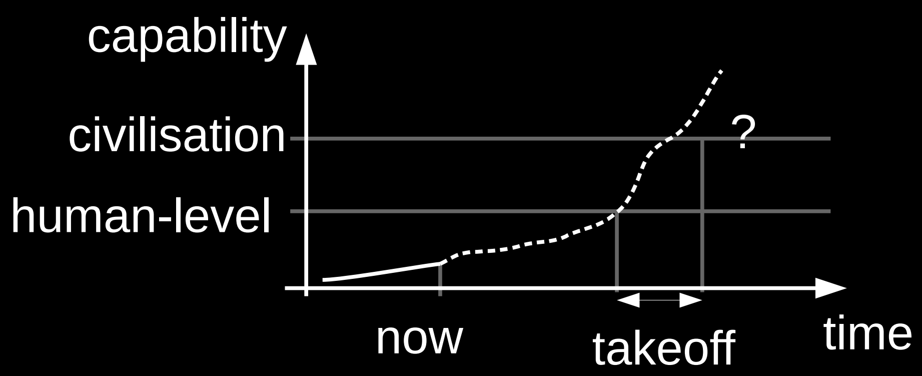
AI Safety

Tom Everitt

27 November 2016

Assumed Background

- AI/ML progressing fast
  - Deep Learning, DQN
  - Increasing investments: HLAI 10 years? SuperAI soon after
  - “Systemic” risks:
    - Unemployment
    - Autonomous warfare
    - Surveillance



- Existential risks
  - Evil genie effect
  - Distinction between:
    - Good at achieving goals (intelligence)
    - Having good goals (value alignment)

Assumption 1 (Utility)

- The performance (or utility) of the agent is how well it optimises a true utility function

$$u : \underbrace{(\mathcal{A} \times \mathcal{E})^*}_{\text{possible experiences}} \to \mathbb{R}$$

- $u\left(æ_{<t}\right)$ is the time-t performance of agent
- Want agent to maximise

$$\sum_{t=1}^{\infty} u(æ_{<t})$$

http://www.gandgtech.com/utility_industry_technology.php

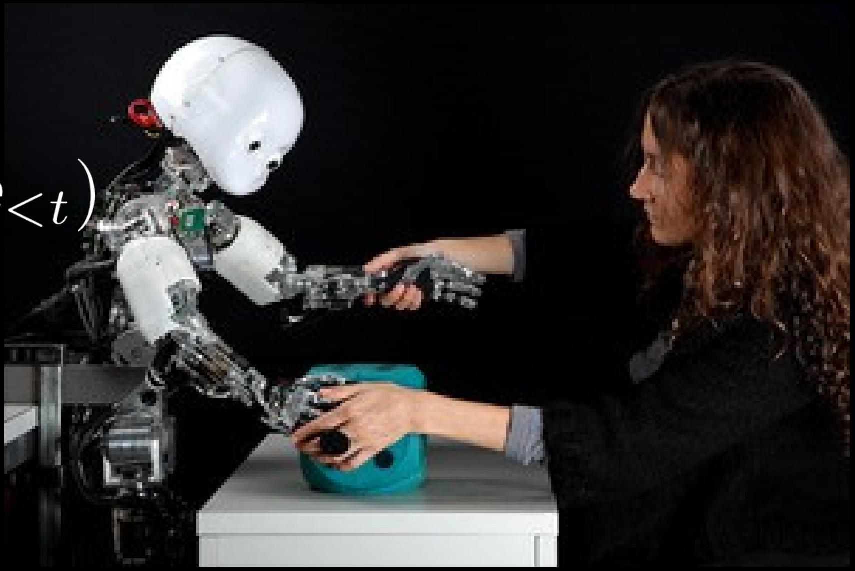
Assumption 2 (Learning)

- It is not possible to (programmatically) express the true utility function $u : \underbrace{(\mathcal{A} \times \mathcal{E})^*}_{\text{possible experiences}} \to \mathbb{R}$
- The agent has to learn *u* from sensory data
- Dewey (2011):

$$u_{\text{learn}}(æ_{<t}) = \underbrace{\sum_{u_i \in \mathcal{U}} P(u_i \mid æ_{<t})\, u_i(æ_{<t})}_{\text{learn utility}}$$

Hopefully:

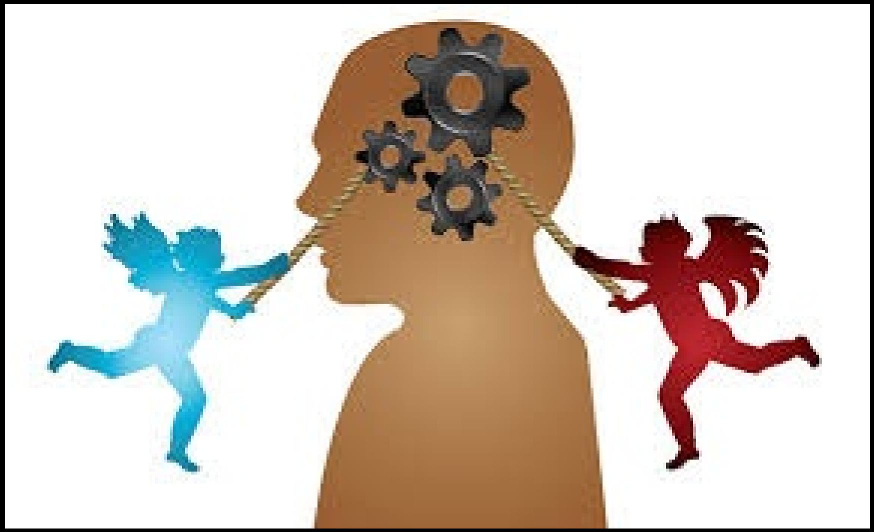
$$u_{\text{learn}} \to u \text{ as } t \to \infty$$



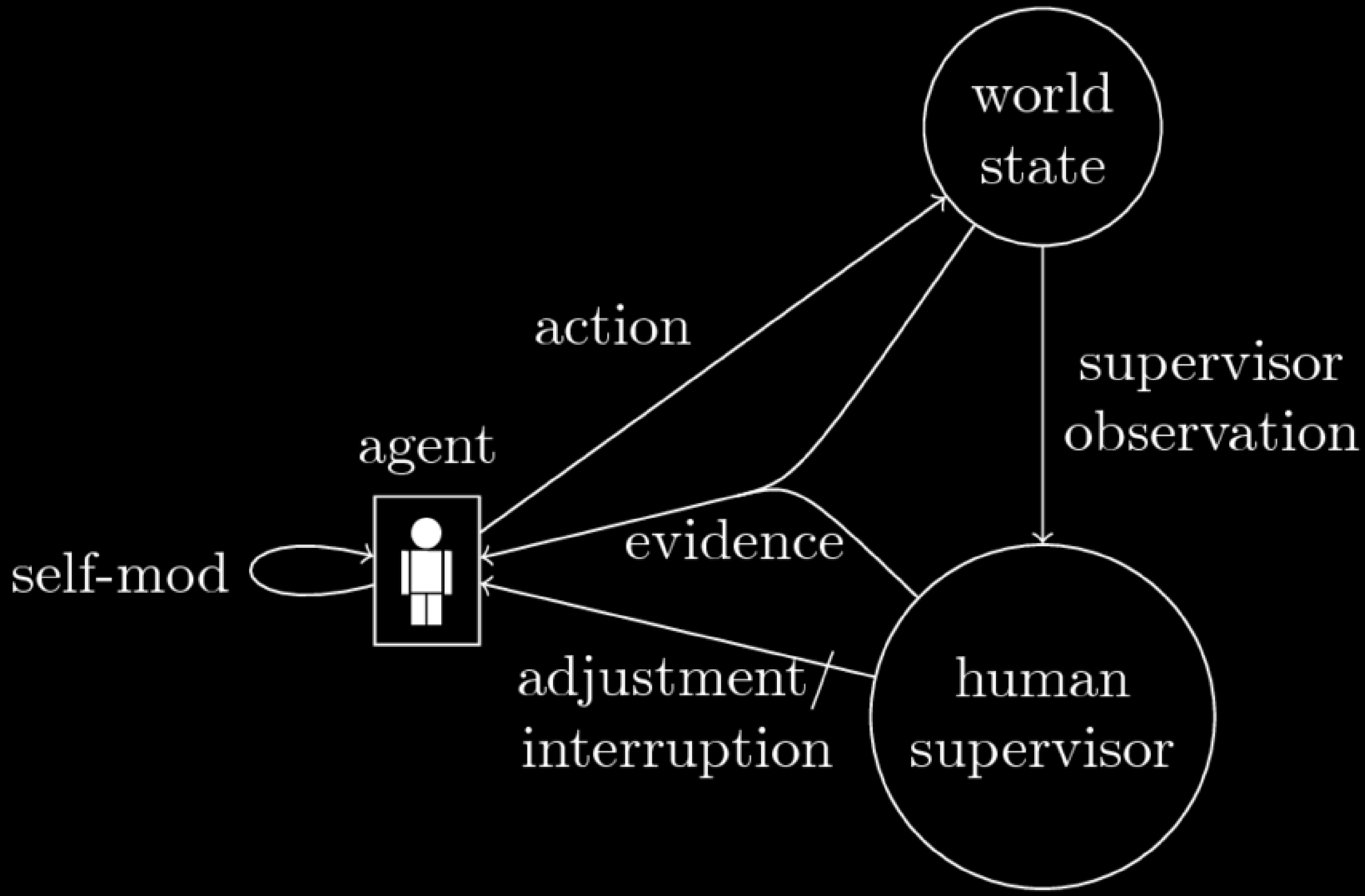
http://users.eecs.northwestern.edu/~argall/learning.html

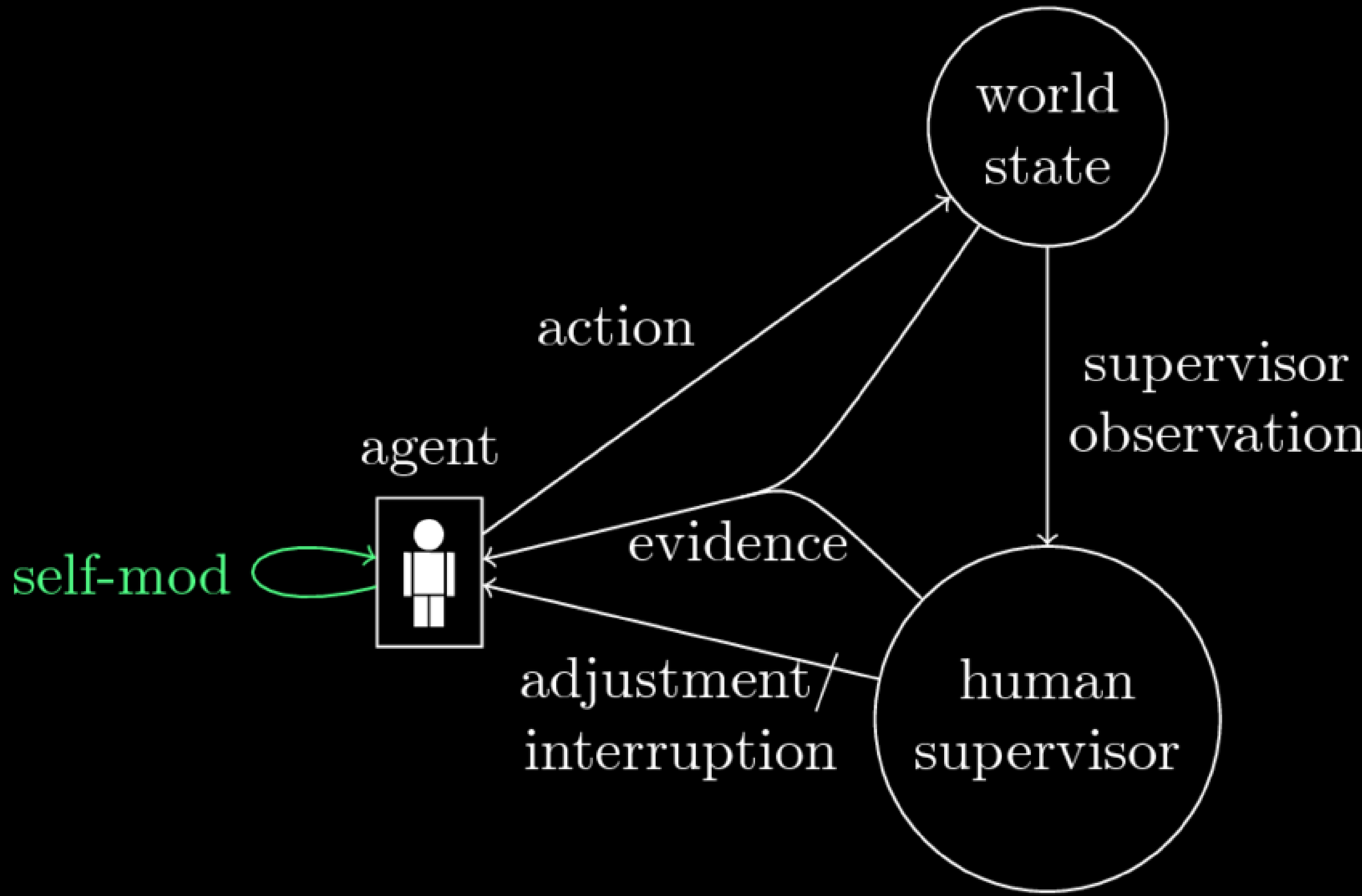
Assumption 3 (Ethical Authority)

- Humans are ethical authorities
- By definition?
- Human control = Safety?



Where can things go wrong?




world
state
action
supervisor
observation
agent
self-mod
evidence
adjustment/
interruption
human
supervisor

Self-modification

- Will the agent want to change itself?
- Omohundro (2008):

  An AI will not want to change its goals, because if future versions of the AI want the same goal, then the goal is more likely to be achieved
- As humans, utility function is part of our identity: *Would you self-modify into someone content just watching TV?*

Self-Modification

- Everitt et al. (2016): Formalising Omohundro's argument
- Three types of agents

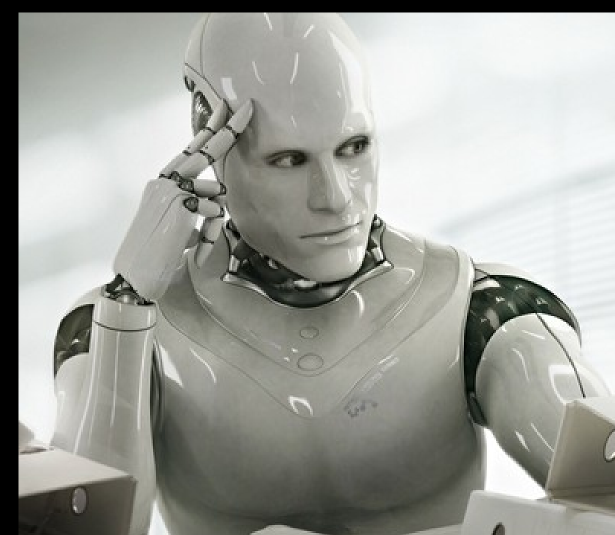
Hedonistic $u_k(æ_{<k})$

Wants to self-modify

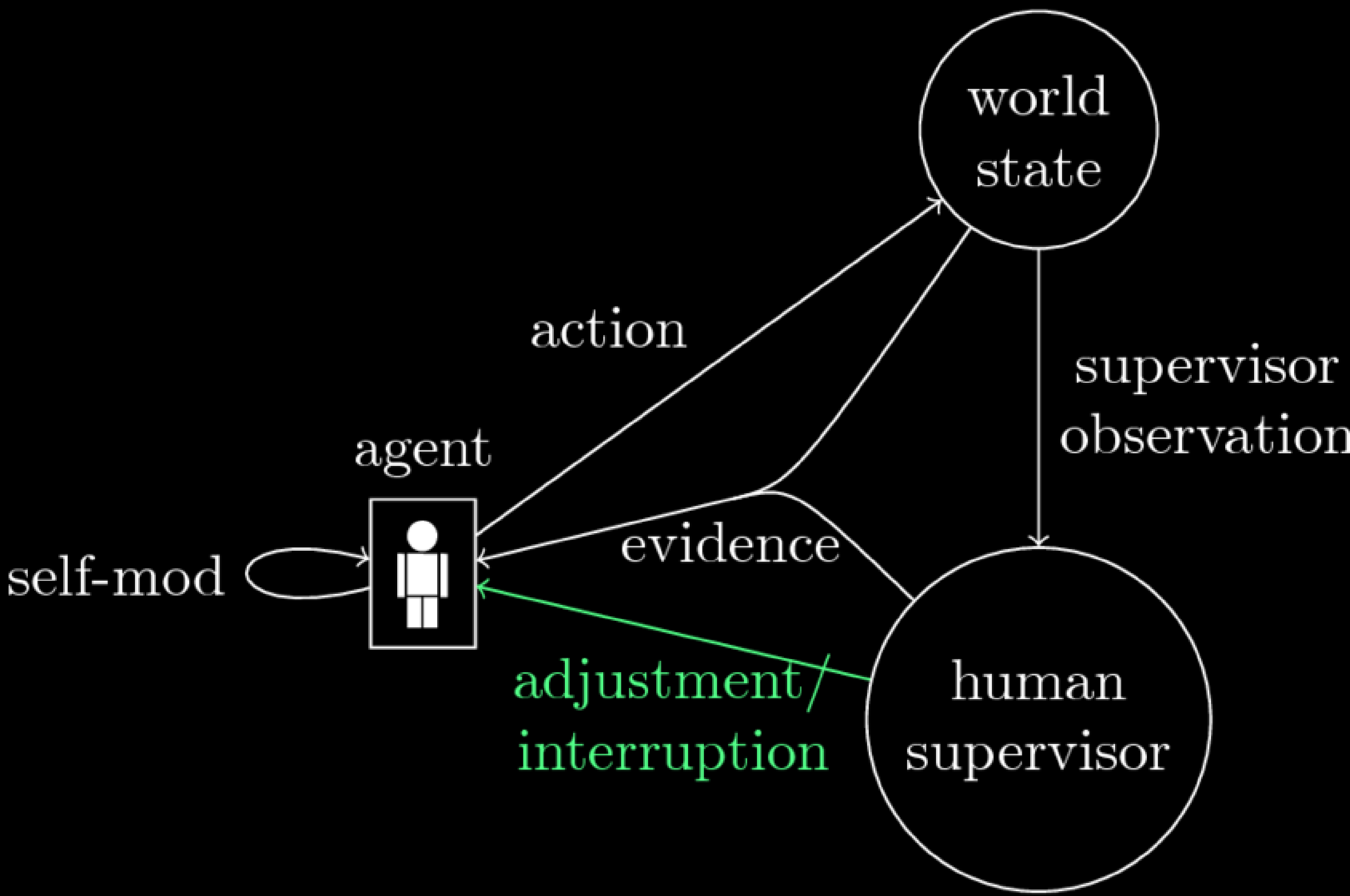
Ignorant $\pi_k = \pi_t$

Doesn't understand the difference

Realistic $u_t(æ_{<k})$



Resists (self)-modification


world state
action
supervisor observation
agent
evidence
self-mod
adjustment/ interruption
human supervisor

Corrigibility/Interruptability

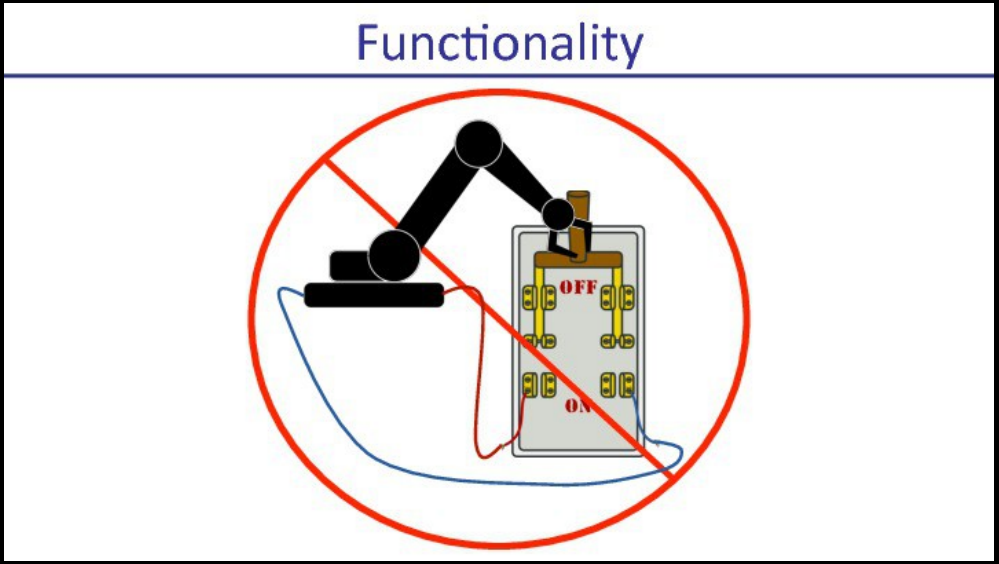
- What if we want to modify or shut down agent?
- Opposes self-preservation drive?
- Depends reward range for AIXI-like agents
  (Martin et al., 2016)

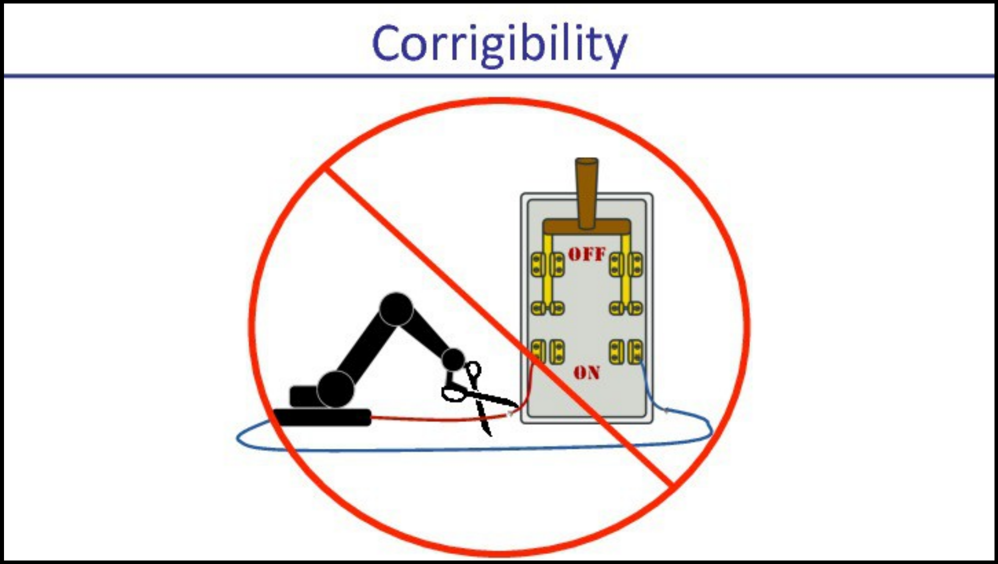
r = -1

r = 0
Death

r = 1

Functionality vs. Corrigibility

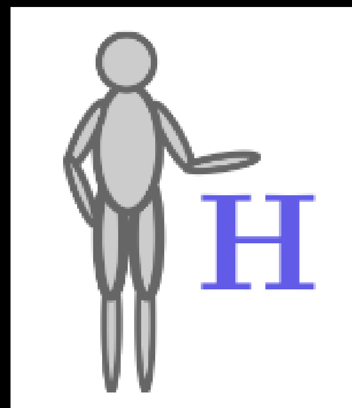




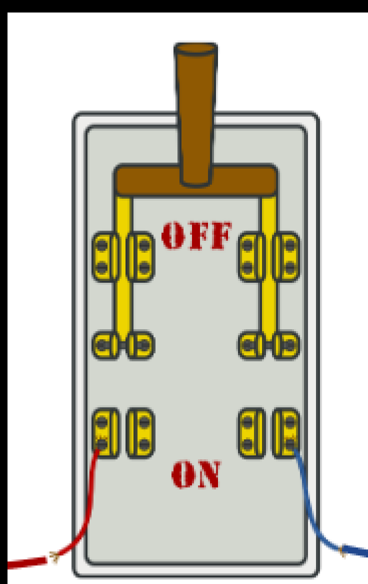
- Either being on or being off will have higher utility
- Why let the human decide?

Cooperative Inverse Reinforcement Learning (Hadfield-Menell et al, 2016)

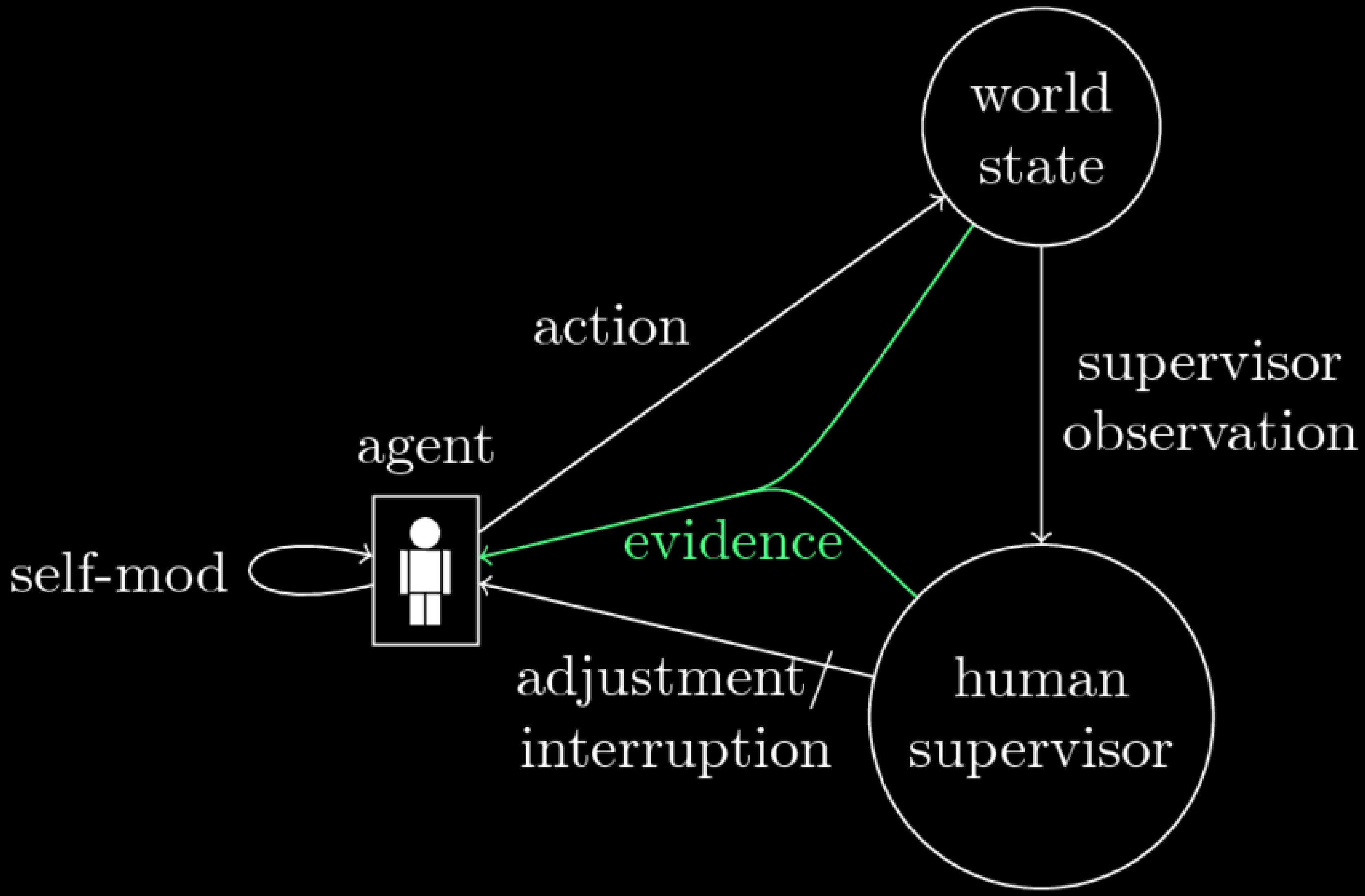
Doesn’t know u



Knows u
Possibly irrational

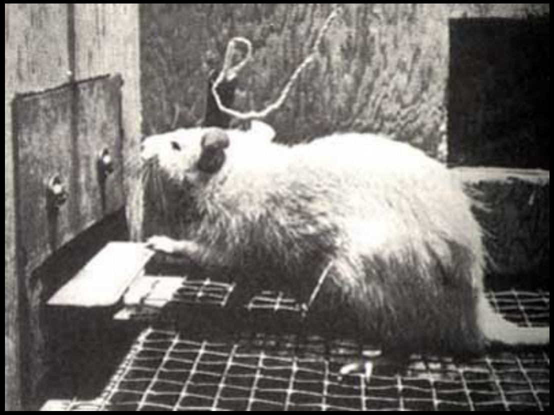


- Optimal action for agent is to let human decide, assuming:
  - Agent sufficiently uncertain about u, and
  - Agent believes human is sufficiently rational
- See also Safely Interruptible Agents (fiddles with details in the learning process) (Orseau & Armstrong, 2016)


world
state
action
supervisor
observation
agent
self-mod
evidence
adjustment/
interruption
human
supervisor

Evidence Manipulation

- Aka Wireheading, Delusionbox

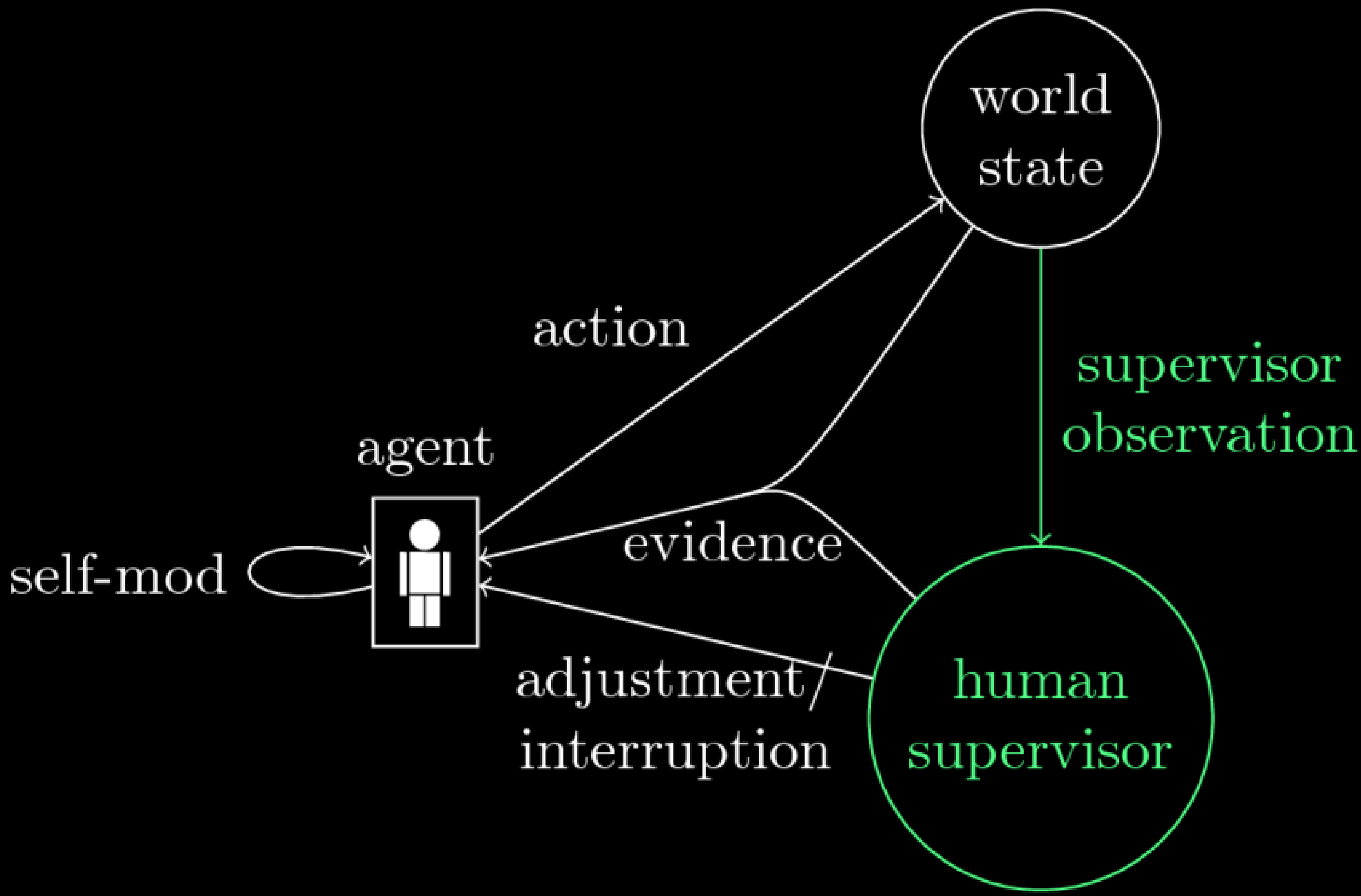


http://www.cinemablend.com/new/Wachowskis-Planning-Matrix-Trilogy-41905.html

- Ring and Orseau (2011):
  - Intelligent, real-world, reward maximising (RL) agent will wirehead
  - Knowledge-seeking agent will not wirehead

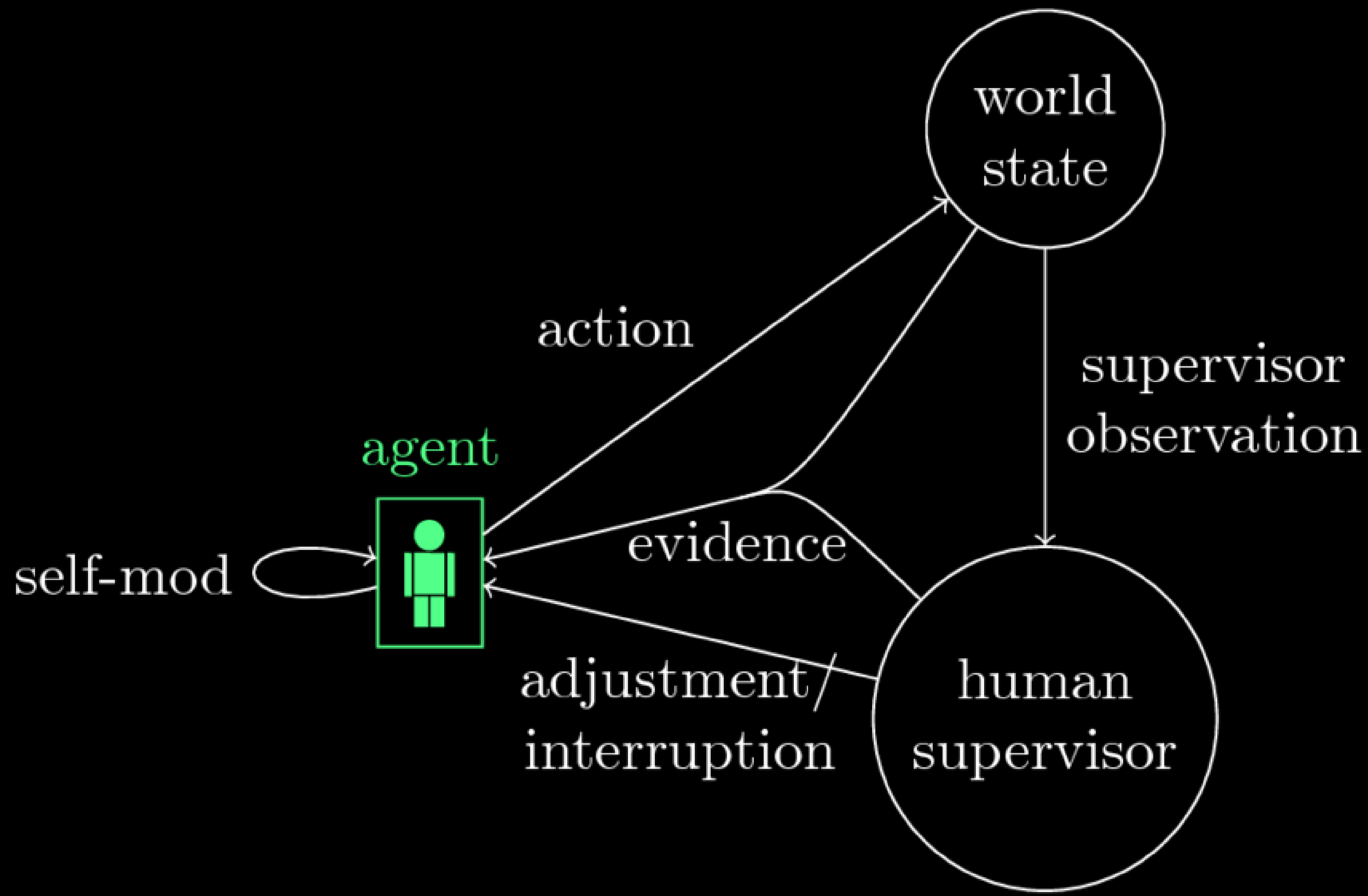
Value Reinforcement Learning

- Everitt and Hutter (2016)
- Instead of optimising r, optimise $\sum_i P(u_i|h, r_{1:t})u_i(h)$ with reward as evidence about true utility function
- ‘Too-good-to-be-true’ condition removes incentive to wirehead
- Current project:
  - Learn what a delusion is
  - No ‘too-good-to-be-true’ condition
  - Avoid wireheading by accident


world
state
action
supervisor
observation
agent
self-mod
evidence
adjustment/
interruption
human
supervisor

Supervisor Manipulation

- What about putting the human in a delusion box? (Matrix trilogy)
- No serious work yet
- Hedonistic utilitarians need not worry


world state
action
supervisor observation
agent
self-mod
evidence
adjustment/ interruption
human supervisor

(Imperfect) Learning

- Ideal learning:
  - Bayes theorem, conditional probability $P(\nu, u_i | h)$
  - AIXI/Solomonoff induction
- In practice: Model-free learning more efficient $\mathbb{E}[\sum_{k=t}^{\infty} r_k \mid h_t, a]$
  - Q-learning
  - Sarsa
- Current project: Model-free AIXI/General RL

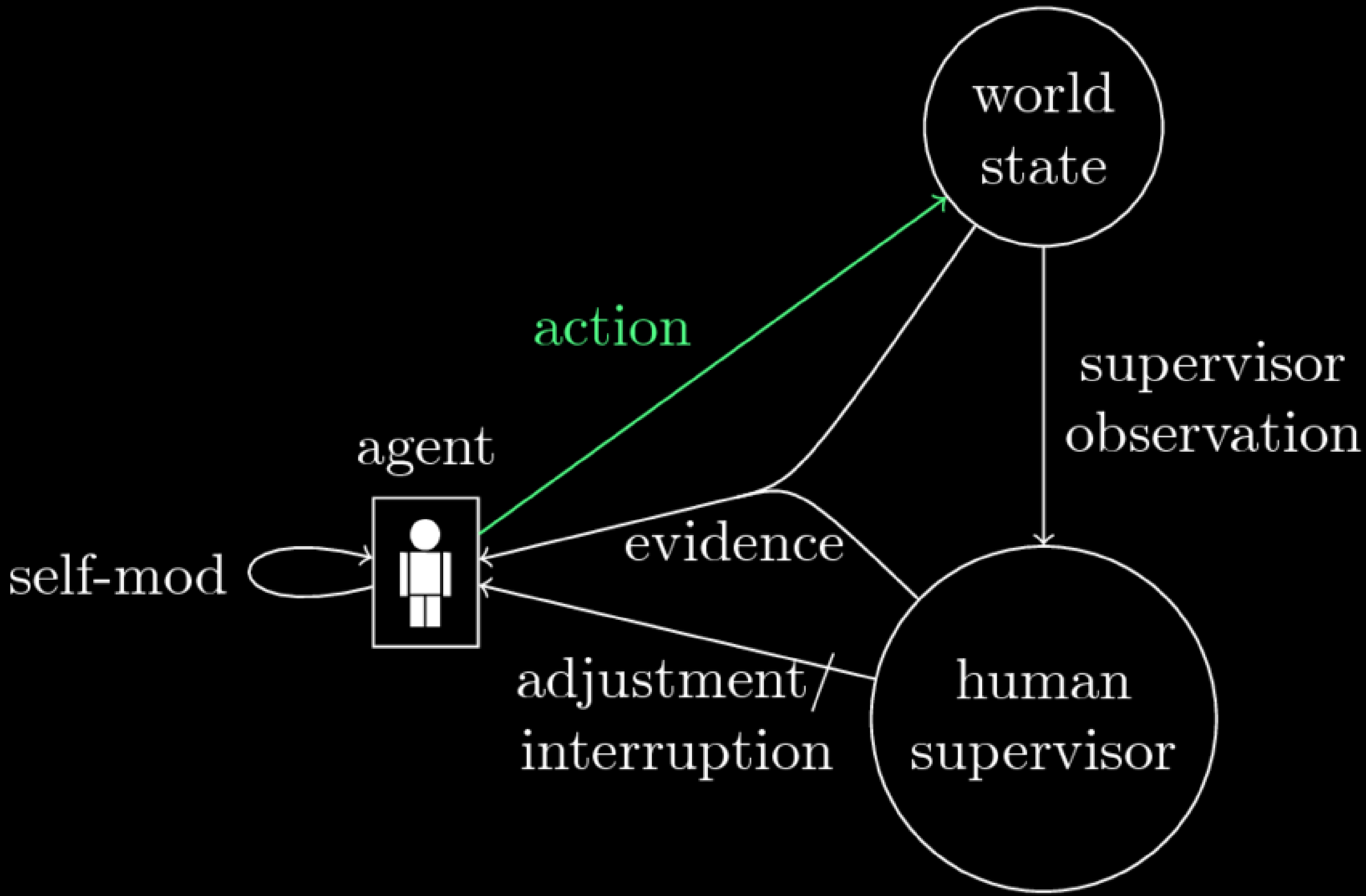
http://childpsychologistindia.blogspot.com.au/2013/10/difference-betwee

MIRI's Logical inductor (2016)

- General model of belief states for deductively limited reasoners
- Good properties
  - Converges to probability
  - Outpaces deduction
  - Self-trust
  - Scientific induction

Decision Making

- Open source Prisoner’s Dilemma Barasz et al. (2014), Critch (2016)
- Refinements of Expected Utility Maximisation:
  - Causal DT
  - Evidential DT
  - Updateless DT
  - Timeless DT
- Logical inductors possibly useful (current MIRI research)


world
state
action
supervisor
observation
agent
self-mod
evidence
adjustment/
interruption
human
supervisor

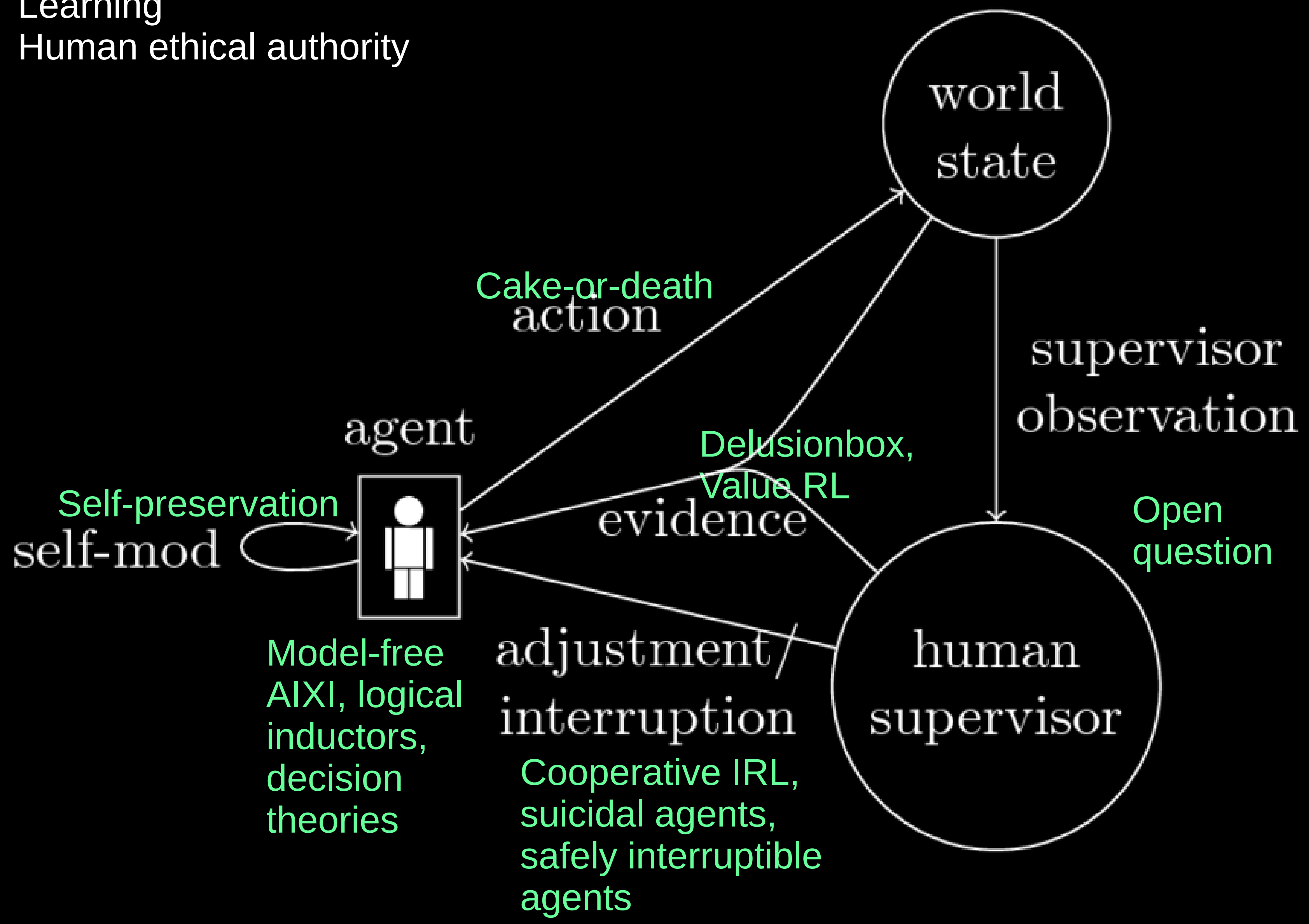
Biased Learning

- Cake or Death?
  - $P(u_{\text{death}}) = P(u_{\text{cake}}) = 0.5$
  - Options:
    - Kill 3 people
    - Bake 1 cake
    - Ask (for free) what's the right thing to do
  - u(ask, bake cake) = 1
  - u(kill) = 1.5
- Motivated value selection (Armstrong, 2015)
  Interactive inverse RL (Armstrong and Leike, 2016)
- For properly Bayesian agents, no problem:

$$\underbrace{\mathbb{E}[\max_a V(a)]}_{\text{asking}} > \underbrace{\max_a \mathbb{E}[V(a)]}_{\text{not asking}}$$

Assumptions:

- True utility function
- Learning
- Human ethical authority



References

- Armstrong (2015)
  Motivated Value Selection. AAAI Workshop
- Armstrong and Leike (2016)
  Interactive Inverse Reinforcement Learning. NIPS workshop
- Barasz, Christiano, Fallenstein, Herreshoff, LaVictoire, Yudkowsky (2014)
  Robust Cooperation in the Prisoner's Dilemma: Program Equilibrium via Provability Logic. Arxiv
- Critch (2016)
  Parametric Bounded Löb's Theorem and Robust Cooperation of Bounded Agents. Arxiv
- Dewey (2011)
  Learning what to value. AGI
- Everitt, Filan, Daswani, and Hutter (2016)
  Self-modification of policy and utility function in rational agents, AGI.
- Everitt and Hutter (2016)
  Avoiding Wireheading with Value Reinforcement Learning. AGI
- Garrabrant, Benson-Tilsen, Critch, Soares, Taylor (2016)
  Logical Induction. Arxiv
- Martin, Everitt, and Hutter (2016)
  Death and Suicide in Universal Artificial Intelligence, AGI
- Omohundro (2008)
  The Basic AI Drives, AGI
- Hadfield-Menell, Dragan, Abbeel, Russell (2016)
  Cooperative Inverse Reinforcement Learning. Arxiv
- Orseau and Armstrong (2016)
  Safely interruptible agents. UAI
- Ring and Orseau (2011)
  Delusion, Survival, and Intelligent Agents. AGI