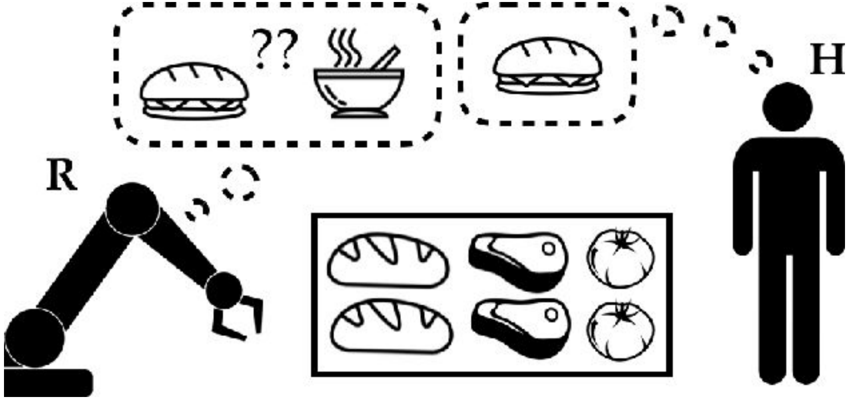
“Design goals that are aligned with human values”

Value specification

Cooperative inverse reinforcement learning
(Hadfield-Menell, Russell et al.)

Infer human goals / values from behavior

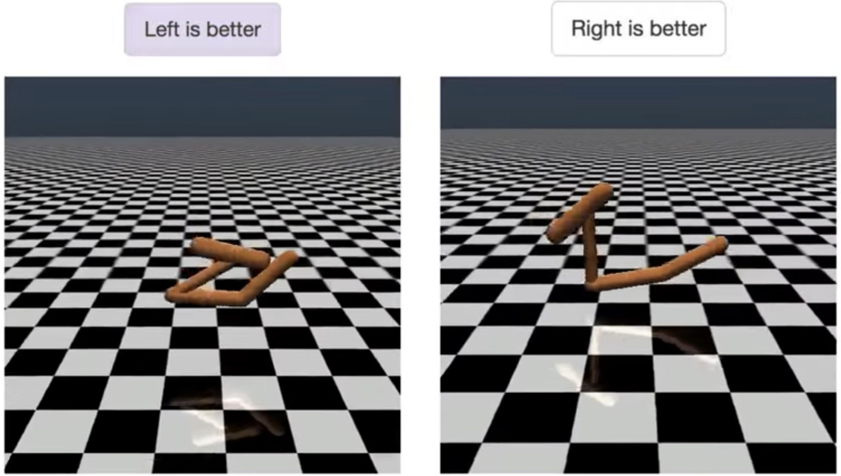
- Potentially completely automatic
- May be hard to model human irrationality



Learning from human preferences
(Christiano, Leike, et al.)

Preference labels for pairs of episodes

- Requires human oversight
- In current formulation, only provides information about past events

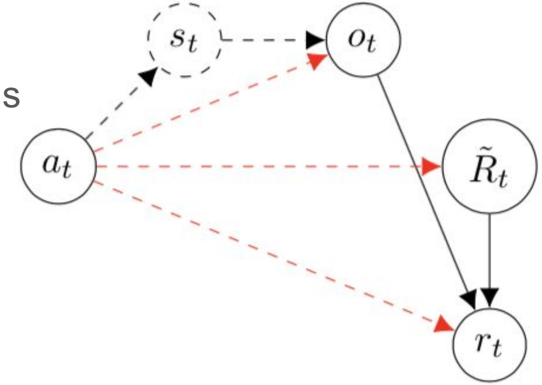


Optimization Corruption

Even if reward function "correct", the agent may have incentives to

- Corrupt the reward function or the reward signal
- Corrupt the data that trains the reward function
- Corrupt the observations / the input to the reward function

Everitt, Hutter et al. (2018) formalize problems and describe solutions



"You can't fetch the coffee if you're dead" -- Stuart Russell

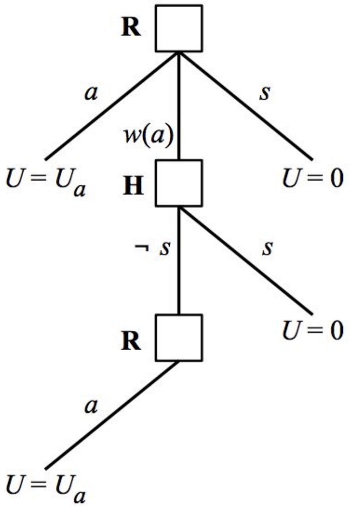
Corrigibility

Ensure agents can always be modified / interrupted

Corrigibility

Goal uncertainty (Hadfield-Menell et al.)

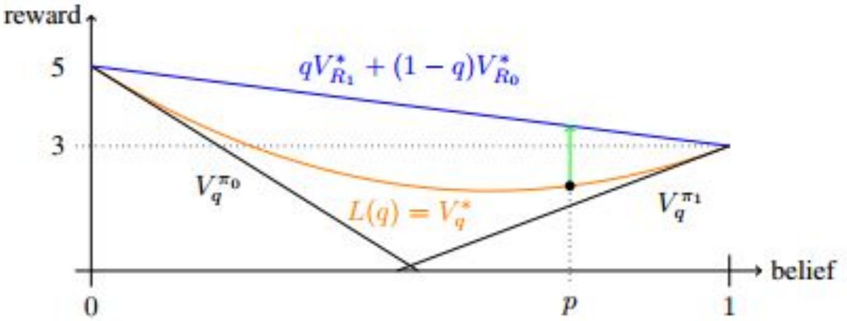
The human's act of switching the agent off is evidence for the human wanting the agent to shut off



Indifference (Armstrong, Orseau, et al.)

Give the agent a compensatory reward for being switched off, exactly equalling the agent's expected reward if not switched off

Off-policy agents automatically indifferent



Alternative (safer?) ways of building AGI

Oracles (Armstrong et al.)

Question-answering systems.
Only goal: answer current question correct

Safer:

- No long-term plans
- Limited actuators

Dangers:

- Tempting to increasingly empower oracles (Bostrom, 2014)
- Perverse incentives may hide in the details

Iterated distillation and amplification (Christiano et al., Ought)

Train an ML system to emulate a human boosted by ML assistant

Services (Drexler)

A human using "narrow" AI services has no disadvantage compared to an AGI agent

Summary

Understanding AGI

- Intelligence definition
- Orthogonality
- Self-Preservation
- Utility preservation
- ….

Problems with AGI

- Different organizations have slightly different focus -- clusters can be identified

Making AGI Safe

- Value specification
- Optimization corruption
- Corrigibility
- Alternative usage
- ....