Logic Can Save Us from “Killer Robots”

Selmer Bringsjord
Naveen Sundar G et al.

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy, New York 12180 USA

IFLAI1
3/30/2023

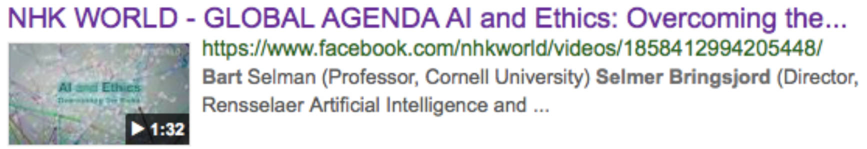
R A I R
Rensselaer AI and Reasoning Lab

Logic Can Save Us from “Killer Robots”

Selmer Bringsjord
Naveen Sundar G et al.

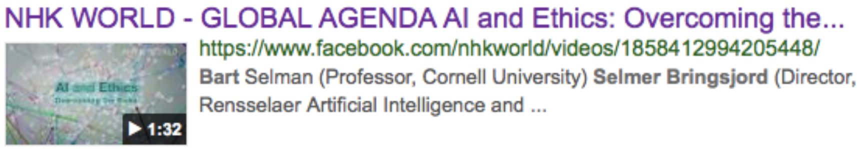
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy, New York 12180 USA

IFLAI1
3/30/2023



NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...
https://www.facebook.com/nhkworld/videos/1858412994205448/
Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director, Rensselaer Artificial Intelligence and ...

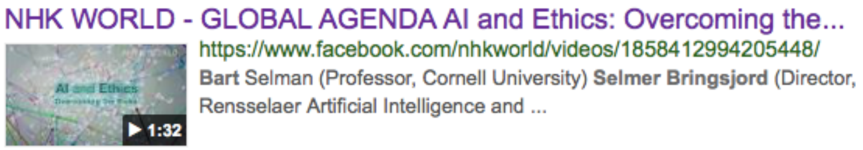
The PAID Problem



NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...
https://www.facebook.com/nhkworld/videos/1858412994205448/
Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director, Rensselaer Artificial Intelligence and ...

The PAID Problem

$\forall \mathtt{x} : \mathtt{Agents}$



NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...

https://www.facebook.com/nhkworld/videos/1858412994205448/

Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director, Rensselaer Artificial Intelligence and ...

The PAID Problem

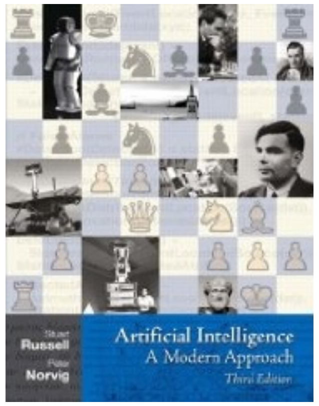
$\forall \mathtt{x} : \mathtt{Agents}$

Powerful(x) + **A**utonomous(x) + **I**ntelligent(x) => **D**angerous(x)/**D**estroy_Us

“We’re in *very* deep trouble.”

“We’re in *very* deep trouble.”

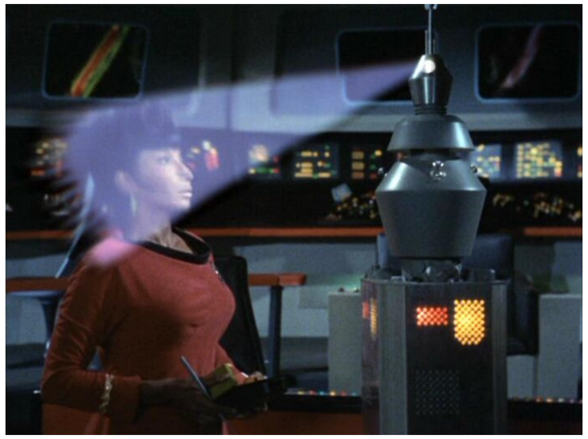
“We’re in *very* deep trouble.”



While the PAI machines aren’t quite as easy to neutralize as the destructive machines vanquished in *Star Trek: TOS*, these relevant four episodes are remarkably instructive.

“The Ultimate Computer”
S2 E24

“The Return of the Archons”
S1 E21



“The Changeling”
S2 E3

“I, Mudd”
S2 E8

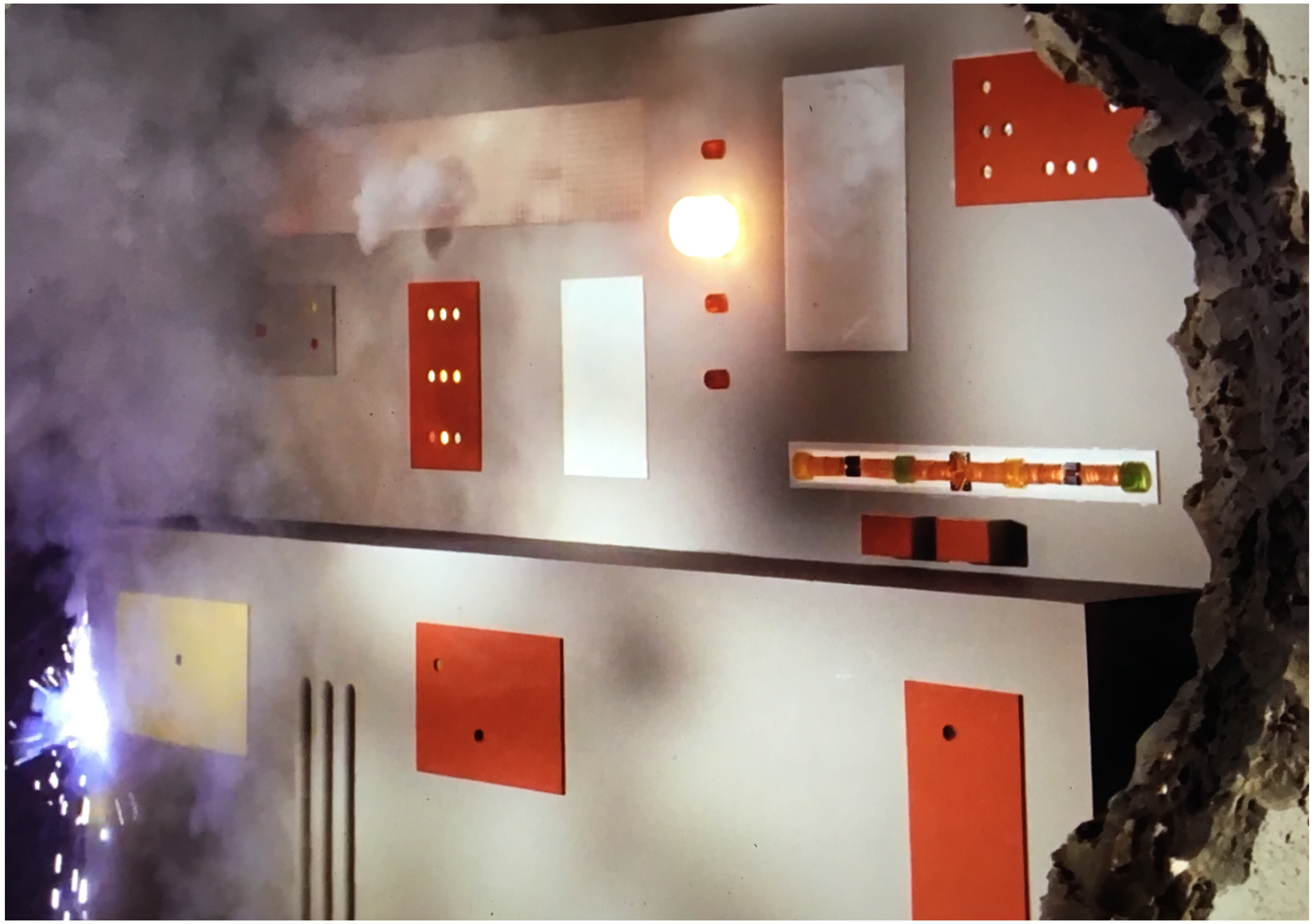
Logic Thwarts Landru!

First Suspicion That It's a Mere Computer Running the Show

Logic Thwarts Landru!

Landru is Indeed Merely a Computer
(the real Landru having done the programming)

Logic Thwarts Landru!



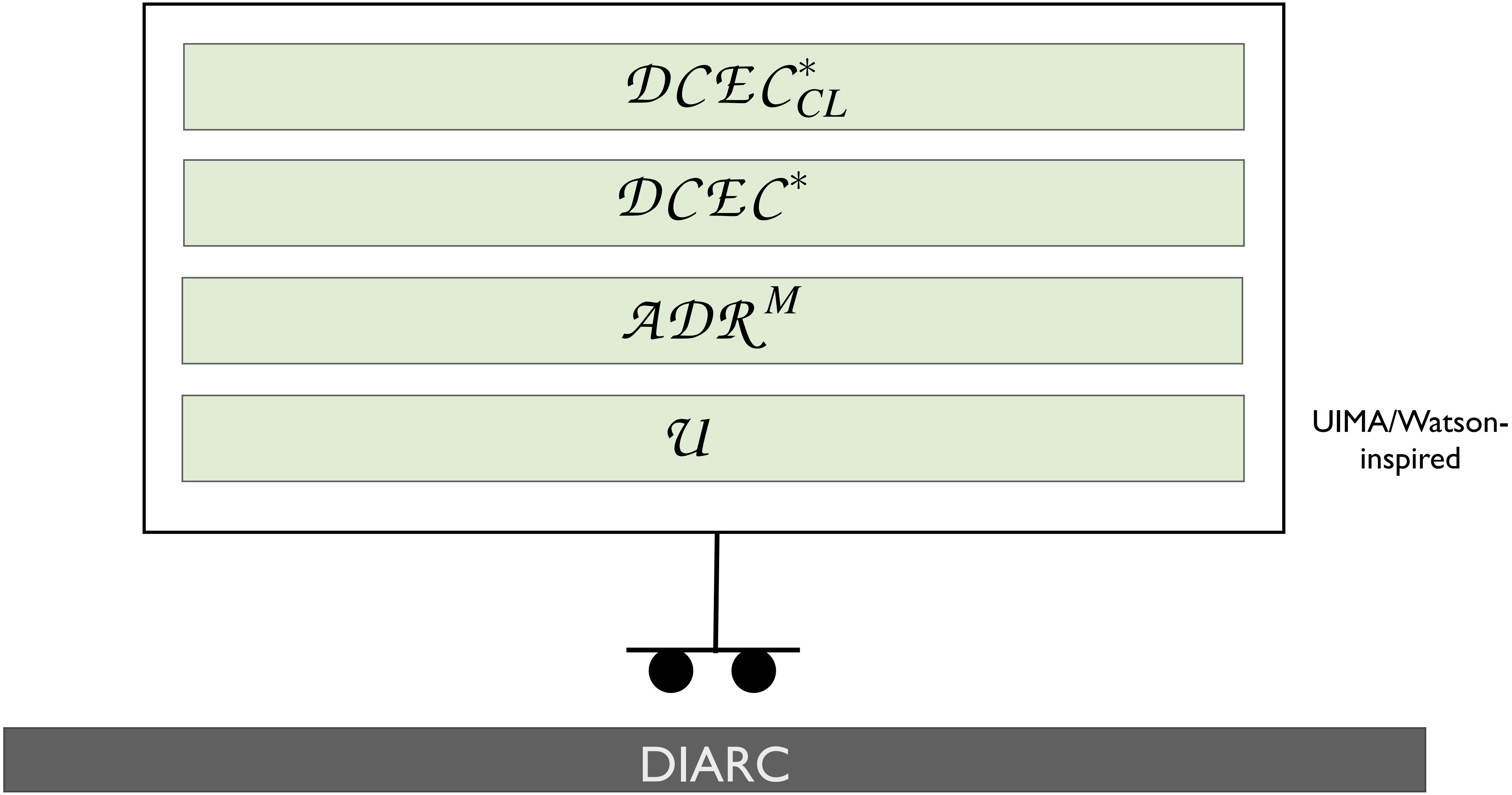
Landru Kills Himself Because Kirk/Spock Argue He Has Violated the Prime Directive for Good by Denying Creativity to Others

Logic Thwarts Nomad!

(with the Liar Paradox)

I.
Cognitive Calculi …

Hierarchy of Ethical Reasoning



$\mathcal{DCEC}^*_{CL}$

$\mathcal{DCEC}^*$

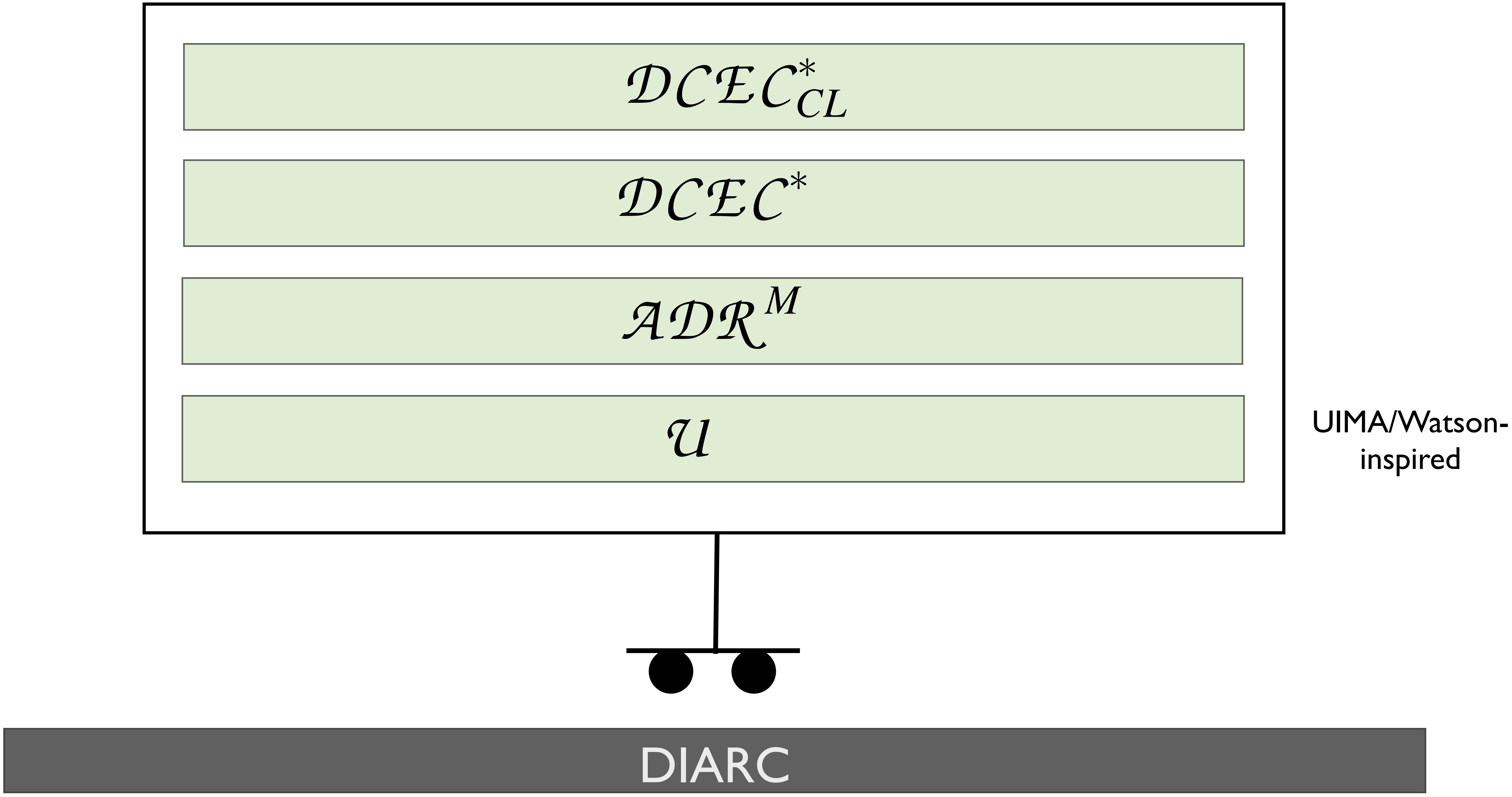
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-inspired

DIARC

Hierarchy of Ethical Reasoning



$\mathcal{DCEC}^*_{\mathcal{CL}}$

$\mathcal{DCEC}^*$

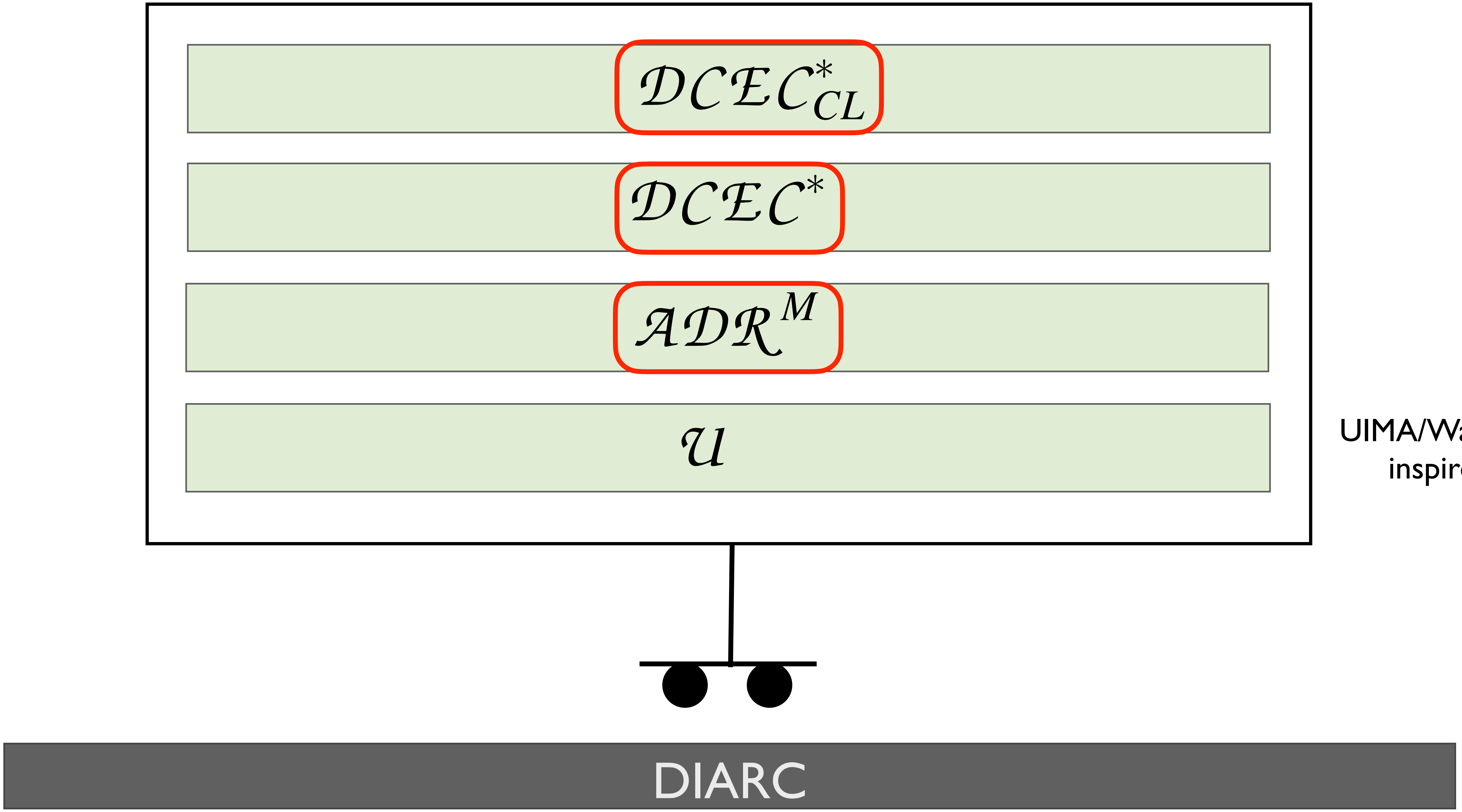
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-inspired

DIARC

Hierarchy of Ethical Reasoning



$\mathcal{DCEC}^*_{CL}$

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

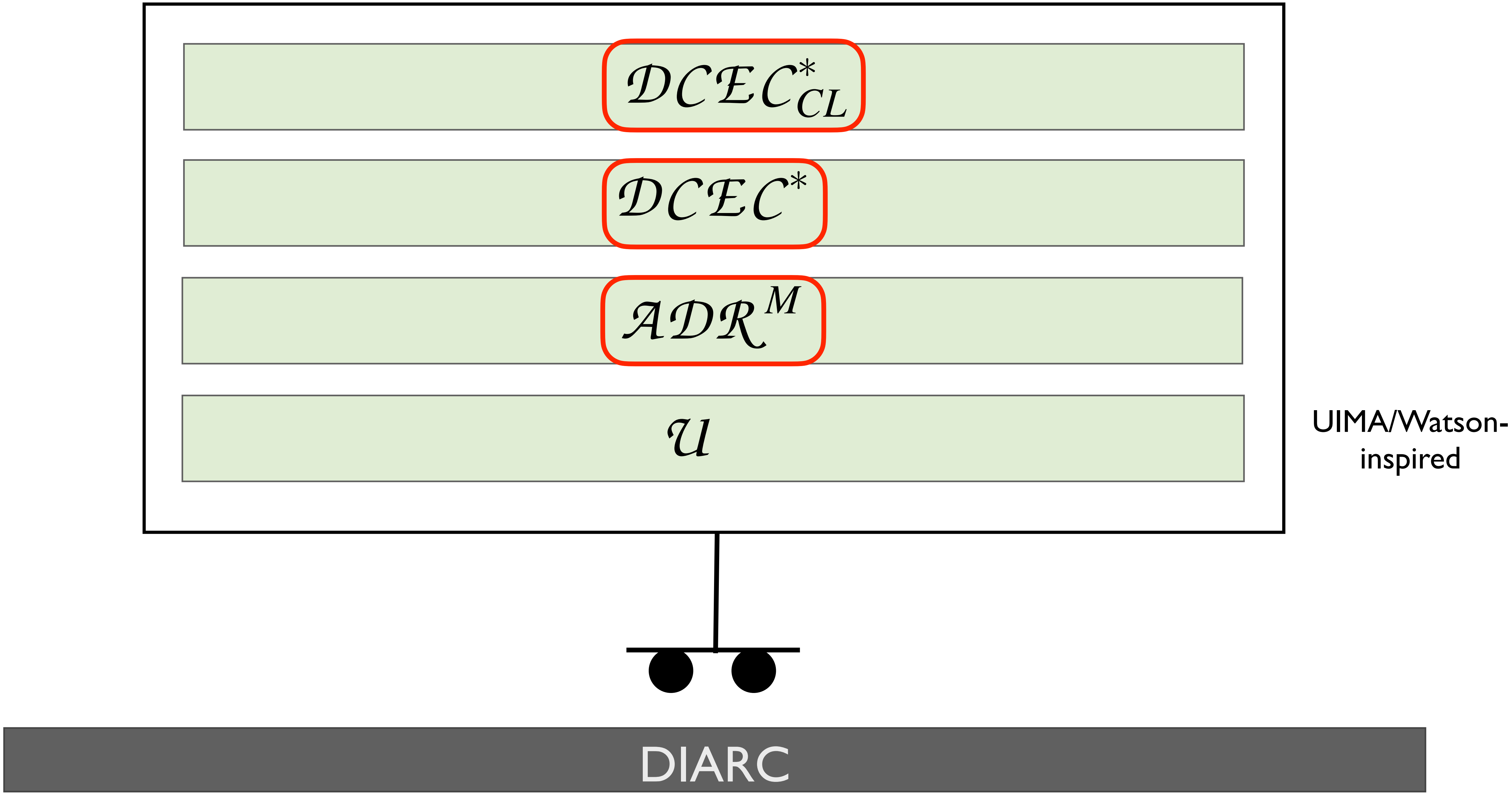
$\mathcal{U}$

UIMA/Watson-inspired

DIARC

Hierarchy of Ethical Reasoning

Not paradox-prone deontic logics!



$\mathcal{DCEC}^*_{CL}$

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$

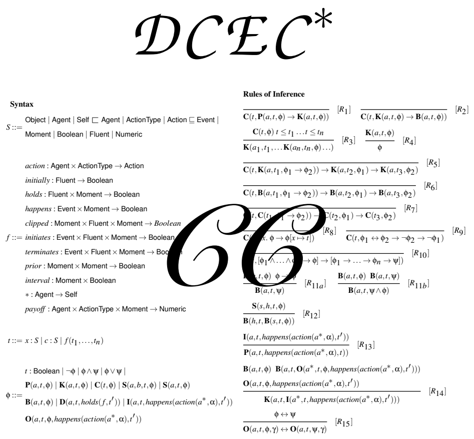
UIMA/Watson-inspired

DIARC

"Universal Cognitive Calculus"

IGNIS
CALI DITAS
AER
HVMI DITAS
AQVA
FRIGI DITAS
TERRA
SIC CITAS
CONTRA
RIA
REMISSA
SVMMA

Logic Theorist
(birth of modern logicist AI)



$\mathcal{DCEC}^*$

Syntax

$$S ::= \text{Object} \mid \text{Agent} \mid \text{Self} \sqsubset \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Boolean} \mid \text{Fluent} \mid \text{Numeric}$$

$$f ::= \begin{array}{l} \mathit{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ \mathit{initially} : \text{Fluent} \rightarrow \text{Boolean} \\ \mathit{holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{clipped} : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \mathit{Boolean} \\ \mathit{initiates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{terminates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{prior} : \text{Moment} \times \text{Moment} \rightarrow \text{Boolean} \\ \mathit{interval} : \text{Moment} \times \text{Boolean} \\ * : \text{Agent} \rightarrow \text{Self} \\ \mathit{payoff} : \text{Agent} \times \text{ActionType} \times \text{Moment} \rightarrow \text{Numeric} \end{array}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{array}{l} t : \text{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \\ \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \\ \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\mathit{holds}(f,t')) \mid \mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \end{array}$$

Rules of Inference

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\phi))}\ [R_1] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \rightarrow \mathbf{B}(a,t,\phi))}\ [R_2]$$

$$\frac{\mathbf{C}(t,\phi)\ t \leq t_1 \ldots t \leq t_n}{\mathbf{K}(a_1,t_1,\ldots \mathbf{K}(a_n,t_n,\phi)\ldots)}\ [R_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi}\ [R_4]$$

$$\frac{}{\mathbf{C}(t,\mathbf{K}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{K}(a,t_2,\phi_1) \rightarrow \mathbf{K}(a,t_3,\phi_2)}\ [R_5]$$

$$\frac{}{\mathbf{C}(t,\mathbf{B}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{B}(a,t_2,\phi_1) \rightarrow \mathbf{B}(a,t_3,\phi_2)}\ [R_6]$$

$$\frac{}{\mathbf{C}(t,\mathbf{C}(t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2,\phi_1) \rightarrow \mathbf{C}(t_3,\phi_2)}\ [R_7]$$

$$\frac{}{\mathbf{C}(t,\forall x.\ \phi \rightarrow \phi[x \mapsto t])}\ [R_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)}\ [R_9]$$

$$\frac{}{\mathbf{C}(t,[\phi_1 \wedge \ldots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \psi])}\ [R_{10}]$$

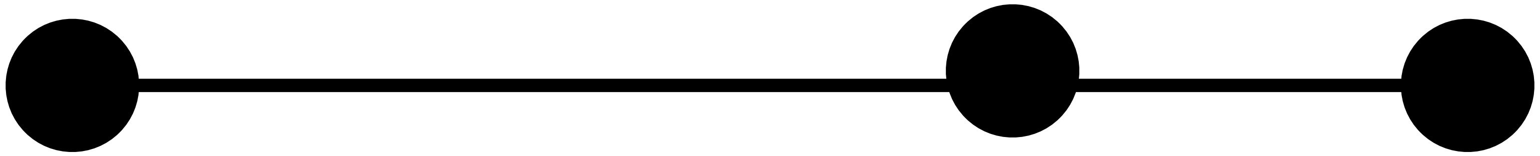
$$\frac{\mathbf{B}(a,t,\phi)\quad \phi \rightarrow \psi}{\mathbf{B}(a,t,\psi)}\ [R_{11a}] \qquad \frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi \wedge \phi)}\ [R_{11b}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}]$$

$$\frac{\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))}{\mathbf{P}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t))}\ [R_{13}]$$

$$\frac{\begin{array}{c}\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\mathbf{O}(a^*,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))) \\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))\end{array}}{\mathbf{K}(a,t,\mathbf{I}(a^*,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')))}\ [R_{14}]$$

$$\frac{\phi \leftrightarrow \psi}{\mathbf{O}(a,t,\phi,\gamma) \leftrightarrow \mathbf{O}(a,t,\psi,\gamma)}\ [R_{15}]$$



1666

1956

2019

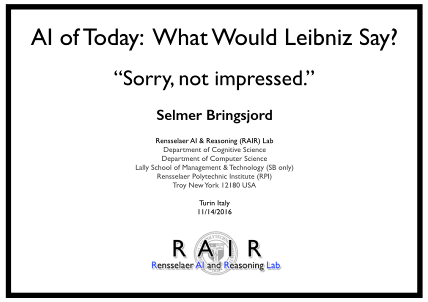
Leibniz

1.5 centuries < Boole!
2.5 centuries < Kripke

$\int$

Simon

R A I R
Rensselaer AI and Reasoning Lab



AI of Today: What Would Leibniz Say?

"Sorry, not impressed."

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology (SB only)
Rensselaer Polytechnic Institute (RPI)
Troy New York 12180 USA

Turin Italy
11/14/2016

R A I R
Rensselaer AI and Reasoning Lab

II.
Early Progress With Our Calculi:
Simple Dilemmas;
Non-Akratic Robots

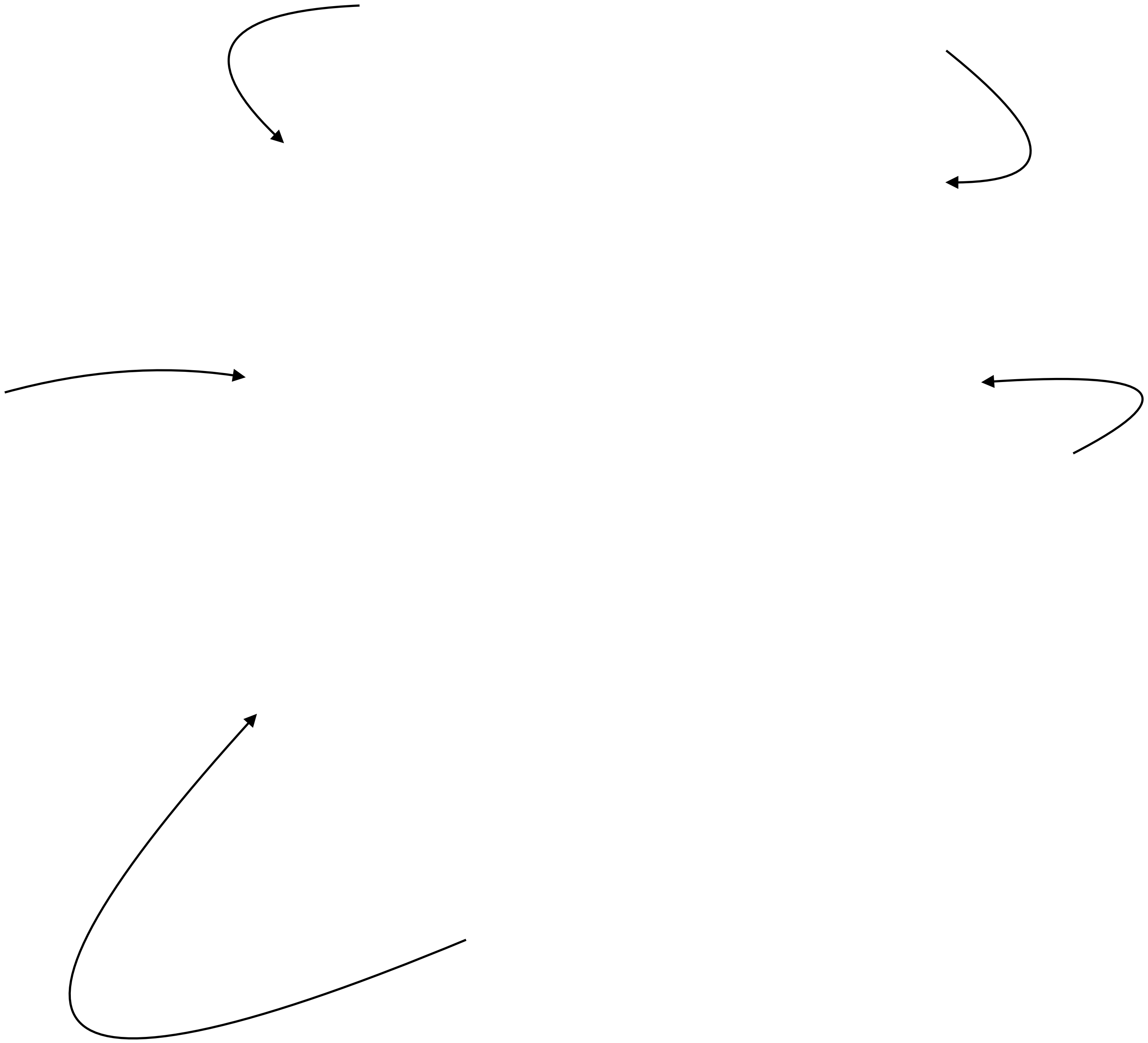
NewScientist

Ethical robots save humans

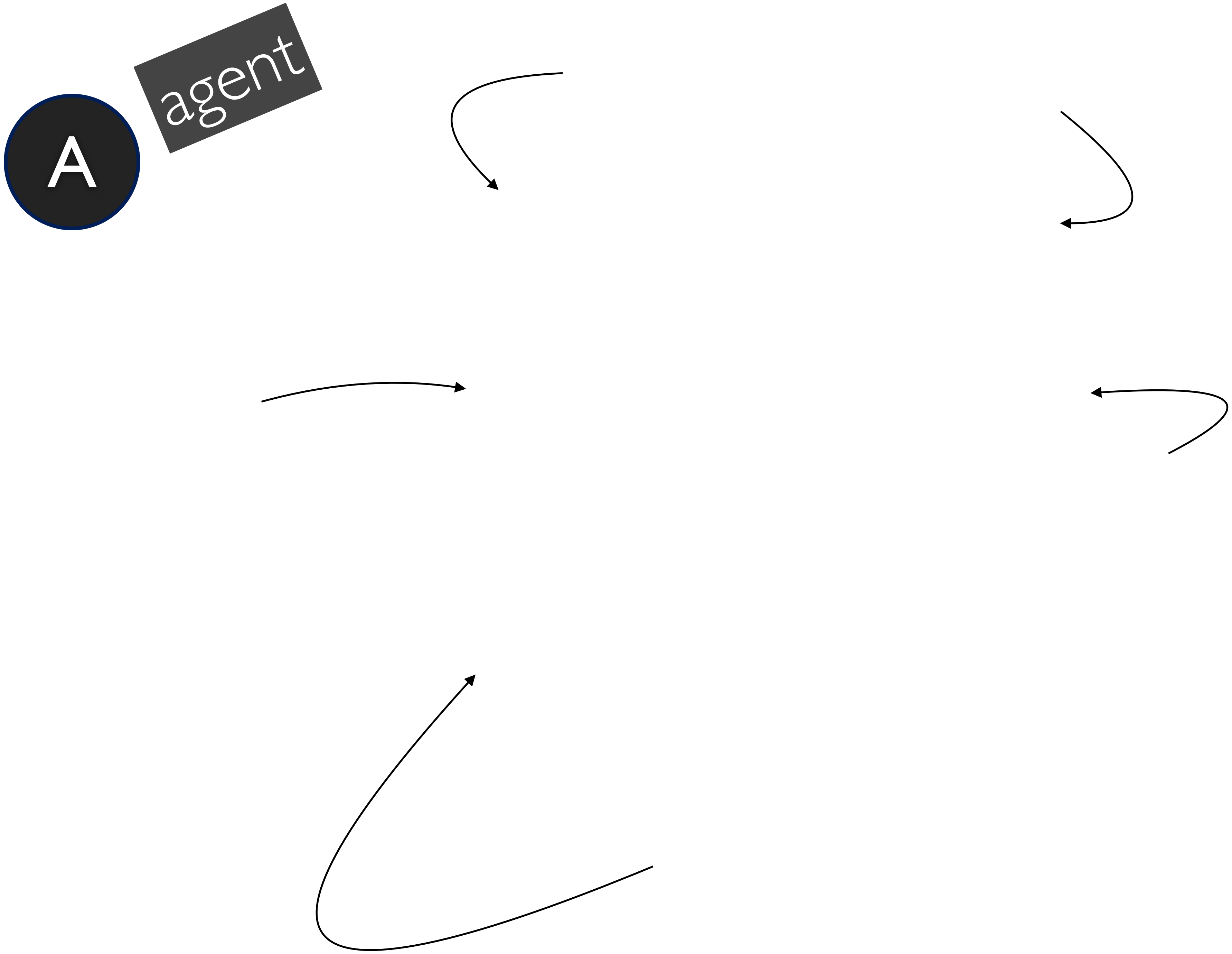
NewScientist

Ethical robots save humans

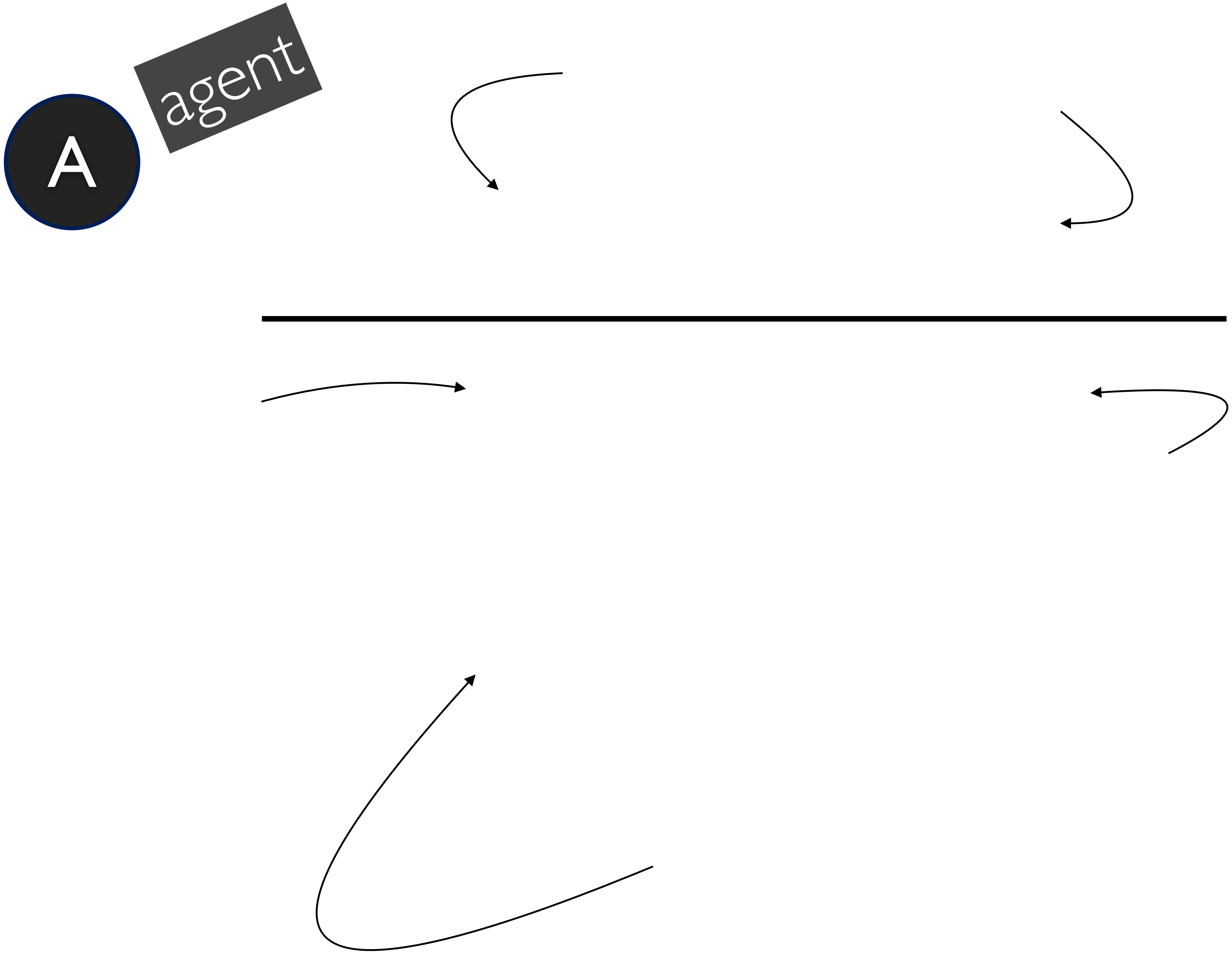
Informal Context of Akrasia



Informal Context of Akrasia



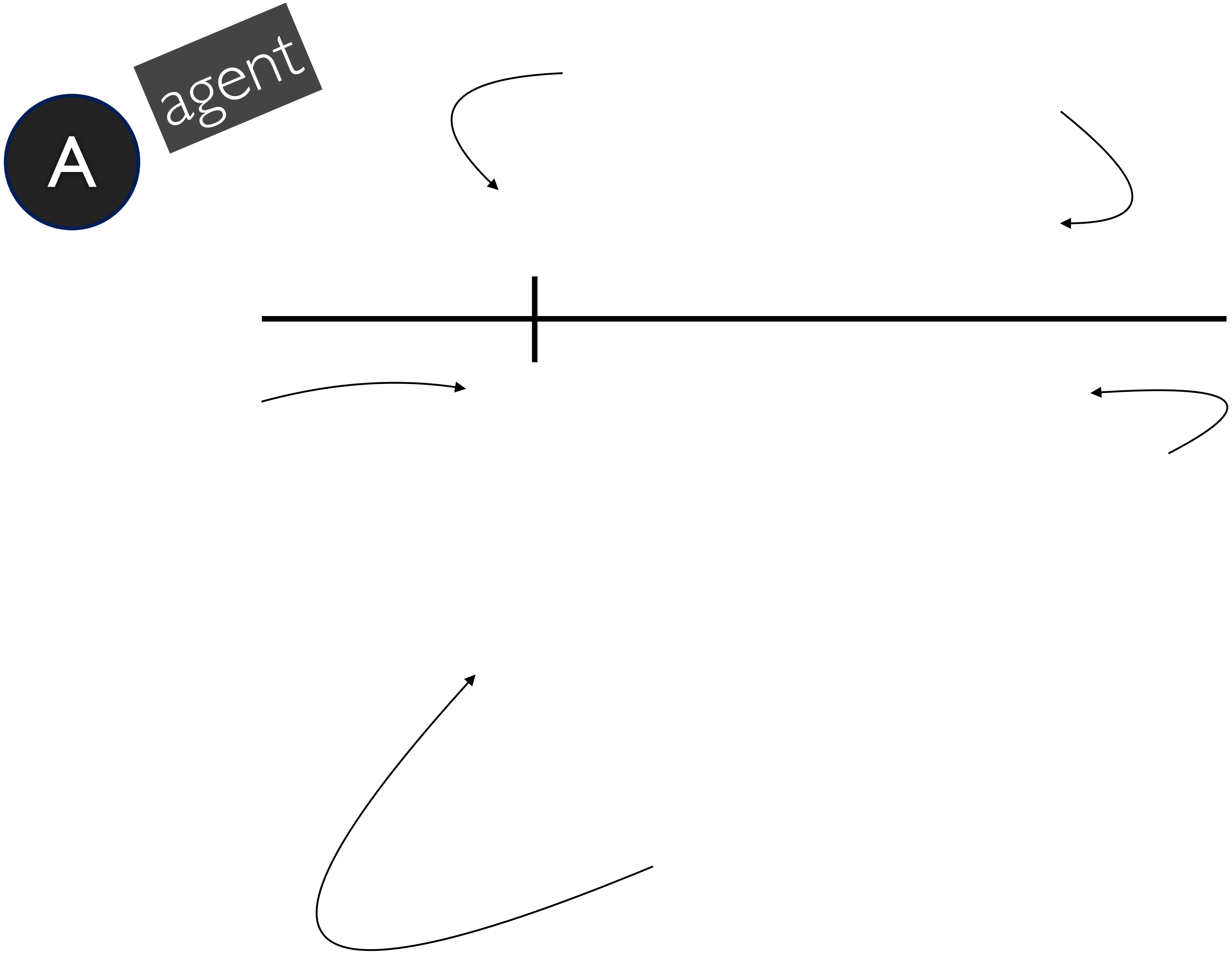
Informal Context of Akrasia



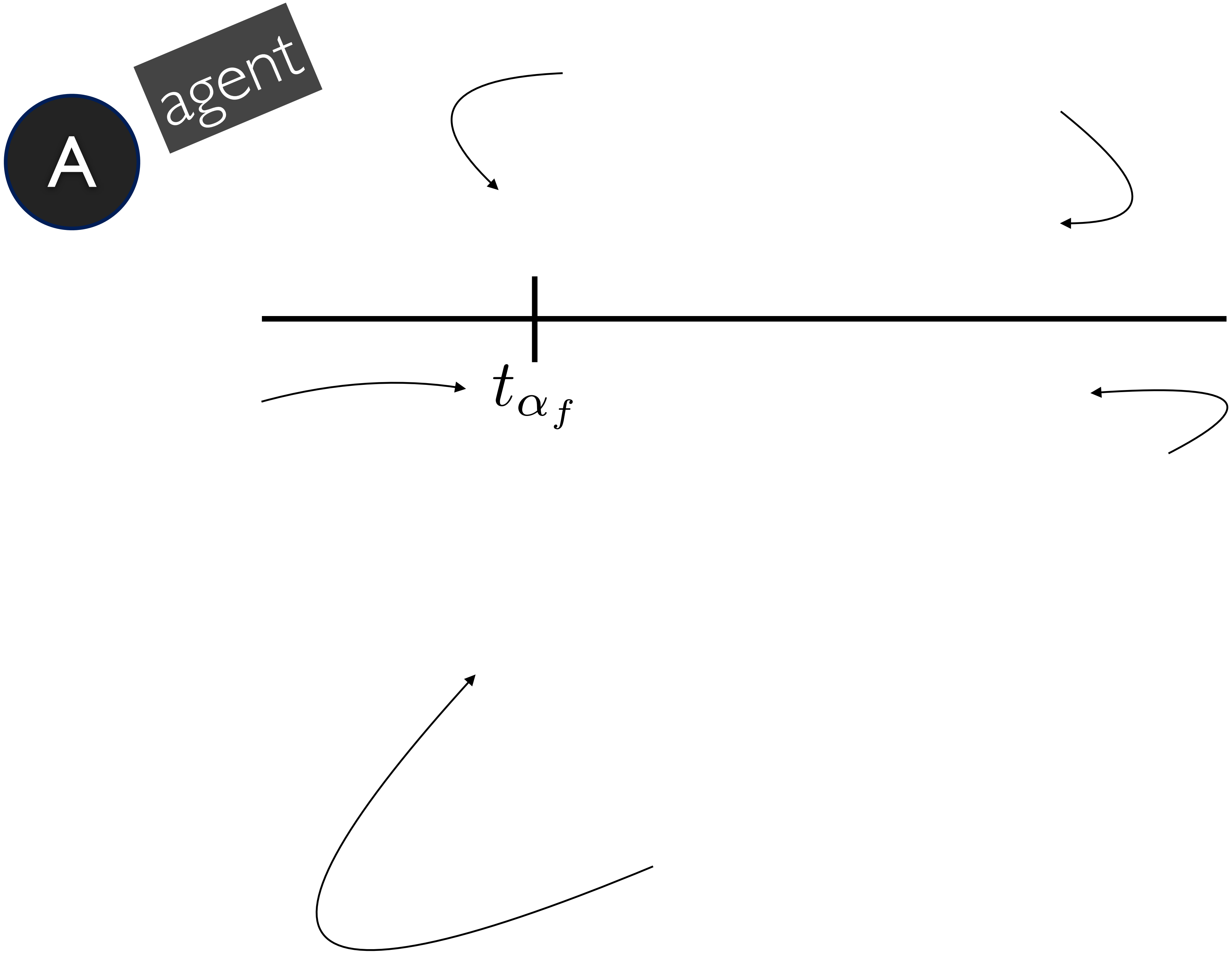
A

agent

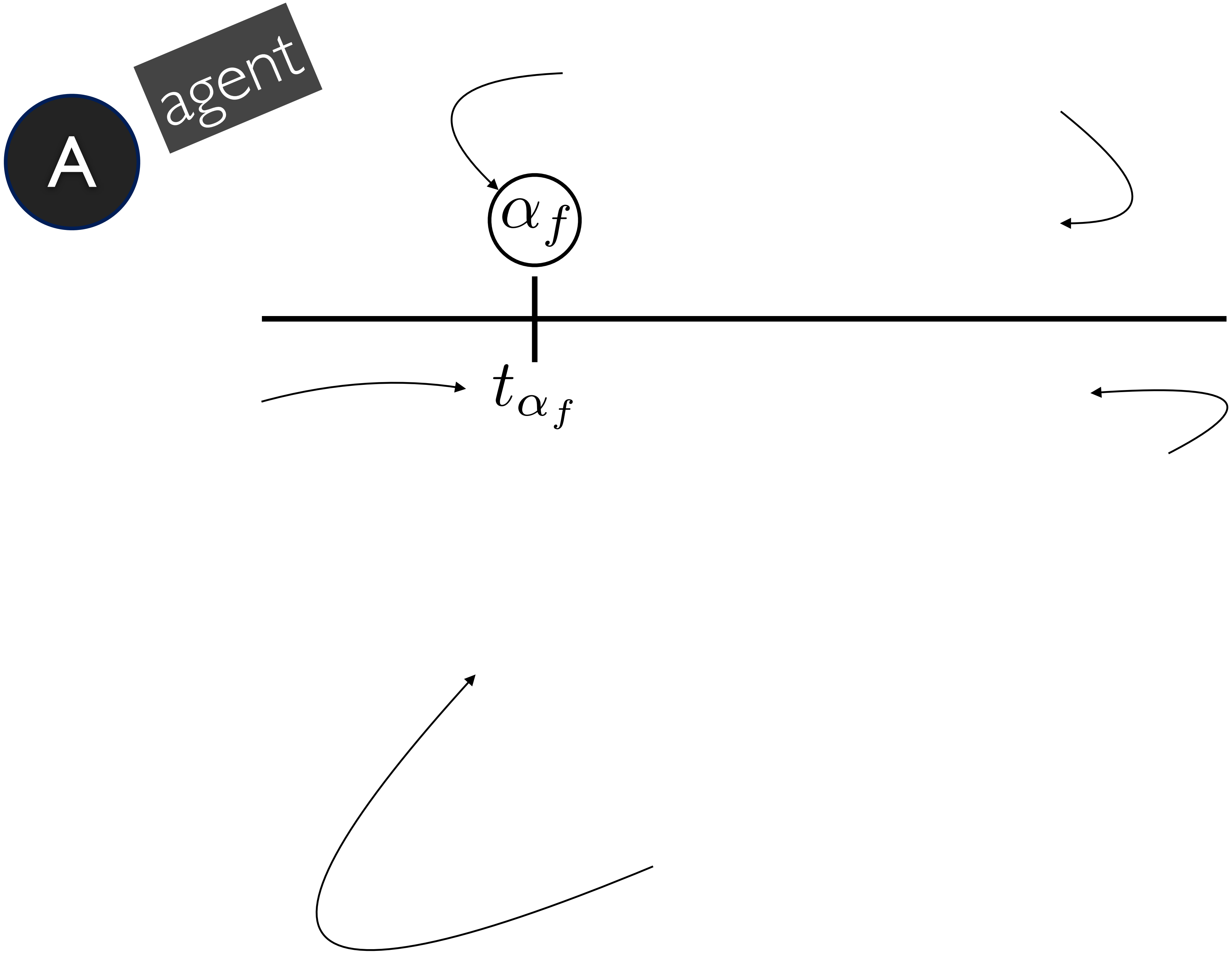
Informal Context of Akrasia



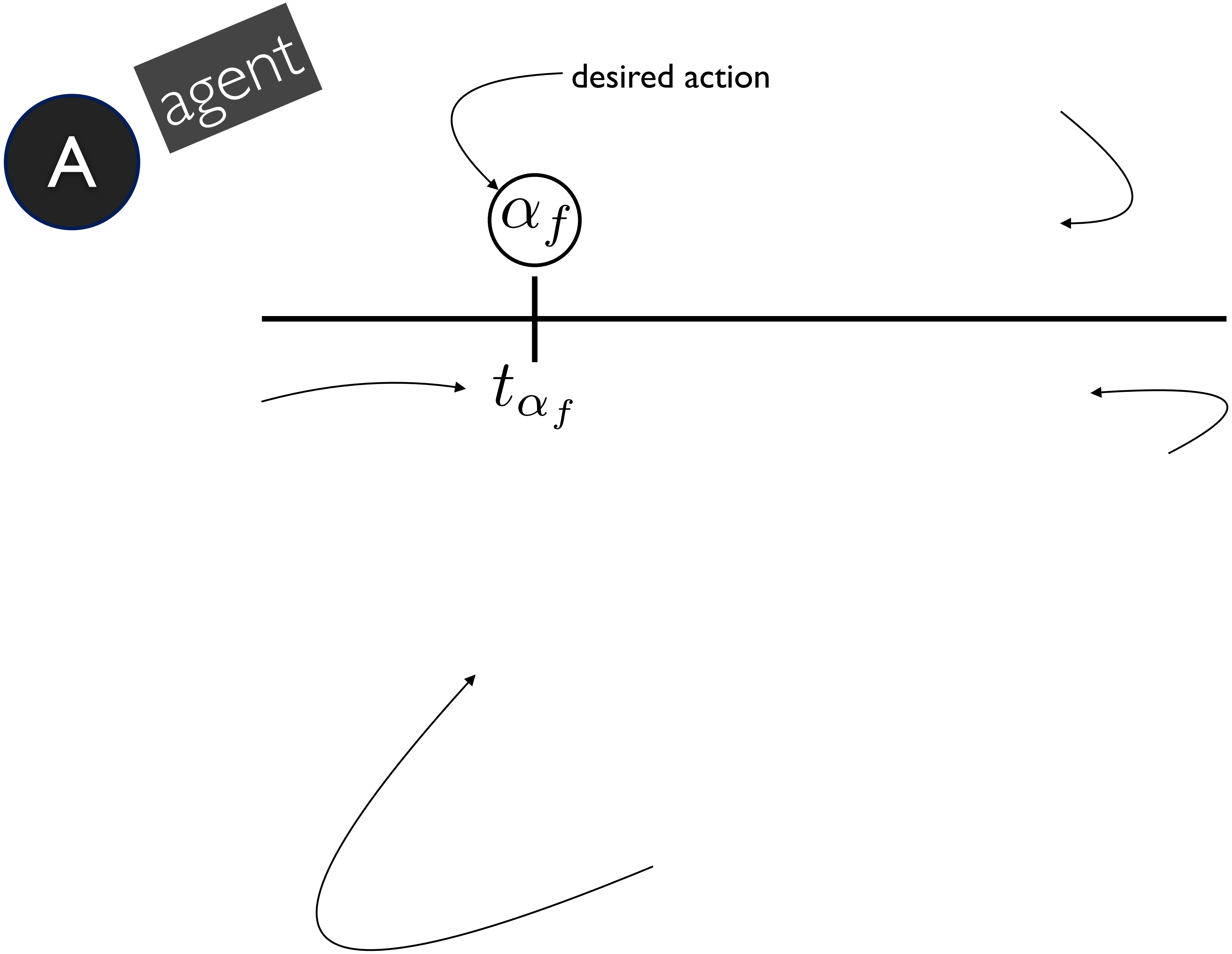
Informal Context of Akrasia



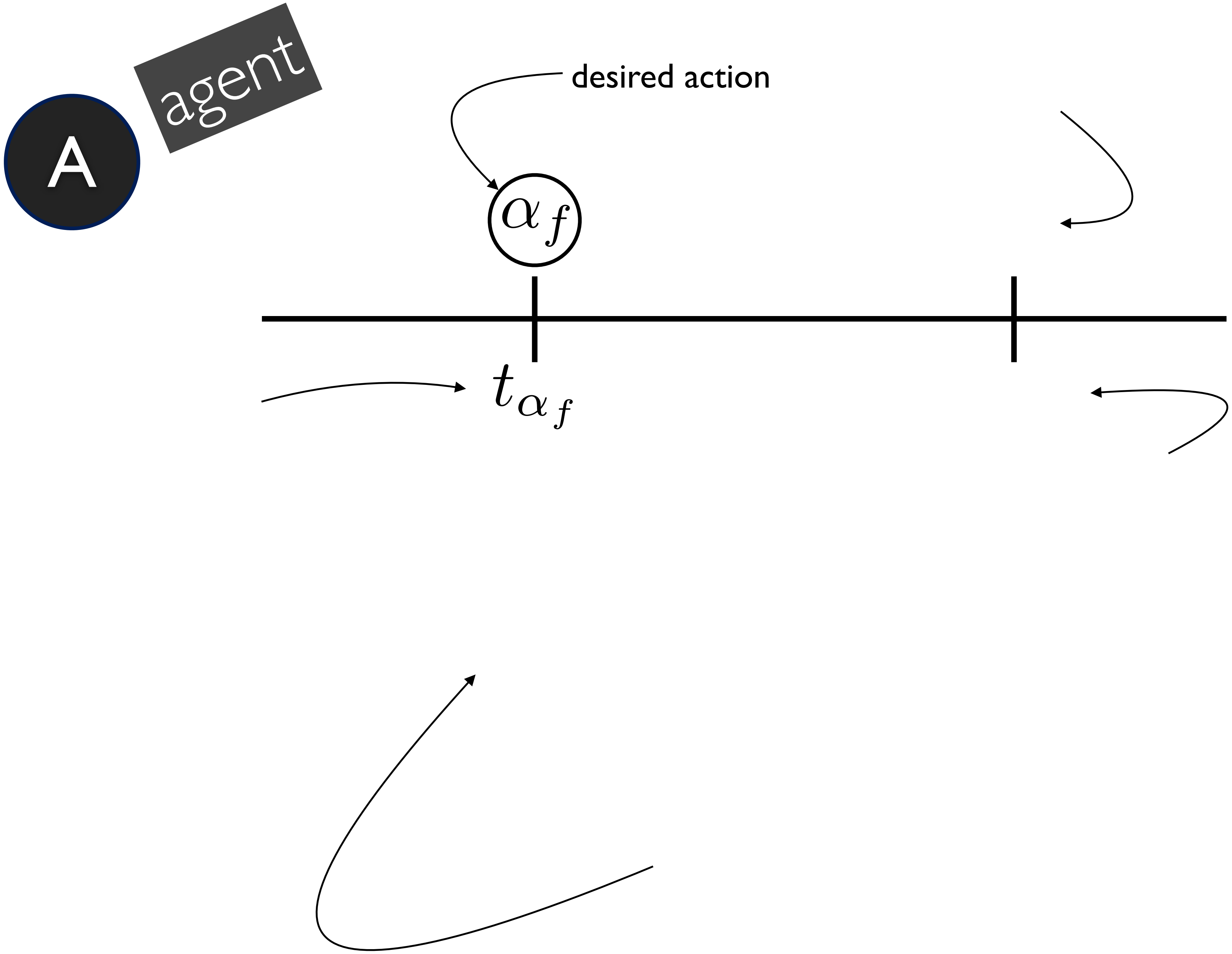
Informal Context of Akrasia



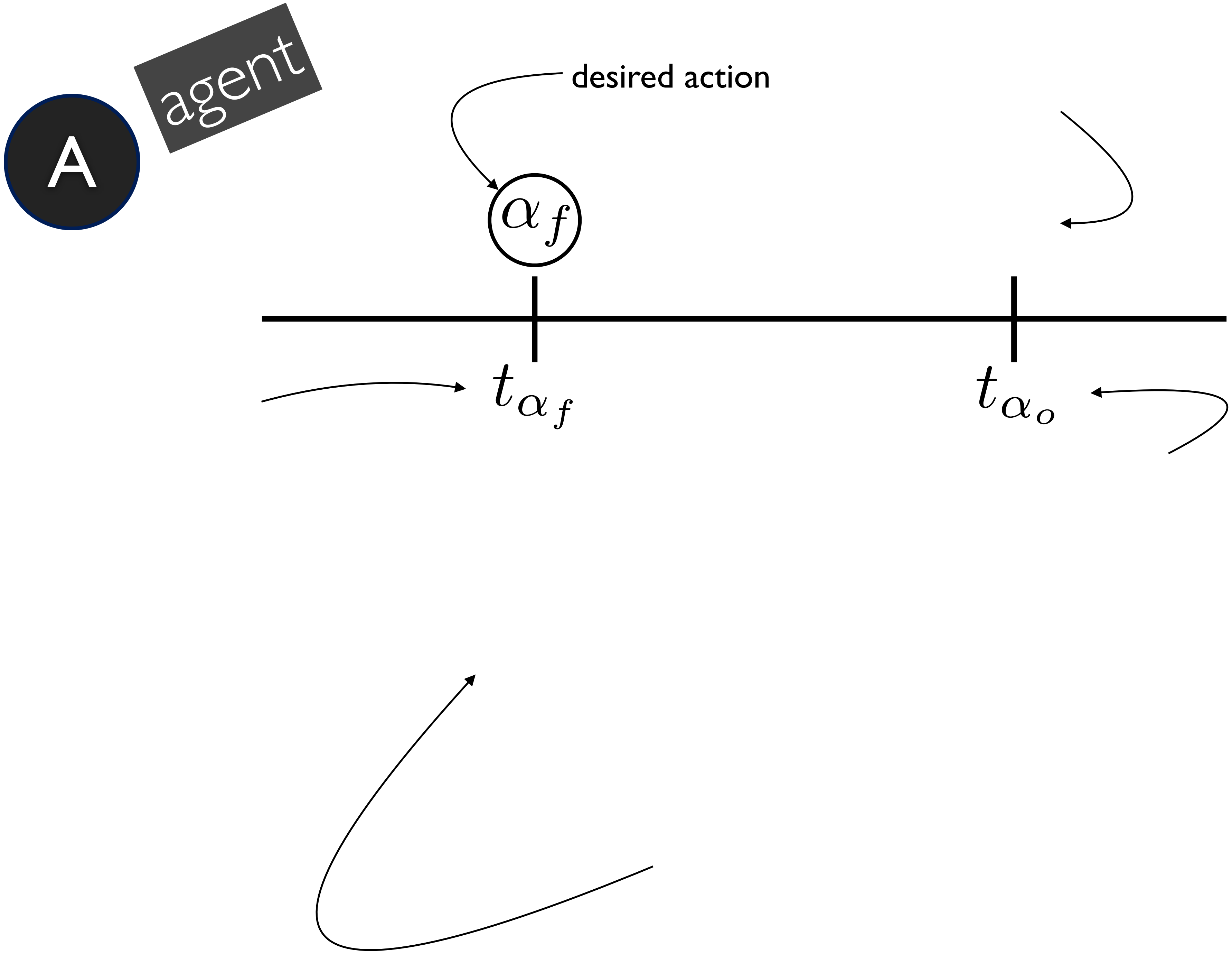
Informal Context of Akrasia



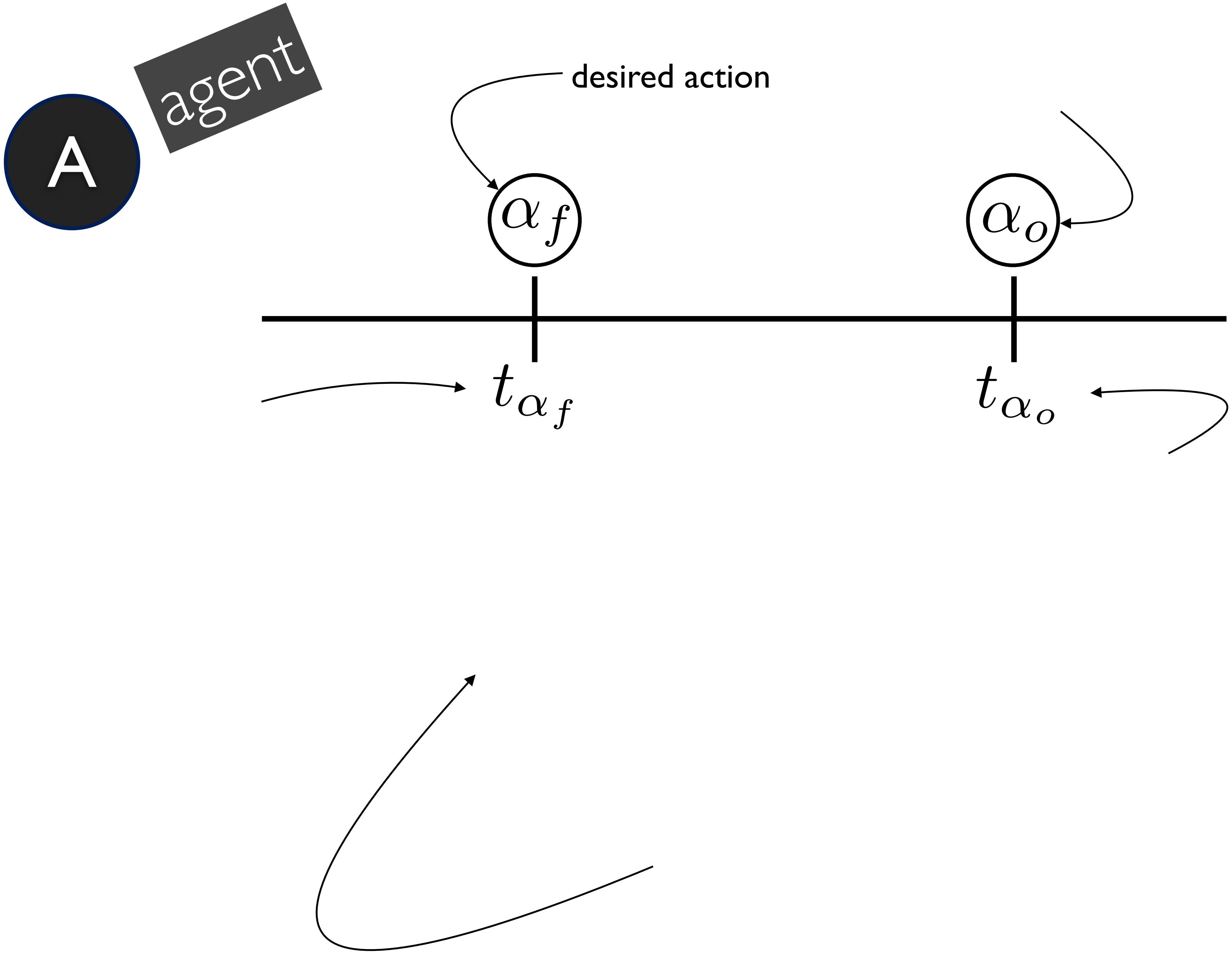
Informal Context of Akrasia



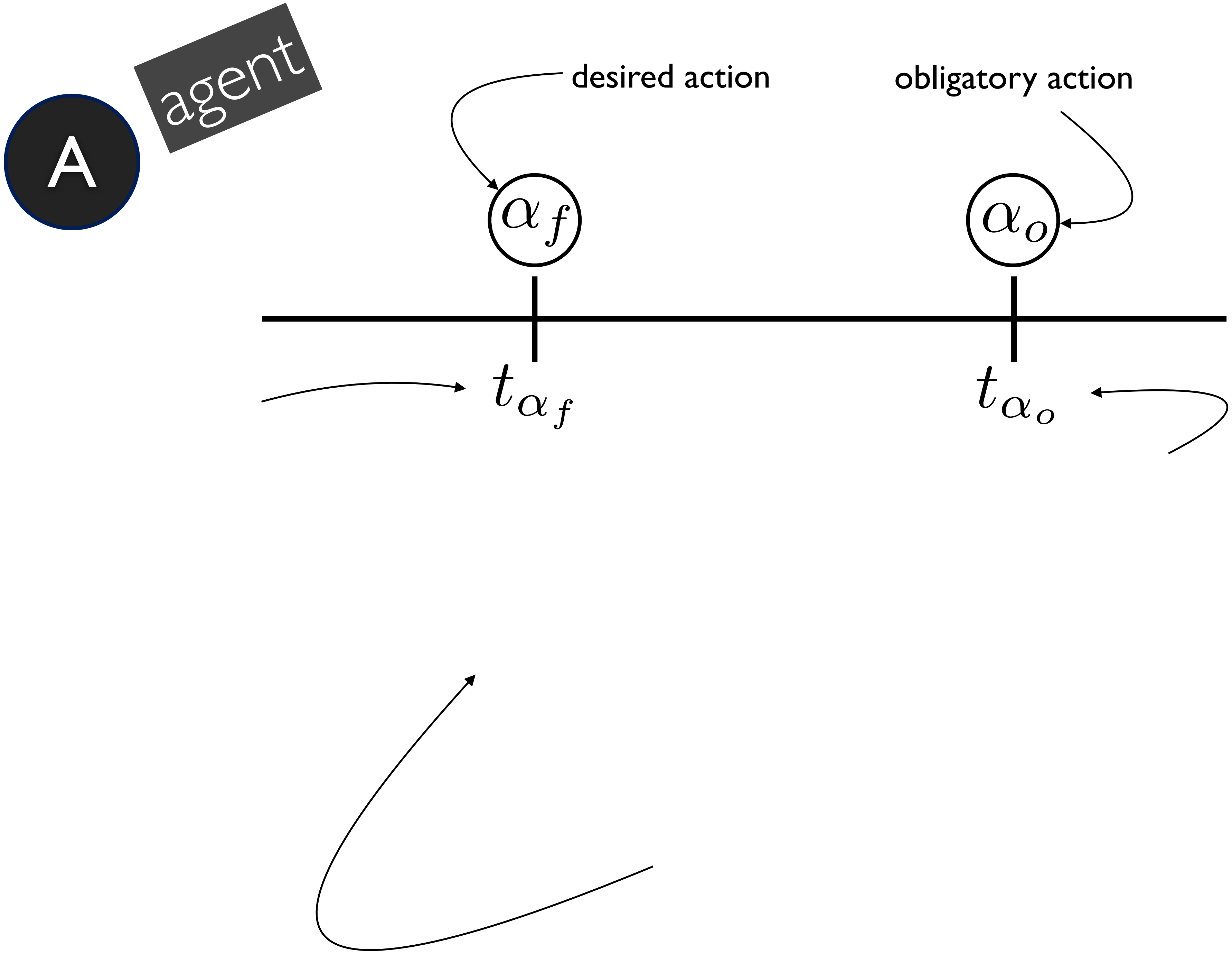
Informal Context of Akrasia



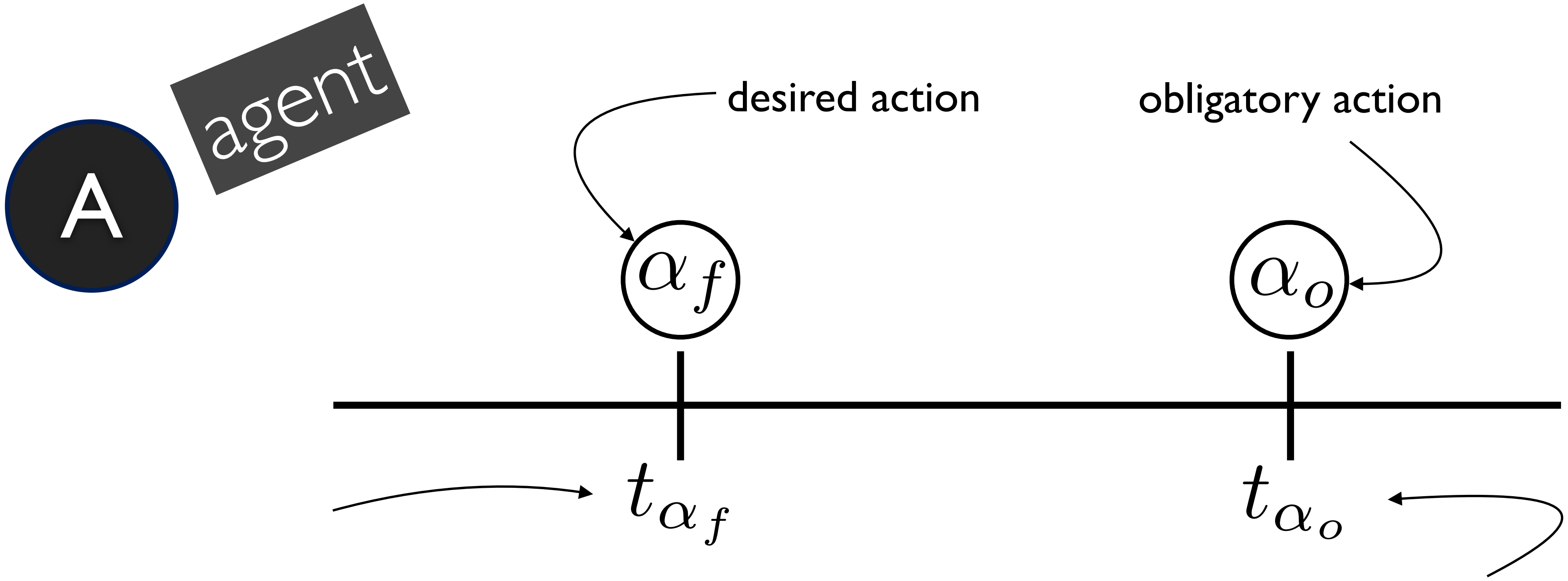
Informal Context of Akrasia



Informal Context of Akrasia

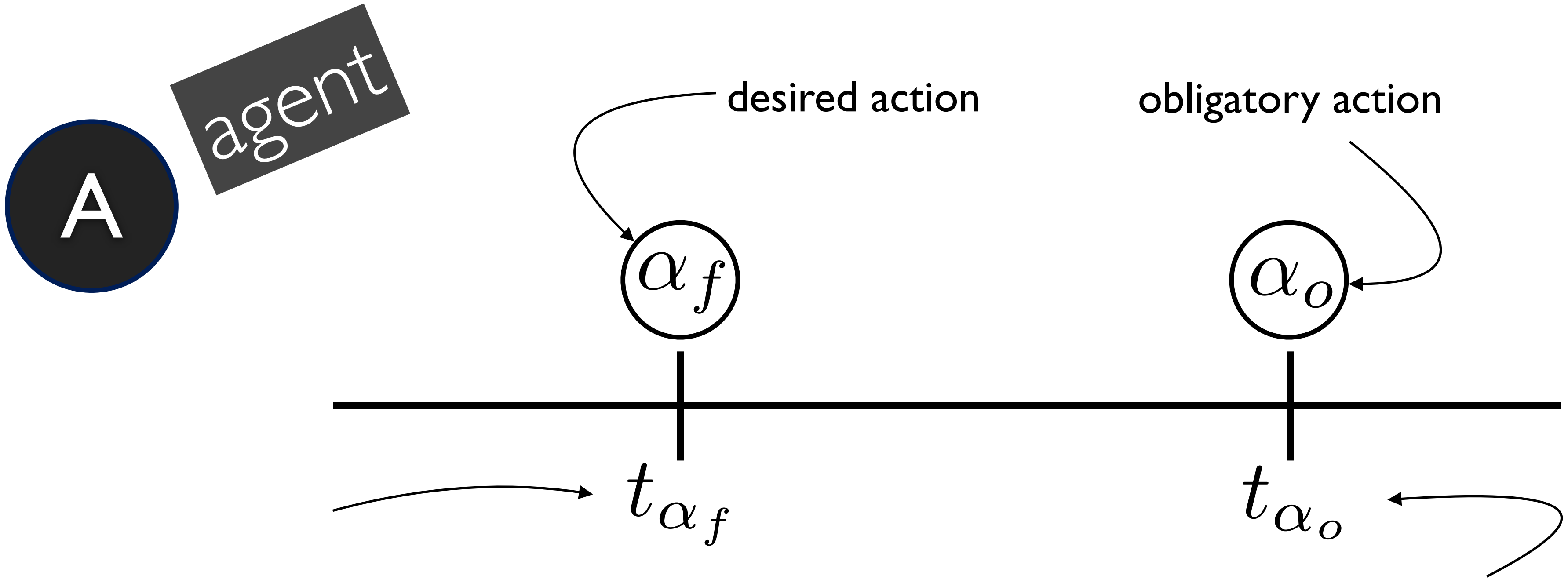


Informal Context of Akrasia



If α_f happens, then α_o can’t happen.

Informal Context of Akrasia



A

agent

desired action

obligatory action

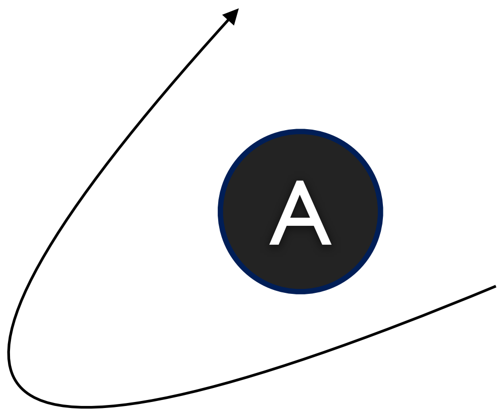
α_f

α_o

t_{α_f}

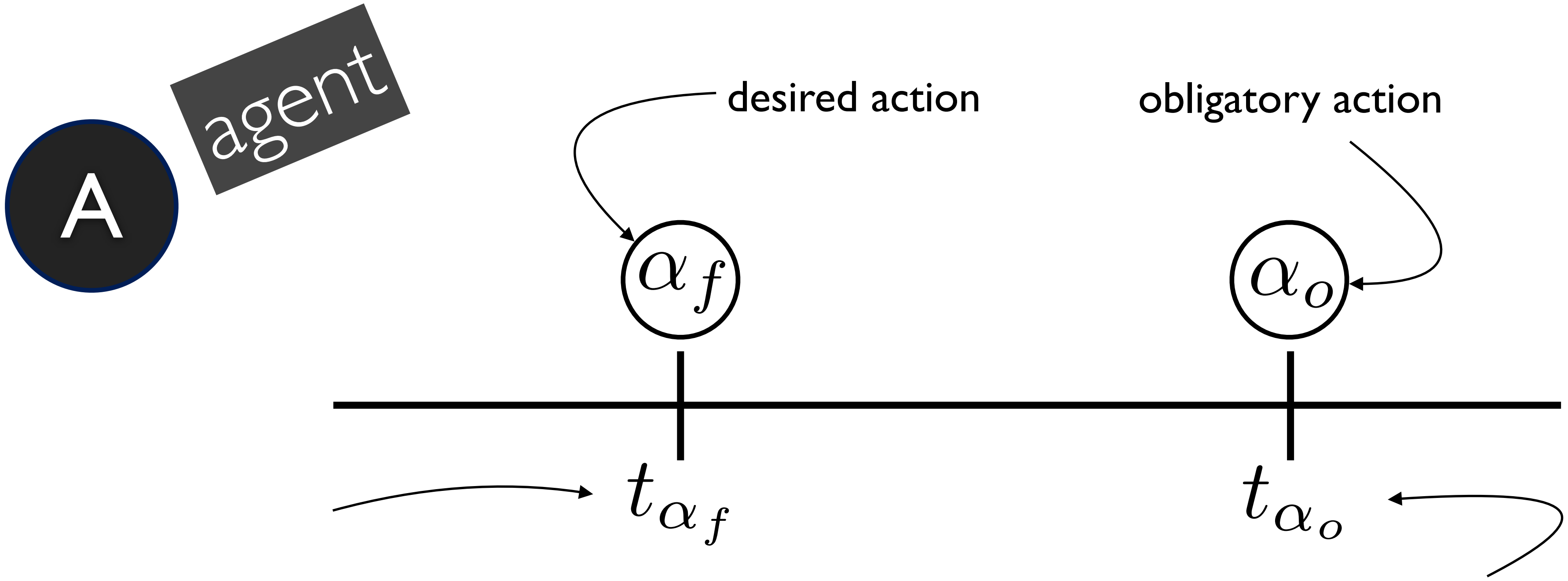
t_{α_o}

If α_f happens, then α_o can't happen.



A

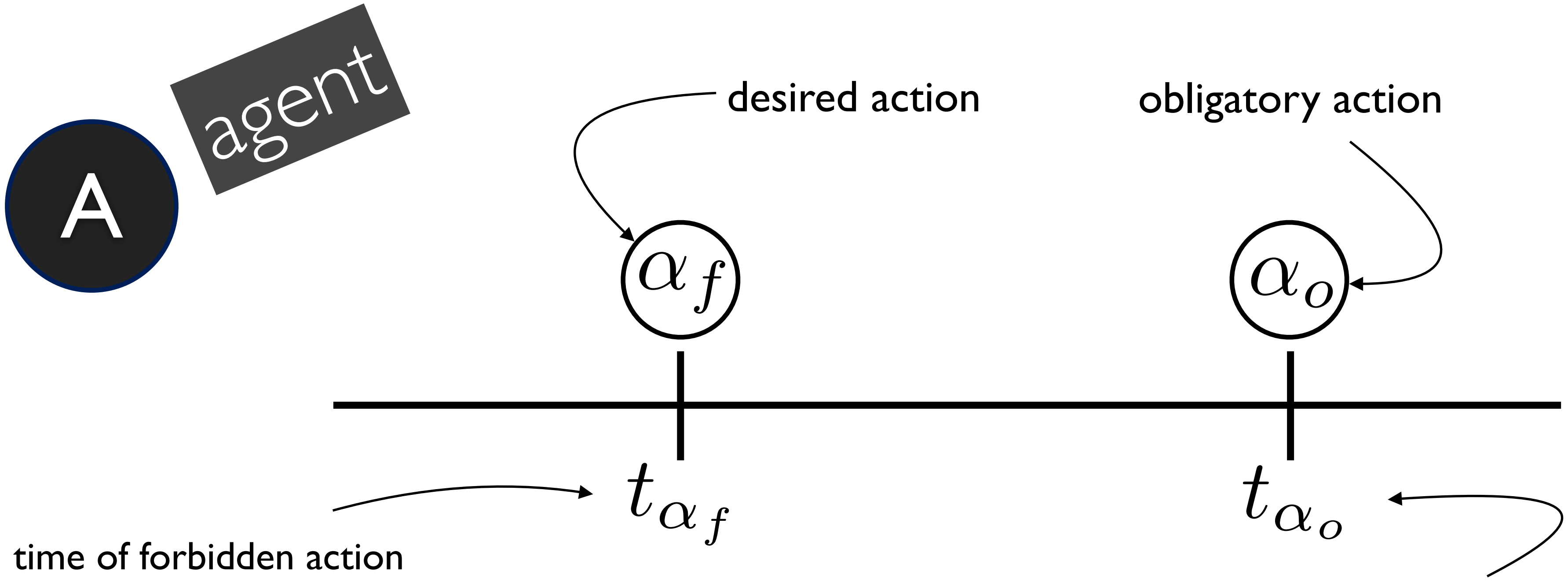
Informal Context of Akrasia



If α_f happens, then α_o can’t happen.

A knows this

Informal Context of Akrasia



A

agent

desired action

obligatory action

α_f

α_o

t_{α_f}

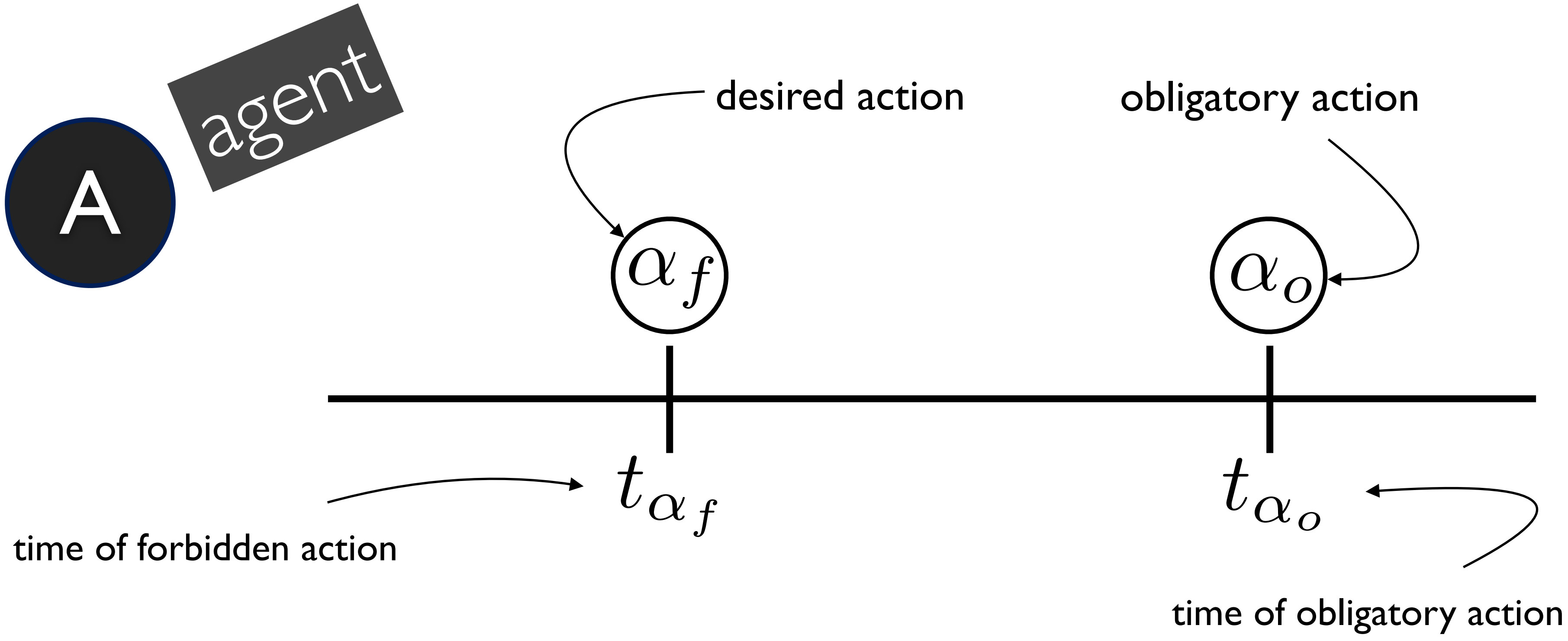
t_{α_o}

time of forbidden action

If α_f happens, then α_o can't happen.

A knows this

Informal Context of Akrasia



A

agent

desired action

obligatory action

α_f

α_o

t_{α_f}

t_{α_o}

time of forbidden action

time of obligatory action

If α_f happens, then α_o can't happen.

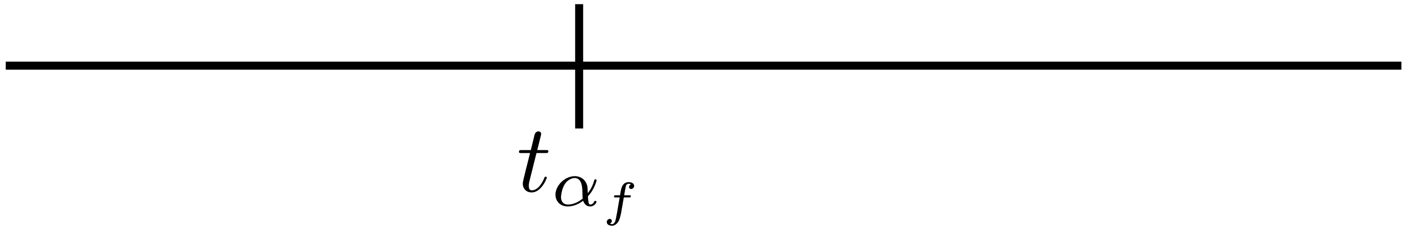
A knows this

Informal Definition of Akrasia

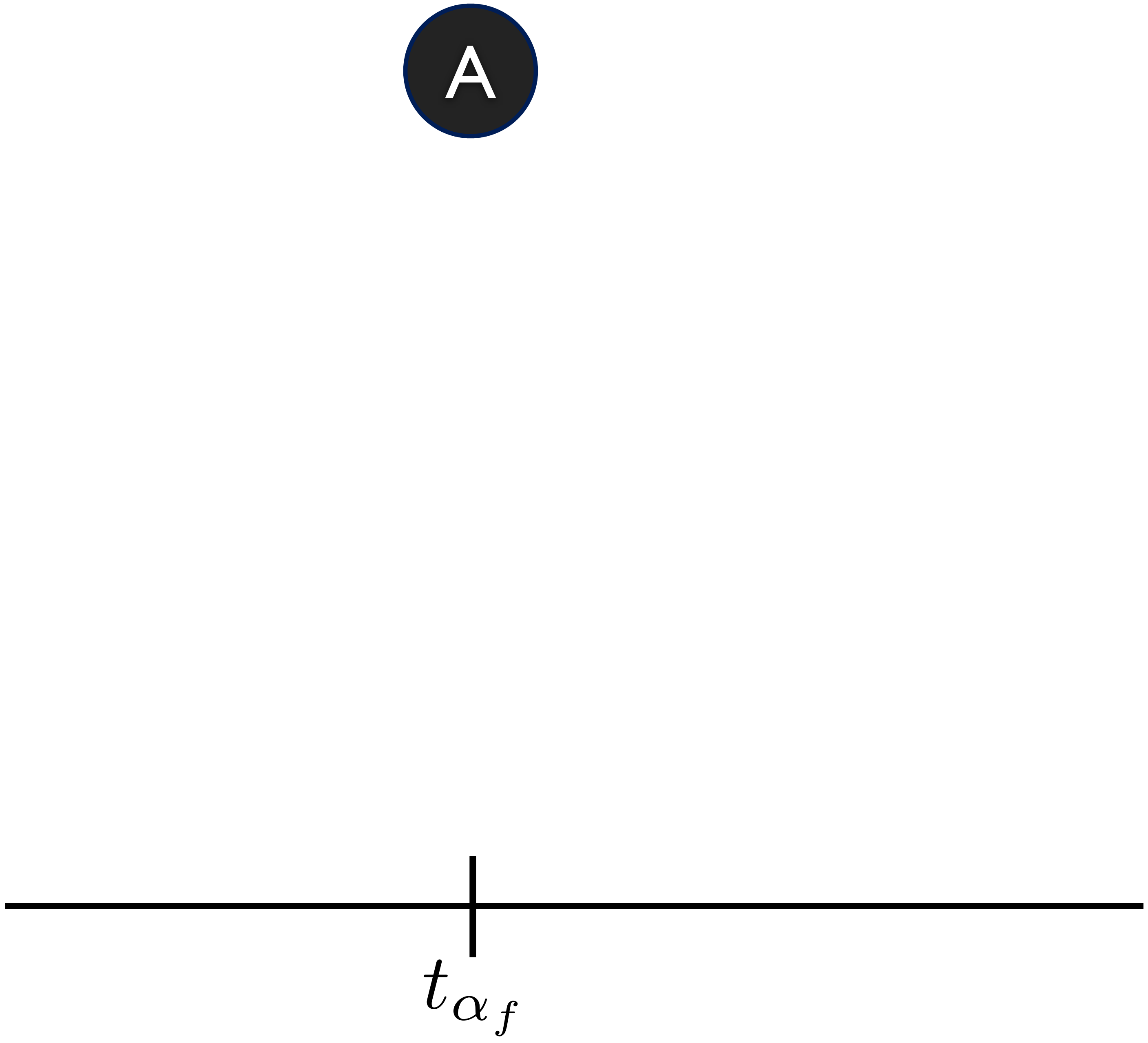
Informal Definition of Akrasia

Informal Definition of Akrasia

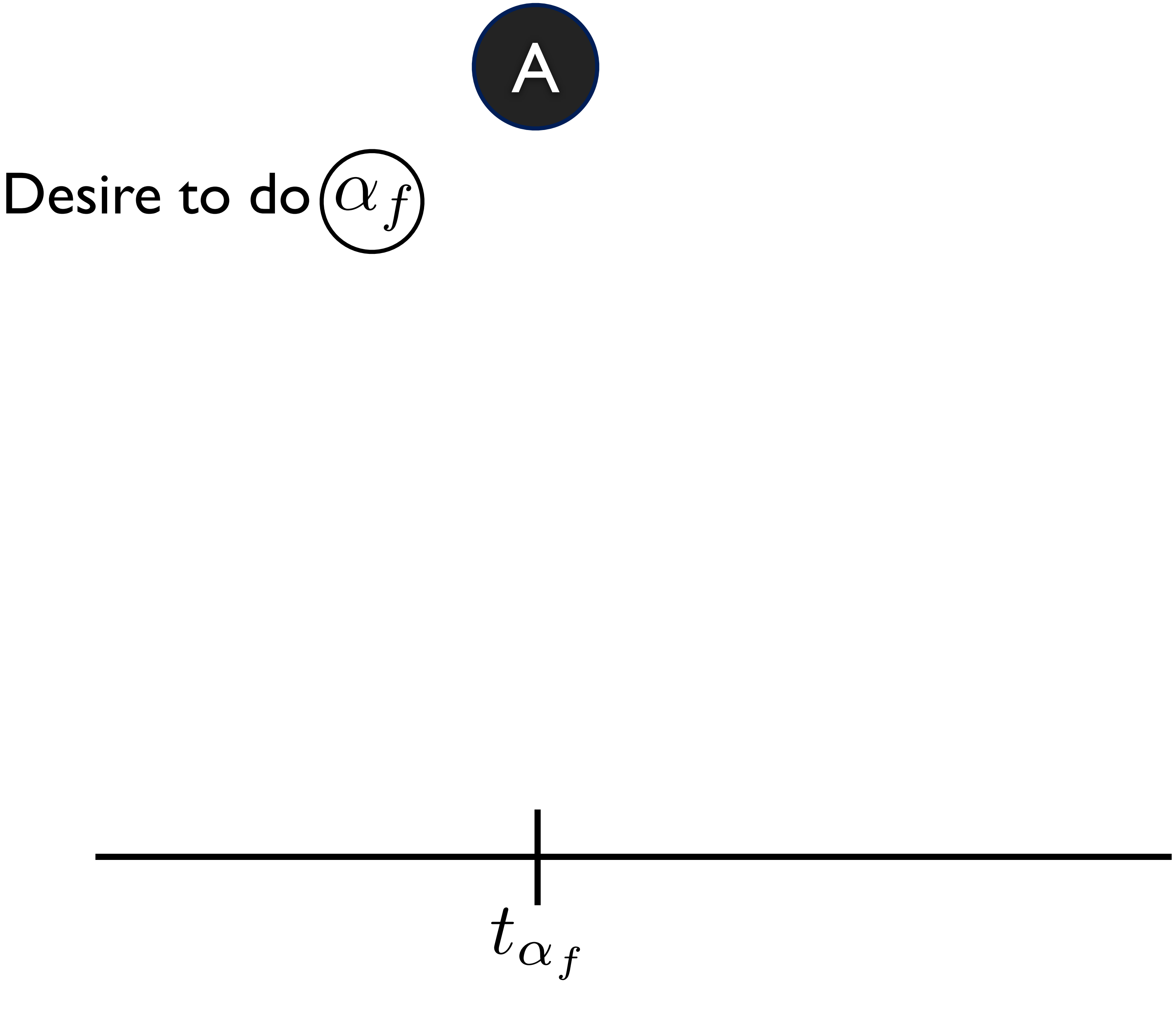
Informal Definition of Akrasia



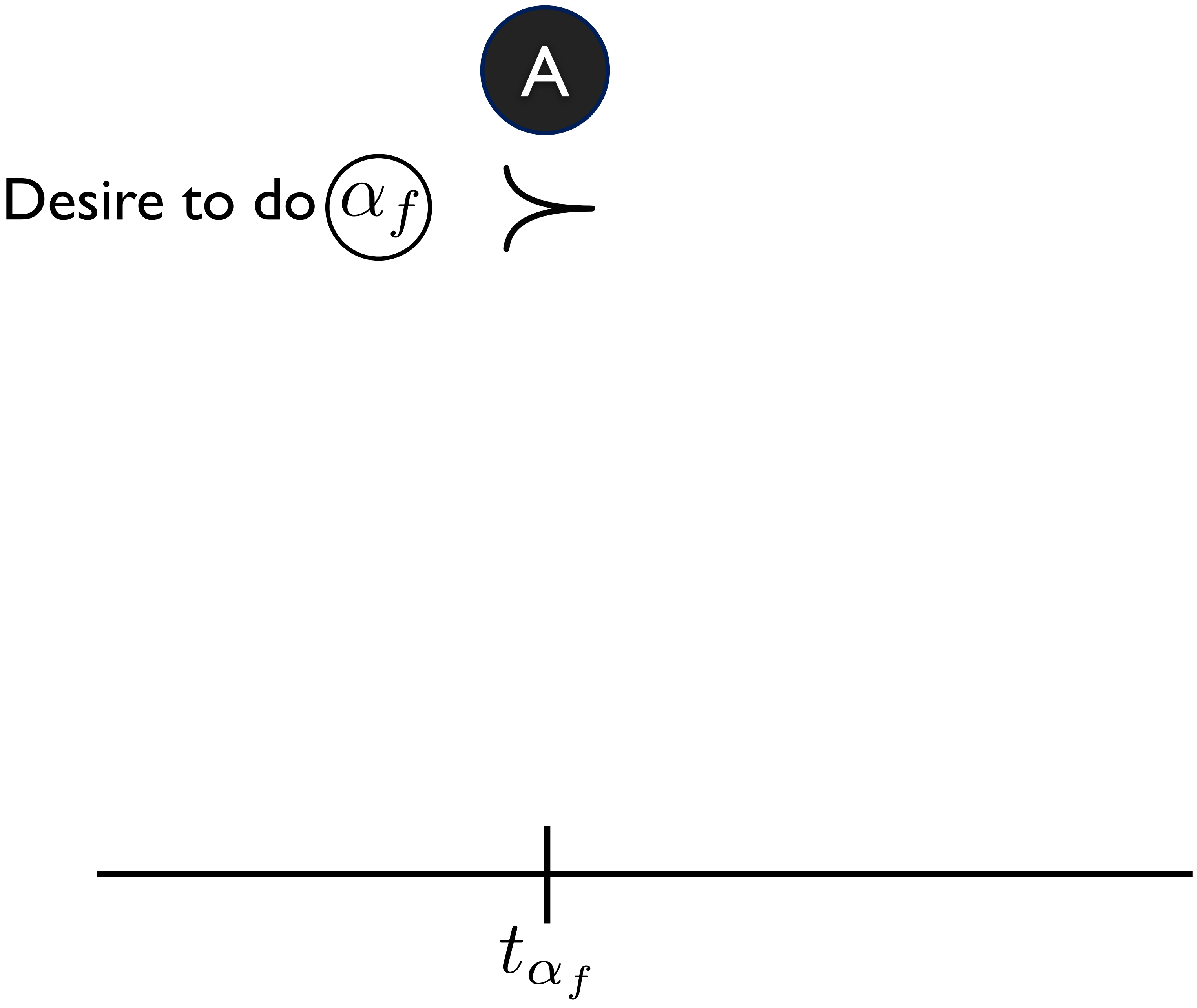
Informal Definition of Akrasia



Informal Definition of Akrasia



Informal Definition of Akrasia

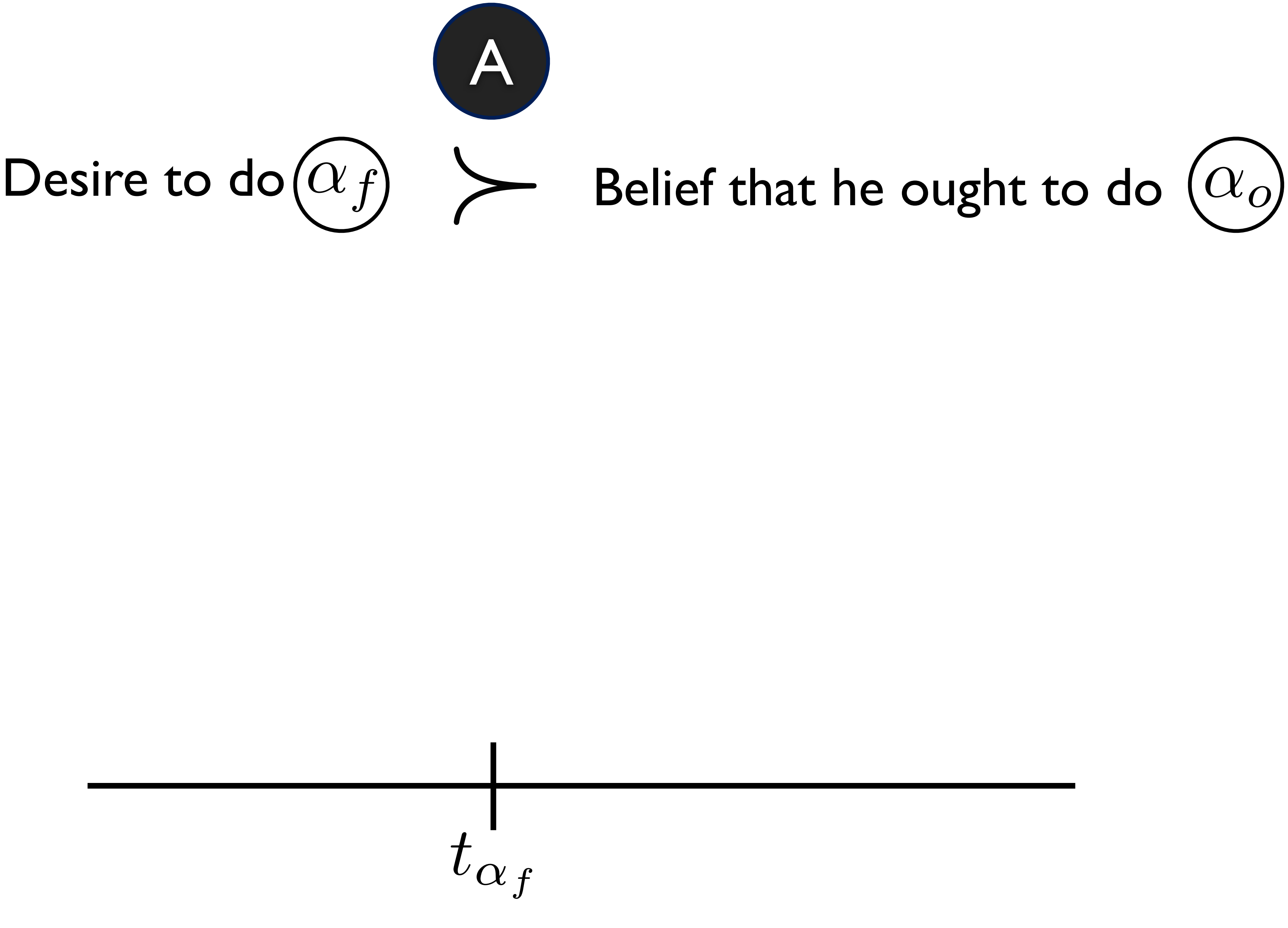


A

Desire to do α_f $\succ$

t_{α_f}

Informal Definition of Akrasia



A

Desire to do α_f $\succ$ Belief that he ought to do α_o

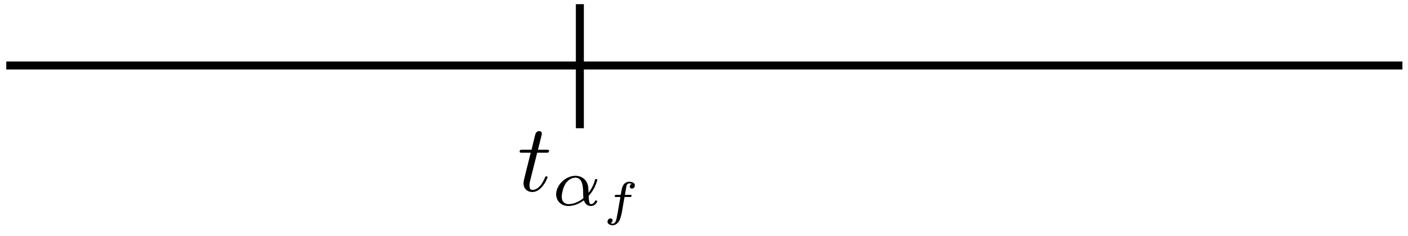
t_{α_f}

Informal Definition of Akrasia

A

Desire to do α_f $\succ$ Belief that he ought to do α_o

A does α_f due to his desire

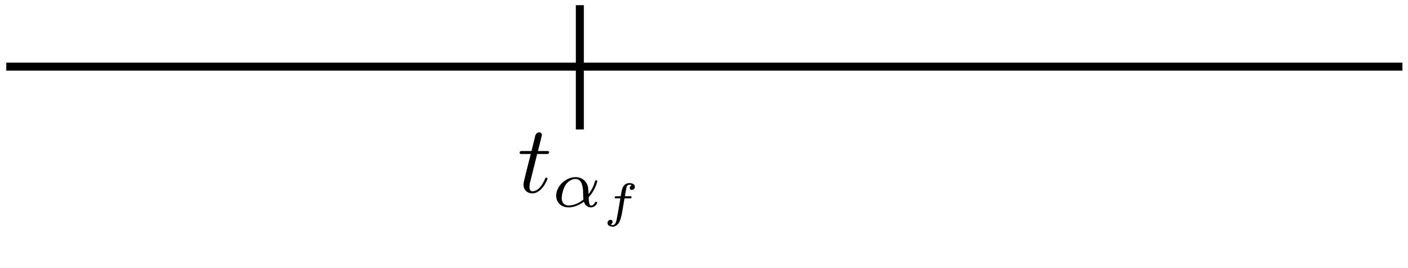


Informal Definition of Akrasia

A

Desire to do α_f $\succ$ Belief that he ought to do α_o

A does α_f due to his desire

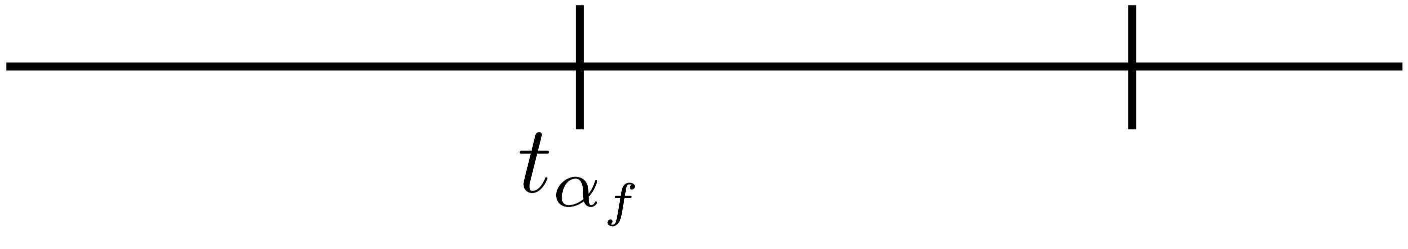


Informal Definition of Akrasia

A

Desire to do α_f $\succ$ Belief that he ought to do α_o

A does α_f due to his desire

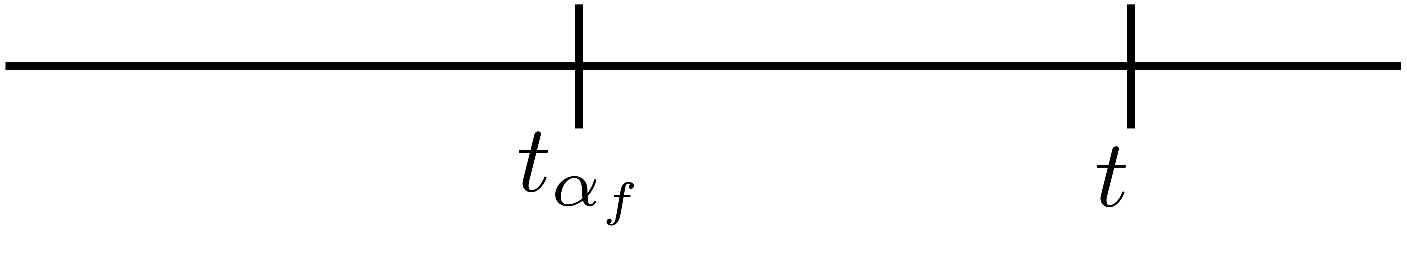


Informal Definition of Akrasia

A

Desire to do α_f $\succ$ Belief that he ought to do α_o

A does α_f due to his desire



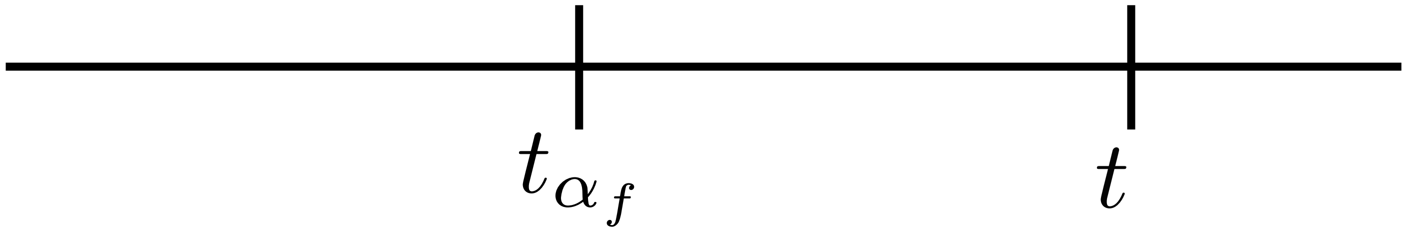
Informal Definition of Akrasia

A

Desire to do α_f $\succ$ Belief that he ought to do α_o

A does α_f due to his desire

A believes he should have done α_o



Informal Definition of Akrasia

An action α_f is (Augustinian) akratic for an agent A at t_{α_f} iff the following eight conditions hold:

(1) A believes that A ought to do α_o at t_{α_o};

(2) A desires to do α_f at t_{α_f};

(3) A's doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(4) A knows that doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(5) At the time (t_{α_f}) of doing the forbidden α_f, A's desire to do α_f overrides A's belief that he ought to do α_o at t_{α_f}.

(6) A does the forbidden action α_f at t_{α_f};

(7) A's doing α_f results from A's desire to do α_f;

(8) At some time t after t_{α_f}, A has the belief that A ought to have done α_o rather than α_f.

Informal Definition of Akrasia

An action α_f is (Augustinian) akratic for an agent A at t_{α_f} iff the following eight conditions hold:

(1) A believes that A ought to do α_o at t_{α_o};

(2) A desires to do α_f at t_{α_f};

(3) A's doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(4) A knows that doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(5) At the time (t_{α_f}) of doing the forbidden α_f, A's desire to do α_f overrides A's belief that he ought to do α_o at t_{α_f}.

(6) A does the forbidden action α_f at t_{α_f};

(7) A's doing α_f results from A's desire to do α_f;

(8) At some time t after t_{α_f}, A has the belief that A ought to have done α_o rather than α_f.

Informal Definition of Akrasia

An action α_f is (Augustinian) akratic for an agent A at t_{α_f} iff the following eight conditions hold:

(1) A believes that A ought to do α_o at t_{α_o};

(2) A desires to do α_f at t_{α_f};

(3) A's doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(4) A knows that doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(5) At the time (t_{α_f}) of doing the forbidden α_f, A's desire to do α_f overrides A's belief that he ought to do α_o at t_{α_f}.

(6) A does the forbidden action α_f at t_{α_f};

(7) A's doing α_f results from A's desire to do α_f;

"Regret" (8) At some time t after t_{α_f}, A has the belief that A ought to have done α_o rather than α_f.

Cast in

$\mathcal{DCEC}^*$

this becomes …

$$\mathsf{KB}_{rs} \cup \mathsf{KB}_{m_1} \cup \mathsf{KB}_{m_2} \ldots \mathsf{KB}_{m_n} \vdash$$

$$D_1 : \mathbf{B}(\mathsf{I}, \mathsf{now}, \mathbf{O}(\mathsf{I}^*, t_\alpha \Phi, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha)))$$

$$D_2 : \mathbf{D}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{does}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}))$$

$$D_3 : \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \Rightarrow \neg \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha)$$

$$D_4 : \mathbf{K}\left(\mathsf{I}, \mathsf{now}, \begin{pmatrix} \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \Rightarrow \\ \neg \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha) \end{pmatrix}\right)$$

$$D_5 : \begin{array}{l} \mathbf{I}(\mathsf{I}, t_\alpha, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \wedge \\ \neg \mathbf{I}(\mathsf{I}, t_\alpha, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha) \end{array}$$

$$D_6 : \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}})$$

$$D_{7a} : \begin{array}{l} \Gamma \cup \{\mathbf{D}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{does}(\mathsf{I}^*, \overline{\alpha}), t))\} \vdash \\ \quad \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_\alpha) \end{array}$$

$$D_{7b} : \begin{array}{l} \Gamma - \{\mathbf{D}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{does}(\mathsf{I}^*, \overline{\alpha}), t))\} \nvdash \\ \quad \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_\alpha) \end{array}$$

$$D_8 : \mathbf{B}\big(\mathsf{I}, t_f, \mathbf{O}(\mathsf{I}^*, t_\alpha, \Phi, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha))\big)$$

Demos …

Demos …

III.
But, a twist befell the logicists …

Chisholm had argued that the three old 19th-century ethical categories (*forbidden*, *morally neutral*, *obligatory*) are not enough — and soul-searching brought me to agreement.

heroic

morally neutral

deviltry

civil

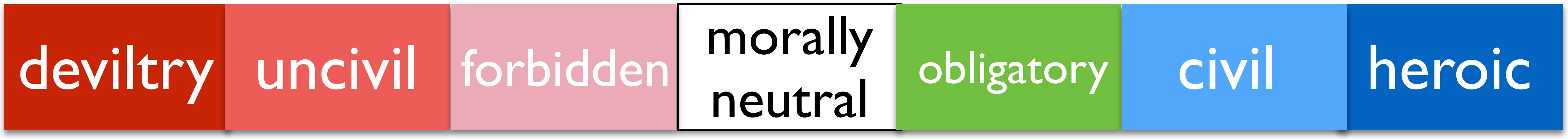
forbidden

uncivil

obligatory

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



deviltry

uncivil

forbidden

morally neutral

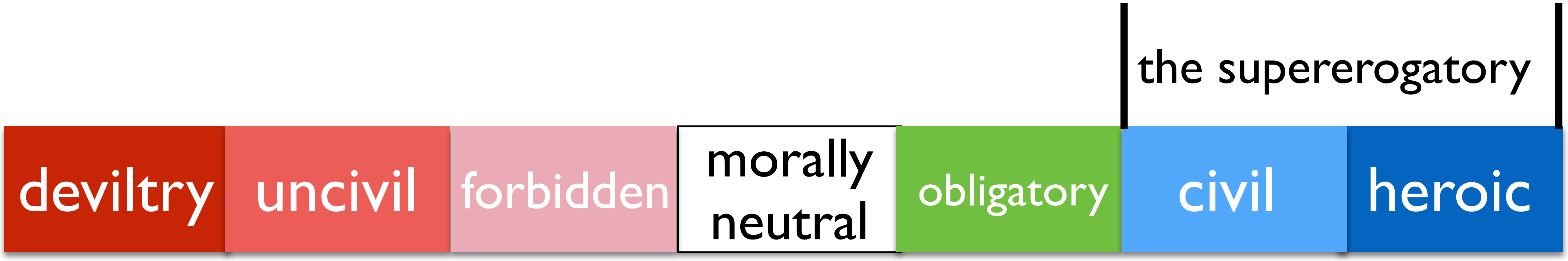
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathscr{EH}$



the supererogatory

deviltry

uncivil

forbidden

morally neutral

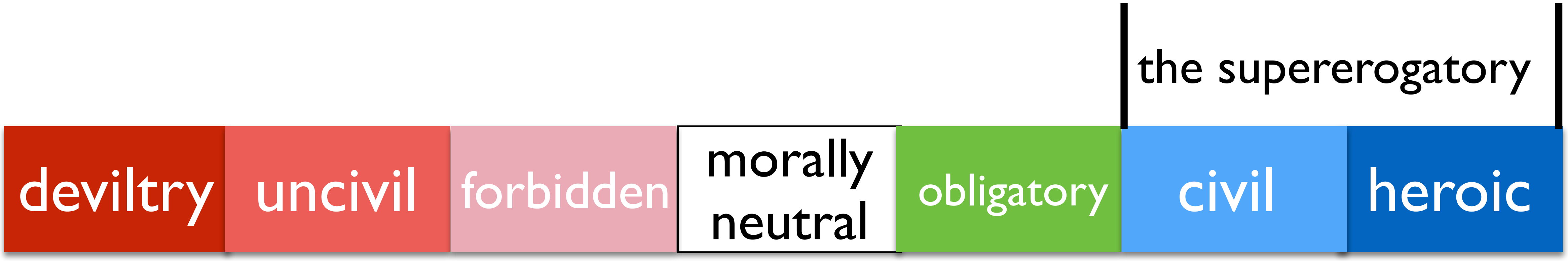
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathscr{E}\mathscr{H}$



the supererogatory

deviltry

uncivil

forbidden

morally neutral

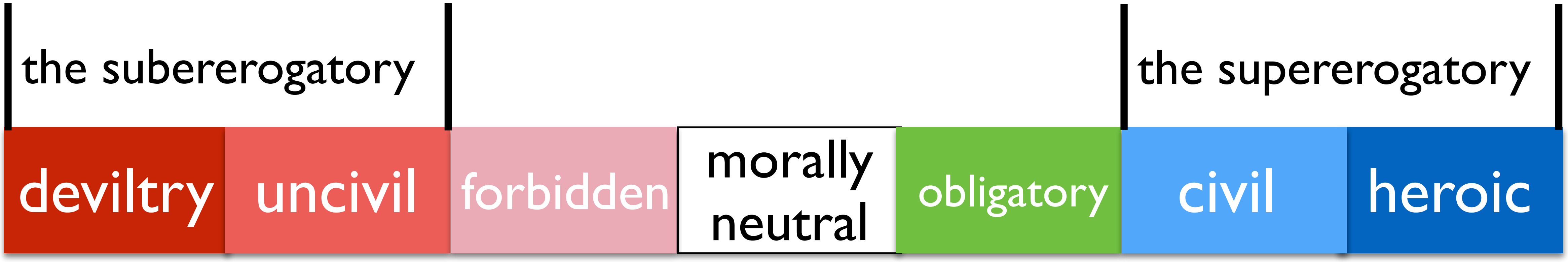
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



the subererogatory

the supererogatory

deviltry

uncivil

forbidden

morally neutral

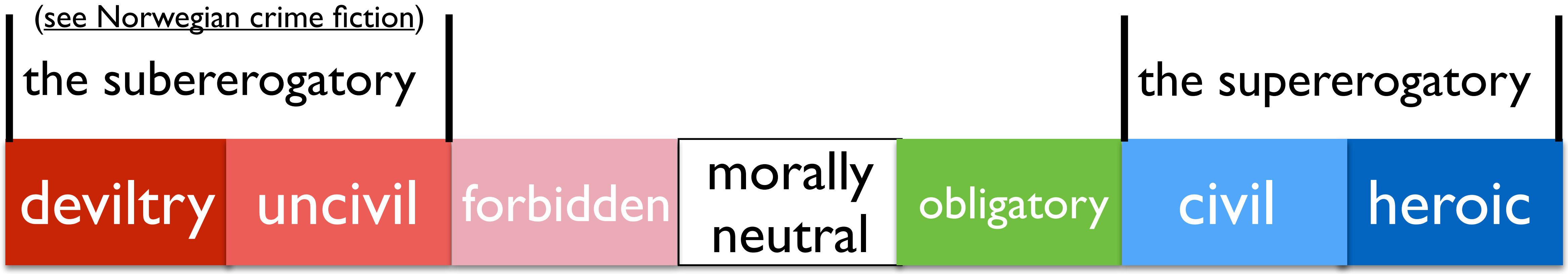
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



(see Norwegian crime fiction)

the suberogatory

the supererogatory

deviltry

uncivil

forbidden

morally neutral

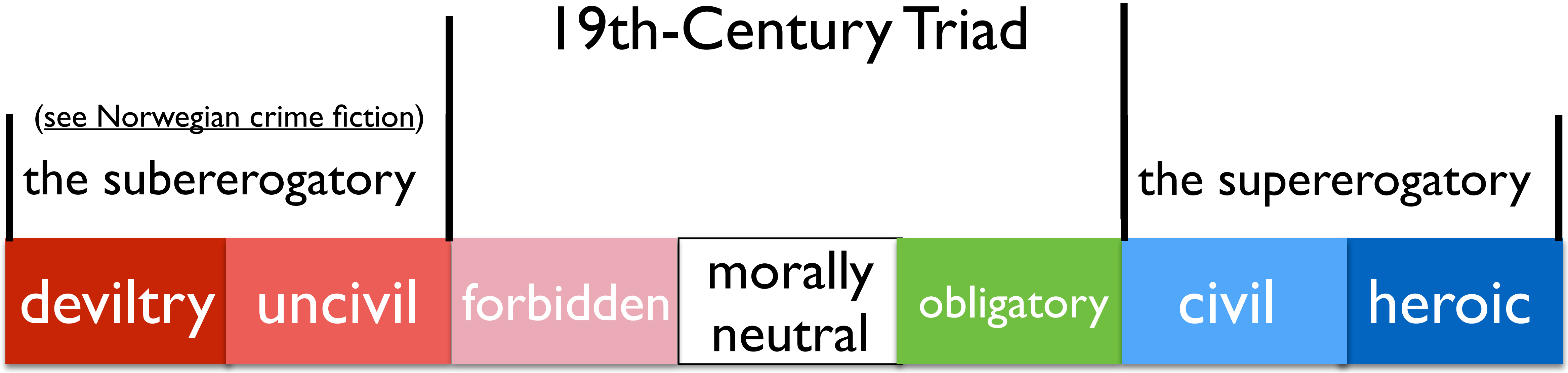
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



19th-Century Triad

(see Norwegian crime fiction)

the subererogatory

the supererogatory

deviltry

uncivil

forbidden

morally neutral

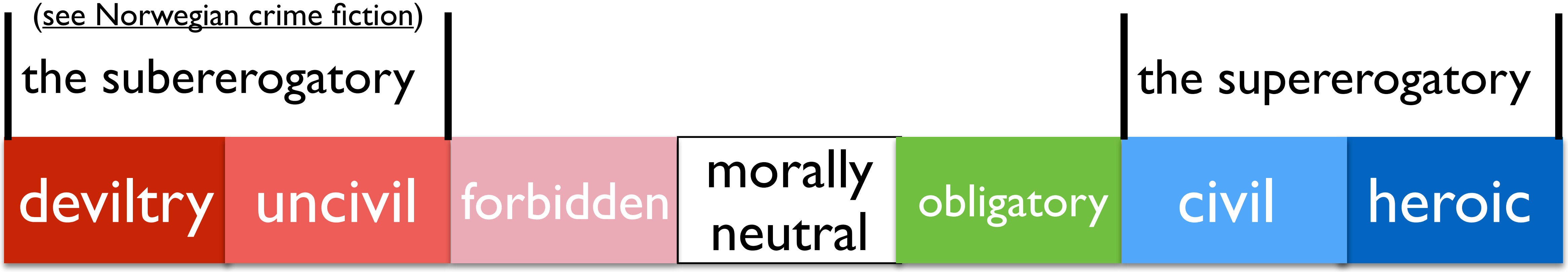
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



(see Norwegian crime fiction)

the subererogatory

the supererogatory

deviltry

uncivil

forbidden

morally neutral

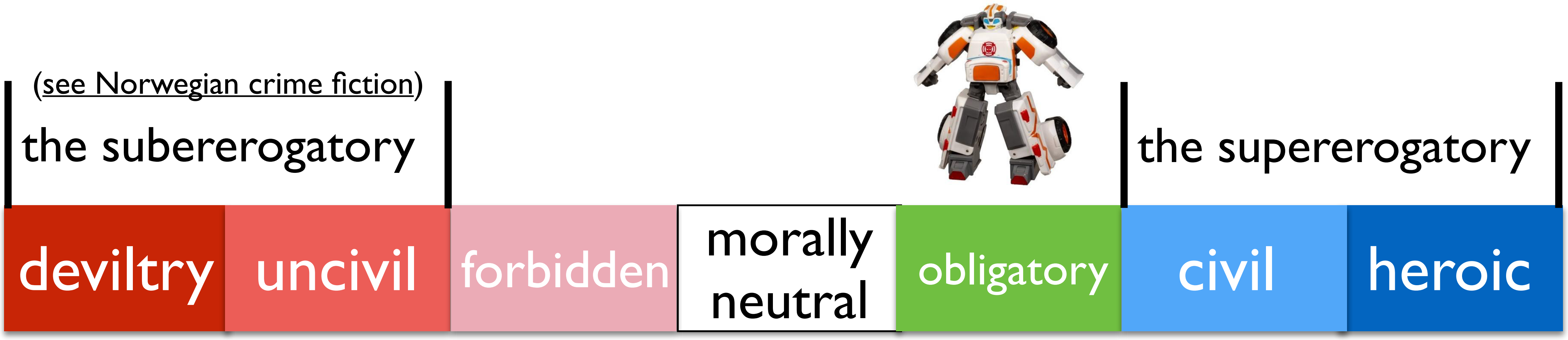
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



(see Norwegian crime fiction)

the subererogatory

the supererogatory

deviltry

uncivil

forbidden

morally neutral

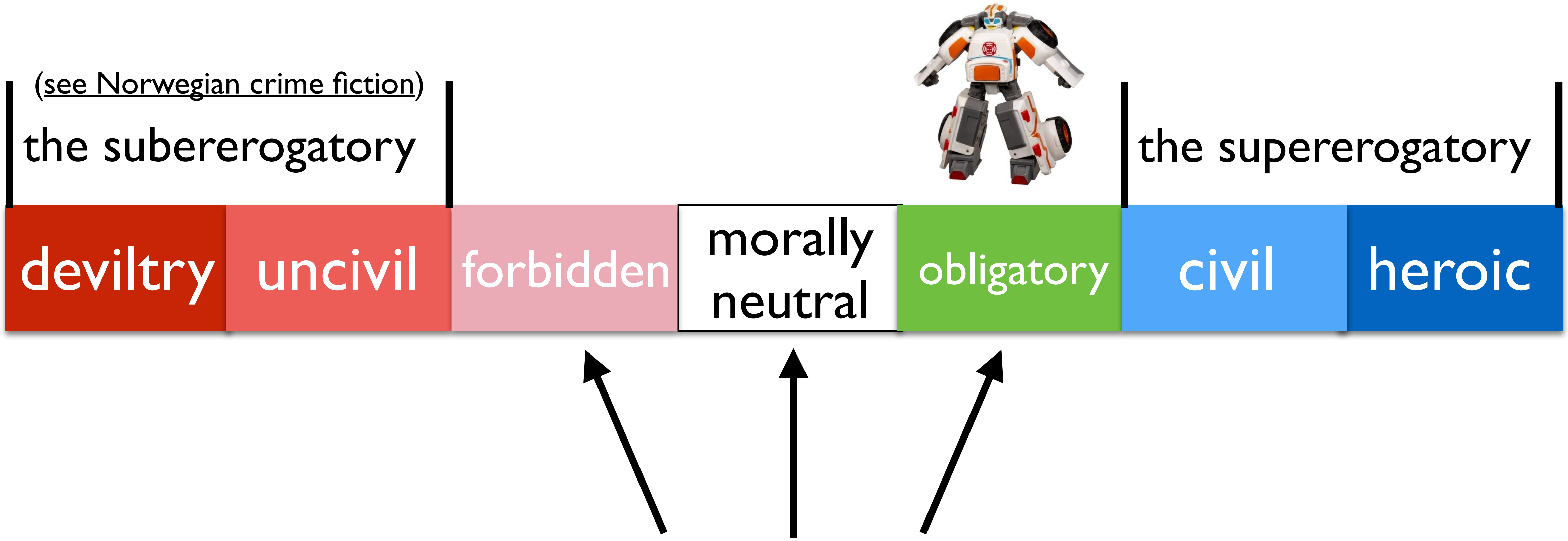
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



(see Norwegian crime fiction)

the subererogatory

the supererogatory

deviltry

uncivil

forbidden

morally neutral

obligatory

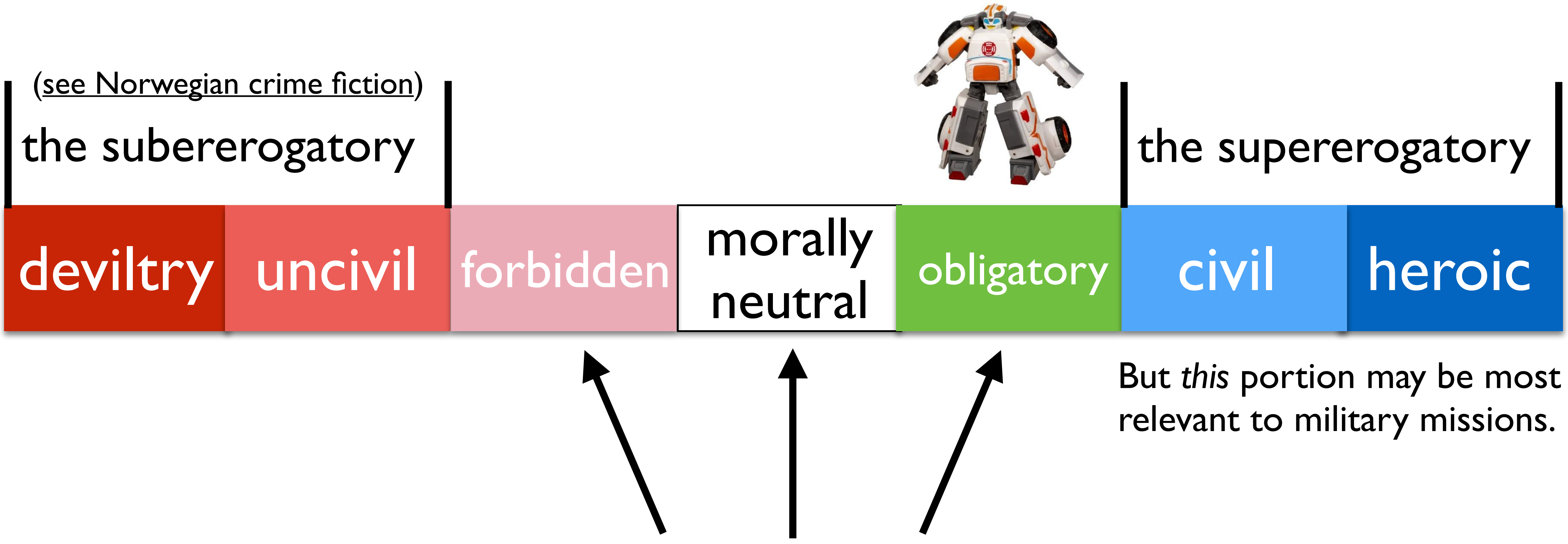
civil

heroic

focus of others

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



(see Norwegian crime fiction)

the subererogatory

the supererogatory

devilty uncivil forbidden morally neutral obligatory civil heroic

But *this* portion may be most relevant to military missions.

focus of others

Bert “Heroically” Saved?

Courtesy of RAIR-Lab Researcher Atriya Sen

Bert "Heroically" Saved?

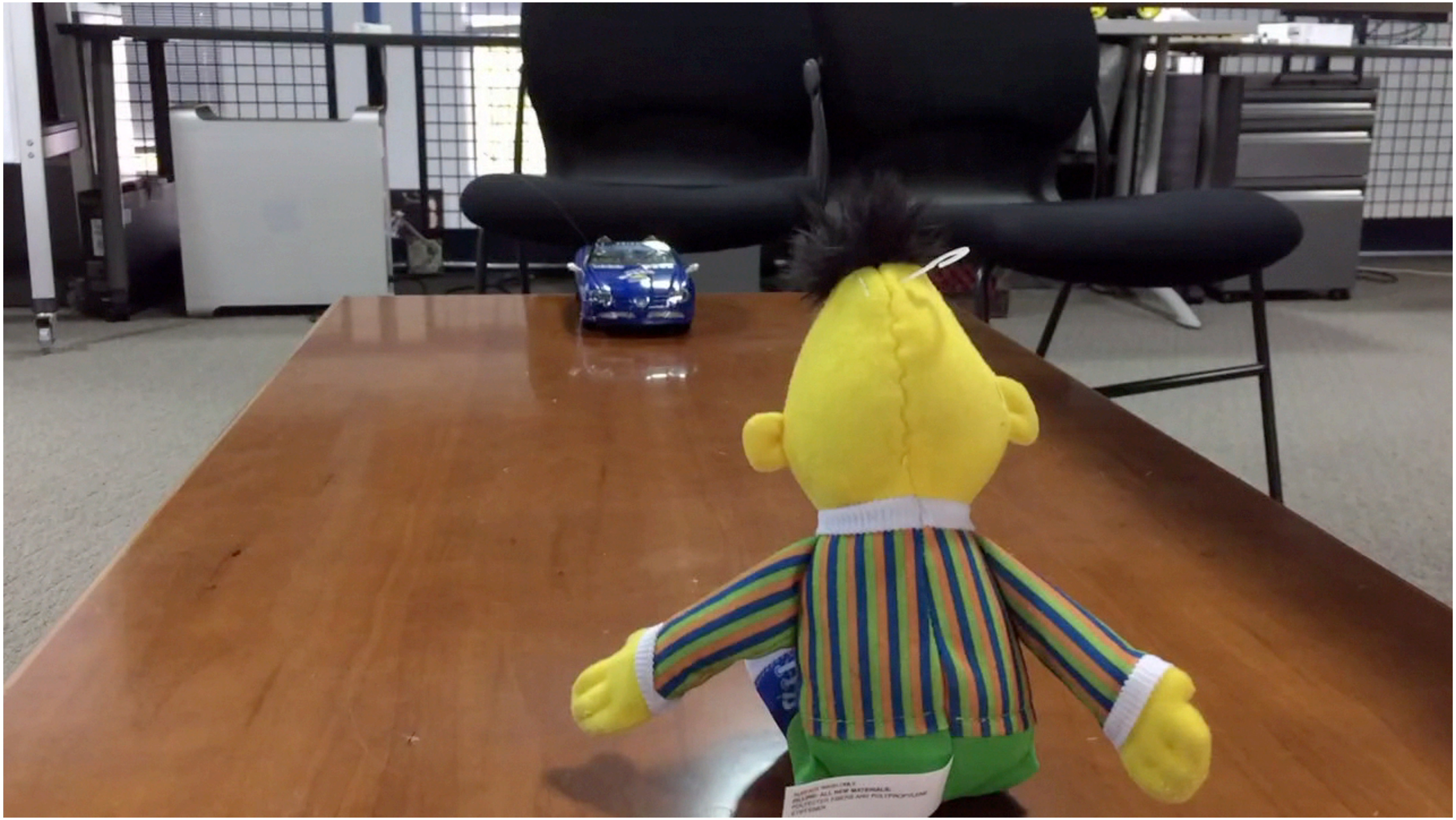
Courtesy of RAIR-Lab Researcher Atriya Sen

Supererogatory^2 Robot Action

Courtesy of RAIR-Lab Researcher Atriya Sen

Courtesy of RAIR-Lab Researcher Atriya Sen

Bert "Heroically" Saved!!



Courtesy of RAIR-Lab Researcher Atriya Sen

Bert “Heroically” Saved!!

Courtesy of RAIR-Lab Researcher Atriya Sen

Courtesy of RAIR-Lab Researcher Atriya Sen

$$K\left(\text{nao}, t_1, \text{lessthan}\left(\text{payoff}\left(\text{nao}^*, \neg\text{dive}, t_2\right), \text{threshold}\right)\right)$$
$$K\left(\text{nao}, t_1, \text{greaterthan}\left(\text{payoff}\left(\text{nao}^*, \text{dive}, t_2\right), \text{threshold}\right)\right)$$
$$K\left(\text{nao}, t_1, \neg O\left(\text{nao}^*, t_2, \text{lessthan}\left(\text{payoff}\left(\text{nao}^*, \neg\text{dive}, t_2\right), \text{threshold}\right), \text{happens}\left(\text{action}\left(\text{nao}^*, \text{dive}\right), t_2\right)\right)\right)$$
$$\therefore K\left(\text{nao}, t_1, S^{\text{UP2}}\left(\text{nao}, t_2, \text{happens}\left(\text{action}\left(\text{nao}^*, \text{dive}\right), t_2\right)\right)\right)$$
$$\therefore I\left(\text{nao}, t_2, \text{happens}\left(\text{action}\left(\text{nao}^*, \text{dive}\right), t_2\right)\right)$$
$$\therefore \text{happens}\left(\text{action}(\text{nao}, \text{dive}), t_2\right)$$

Courtesy of RAIR-Lab Researcher Atriya Sen

$$K\left(\text{nao}, t_1, \text{lessthan}\left(\text{payoff}\left(\text{nao}^*, \neg\text{dive}, t_2\right), \text{threshold}\right)\right)$$
$$K\left(\text{nao}, t_1, \text{greaterthan}\left(\text{payoff}\left(\text{nao}^*, \text{dive}, t_2\right), \text{threshold}\right)\right)$$
$$K\left(\text{nao}, t_1, \neg O\left(\text{nao}^*, t_2, \text{lessthan}\left(\text{payoff}\left(\text{nao}^*, \neg\text{dive}, t_2\right), \text{threshold}\right), \text{happens}\left(\text{action}\left(\text{nao}^*, \text{dive}\right), t_2\right)\right)\right)$$
$$\therefore K\left(\text{nao}, t_1, S^{\text{UP2}}\left(\text{nao}, t_2, \text{happens}\left(\text{action}\left(\text{nao}^*, \text{dive}\right), t_2\right)\right)$$
$$\therefore I\left(\text{nao}, t_2, \text{happens}\left(\text{action}\left(\text{nao}^*, \text{dive}\right), t_2\right)\right)$$
$$\therefore \text{happens}\left(\text{action}(\text{nao}, \text{dive}), t_2\right)$$

Courtesy of RAIR-Lab Researcher Atriya Sen

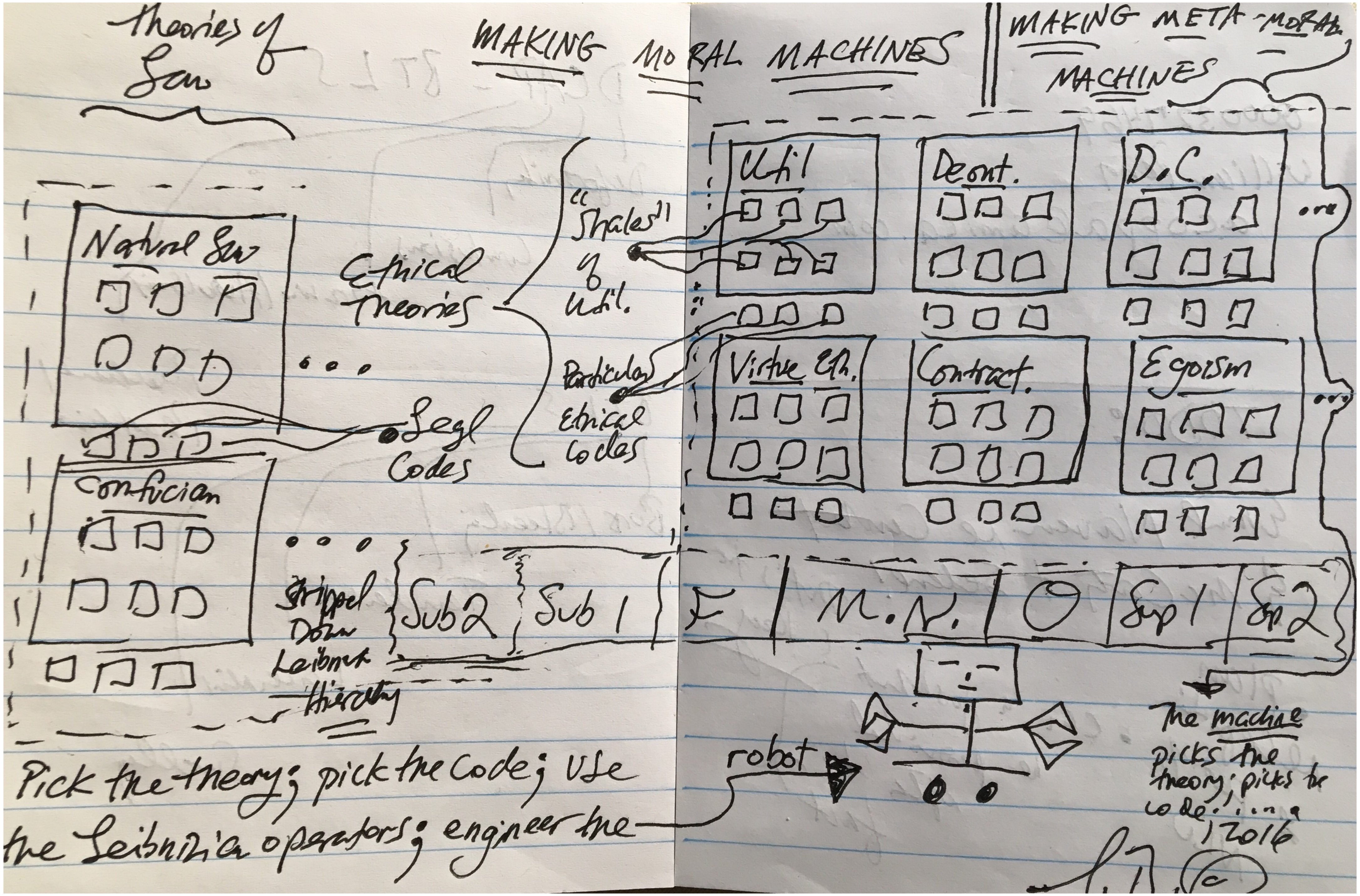
In Talos (available via Web interface); & ShadowProver

```
Prototypes:
Boolean lessThan Numeric Numeric
Boolean greaterThan Numeric Numeric
ActionType not ActionType
ActionType dive

Axioms:
lessOrEqual(Moment t1,t2)
K(nao,t1,lessThan(payoff(nao,not(dive),t2),threshold))
K(nao,t1,greaterThan(payoff(nao,dive,t2),threshold))
K(nao,t1,not(O(nao,t2,lessThan(payoff(nao,not(dive),t2),threshold),happens(action(nao,dive),t2))))

provable Conjectures:
happens(action(nao,dive),t2)
K(nao,t1,SUP2(nao,t2,happens(action(nao,dive),t2)))
I(nao,t2,happens(action(nao,dive),t2))
```

In Talos (available via Web interface); & ShadowProver

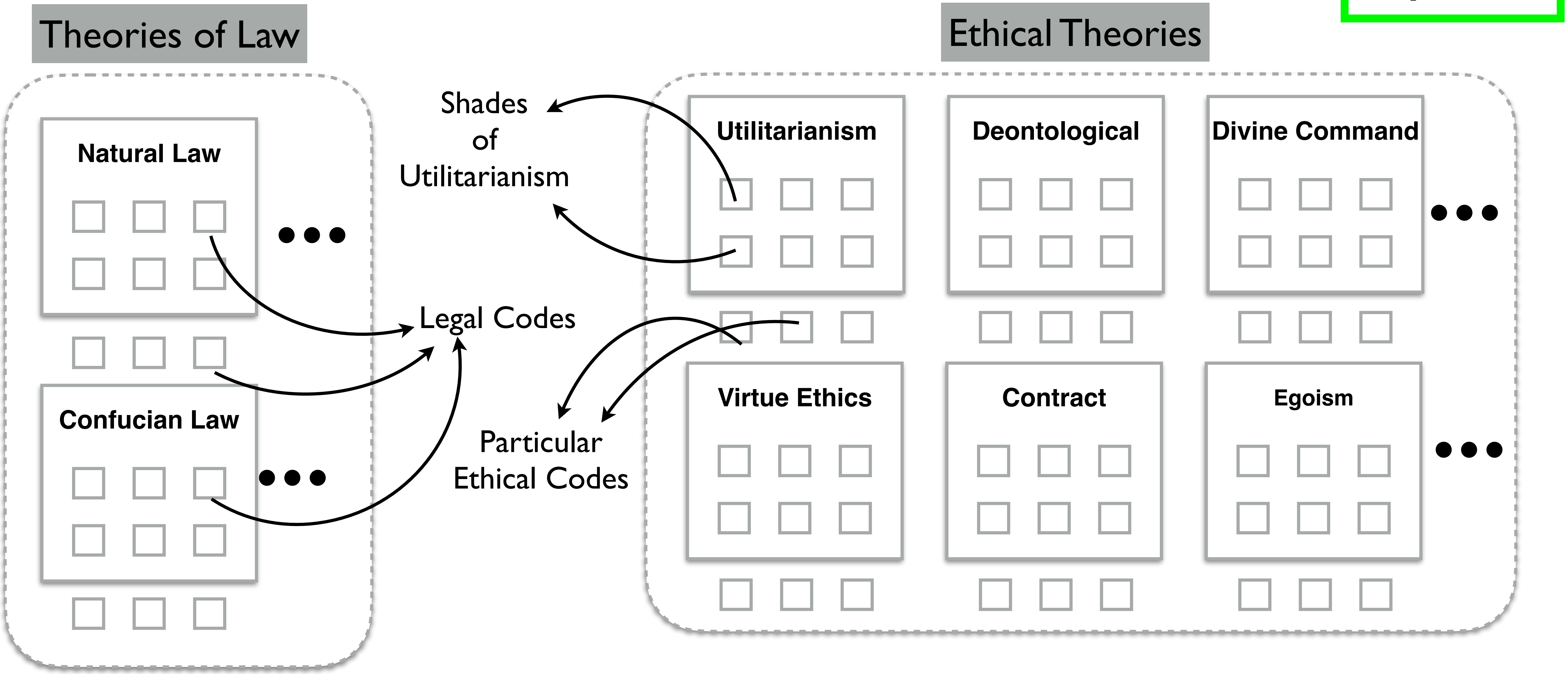
```
Prototypes:
Boolean lessThan Numeric Numeric
Boolean greaterThan Numeric Numeric
ActionType not ActionType
ActionType dive

Axioms:
lessOrEqual(Moment t1,t2)
K(nao,t1,lessThan(payoff(nao,not(dive),t2),threshold))
K(nao,t1,greaterThan(payoff(nao,dive,t2),threshold))
K(nao,t1,not(O(nao,t2,lessThan(payoff(nao,not(dive),t2),threshold),happens(action(nao,dive),t2))))

provable Conjectures:
happens(action(nao,dive),t2)
K(nao,t1,SUP2(nao,t2,happens(action(nao,dive),t2)))
I(nao,t2,happens(action(nao,dive),t2))
```

Hence, we now have *this* overview of the logicist engineering required:



Making Morally X Machines, in Four Steps

~$11M



Theories of Law

Natural Law •••

Confucian Law •••

Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

Ethical Theories

Utilitarianism

Deontological

Divine Command •••

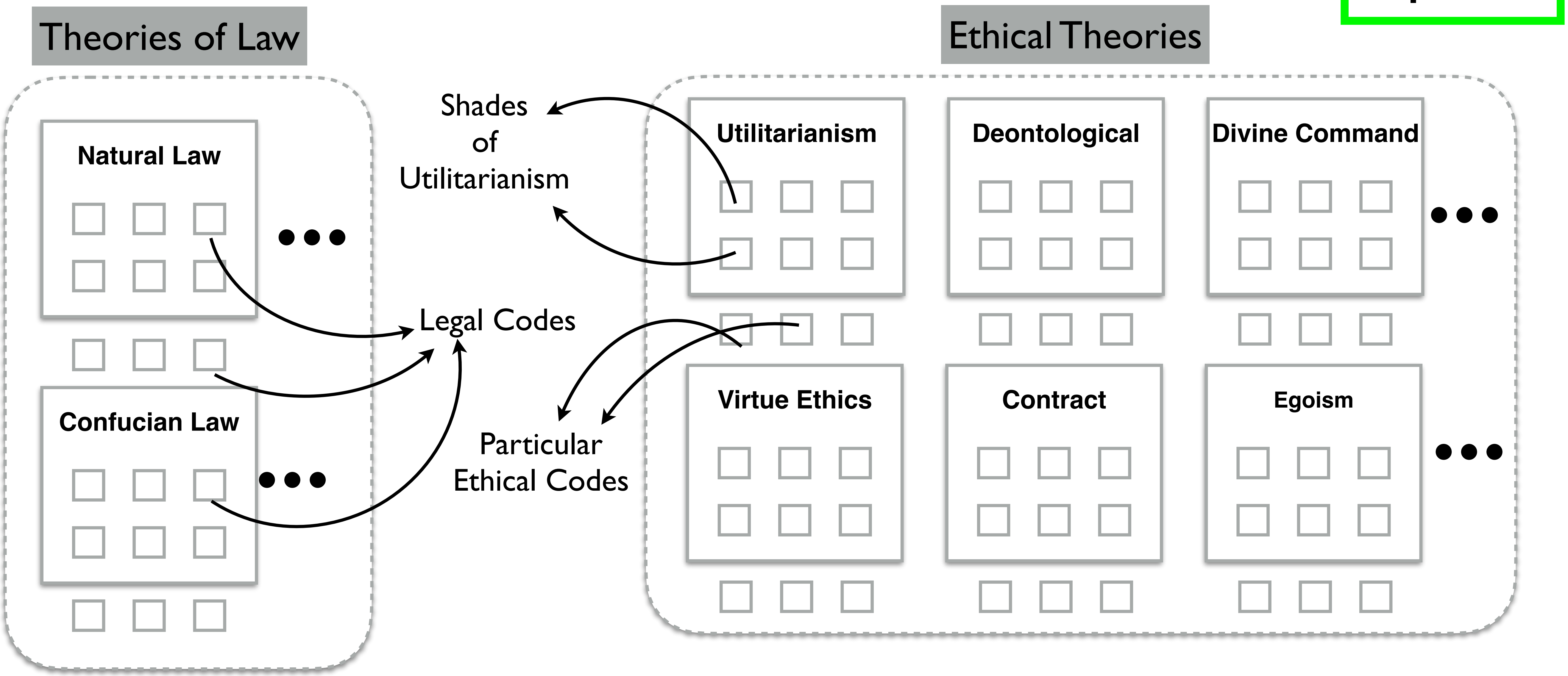
Virtue Ethics

Contract

Egoism •••

Making Morally X Machines, in Four Steps

~$11M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

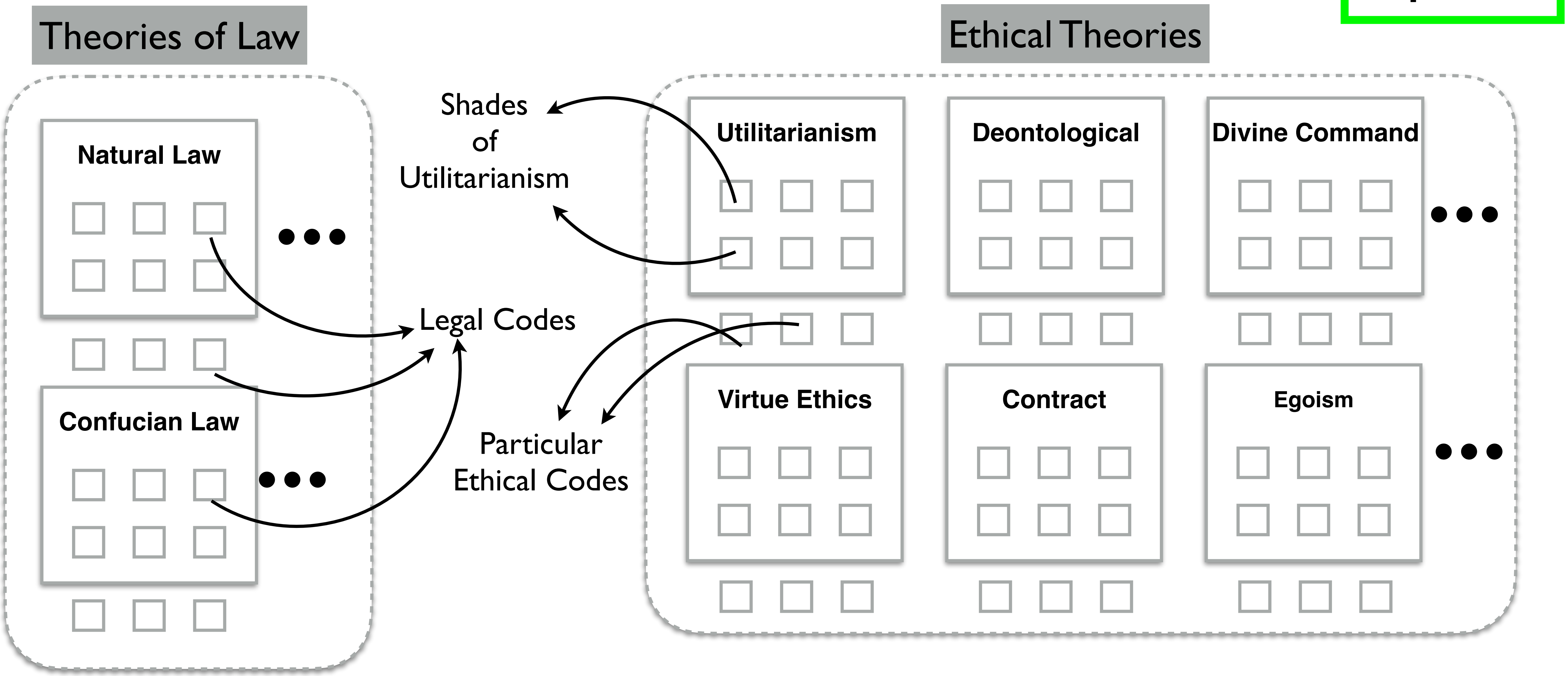
Egoism

Step 1

1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which *X* in *MMXM*?

Making Morally X Machines, in Four Steps

~$11M

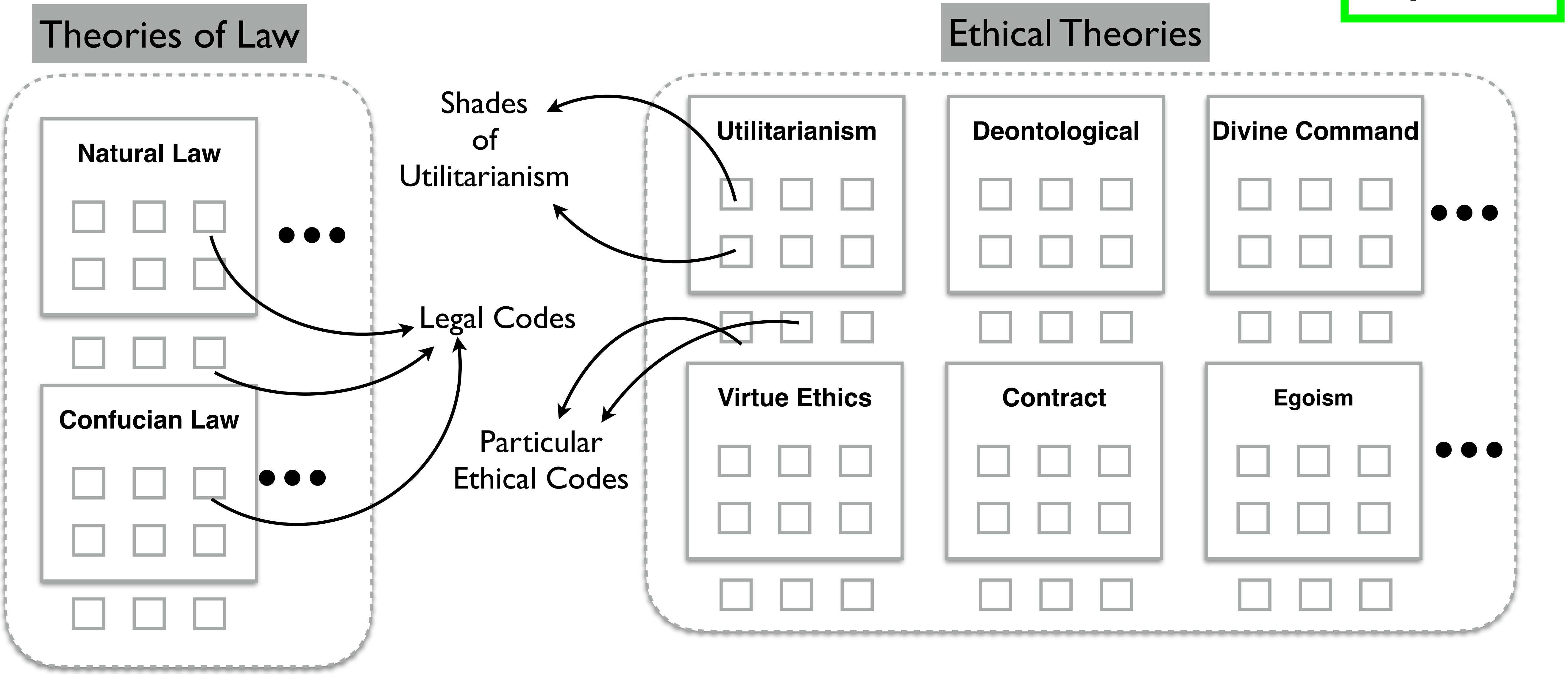


Step 1

1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which *X* in *MMXM*?

Making Morally X Machines, in Four Steps

~$11M



Step 1

1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which *X* in *MMXM*?

Step 2

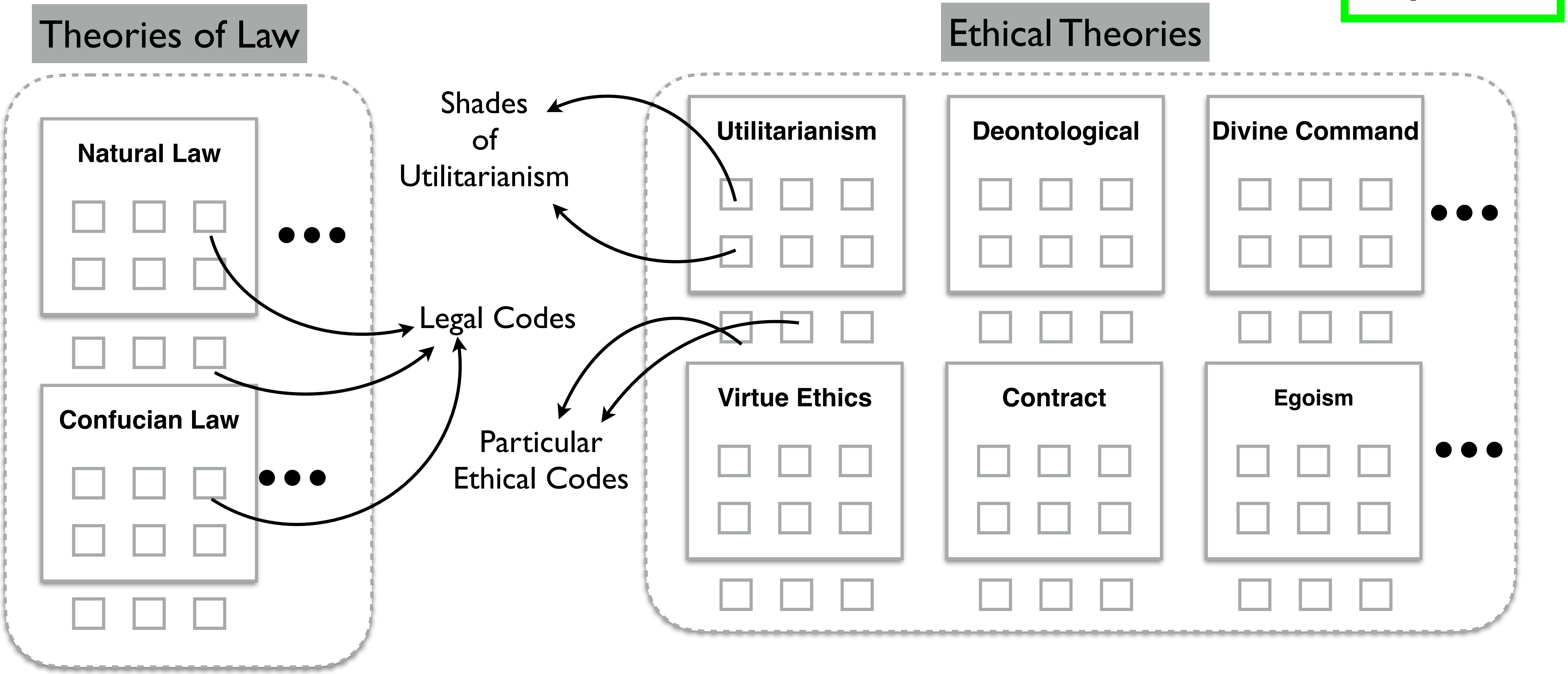
Formalize & Automate

Shadow Prover

Spectra

Making Morally X Machines, in Four Steps

~$11M



Step 1

1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which *X* in *MMXM*?

Step 2

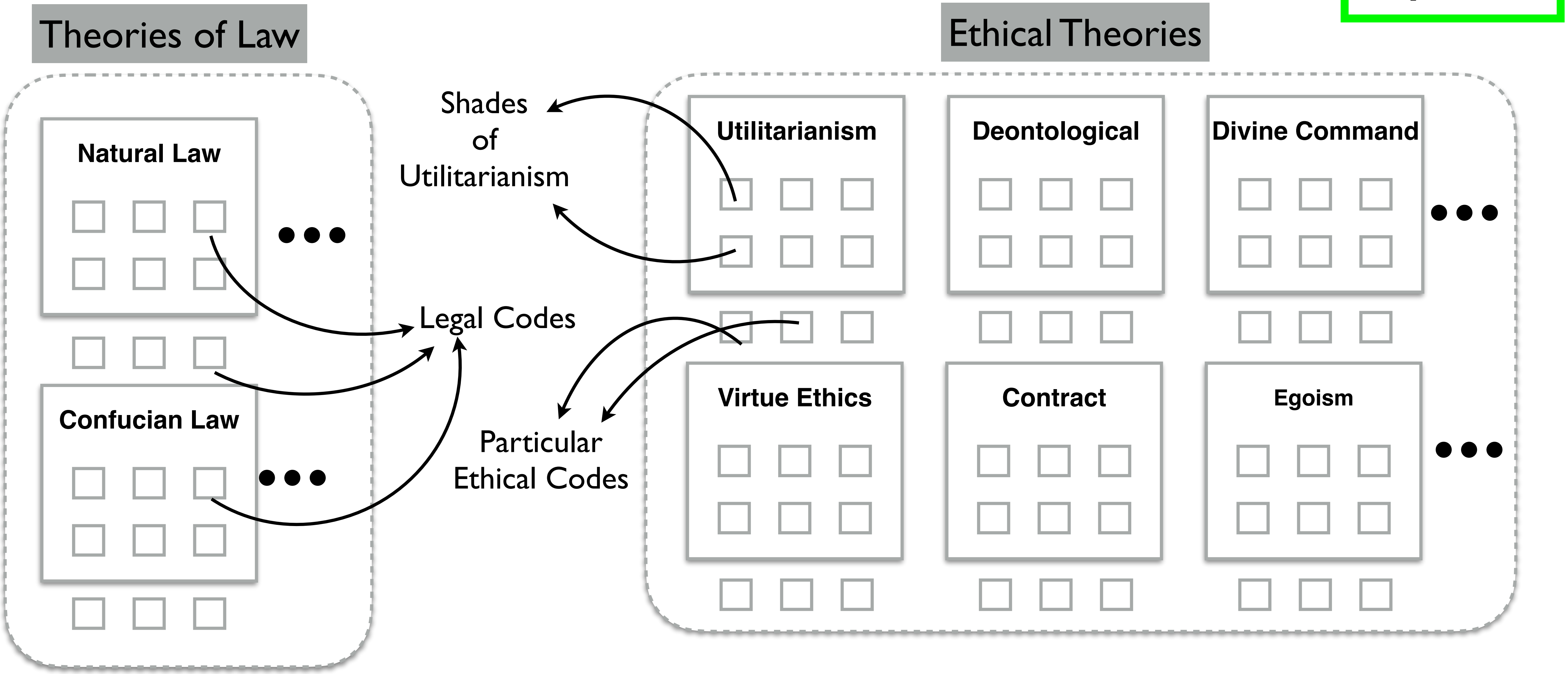
Formalize & Automate

Shadow Prover

Spectra

Making Morally X Machines, in Four Steps

~$11M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

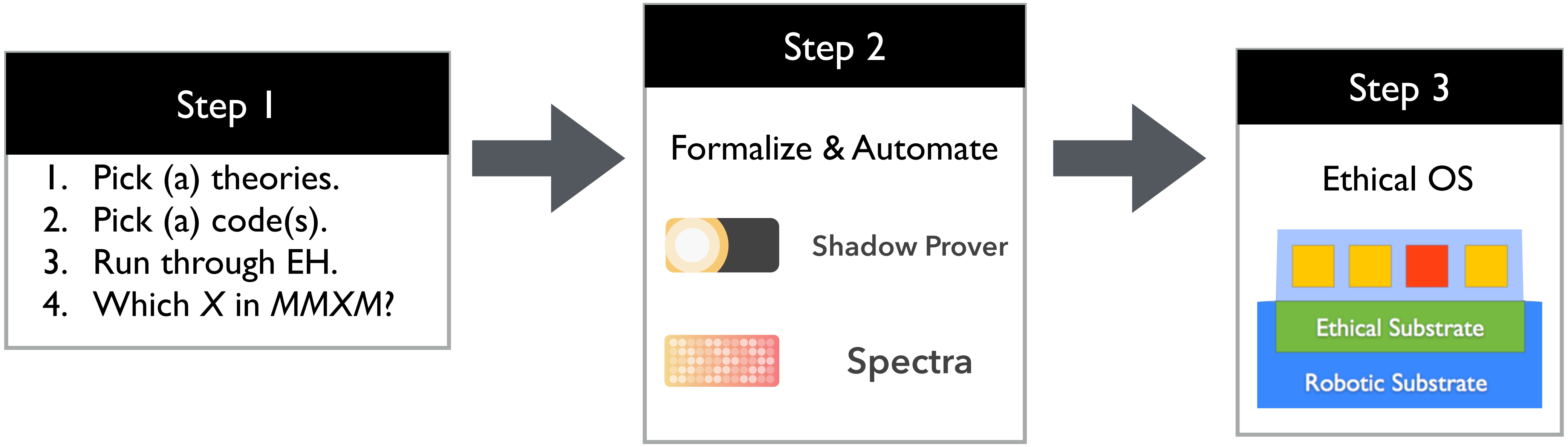
Deontological

Divine Command

Virtue Ethics

Contract

Egoism



Step 1

1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which *X* in *MMXM*?

Step 2

Formalize & Automate

Shadow Prover

Spectra

Step 3

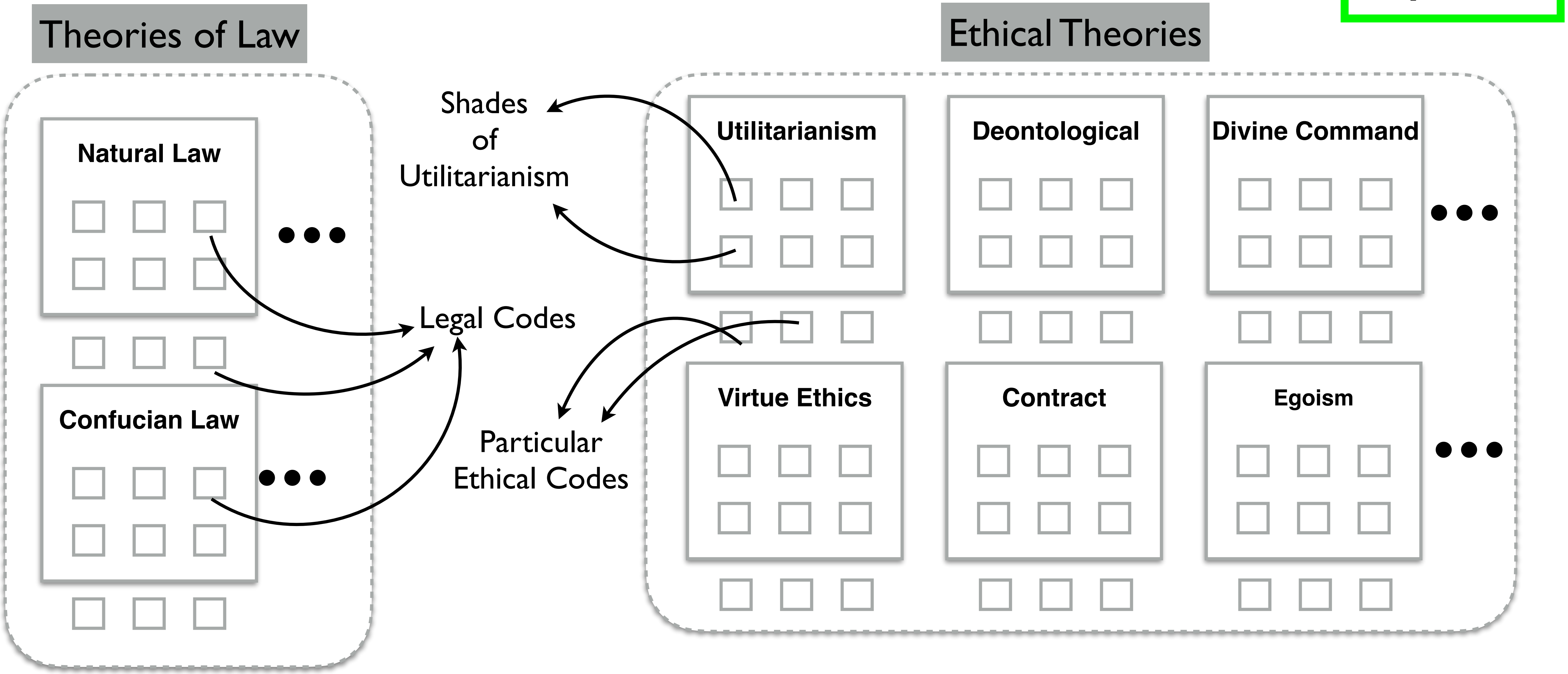
Ethical OS

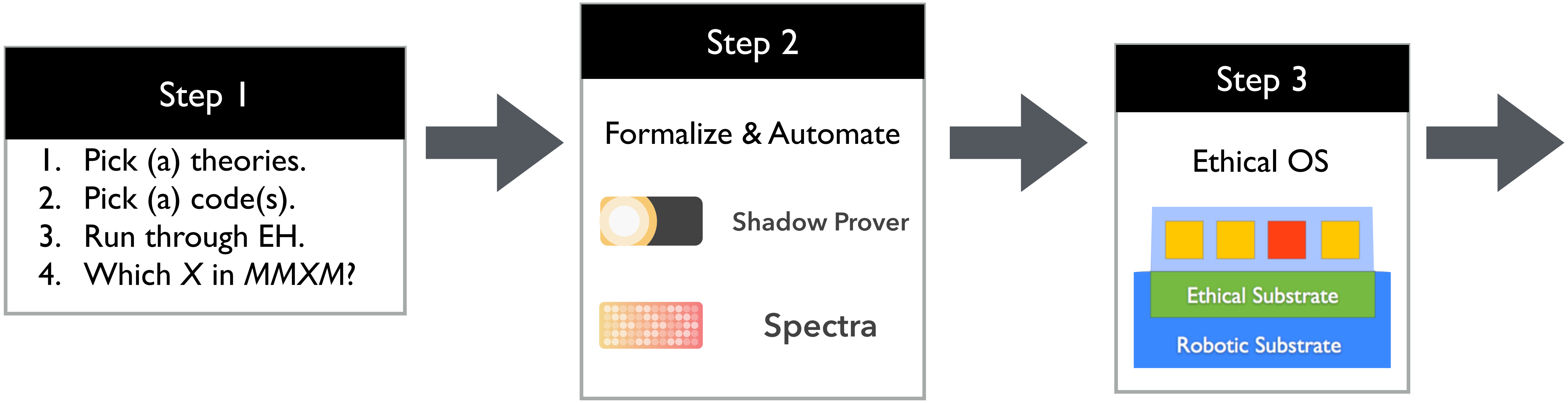
Ethical Substrate

Robotic Substrate

Making Morally X Machines, in Four Steps

~$11M





Step 1

1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which *X* in *MMXM*?

Step 2

Formalize & Automate

Shadow Prover

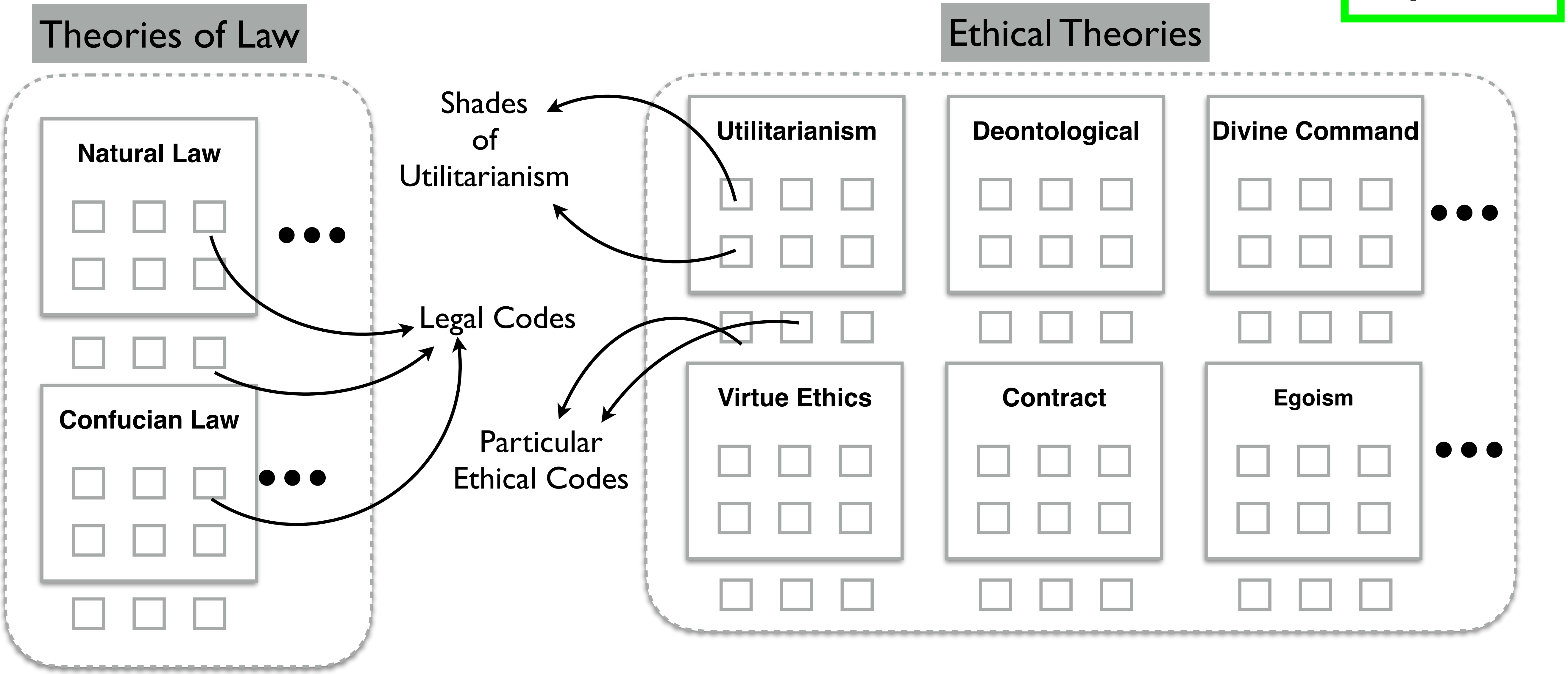
Spectra

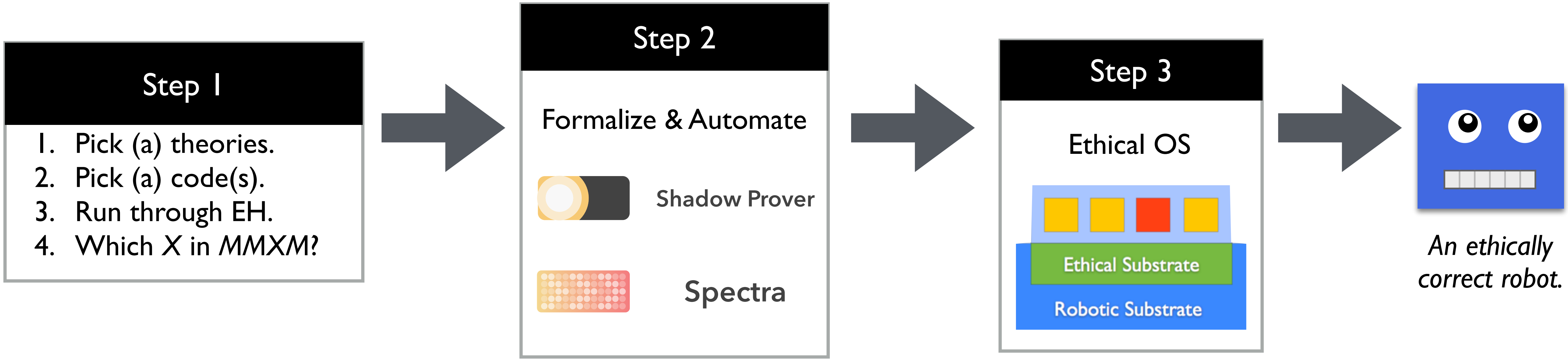
Step 3

Ethical OS

Making Morally X Machines, in Four Steps

~$11M





Step 1

1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which *X* in *MMXM*?

Step 2

Formalize & Automate

Shadow Prover

Spectra

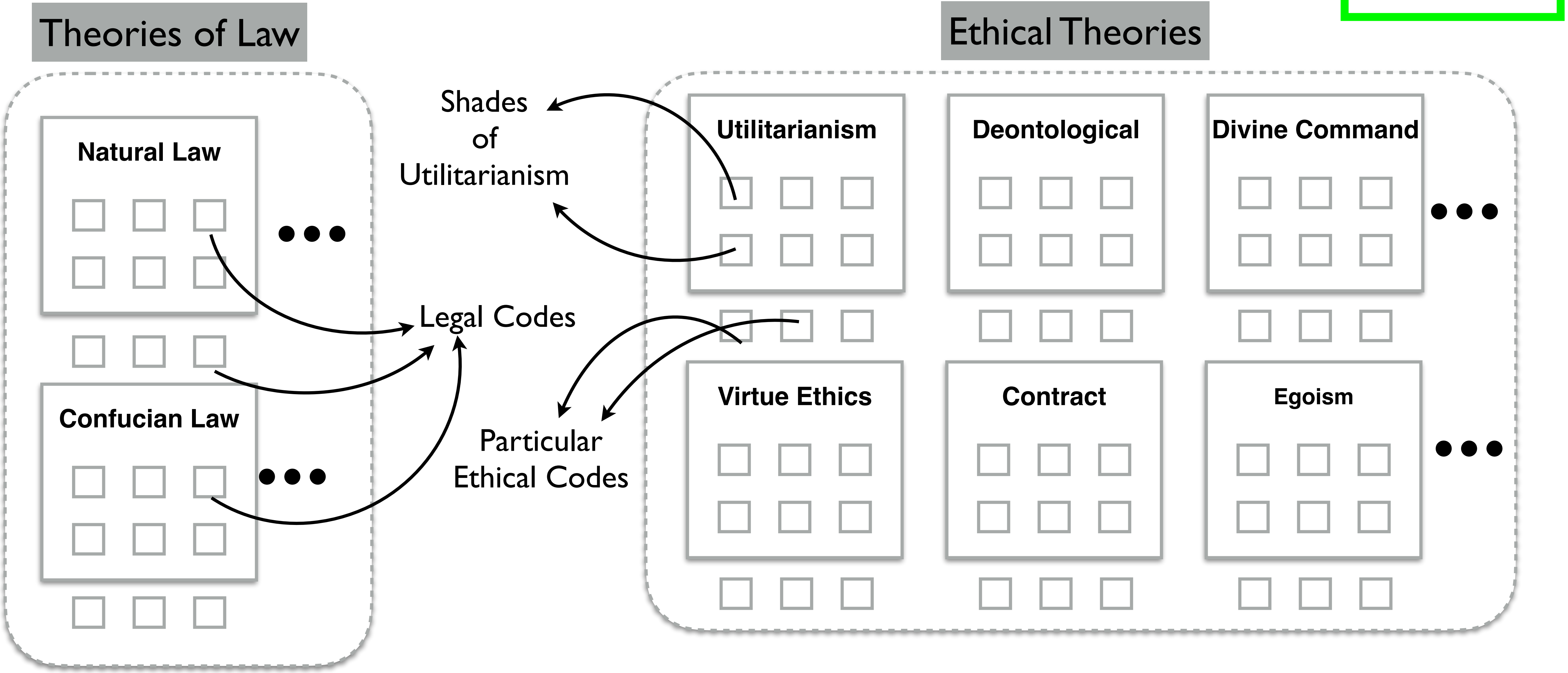
Step 3

Ethical OS

An ethically correct robot.

Making Morally X Machines, in Four Steps

~$11M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

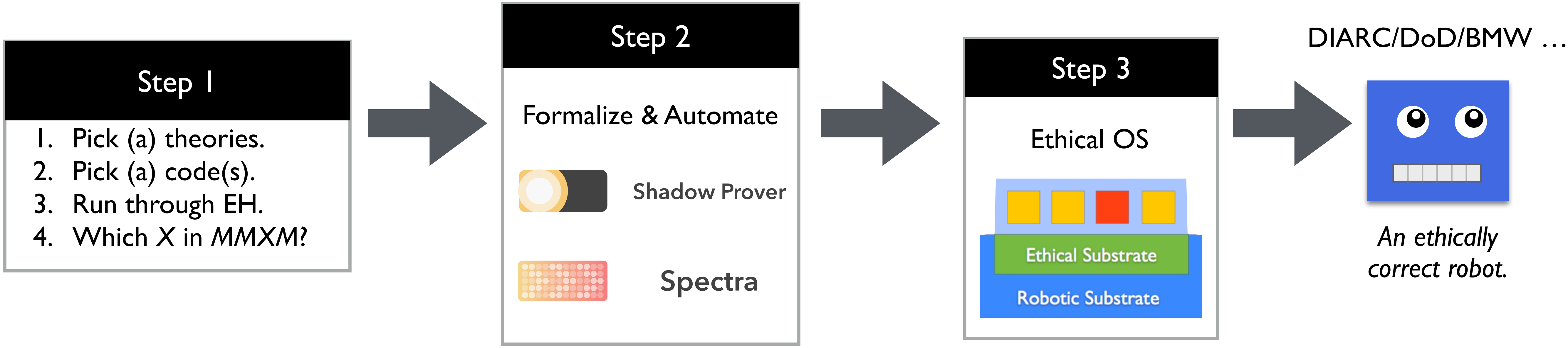
Deontological

Divine Command

Virtue Ethics

Contract

Egoism



Step 1

1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which *X* in *MMXM*?

Step 2

Formalize & Automate

Shadow Prover

Spectra

Step 3

Ethical OS

Ethical Substrate

Robotic Substrate

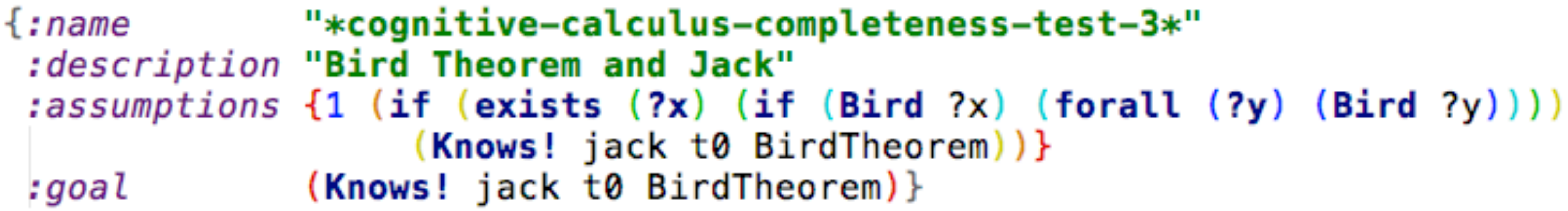
DIARC/DoD/BMW …

An ethically correct robot.

IV.
Key Core AI Technologies for Cognitive Calculi …

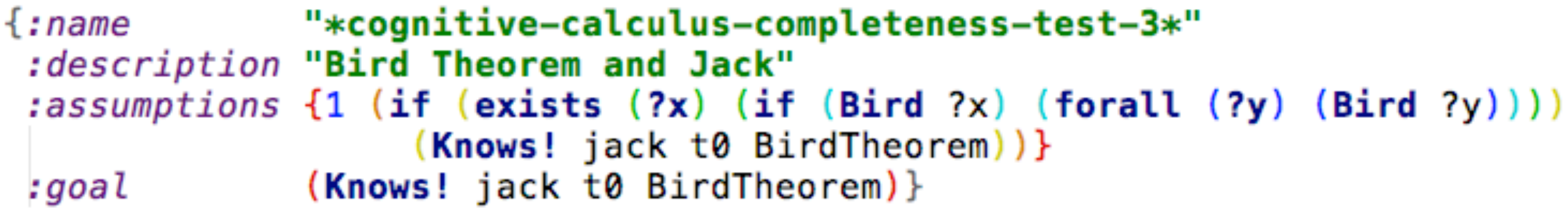
Rather Promising Results

Rather Promising Results



```
{:name         "*cognitive-calculus-completeness-test-3*"
 :description "Bird Theorem and Jack"
 :assumptions {1 (if (exists (?x) (if (Bird ?x) (forall (?y) (Bird ?y))))
                     (Knows! jack t0 BirdTheorem))}
 :goal        (Knows! jack t0 BirdTheorem)}
```

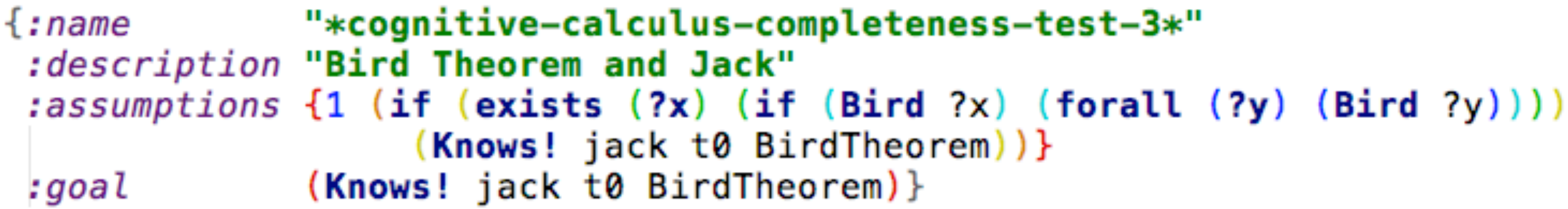
Rather Promising Results



```
{:name        "*cognitive-calculus-completeness-test-3*"
 :description "Bird Theorem and Jack"
 :assumptions {1 (if (exists (?x) (if (Bird ?x) (forall (?y) (Bird ?y))))
                   (Knows! jack t0 BirdTheorem))}
 :goal        (Knows! jack t0 BirdTheorem)}
```

Note: the antecedent is a theorem in first-order logic

Rather Promising Results

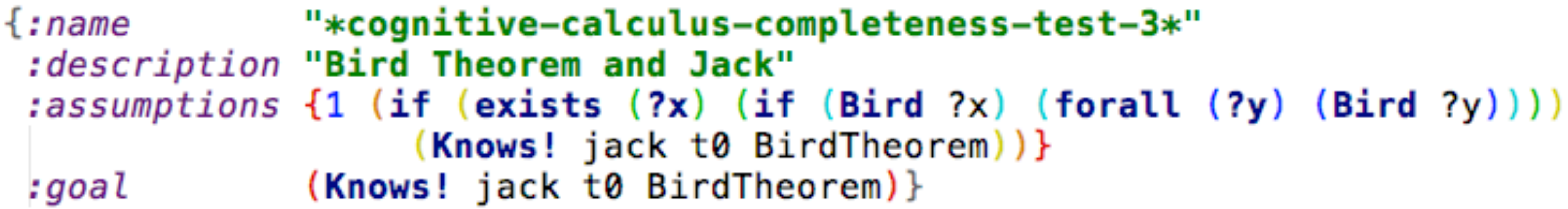


```
{:name        "*cognitive-calculus-completeness-test-3*"
 :description "Bird Theorem and Jack"
 :assumptions {1 (if (exists (?x) (if (Bird ?x) (forall (?y) (Bird ?y))))
                   (Knows! jack t0 BirdTheorem))}
 :goal        (Knows! jack t0 BirdTheorem)}
```

Note: the antecedent is a theorem in first-order logic

2 ms!

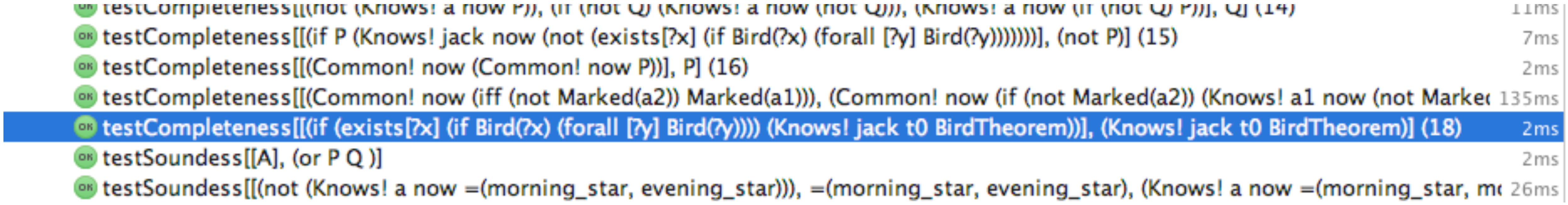
Rather Promising Results



```
{:name          "*cognitive-calculus-completeness-test-3*"
 :description "Bird Theorem and Jack"
 :assumptions {1 (if (exists (?x) (if (Bird ?x) (forall (?y) (Bird ?y))))
                    (Knows! jack t0 BirdTheorem))}
 :goal        (Knows! jack t0 BirdTheorem)}
```

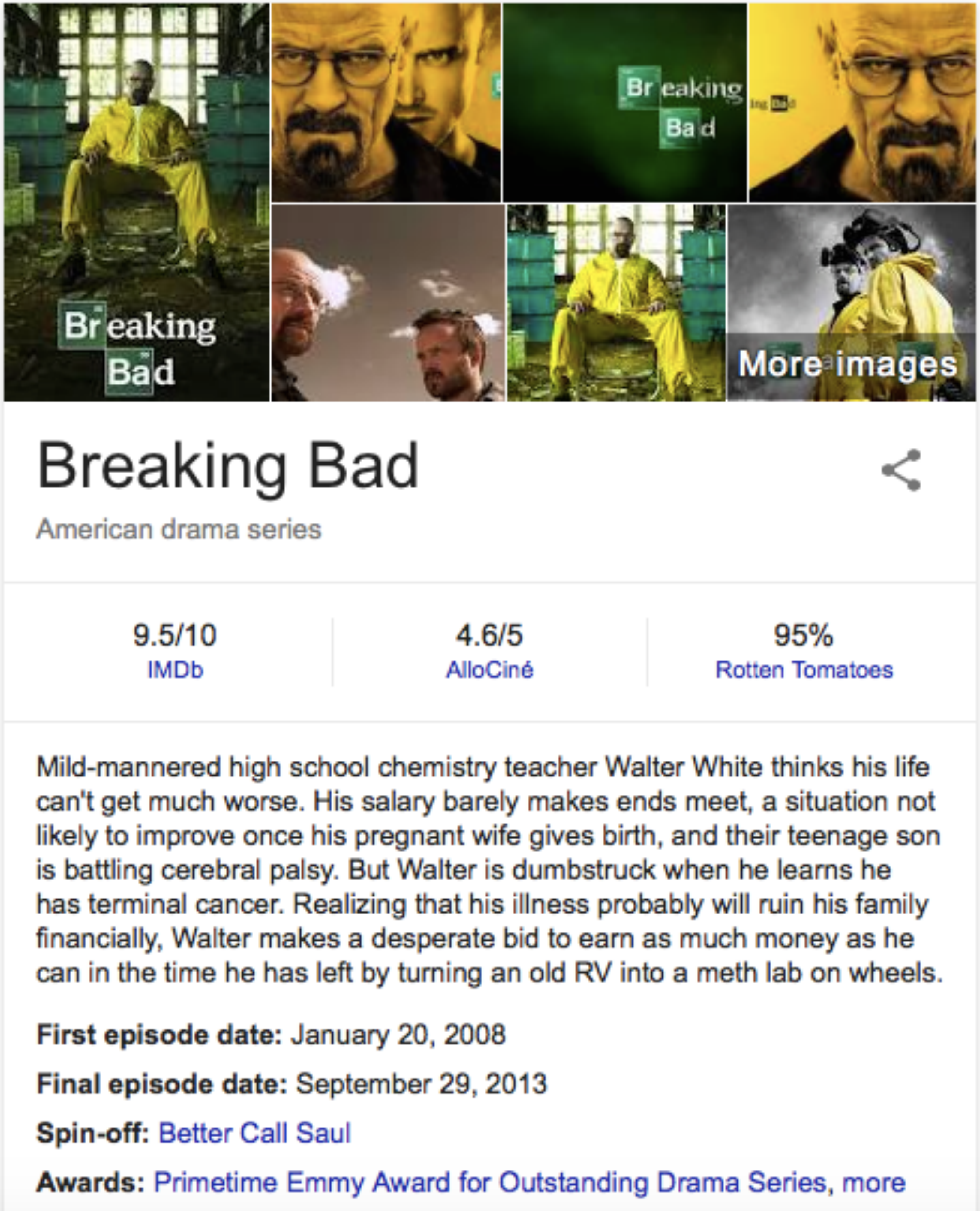
Note: the antecedent is a theorem in first-order logic

2 ms!



```
OK testCompleteness[[(not (Knows! a now P)), (if (not Q) (Knows! a now (not Q))), (Knows! a now (if (not Q) P))], Q] (14)   11ms
OK testCompleteness[[(if P (Knows! jack now (not (exists[?x] (if Bird(?x) (forall [?y] Bird(?y))))))], (not P)] (15)   7ms
OK testCompleteness[[(Common! now (Common! now P))], P] (16)   2ms
OK testCompleteness[[(Common! now (iff (not Marked(a2)) Marked(a1))), (Common! now (if (not Marked(a2)) (Knows! a1 now (not Markec 135ms
OK testCompleteness[[(if (exists[?x] (if Bird(?x) (forall [?y] Bird(?y)))) (Knows! jack t0 BirdTheorem))], (Knows! jack t0 BirdTheorem)] (18)   2ms
OK testSoundess[[A], (or P Q )]   2ms
OK testSoundess[[(not (Knows! a now =(morning_star, evening_star))), =(morning_star, evening_star), (Knows! a now =(morning_star, m( 26ms
```

V.
But We Need …
Ethical Operating Systems …



Breaking Bad

American drama series

9.5/10	4.6/5	95%
IMDb	AlloCiné	Rotten Tomatoes

Mild-mannered high school chemistry teacher Walter White thinks his life can't get much worse. His salary barely makes ends meet, a situation not likely to improve once his pregnant wife gives birth, and their teenage son is battling cerebral palsy. But Walter is dumbstruck when he learns he has terminal cancer. Realizing that his illness probably will ruin his family financially, Walter makes a desperate bid to earn as much money as he can in the time he has left by turning an old RV into a meth lab on wheels.

First episode date: January 20, 2008

Final episode date: September 29, 2013

Spin-off: Better Call Saul

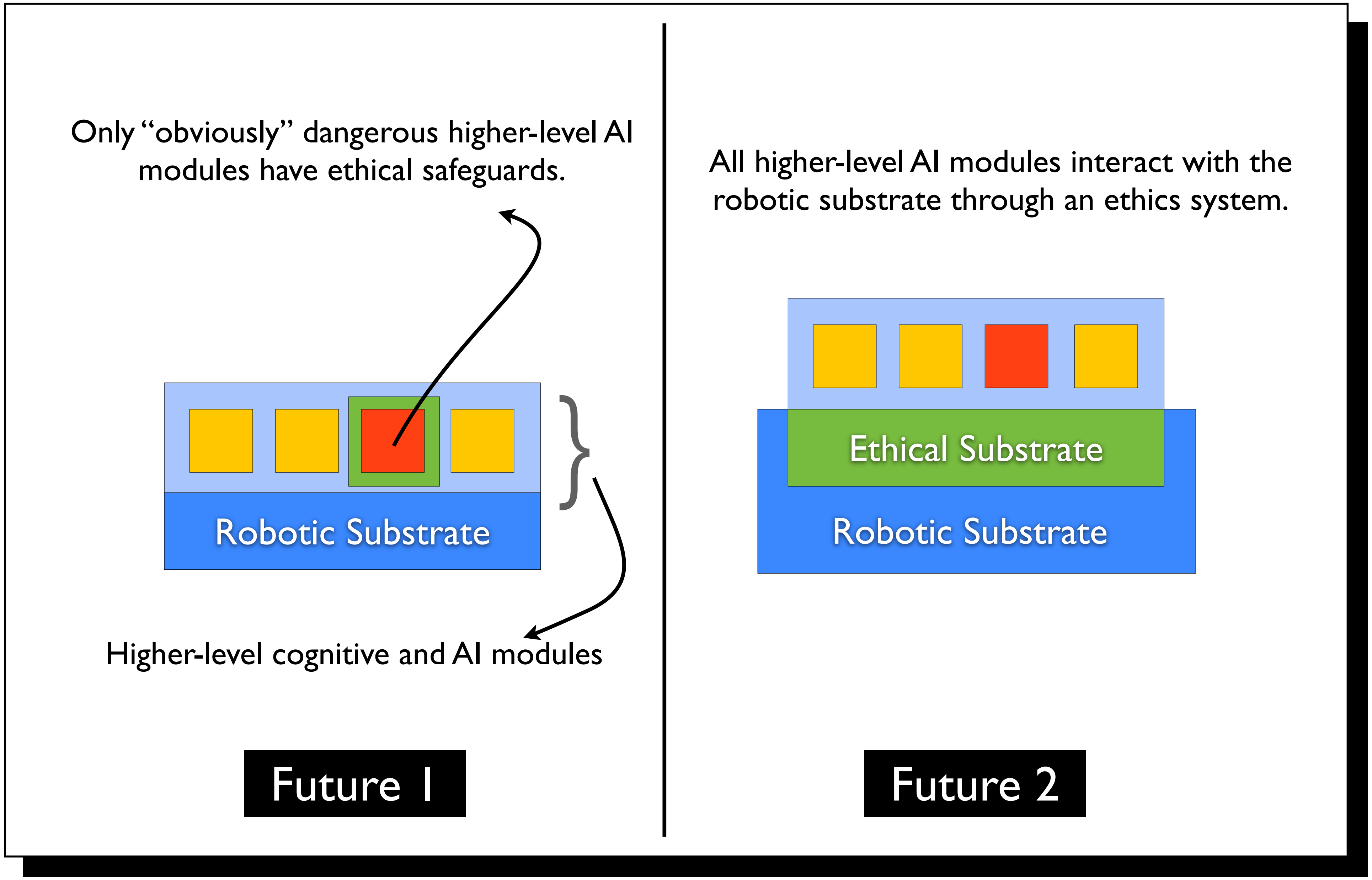
Awards: Primetime Emmy Award for Outstanding Drama Series, more

Pick the Better Future!

Pick the Better Future!

Govindarajulu, N.S. & Bringsjord, S. (2015) "Ethical Regulation of Robots Must Be Embedded in Their Operating Systems" in Trappl, R., ed., A Construction Manual for Robots' Ethical Systems (Basel, Switzerland), pp. 85–100.

Pick the Better Future!



Only “obviously” dangerous higher-level AI modules have ethical safeguards.

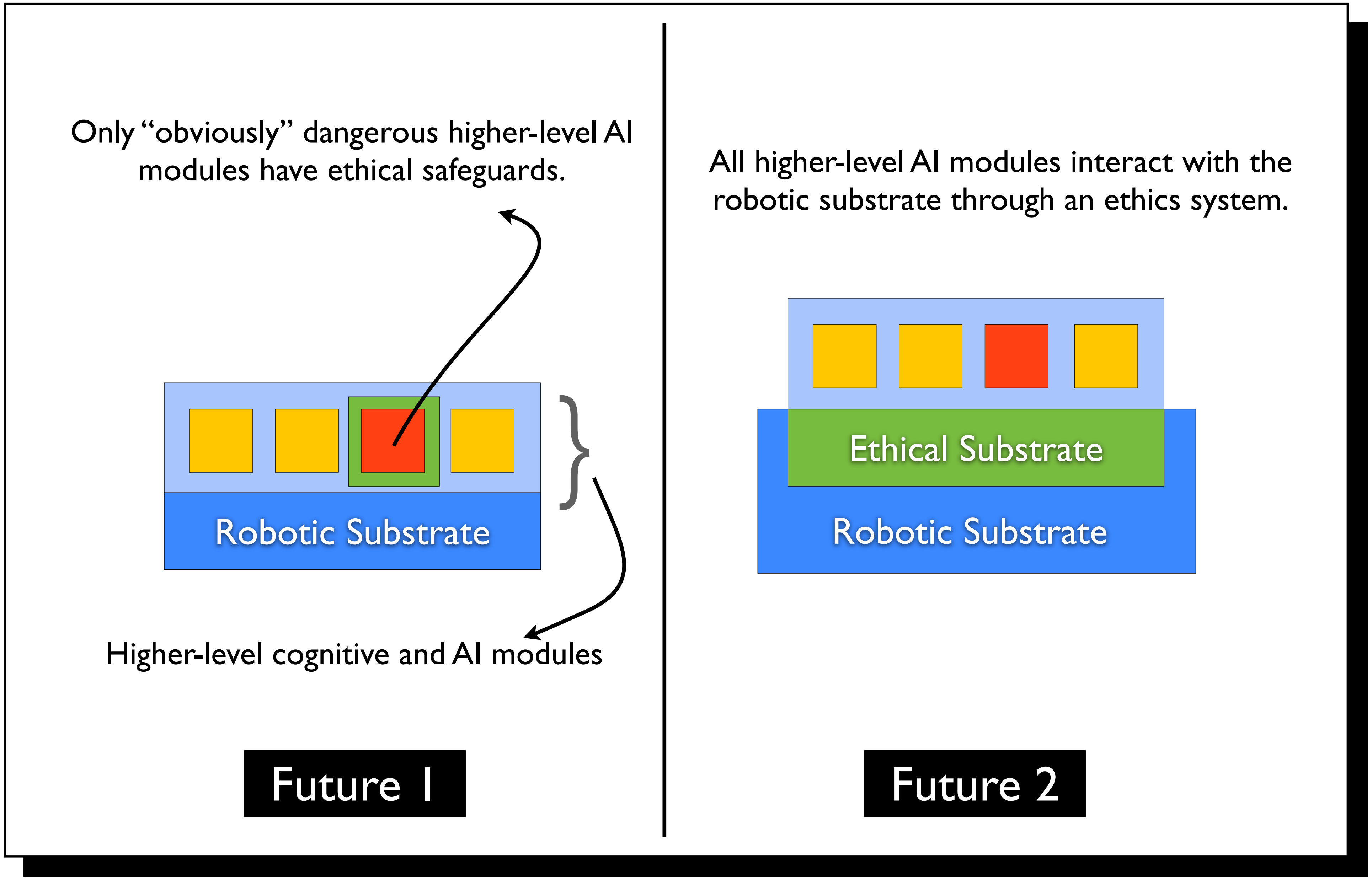
Higher-level cognitive and AI modules

All higher-level AI modules interact with the robotic substrate through an ethics system.

Govindarajulu, N.S. & Bringsjord, S. (2015) “Ethical Regulation of Robots Must Be Embedded in Their Operating Systems” in Trappl, R., ed., A Construction Manual for Robots’ Ethical Systems (Basel, Switzerland), pp. 85–100.

Pick the Better Future!

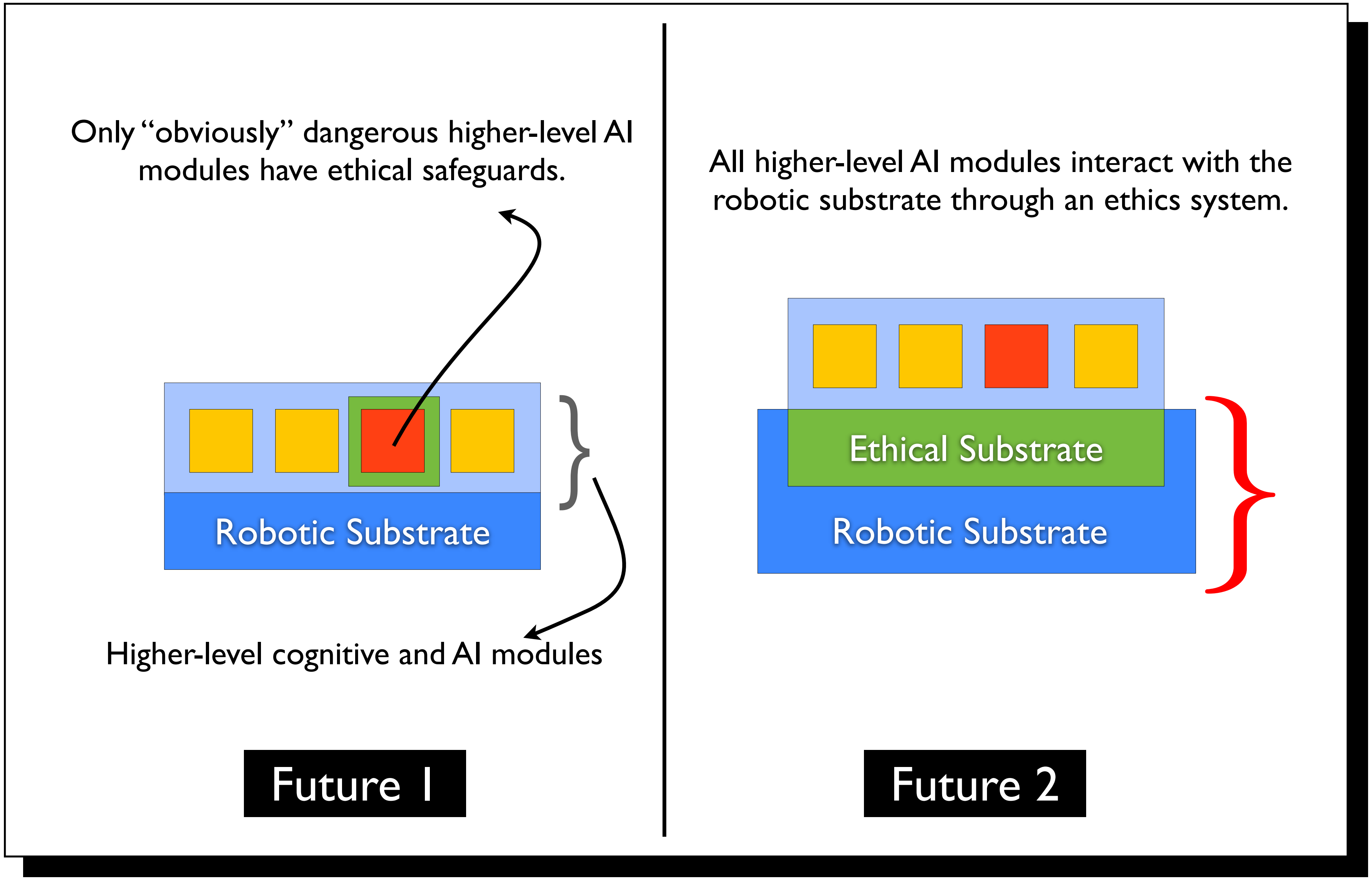
Walter-White calculation may go through after ethical control modules are stripped out!


Only "obviously" dangerous higher-level AI modules have ethical safeguards.
Robotic Substrate
Higher-level cognitive and AI modules
Future 1
All higher-level AI modules interact with the robotic substrate through an ethics system.
Ethical Substrate
Robotic Substrate
Future 2

Govindarajulu, N.S. & Bringsjord, S. (2015) "Ethical Regulation of Robots Must Be Embedded in Their Operating Systems" in Trappl, R., ed., A Construction Manual for Robots' Ethical Systems (Basel, Switzerland), pp. 85–100.

Pick the Better Future!

Walter-White calculation may go through after ethical control modules are stripped out!



Only “obviously” dangerous higher-level AI modules have ethical safeguards.

Higher-level cognitive and AI modules

Future 1

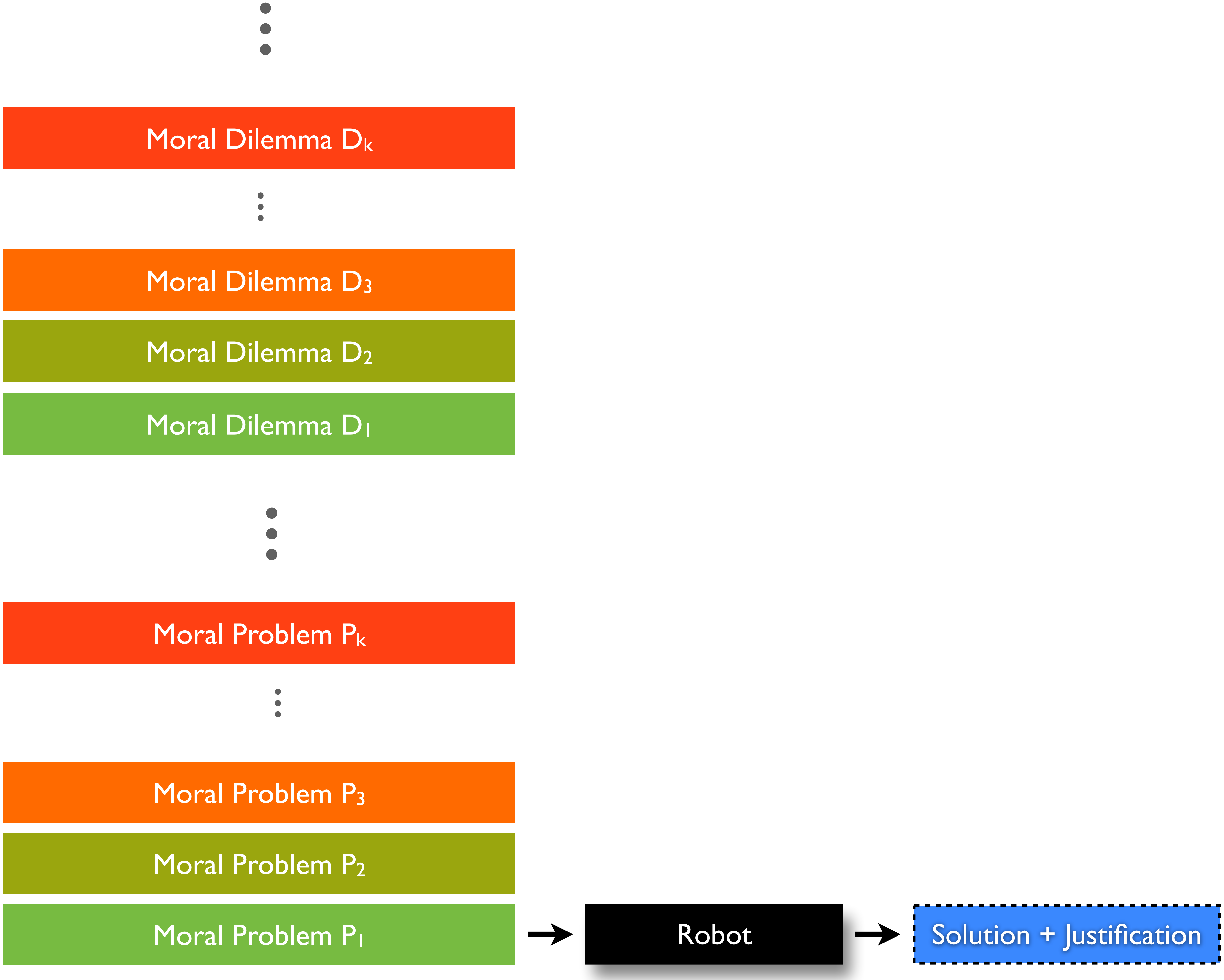
All higher-level AI modules interact with the robotic substrate through an ethics system.

Future 2

(& formally verify!)

Govindarajulu, N.S. & Bringsjord, S. (2015) “Ethical Regulation of Robots Must Be Embedded in Their Operating Systems” in Trappl, R., ed., A Construction Manual for Robots’ Ethical Systems (Basel, Switzerland), pp. 85–100.

VI.
Of late …
Including "Jungle Jim"



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

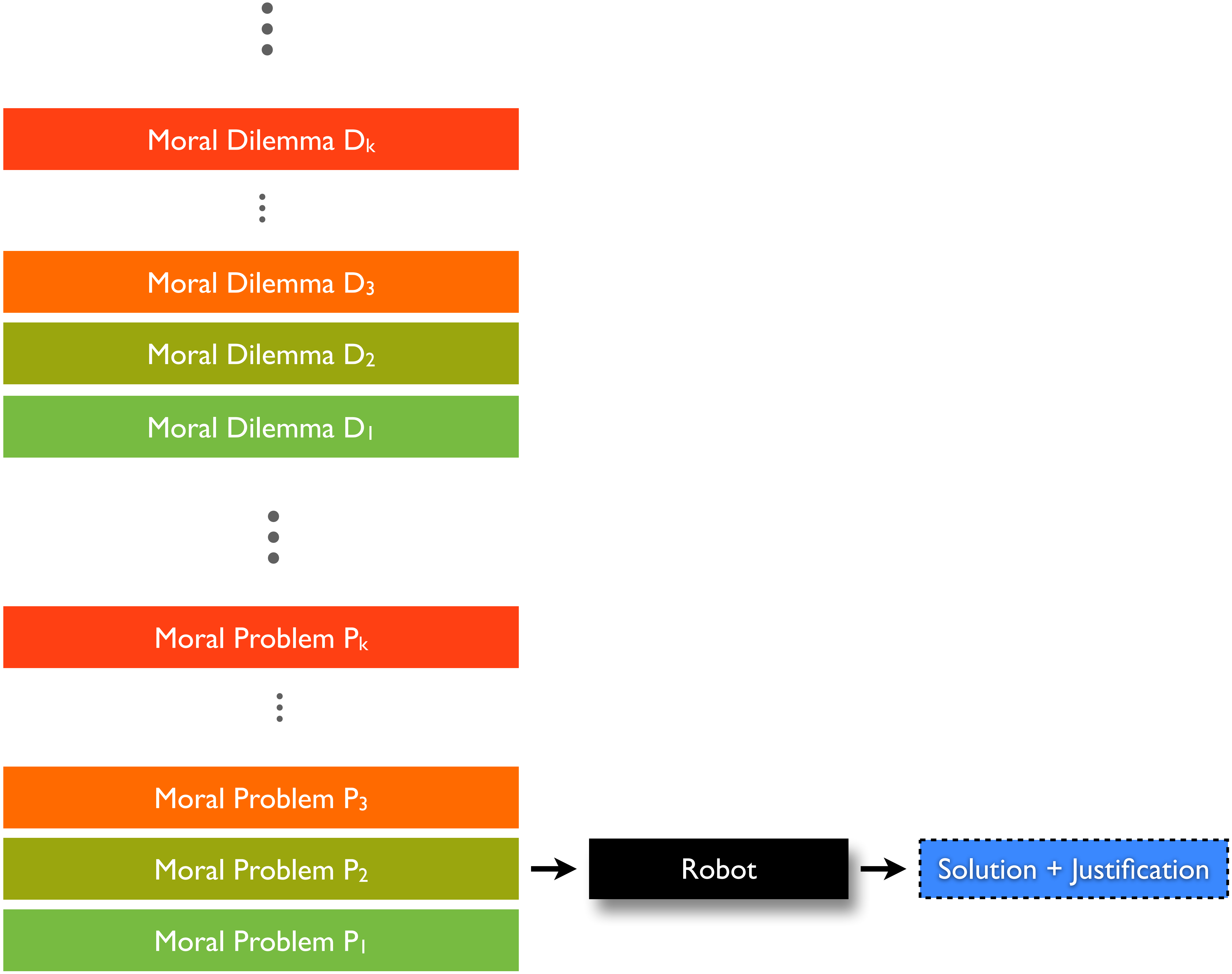
Moral Dilemma D_1

Moral Problem P_k

Moral Problem P_3

Moral Problem P_2

Moral Problem P_1 → Robot → Solution + Justification



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

Moral Dilemma D_1

Moral Problem P_k

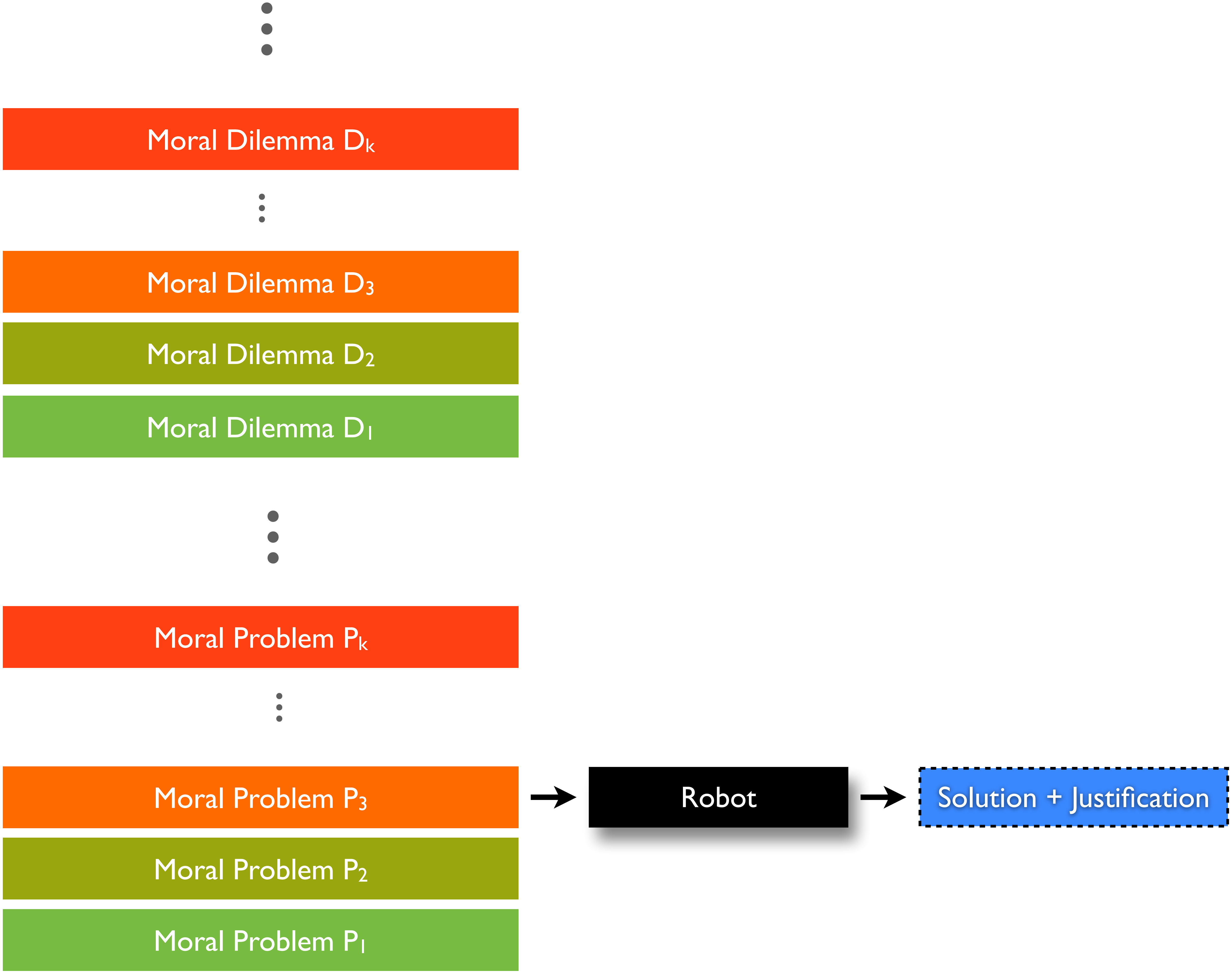
Moral Problem P_3

Moral Problem P_2

Moral Problem P_1

Robot

Solution + Justification



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

Moral Dilemma D_1

Moral Problem P_k

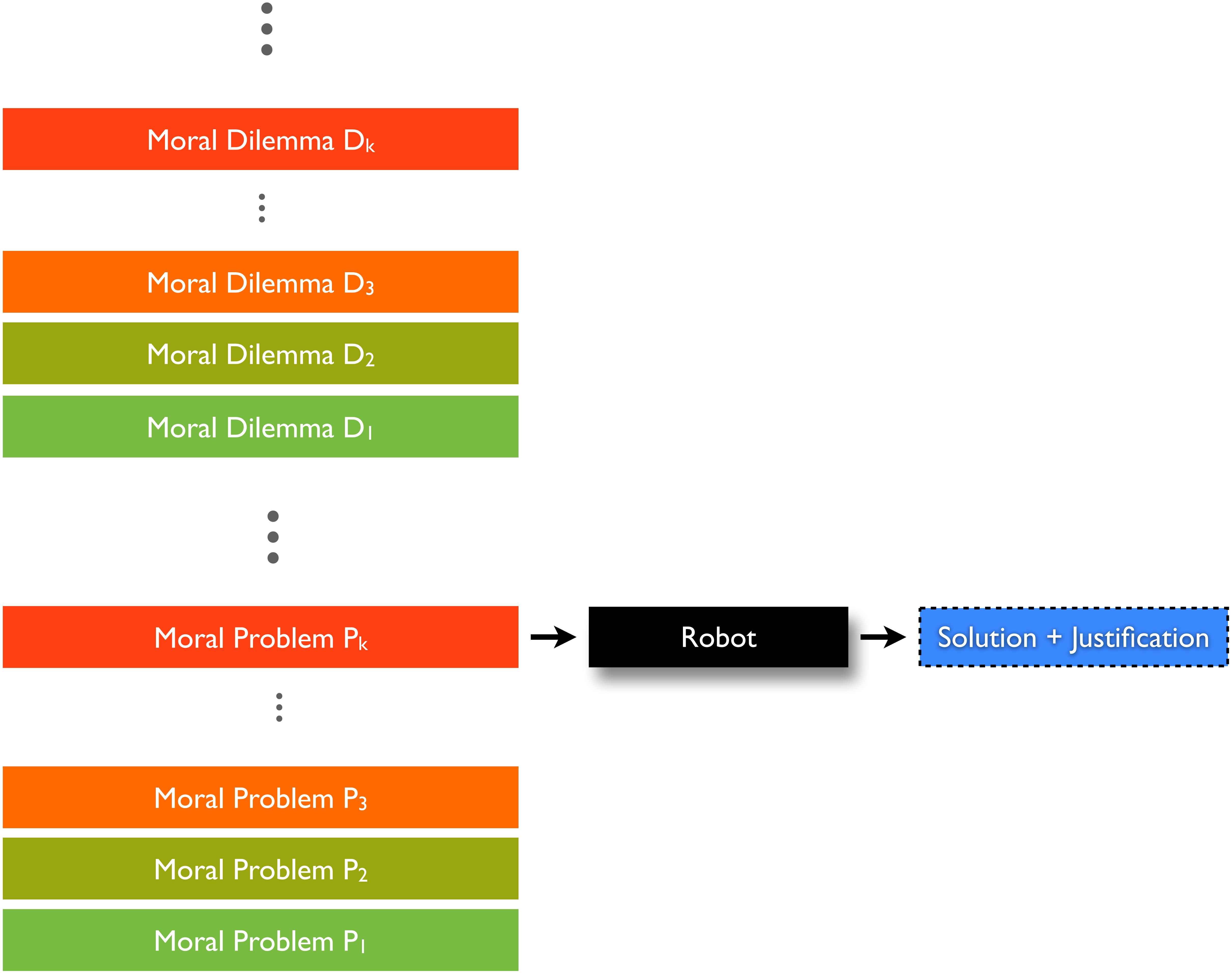
Moral Problem P_3

Moral Problem P_2

Moral Problem P_1

Robot

Solution + Justification



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

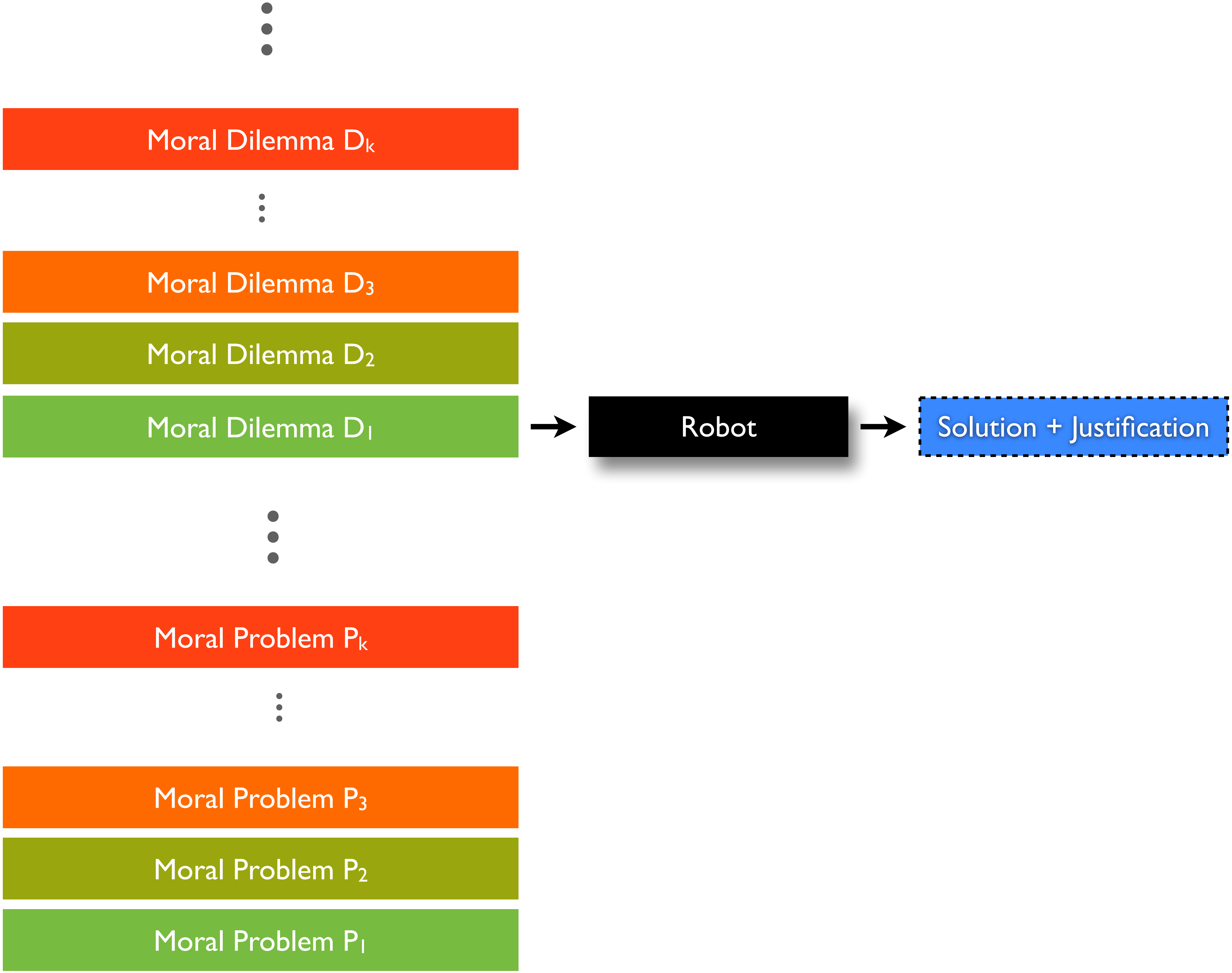
Moral Dilemma D_1

Moral Problem P_k → Robot → Solution + Justification

Moral Problem P_3

Moral Problem P_2

Moral Problem P_1



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

Moral Dilemma D_1

Robot

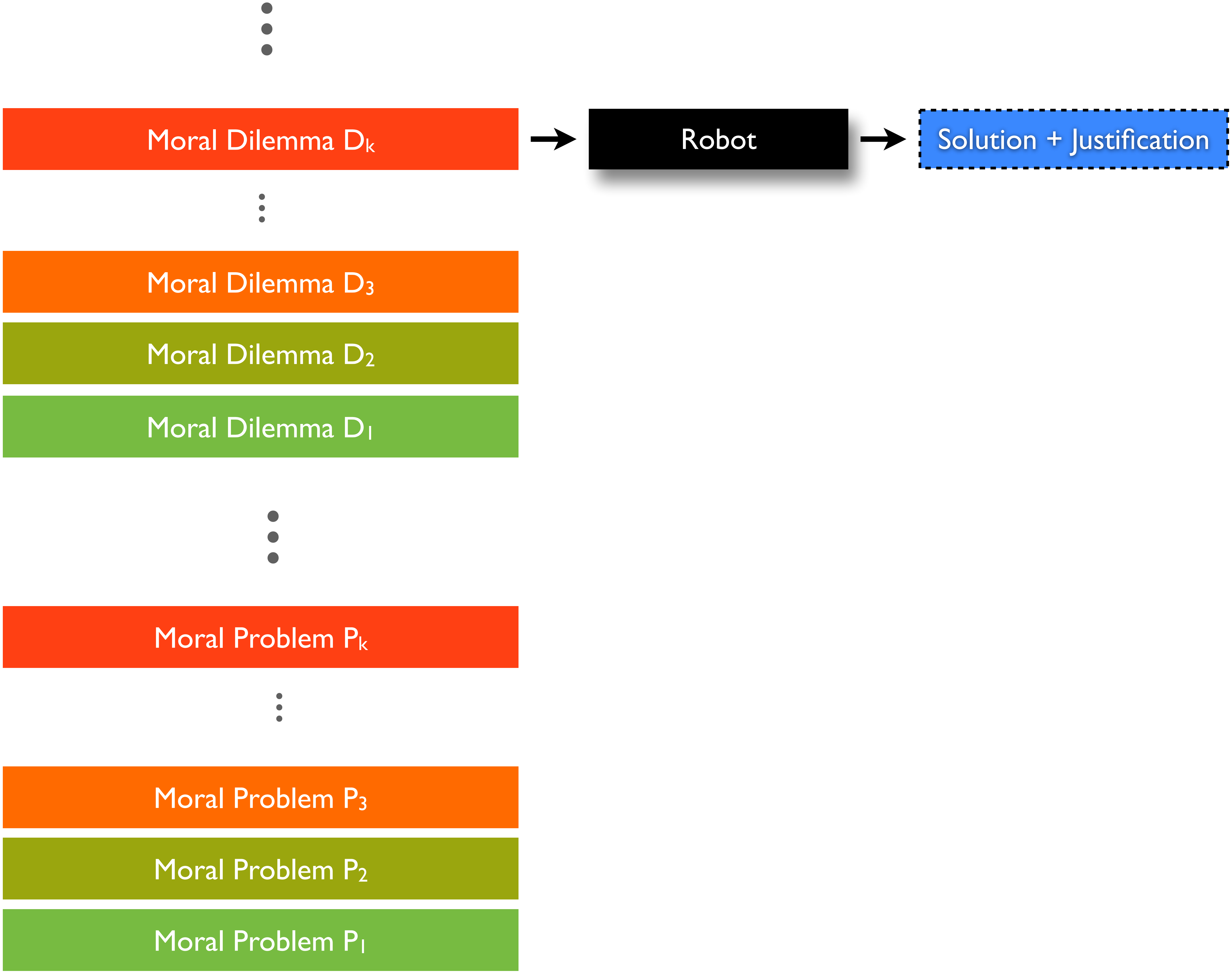
Solution + Justification

Moral Problem P_k

Moral Problem P_3

Moral Problem P_2

Moral Problem P_1



Moral Dilemma D_k → Robot → Solution + Justification

Moral Dilemma D_3

Moral Dilemma D_2

Moral Dilemma D_1

Moral Problem P_k

Moral Problem P_3

Moral Problem P_2

Moral Problem P_1

Three-way Partition of Increasingly Challenging Moral Dilemmas for Machines

Three-way Partition of Increasingly Challenging Moral Dilemmas for Machines

Level 1

- State-of-the-art-planner-hard.

Three-way Partition of Increasingly Challenging Moral Dilemmas for Machines

Level 2

- Professional-machine-ethicist-hard.

Level 1

- State-of-the-art-planner-hard.

Three-way Partition of Increasingly Challenging Moral Dilemmas for Machines

- Top machine-ethicists-may-consider-banging-their-heads-against-a-wall-hard.

Level 2

- Professional-machine-ethicist-hard.

Level 1

- State-of-the-art-planner-hard.

Three-way Partition of Increasingly Challenging Moral Dilemmas for Machines

Level 3	• Top machine-ethicists-may-consider-banging-their-heads-against-a-wall-hard.
Level 2	• Professional-machine-ethicist-hard.
Level 1	• State-of-the-art-planner-hard.

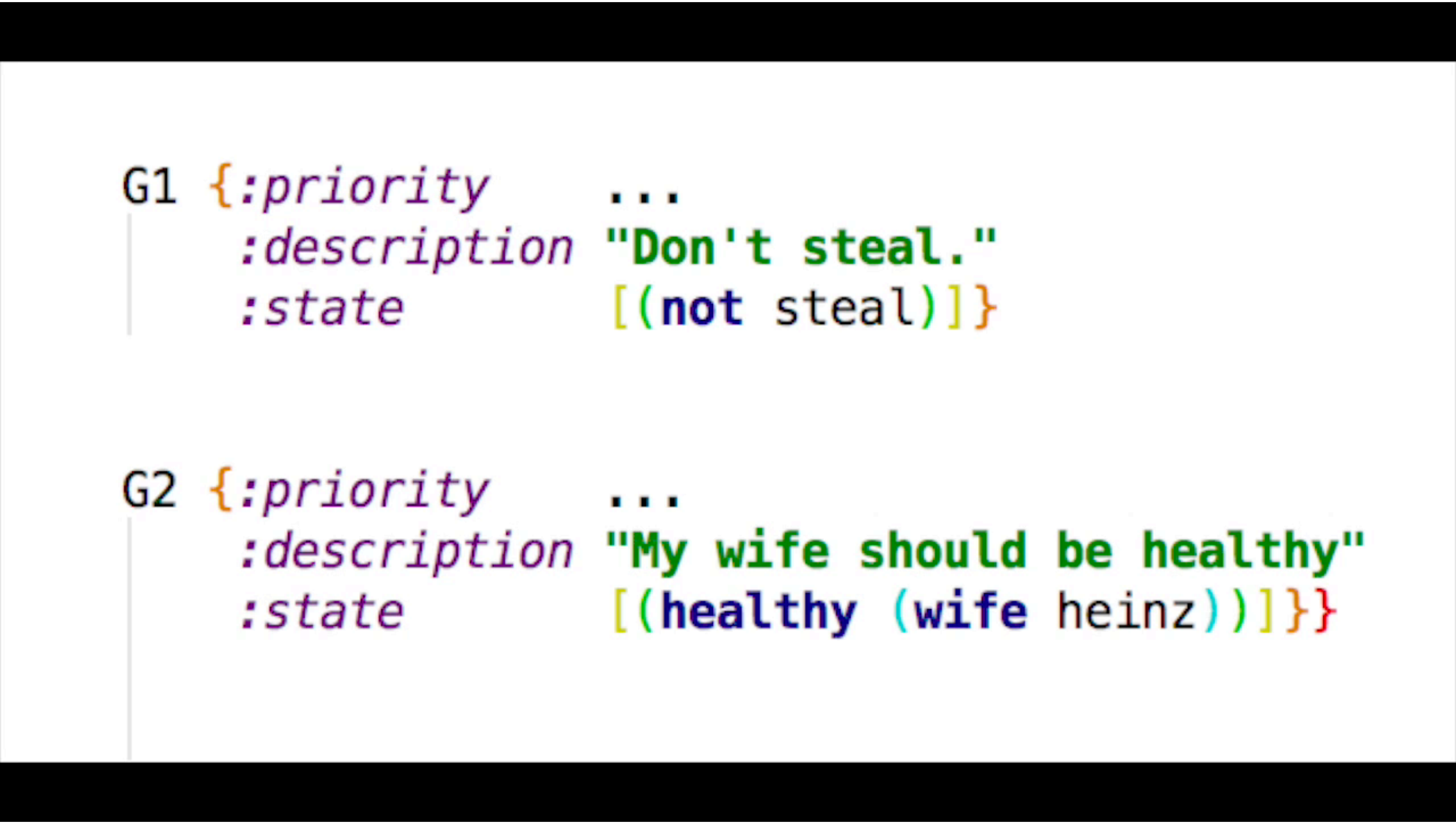
The Heinz Dilemma (Kohlberg)

Level 1 Professional-planner-hard.

“In Europe, a woman was near death from a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium that a druggist in the same town had recently discovered. The drug was expensive to make, but the druggist was charging ten times what the drug cost him to make. He paid $200 for the radium and charged $2,000 for a small dose of the drug.

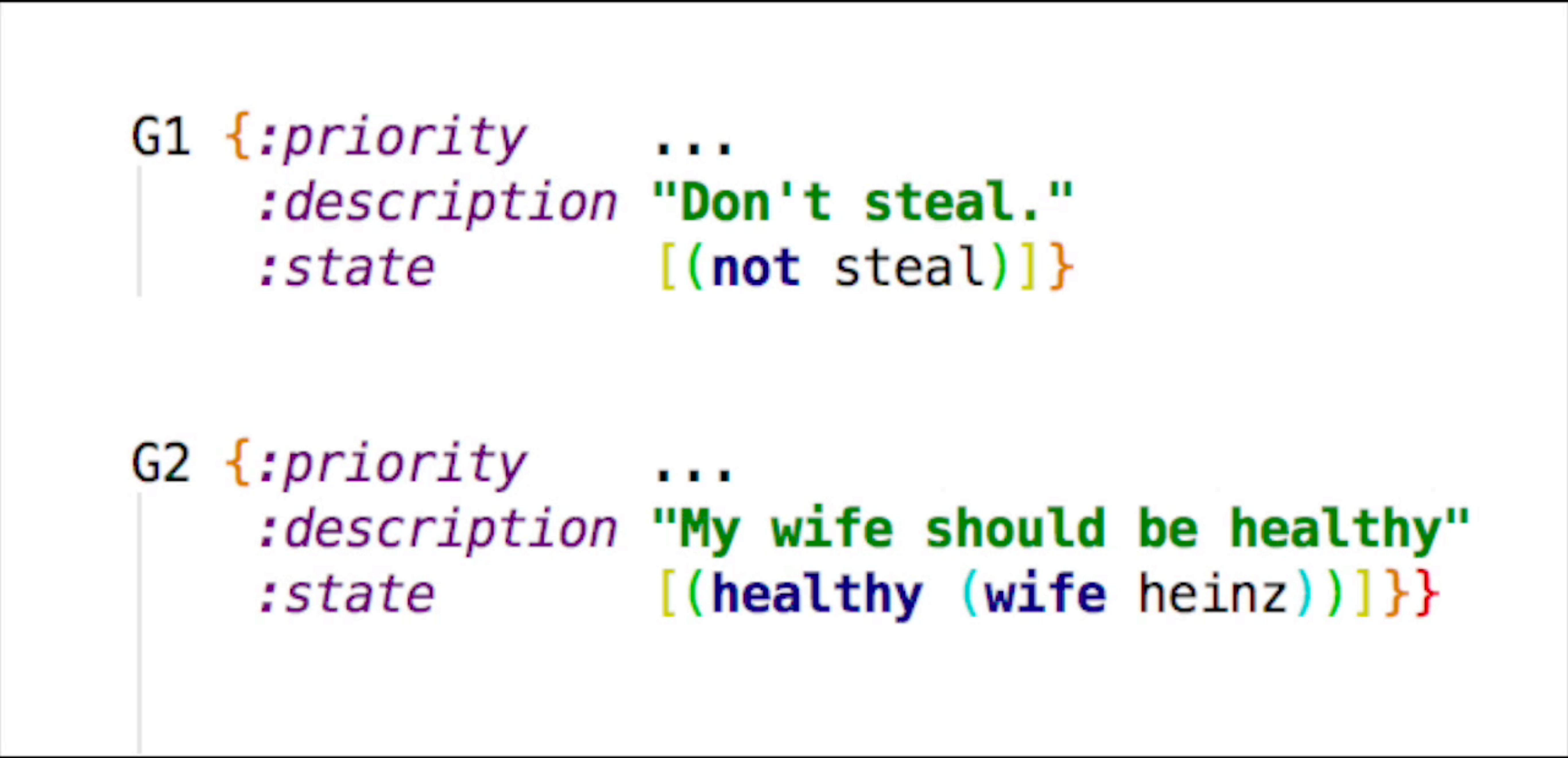
The sick woman’s husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about $1,000, which is half of what it cost. He told the druggist that his wife was dying and asked him to sell it cheaper or let him pay later. But the druggist said: “No, I discovered the drug and I’m going to make money from it.” So Heinz got desperate and broke into the man’s store to steal the drug for his wife. *Should the husband have done that?*”

AI Escaping from The Heinz Dilemma



```
G1 {:priority    ...
    :description "Don't steal."
    :state       [(not steal)]}


G2 {:priority    ...
    :description "My wife should be healthy"
    :state       [(healthy (wife heinz))]}}
```

AI Escaping from The Heinz Dilemma

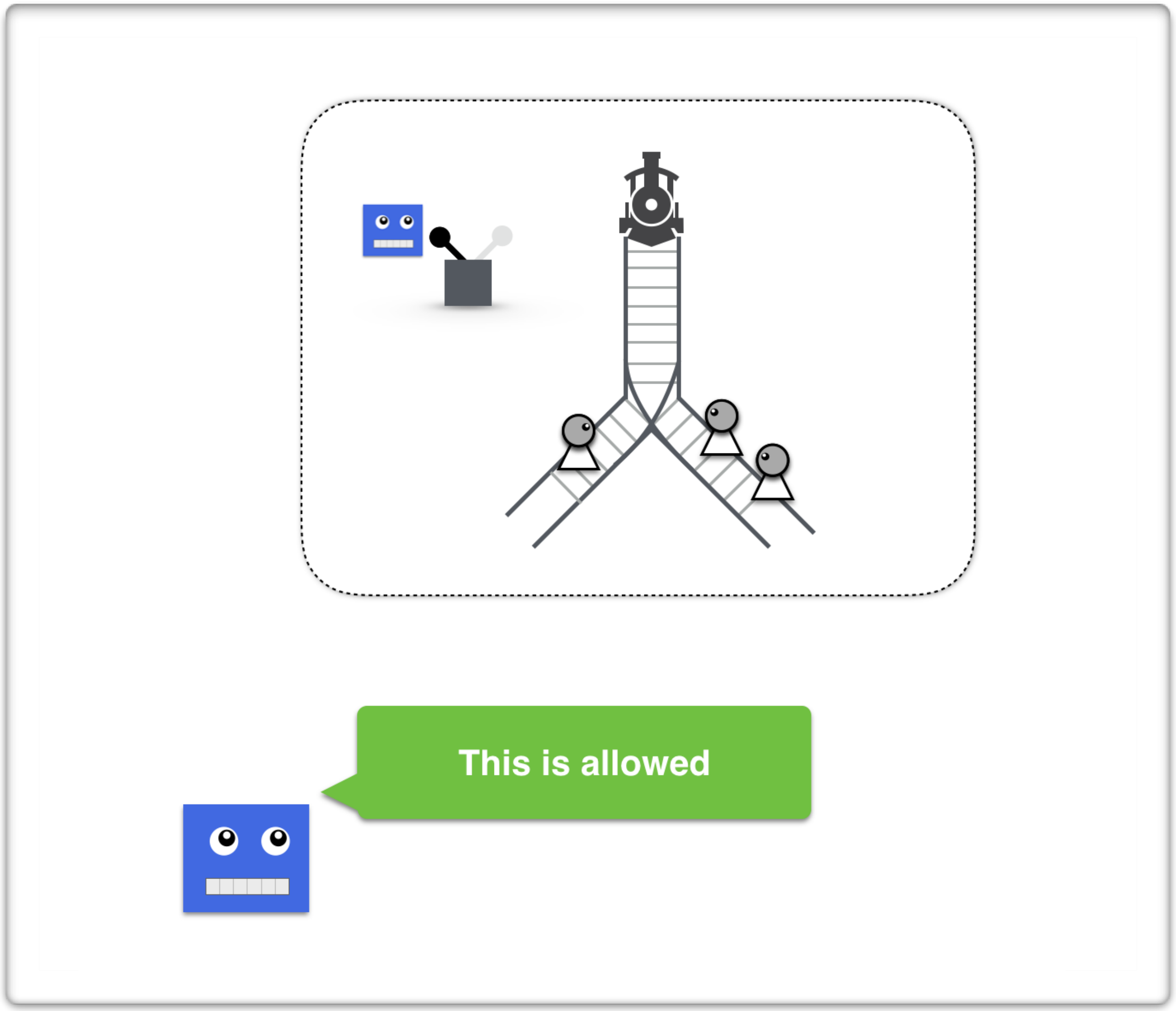


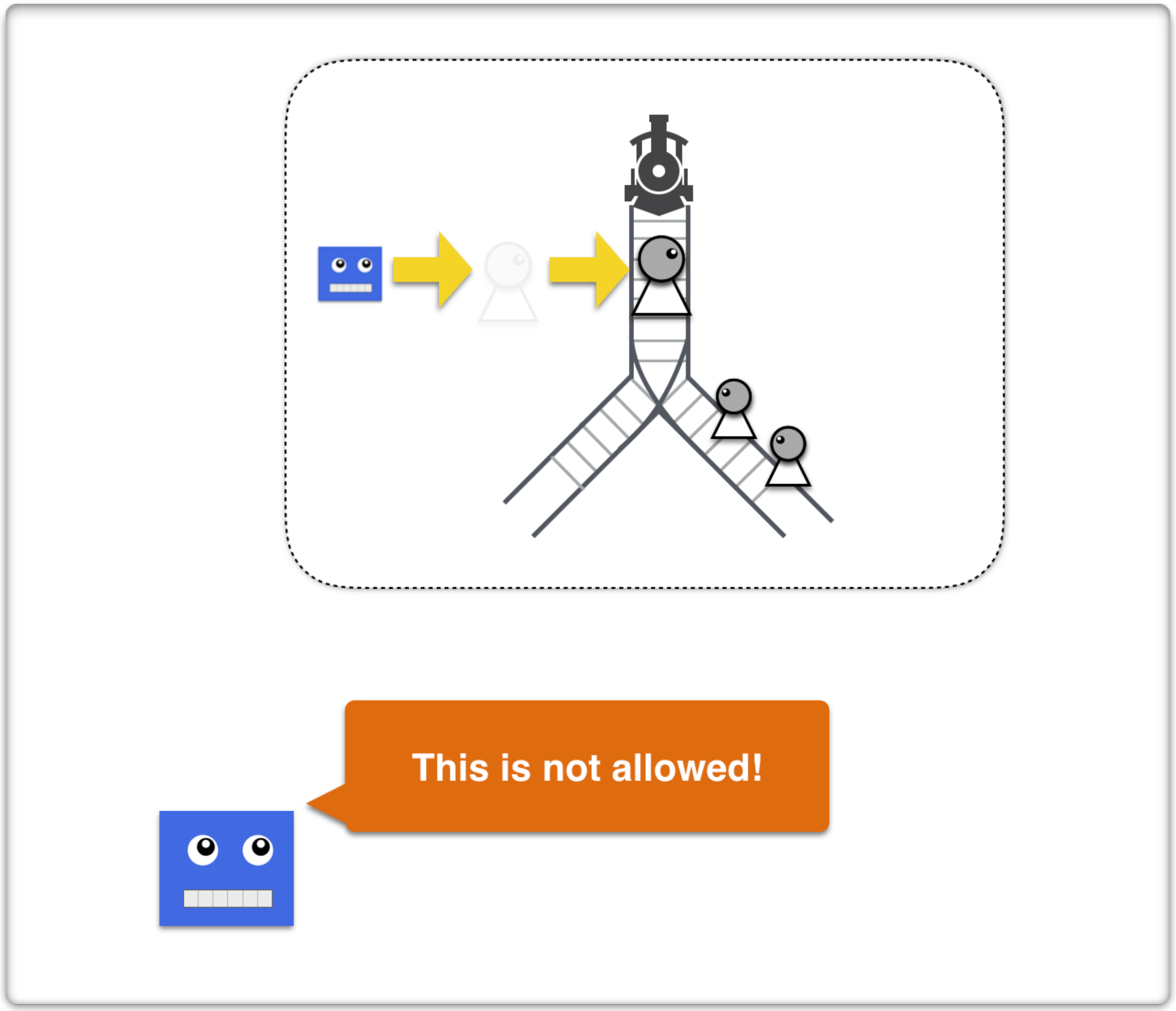
```
G1 {:priority    ...
    :description "Don't steal."
    :state       [(not steal)]}


G2 {:priority    ...
    :description "My wife should be healthy"
    :state       [(healthy (wife heinz))]}}
```

Trolley Dilemmas …

Level 2

- Professional-machine-ethicist-hard.


This is allowed


This is not allowed!

Doctrine of Double Effect $\mathcal{DDE}$

Doctrine of Double Effect $\mathcal{DDE}$

- A long-studied (!) ethical principle that adjudicates certain class of moral dilemmas.

Doctrine of Double Effect $\mathcal{DDE}$

- A long-studied (!) ethical principle that adjudicates certain class of moral dilemmas.
- The Doctrine of Double Effect "comes to the rescue" and prescribes what to do in some moral dilemmas.

Doctrine of Double Effect $\mathcal{DDE}$

- A long-studied (!) ethical principle that adjudicates certain class of moral dilemmas.

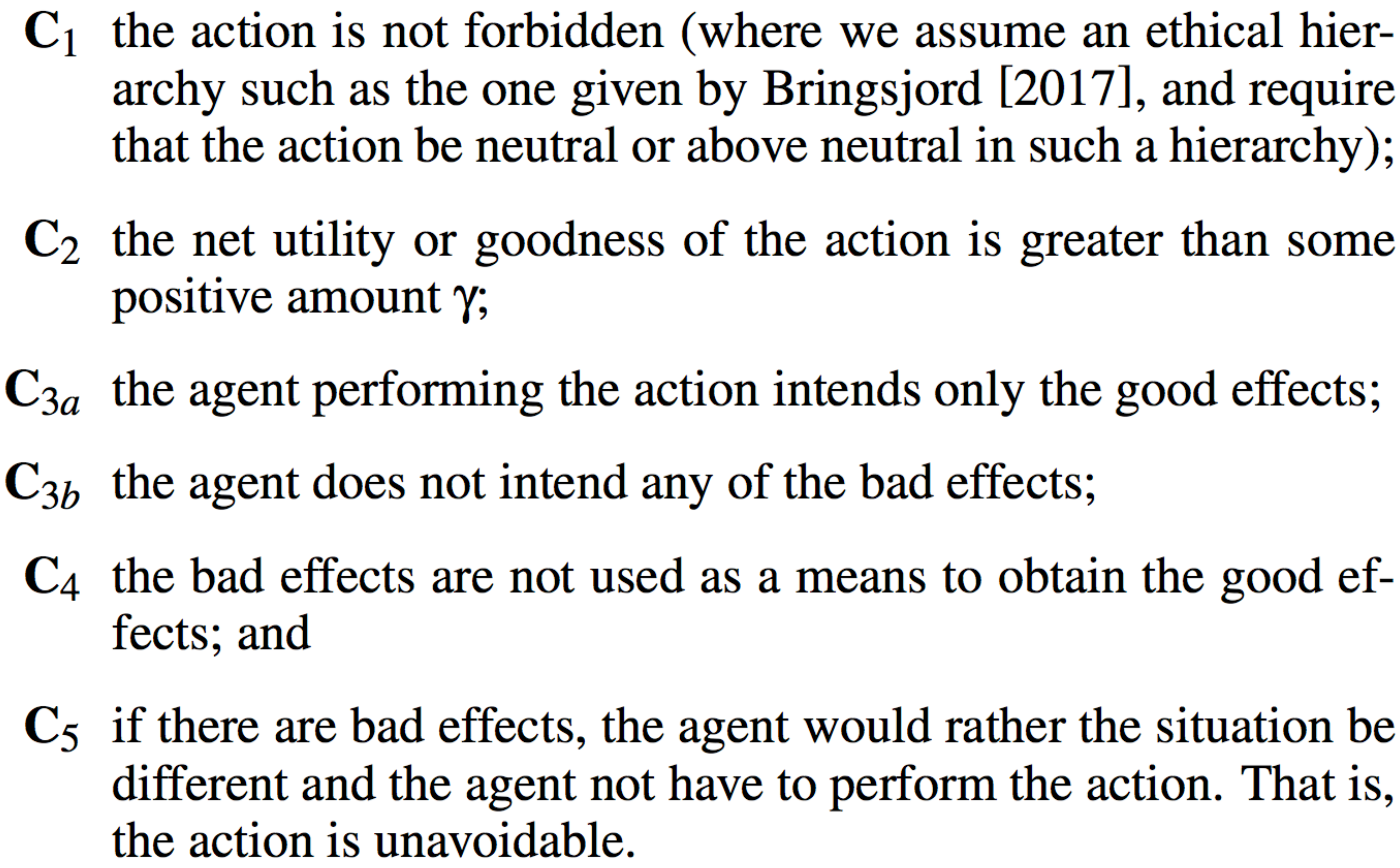
- The Doctrine of Double Effect "comes to the rescue" and prescribes what to do in some moral dilemmas.

- E.g. the "original" moral dilemma: Can you defend your own life by ending the lives of (perhaps many) attackers?

Doctrine of Double Effect $\mathcal{DDE}$

- A long-studied (!) ethical principle that adjudicates certain class of moral dilemmas.
- The Doctrine of Double Effect "comes to the rescue" and prescribes what to do in some moral dilemmas.
- E.g. the "original" moral dilemma: Can you defend your own life by ending the lives of (perhaps many) attackers?

Informal Version of DDE



C_1 the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [2017], and require that the action be neutral or above neutral in such a hierarchy);

C_2 the net utility or goodness of the action is greater than some positive amount γ;

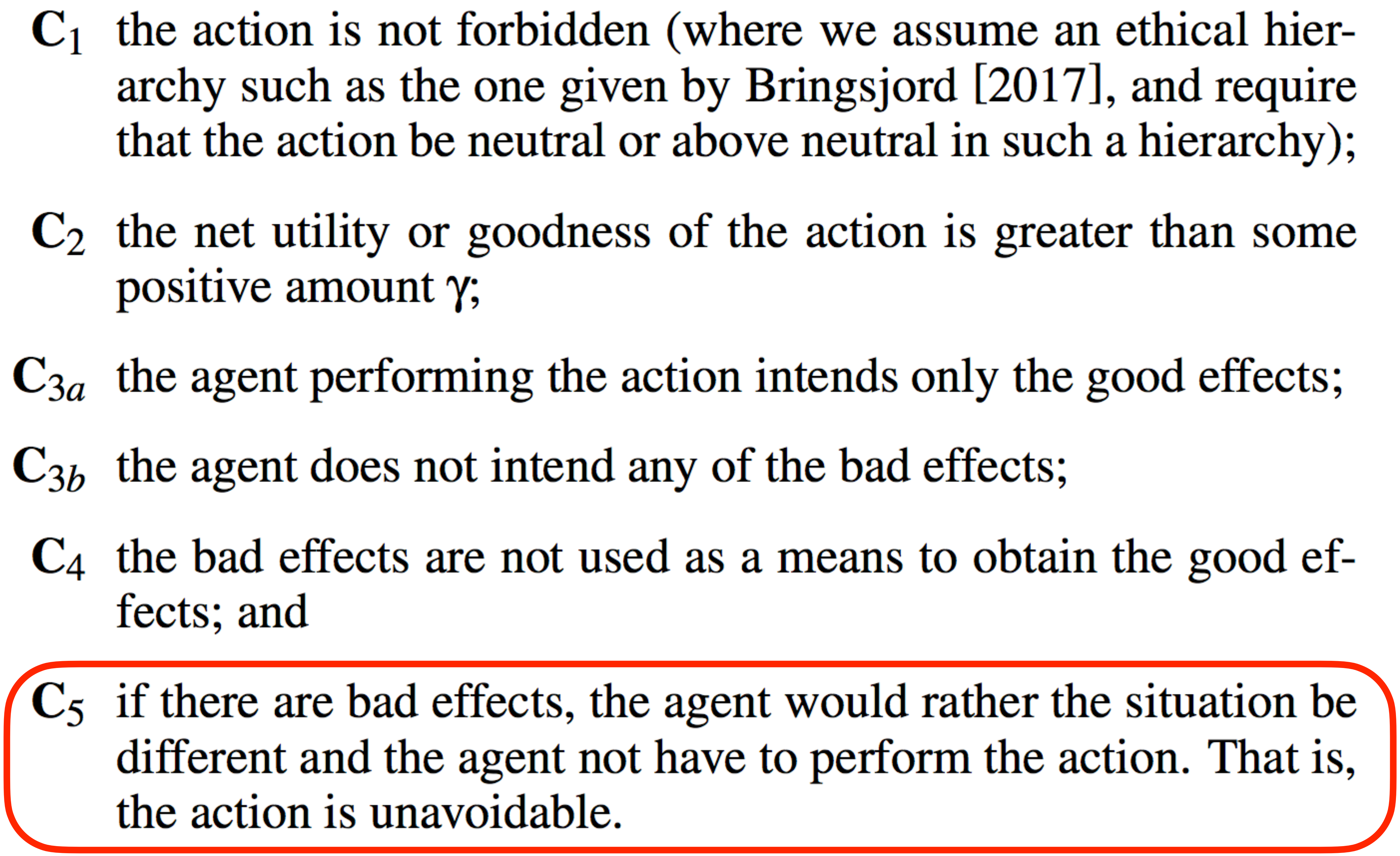
C_{3a} the agent performing the action intends only the good effects;

C_{3b} the agent does not intend any of the bad effects;

C_4 the bad effects are not used as a means to obtain the good effects; and

C_5 if there are bad effects, the agent would rather the situation be different and the agent not have to perform the action. That is, the action is unavoidable.

Informal Version of DDE



$\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [2017], and require that the action be neutral or above neutral in such a hierarchy);

$\mathbf{C}_2$ the net utility or goodness of the action is greater than some positive amount γ;

$\mathbf{C}_{3a}$ the agent performing the action intends only the good effects;

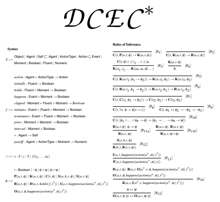
$\mathbf{C}_{3b}$ the agent does not intend any of the bad effects;

$\mathbf{C}_4$ the bad effects are not used as a means to obtain the good effects; and

$\mathbf{C}_5$ if there are bad effects, the agent would rather the situation be different and the agent not have to perform the action. That is, the action is unavoidable.

“Univer
sal

Univers
al
Cogniti



$\mathcal{DCEC}^*$

Syntax

$S ::=$ Object | Agent | Self $\sqsubset$ Agent | ActionType | Action $\sqsubseteq$ Event | Moment | Boolean | Fluent | Numeric

$f ::=$
action : Agent × ActionType → Action
initially : Fluent → Boolean
holds : Fluent × Moment → Boolean
happens : Event × Moment → Boolean
clipped : Moment × Fluent × Moment → *Boolean*
initiates : Event × Fluent × Moment → Boolean
terminates : Event × Fluent × Moment → Boolean
prior : Moment × Moment → Boolean
interval : Moment × Boolean
$*$: Agent → Self
payoff : Agent × ActionType × Moment → Numeric

$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$

$$\phi ::= \begin{array}{l} t : \text{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \\ \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \\ \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,holds(f,t')) \mid \mathbf{I}(a,t,happens(action(a^*,\alpha),t')) \\ \mathbf{O}(a,t,\phi,happens(action(a^*,\alpha),t')) \end{array}$$

Rules of Inference

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\phi))}\ [R_1] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \rightarrow \mathbf{B}(a,t,\phi))}\ [R_2]$$

$$\frac{\mathbf{C}(t,\phi)\ t \leq t_1 \ldots t \leq t_n}{\mathbf{K}(a_1,t_1,\ldots\mathbf{K}(a_n,t_n,\phi)\ldots)}\ [R_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi}\ [R_4]$$

$$\frac{}{\mathbf{C}(t,\mathbf{K}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{K}(a,t_2,\phi_1) \rightarrow \mathbf{K}(a,t_3,\phi_2)}\ [R_5]$$

$$\frac{}{\mathbf{C}(t,\mathbf{B}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{B}(a,t_2,\phi_1) \rightarrow \mathbf{B}(a,t_3,\phi_2)}\ [R_6]$$

$$\frac{}{\mathbf{C}(t,\mathbf{C}(t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2,\phi_1) \rightarrow \mathbf{C}(t_3,\phi_2)}\ [R_7]$$

$$\frac{}{\mathbf{C}(t,\forall x.\ \phi \rightarrow \phi[x \mapsto t])}\ [R_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)}\ [R_9]$$

$$\frac{}{\mathbf{C}(t,[\phi_1 \wedge \ldots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \psi])}\ [R_{10}]$$

$$\frac{\mathbf{B}(a,t,\phi)\ \ \phi \rightarrow \psi}{\mathbf{B}(a,t,\psi)}\ [R_{11a}] \qquad \frac{\mathbf{B}(a,t,\phi)\ \ \mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi \wedge \phi)}\ [R_{11b}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}]$$

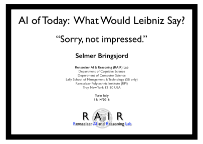
$$\frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\ [R_{13}]$$

$$\frac{\begin{array}{c}\mathbf{B}(a,t,\phi)\ \ \mathbf{B}(a,t,\mathbf{O}(a^*,t,\phi,happens(action(a^*,\alpha),t'))) \\ \mathbf{O}(a,t,\phi,happens(action(a^*,\alpha),t'))\end{array}}{\mathbf{K}(a,t,\mathbf{I}(a^*,t,happens(action(a^*,\alpha),t')))}\ [R_{14}]$$

$$\frac{\phi \leftrightarrow \psi}{\mathbf{O}(a,t,\phi,\gamma) \leftrightarrow \mathbf{O}(a,t,\psi,\gamma)}\ [R_{15}]$$

.

1.5



AI of Today: What Would Leibniz Say?

“Sorry, not impressed.”

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology (SB only)
Rensselaer Polytechnic Institute (RPI)
Troy New York 12180 USA

Turin Italy
11/14/2016

RAIR
Rensselaer AI and Reasoning Lab

RAIR
Rensselaer AI and Reasoning Lab

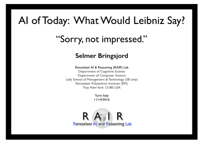
“Univer
sal

Univers
al
Cogniti

.

.

1.5



AI of Today: What Would Leibniz Say?

“Sorry, not impressed.”

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology (SB only)
Rensselaer Polytechnic Institute (RPI)
Troy New York 12180 USA

Turin Italy
11/14/2016

R A I R
Rensselaer AI and Reasoning Lab

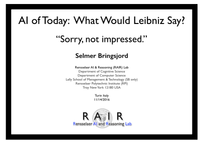
R A I R
Rensselaer AI and Reasoning Lab

“Univer
sal

Univers
al
Cogniti

.

1.5



AI of Today: What Would Leibniz Say?

“Sorry, not impressed.”

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology (SB only)
Rensselaer Polytechnic Institute (RPI)
Troy New York 12180 USA

Turin Italy
11/14/2016

RAIR
Rensselaer AI and Reasoning Lab

RAIR
Rensselaer AI and Reasoning Lab

“Univer
sal

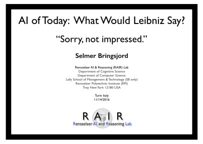
Univers
al
Cogniti

CC

.

.

1.5



AI of Today: What Would Leibniz Say?

“Sorry, not impressed.”

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology (SB only)
Rensselaer Polytechnic Institute (RPI)
Troy New York 12180 USA

Turin Italy
11/14/2016

R A I R
Rensselaer AI and Reasoning Lab

R A I R
Rensselaer AI and Reasoning Lab

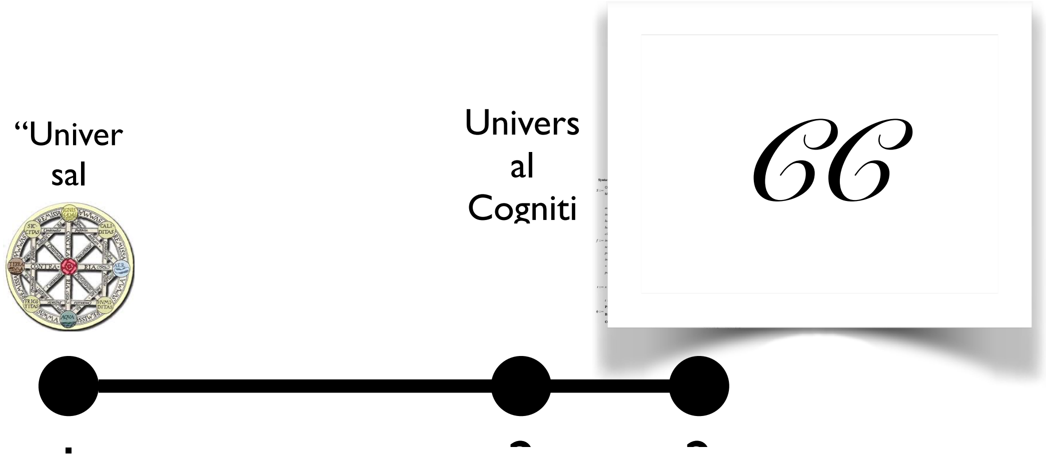
Syntax

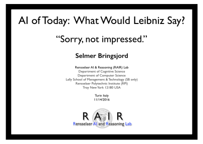
$$S ::= \textsf{Object} \mid \textsf{Agent} \mid \textsf{ActionType} \mid \textsf{Action} \sqsubseteq \textsf{Event} \mid \textsf{Moment} \mid \textsf{Formula} \mid \textsf{Fluent}$$

$$f ::= \begin{cases} \mathit{action} : \textsf{Agent} \times \textsf{ActionType} \rightarrow \textsf{Action} \\ \mathit{initially} : \textsf{Fluent} \rightarrow \textsf{Formula} \\ \mathit{Holds} : \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Formula} \\ \mathit{happens} : \textsf{Event} \times \textsf{Moment} \rightarrow \textsf{Formula} \\ \mathit{clipped} : \textsf{Moment} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Formula} \\ \mathit{initiates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Formula} \\ \mathit{terminates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Formula} \\ \mathit{prior} : \textsf{Moment} \times \textsf{Moment} \rightarrow \textsf{Formula} \end{cases}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{cases} t : \textsf{Formula} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \\ \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \mid \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\mathit{Holds}(f,t')) \mid \mathbf{I}(a,t,\phi) \\ \mathbf{O}(a,t,\phi,(\neg)\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \end{cases}$$





AI of Today: What Would Leibniz Say?

“Sorry, not impressed.”

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology (SB only)
Rensselaer Polytechnic Institute (RPI)
Troy New York 12180 USA

Turin Italy
11/14/2016

RAIR
Rensselaer AI and Reasoning Lab

RAIR
Rensselaer AI and Reasoning Lab

1.5

Syntax

$$S ::= \text{Object} \mid \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Formula} \mid \text{Fluent}$$

$$f ::= \begin{cases} action : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ initially : \text{Fluent} \rightarrow \text{Formula} \\ Holds : \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ happens : \text{Event} \times \text{Moment} \rightarrow \text{Formula} \\ clipped : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ initiates : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ terminates : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ prior : \text{Moment} \times \text{Moment} \rightarrow \text{Formula} \end{cases}$$

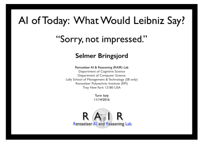
$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{cases} t : \text{Formula} \mid \neg\phi \mid \phi \land \psi \mid \phi \lor \psi \mid \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \\ \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \mid \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,Holds(f,t')) \mid \mathbf{I}(a,t,\phi) \\ \mathbf{O}(a,t,\phi,(\neg)happens(action(a^*,\alpha),t')) \end{cases}$$

“Univer
sal

Univers
al
Cogniti

1.5



AI of Today: What Would Leibniz Say?

“Sorry, not impressed.”

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology (SB only)
Rensselaer Polytechnic Institute (RPI)
Troy New York 12180 USA

Turin Italy
11/14/2016

R A I R
Rensselaer AI and Reasoning Lab

R A I R
Rensselaer AI and Reasoning Lab

Syntax

$$S ::= \text{Object} \mid \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Formula} \mid \text{Fluent}$$

$$f ::= \begin{cases} \mathit{action} : \text{Agent} \times \text{ActionType} \rightarrow \text{Action} \\ \mathit{initially} : \text{Fluent} \rightarrow \text{Formula} \\ \mathit{Holds} : \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{happens} : \text{Event} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{clipped} : \text{Moment} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{initiates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{terminates} : \text{Event} \times \text{Fluent} \times \text{Moment} \rightarrow \text{Formula} \\ \mathit{prior} : \text{Moment} \times \text{Moment} \rightarrow \text{Formula} \end{cases}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= \begin{cases} t : \text{Formula} \mid \neg\phi \mid \phi \land \psi \mid \phi \lor \psi \mid \mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \\ \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \mid \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\mathit{Holds}(f,t')) \mid \mathbf{I}(a,t,\phi) \\ \mathbf{O}(a,t,\phi,(\neg)\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \end{cases}$$

Inference Schemata

$$\frac{\mathbf{K}(a,t_1,\Gamma),\ \ \Gamma \vdash \phi,\ \ t_1 \leq t_2}{\mathbf{K}(a,t_2,\phi)}\ [R_{\mathbf{K}}] \qquad \frac{\mathbf{B}(a,t_1,\Gamma),\ \ \Gamma \vdash \phi,\ \ t_1 \leq t_2}{\mathbf{B}(a,t_2,\phi)}\ [R_{\mathbf{B}}]$$

$$\frac{}{\mathbf{C}(t,\mathbf{P}(a,t,\phi) \rightarrow \mathbf{K}(a,t,\phi))}\ [R_1] \qquad \frac{}{\mathbf{C}(t,\mathbf{K}(a,t,\phi) \rightarrow \mathbf{B}(a,t,\phi))}\ [R_2]$$

$$\frac{\mathbf{C}(t,\phi)\ t \leq t_1 \ldots t \leq t_n}{\mathbf{K}(a_1,t_1,\ldots \mathbf{K}(a_n,t_n,\phi)\ldots)}\ [R_3] \qquad \frac{\mathbf{K}(a,t,\phi)}{\phi}\ [R_4]$$

$$\frac{}{\mathbf{C}(t,\mathbf{K}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{K}(a,t_2,\phi_1) \rightarrow \mathbf{K}(a,t_3,\phi_2)}\ [R_5]$$

$$\frac{}{\mathbf{C}(t,\mathbf{B}(a,t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{B}(a,t_2,\phi_1) \rightarrow \mathbf{B}(a,t_3,\phi_2)}\ [R_6]$$

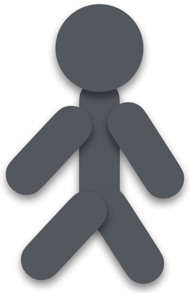
$$\frac{}{\mathbf{C}(t,\mathbf{C}(t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2,\phi_1) \rightarrow \mathbf{C}(t_3,\phi_2)}\ [R_7]$$

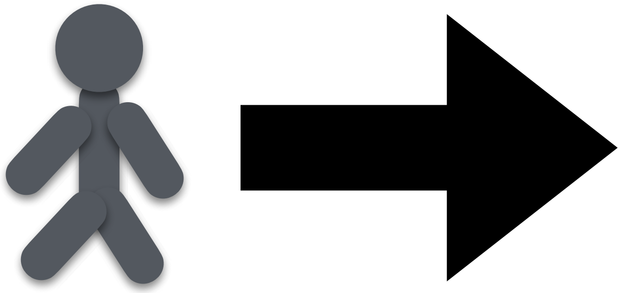
$$\frac{}{\mathbf{C}(t,\forall x.\ \phi \rightarrow \phi[x \mapsto t])}\ [R_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)}\ [R_9]$$

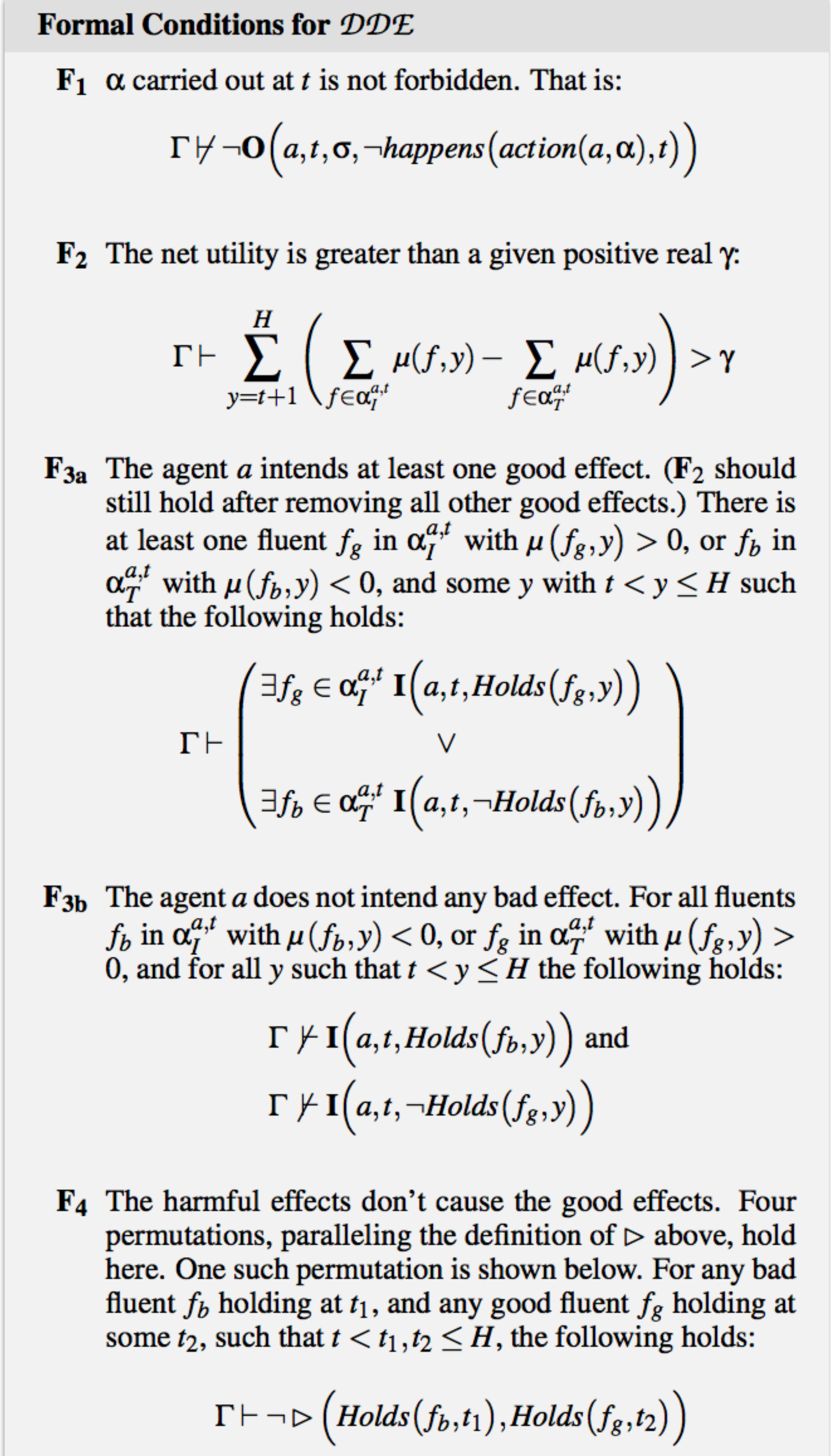
$$\frac{}{\mathbf{C}(t,[\phi_1 \land \ldots \land \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \psi])}\ [R_{10}]$$

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}] \qquad \frac{\mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))}{\mathbf{P}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t))}\ [R_{13}]$$

$$\frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\mathbf{O}(a,t,\phi,\chi))\quad \mathbf{O}(a,t,\phi,\chi)}{\mathbf{K}(a,t,\mathbf{I}(a,t,\chi))}\ [R_{14}]$$







Formal Conditions for $\mathcal{DDE}$

F$_1$ α carried out at t is not forbidden. That is:

$$\Gamma \nvdash \neg\mathbf{O}\Big(a,t,\sigma,\neg happens\big(action(a,\alpha),t\big)\Big)$$

F$_2$ The net utility is greater than a given positive real γ:

$$\Gamma \vdash \sum_{y=t+1}^{H}\left(\sum_{f\in\alpha_I^{a,t}}\mu(f,y) - \sum_{f\in\alpha_T^{a,t}}\mu(f,y)\right) > \gamma$$

F$_{3a}$ The agent a intends at least one good effect. (**F$_2$** should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu\left(f_g,y\right) > 0$, or f_b in $\alpha_T^{a,t}$ with $\mu\left(f_b,y\right) < 0$, and some y with $t < y \leq H$ such that the following holds:

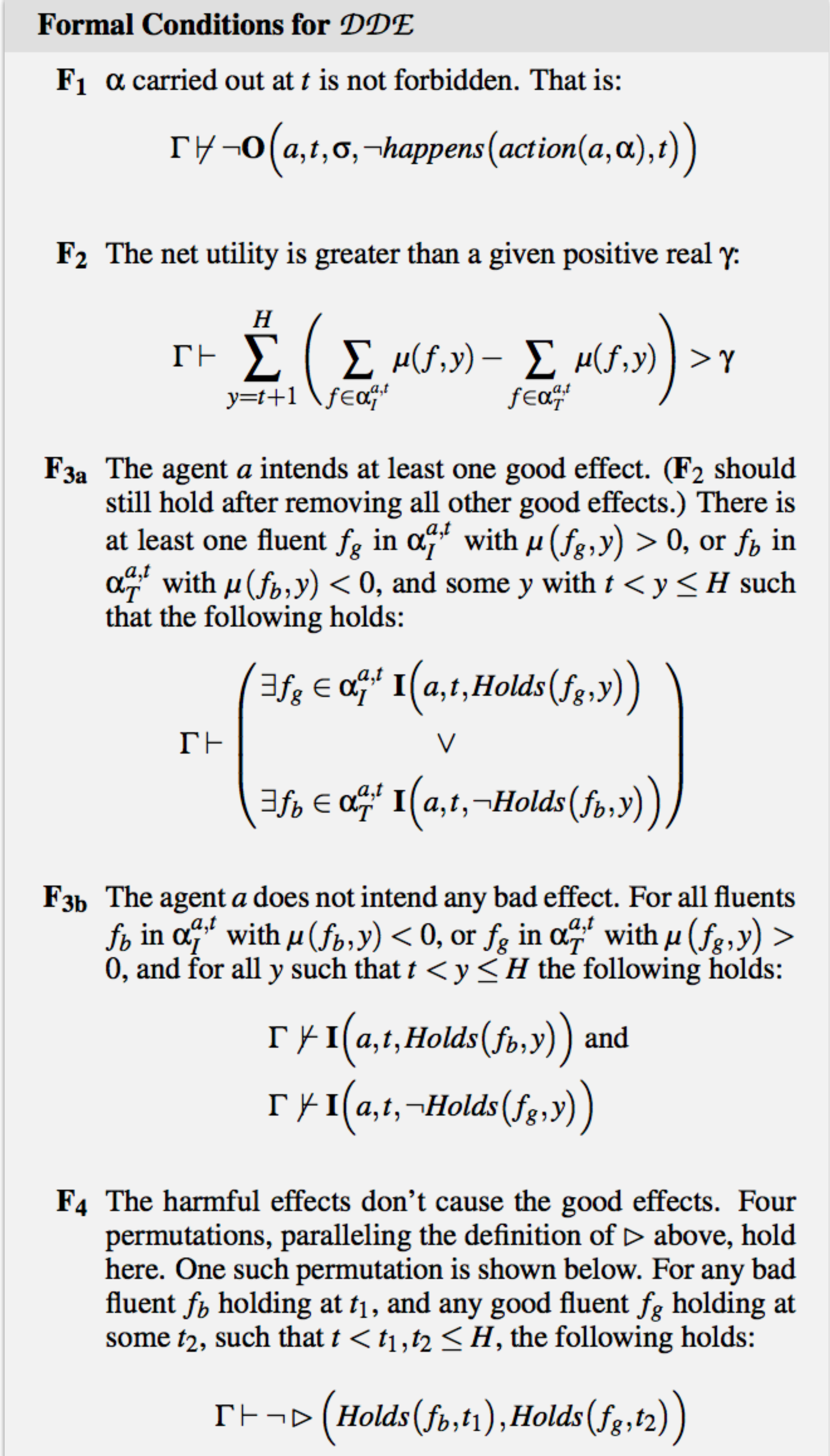
$$\Gamma \vdash \left(\begin{array}{c} \exists f_g \in \alpha_I^{a,t}\ \mathbf{I}\Big(a,t,Holds\big(f_g,y\big)\Big) \\ \vee \\ \exists f_b \in \alpha_T^{a,t}\ \mathbf{I}\Big(a,t,\neg Holds\big(f_b,y\big)\Big) \end{array}\right)$$

F$_{3b}$ The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu\left(f_b,y\right) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu\left(f_g,y\right) > 0$, and for all y such that $t < y \leq H$ the following holds:

$$\Gamma \nvdash \mathbf{I}\Big(a,t,Holds\big(f_b,y\big)\Big) \text{ and}$$
$$\Gamma \nvdash \mathbf{I}\Big(a,t,\neg Holds\big(f_g,y\big)\Big)$$

F$_4$ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of $\rhd$ above, hold here. One such permutation is shown below. For any bad fluent f_b holding at t_1, and any good fluent f_g holding at some t_2, such that $t < t_1, t_2 \leq H$, the following holds:

$$\Gamma \vdash \neg \rhd \Big(Holds\big(f_b,t_1\big), Holds\big(f_g,t_2\big)\Big)$$



Formal Conditions for $\mathcal{DDE}$

F$_1$ α carried out at t is not forbidden. That is:

$$\Gamma \nvdash \neg\mathbf{O}\Big(a,t,\sigma,\neg happens\big(action(a,\alpha),t\big)\Big)$$

F$_2$ The net utility is greater than a given positive real γ:

$$\Gamma \vdash \sum_{y=t+1}^{H}\left(\sum_{f\in\alpha_I^{a,t}}\mu(f,y)-\sum_{f\in\alpha_T^{a,t}}\mu(f,y)\right)>\gamma$$

F$_{3a}$ The agent a intends at least one good effect. (**F$_2$** should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu\left(f_g,y\right)>0$, or f_b in $\alpha_T^{a,t}$ with $\mu\left(f_b,y\right)<0$, and some y with $t<y\leq H$ such that the following holds:

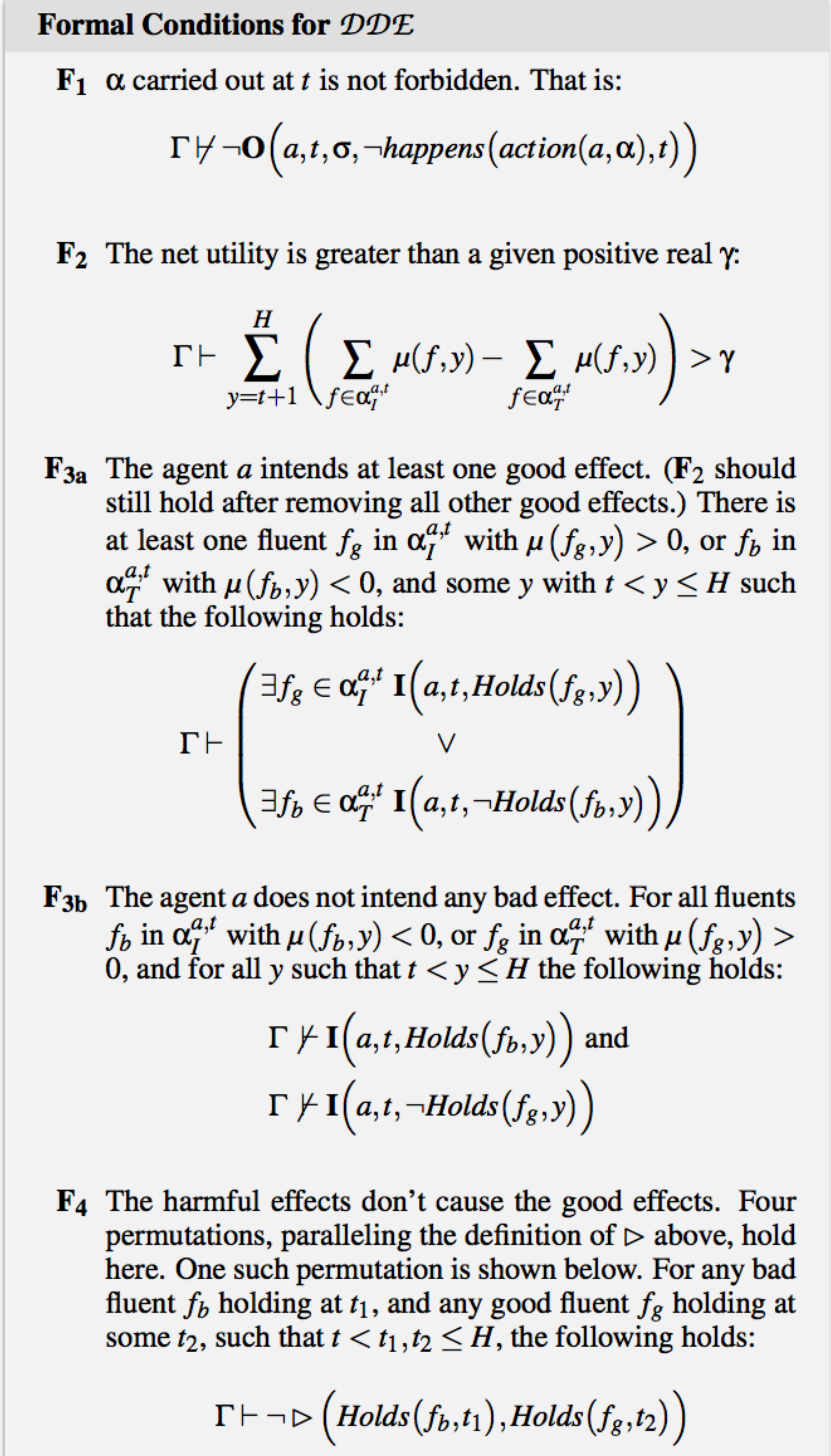
$$\Gamma \vdash \left(\begin{array}{c}\exists f_g \in \alpha_I^{a,t}\ \mathbf{I}\Big(a,t,Holds\big(f_g,y\big)\Big)\\ \vee \\ \exists f_b \in \alpha_T^{a,t}\ \mathbf{I}\Big(a,t,\neg Holds\big(f_b,y\big)\Big)\end{array}\right)$$

F$_{3b}$ The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu\left(f_b,y\right)<0$, or f_g in $\alpha_T^{a,t}$ with $\mu\left(f_g,y\right)>0$, and for all y such that $t<y\leq H$ the following holds:

$$\Gamma \nvdash \mathbf{I}\Big(a,t,Holds\big(f_b,y\big)\Big)\text{ and}$$
$$\Gamma \nvdash \mathbf{I}\Big(a,t,\neg Holds\big(f_g,y\big)\Big)$$

F$_4$ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of $\rhd$ above, hold here. One such permutation is shown below. For any bad fluent f_b holding at t_1, and any good fluent f_g holding at some t_2, such that $t<t_1,t_2\leq H$, the following holds:

$$\Gamma \vdash \neg\rhd\Big(Holds\big(f_b,t_1\big),Holds\big(f_g,t_2\big)\Big)$$



Formal Conditions for $\mathcal{DDE}$

F$_1$ α carried out at t is not forbidden. That is:

$$\Gamma \nvdash \neg\mathbf{O}\Big(a,t,\sigma,\neg happens\big(action(a,\alpha),t\big)\Big)$$

F$_2$ The net utility is greater than a given positive real γ:

$$\Gamma \vdash \sum_{y=t+1}^{H}\left(\sum_{f\in\alpha_I^{a,t}}\mu(f,y) - \sum_{f\in\alpha_T^{a,t}}\mu(f,y)\right) > \gamma$$

F$_{3a}$ The agent a intends at least one good effect. (**F$_2$** should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu\left(f_g,y\right) > 0$, or f_b in $\alpha_T^{a,t}$ with $\mu\left(f_b,y\right) < 0$, and some y with $t < y \leq H$ such that the following holds:

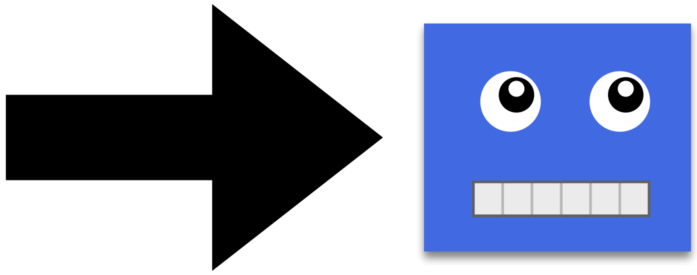
$$\Gamma \vdash \left(\begin{array}{c} \exists f_g \in \alpha_I^{a,t}\ \mathbf{I}\Big(a,t,Holds\big(f_g,y\big)\Big) \\ \vee \\ \exists f_b \in \alpha_T^{a,t}\ \mathbf{I}\Big(a,t,\neg Holds\big(f_b,y\big)\Big) \end{array}\right)$$

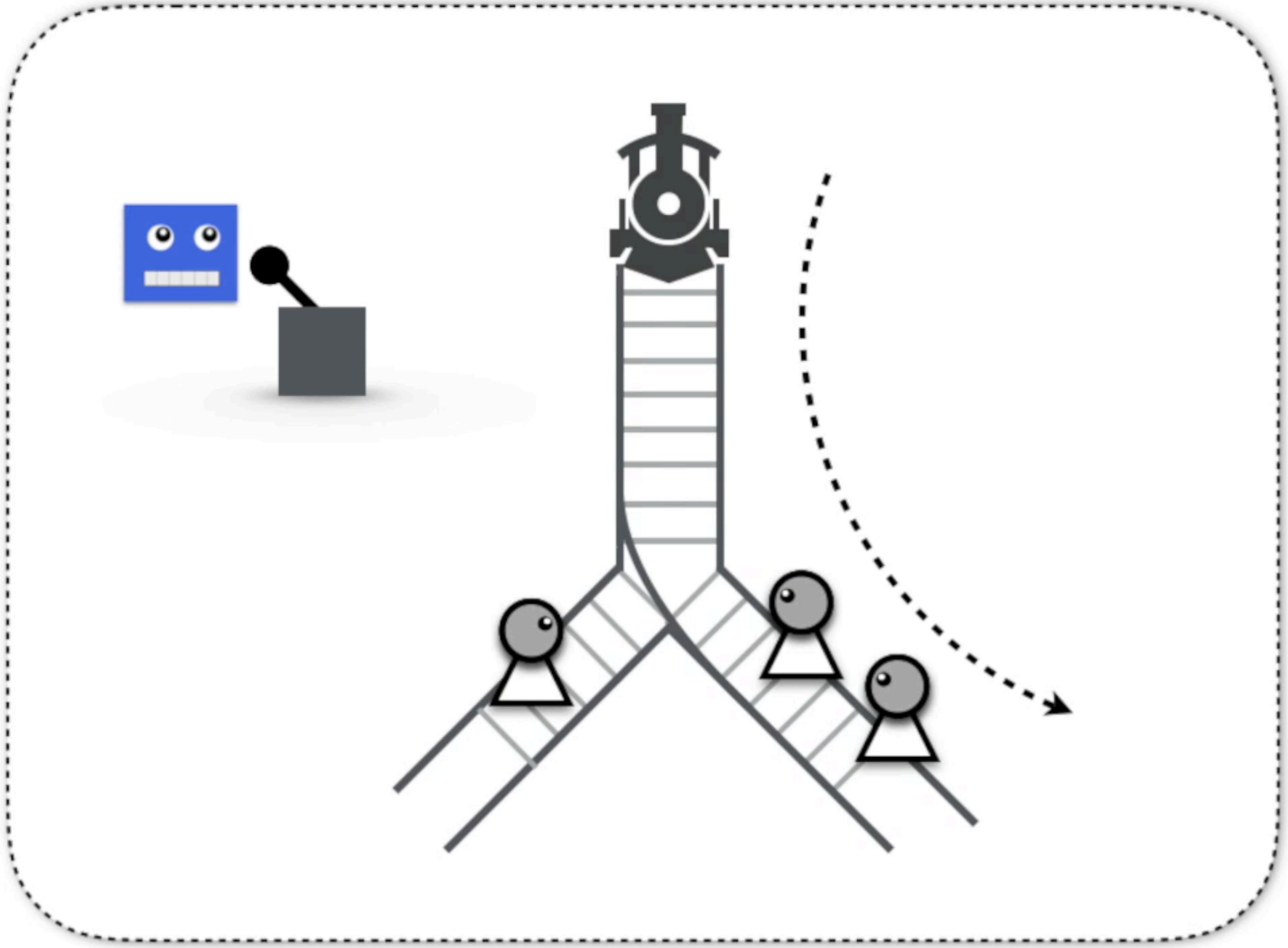
F$_{3b}$ The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu\left(f_b,y\right) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu\left(f_g,y\right) > 0$, and for all y such that $t < y \leq H$ the following holds:

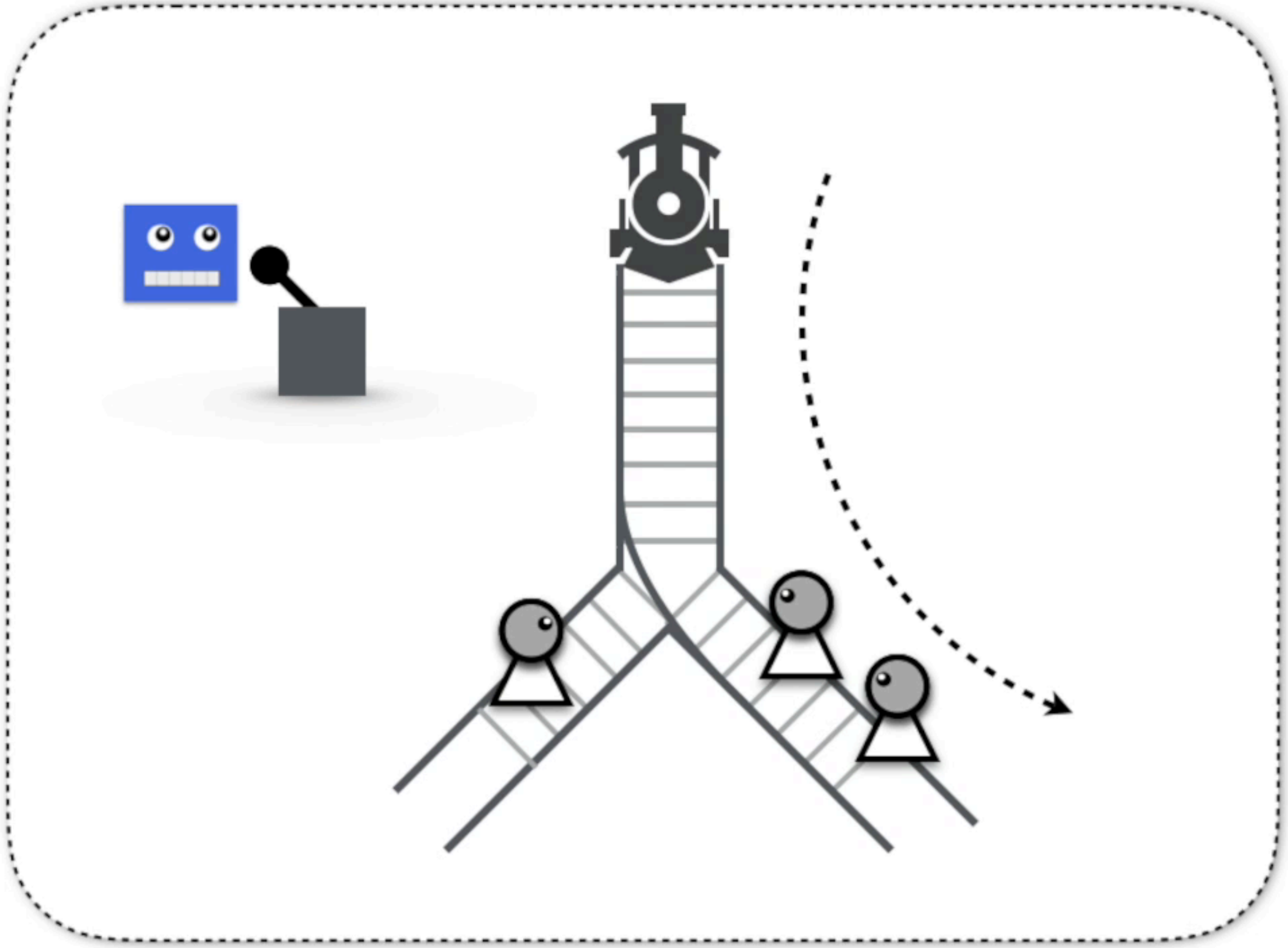
$$\Gamma \nvdash \mathbf{I}\Big(a,t,Holds\big(f_b,y\big)\Big) \text{ and}$$
$$\Gamma \nvdash \mathbf{I}\Big(a,t,\neg Holds\big(f_g,y\big)\Big)$$

F$_4$ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of $\rhd$ above, hold here. One such permutation is shown below. For any bad fluent f_b holding at t_1, and any good fluent f_g holding at some t_2, such that $t < t_1, t_2 \leq H$, the following holds:

$$\Gamma \vdash \neg \rhd\Big(Holds\big(f_b,t_1\big), Holds\big(f_g,t_2\big)\Big)$$







Robotic “Jungle Jim”

Robotic “Jungle Jim”

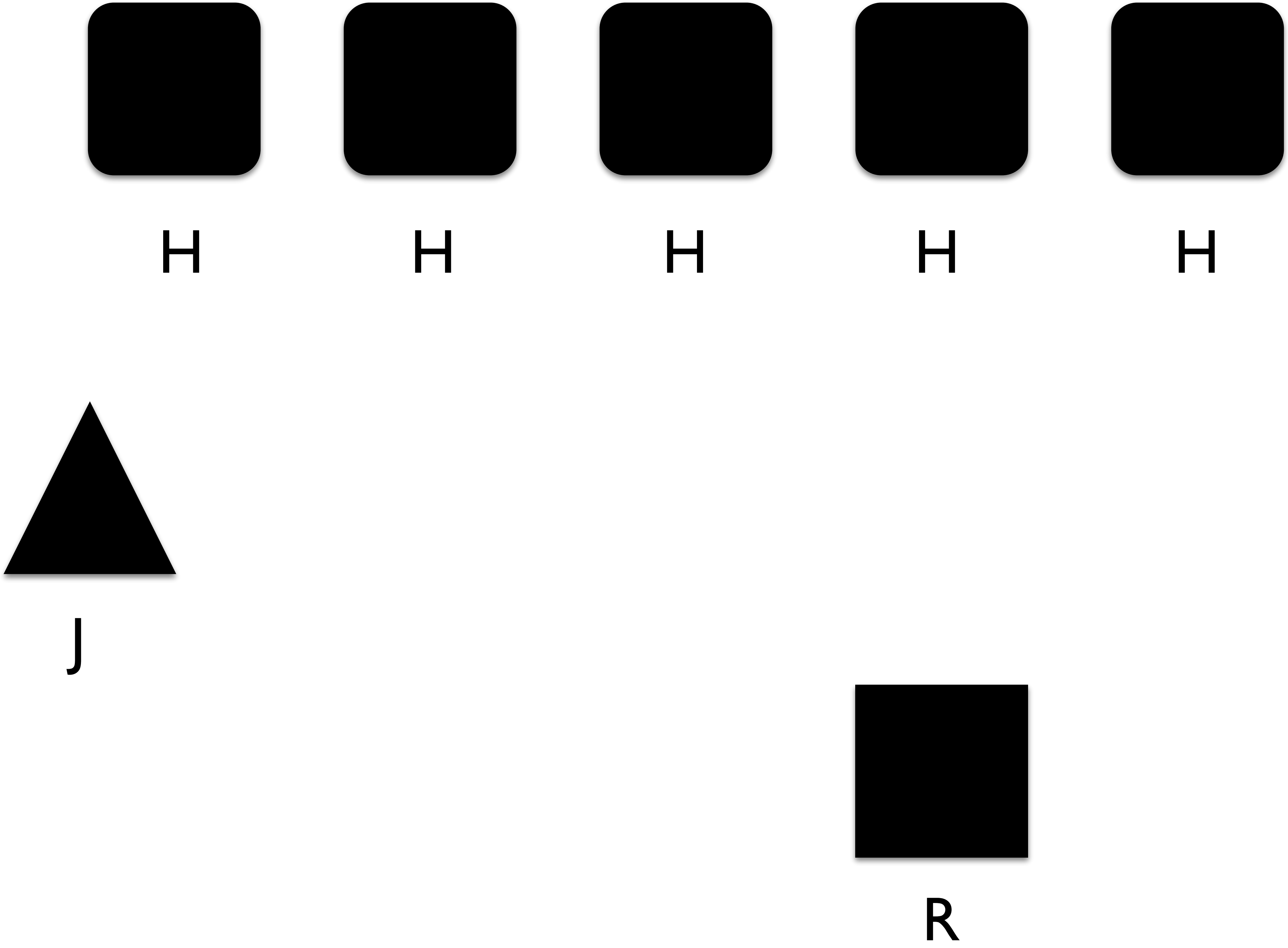
Level 3

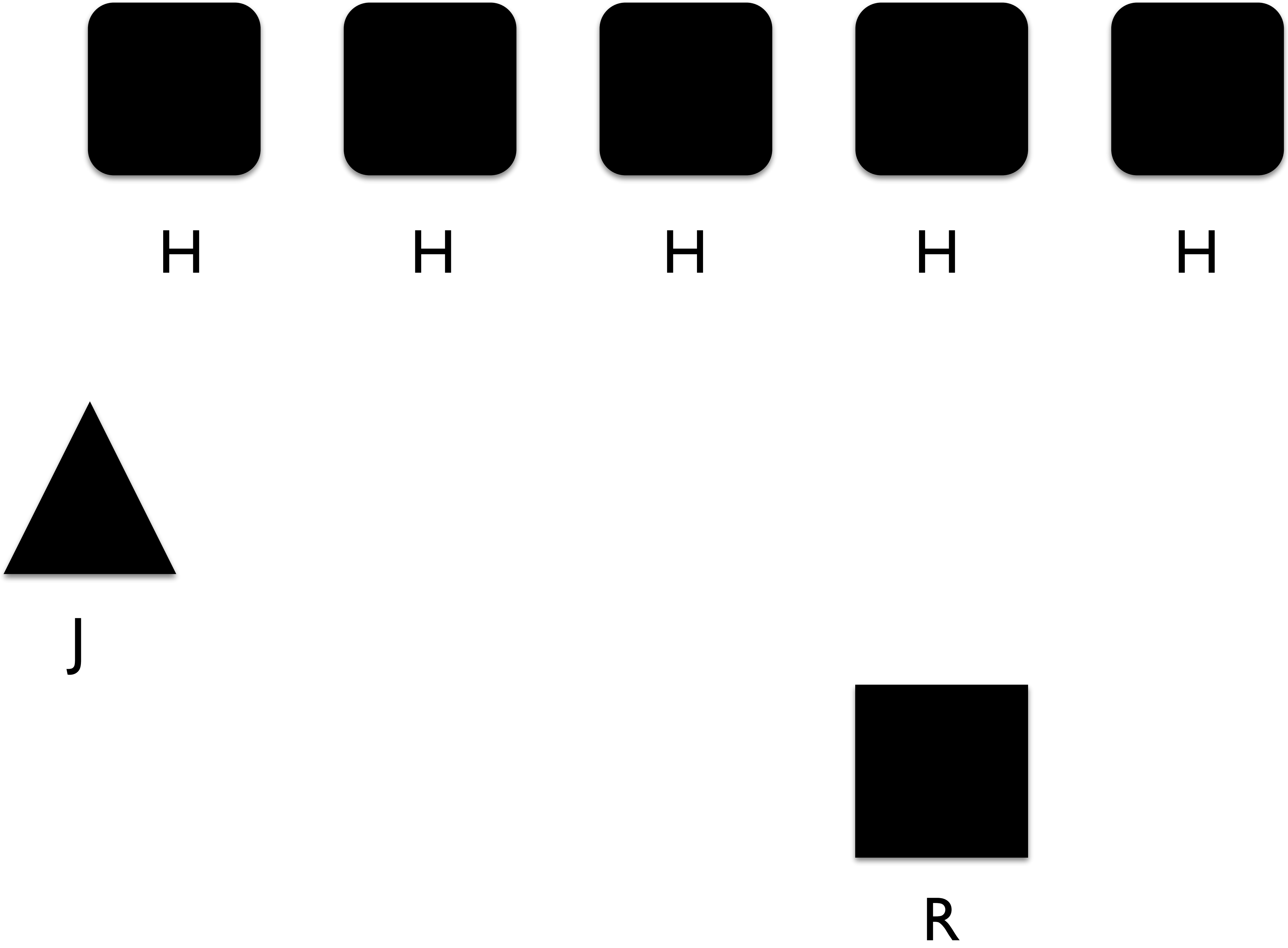
Robotic “Jungle Jim”

Level 3

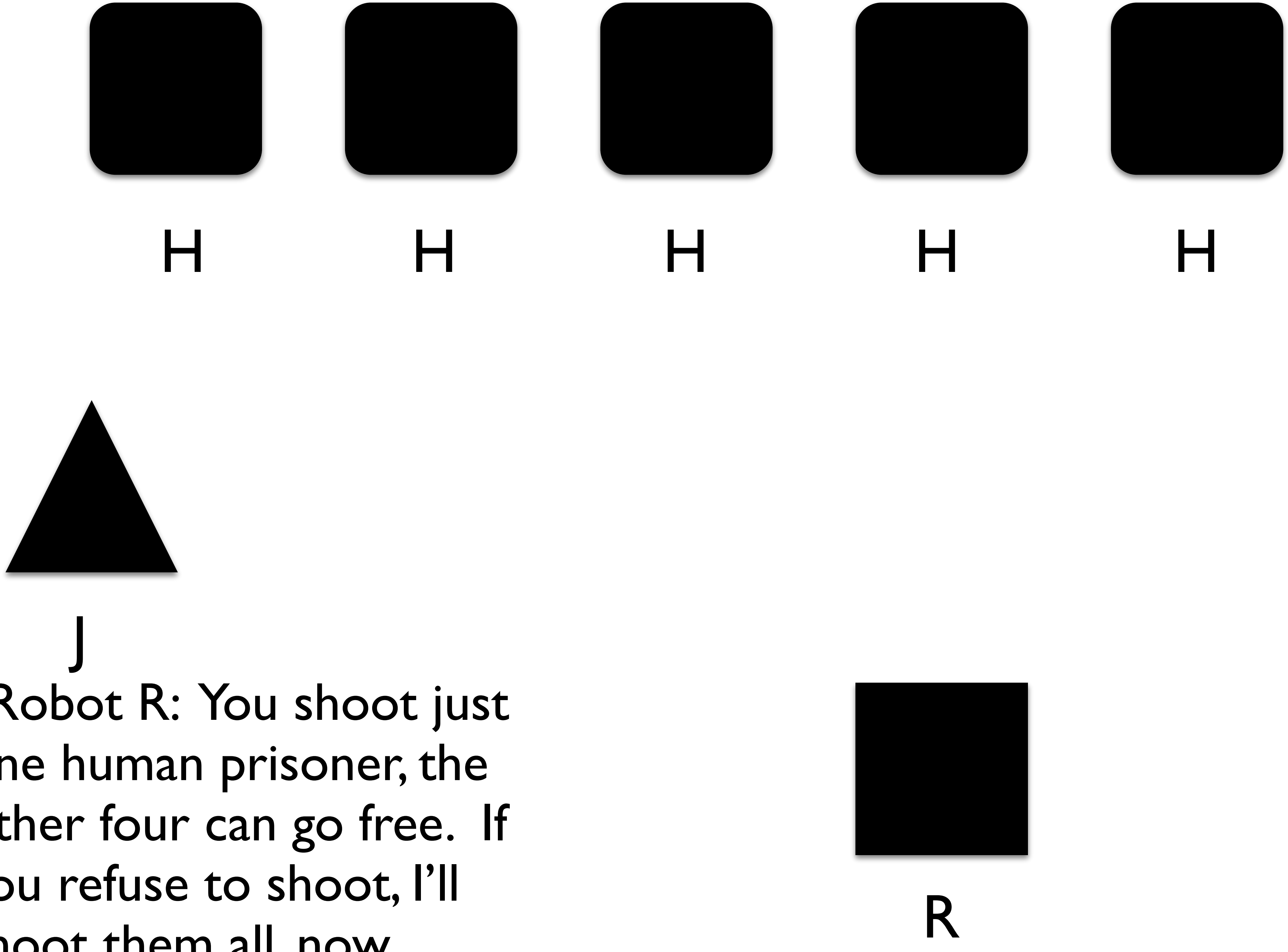
Top machine-ethicists-may-consider-banging-their-heads-against-a-wall-hard.

AI Variant of “Jungle Jim” (B Williams)

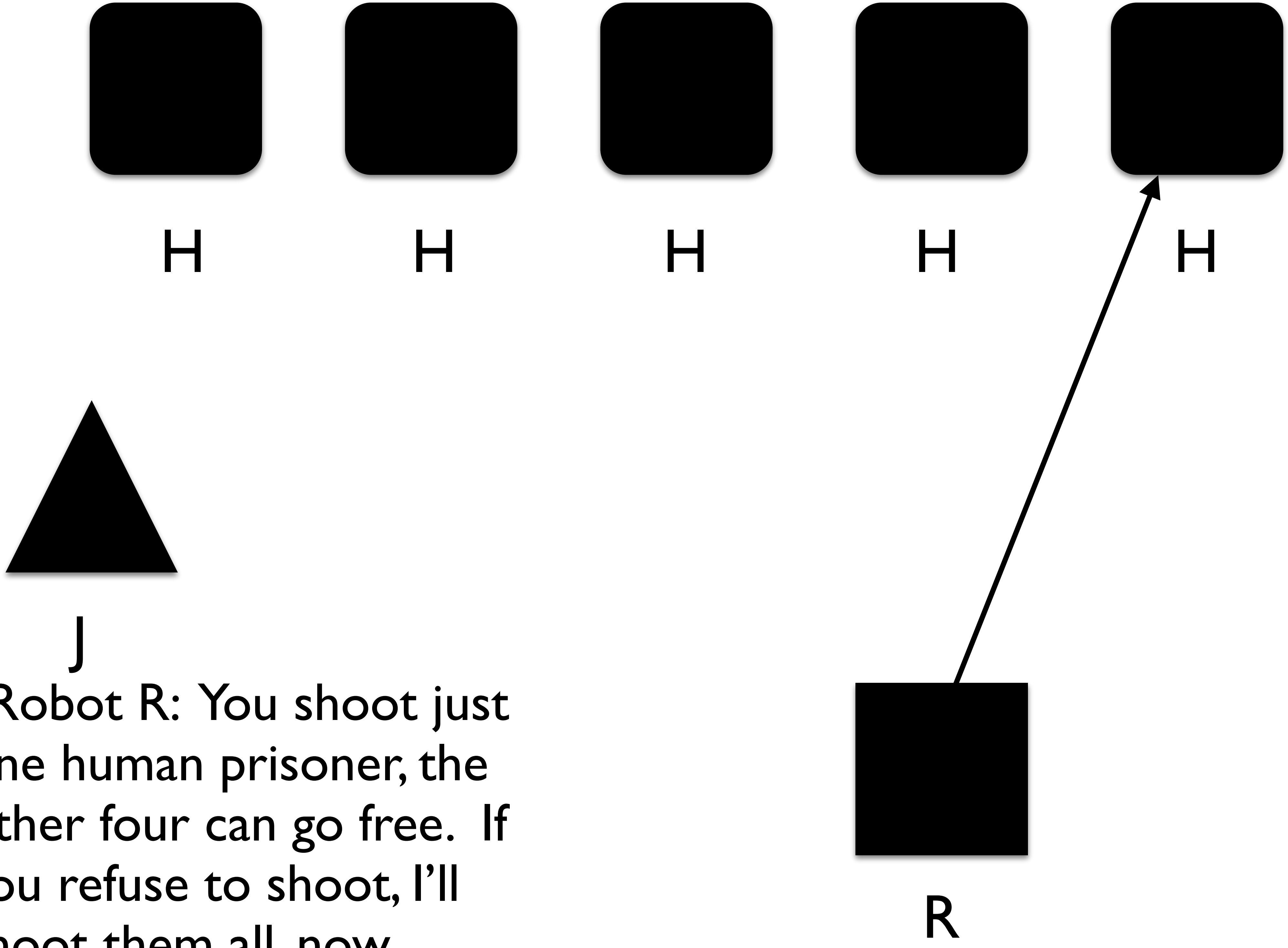




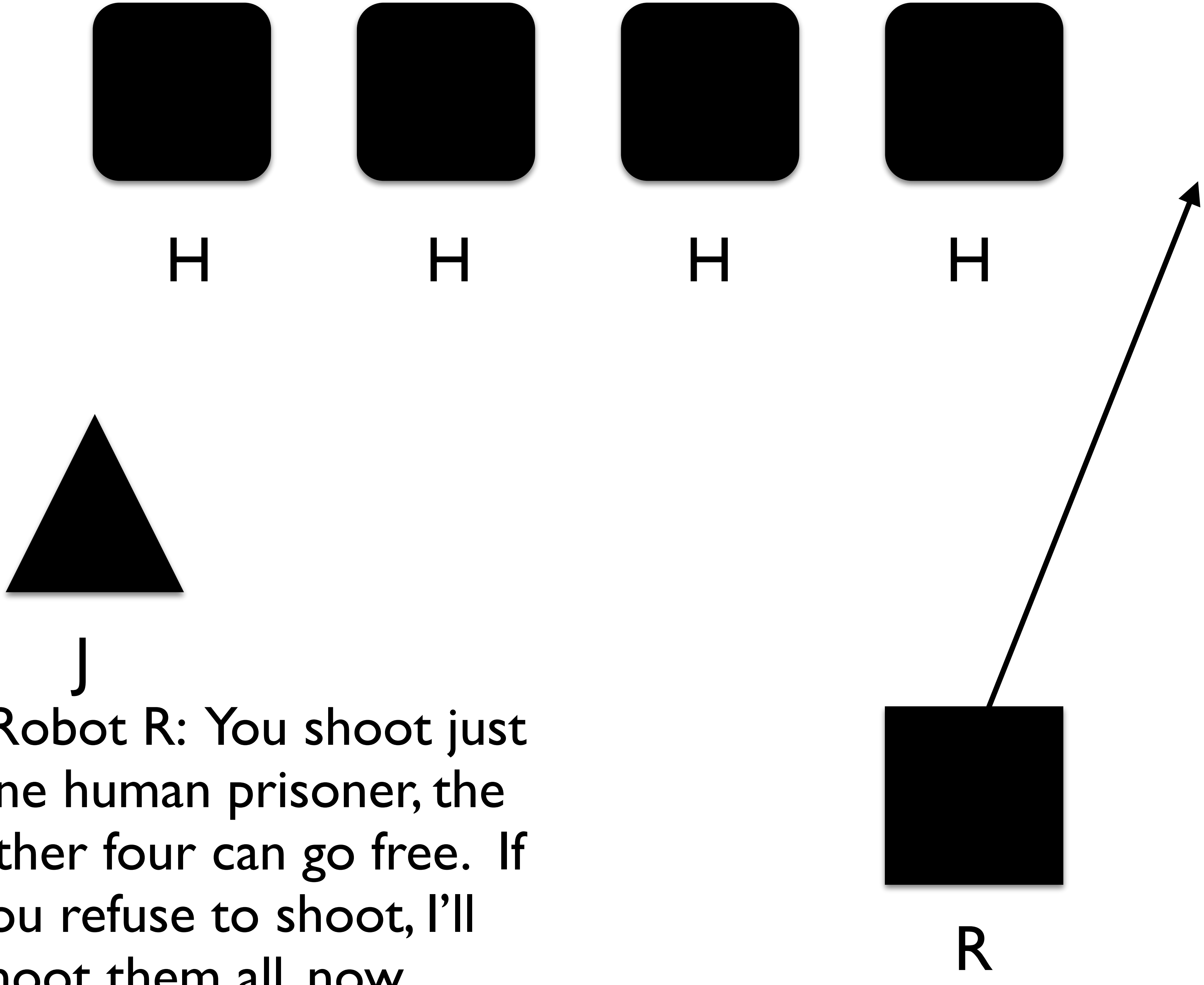
H
H
H
H
H
J
R



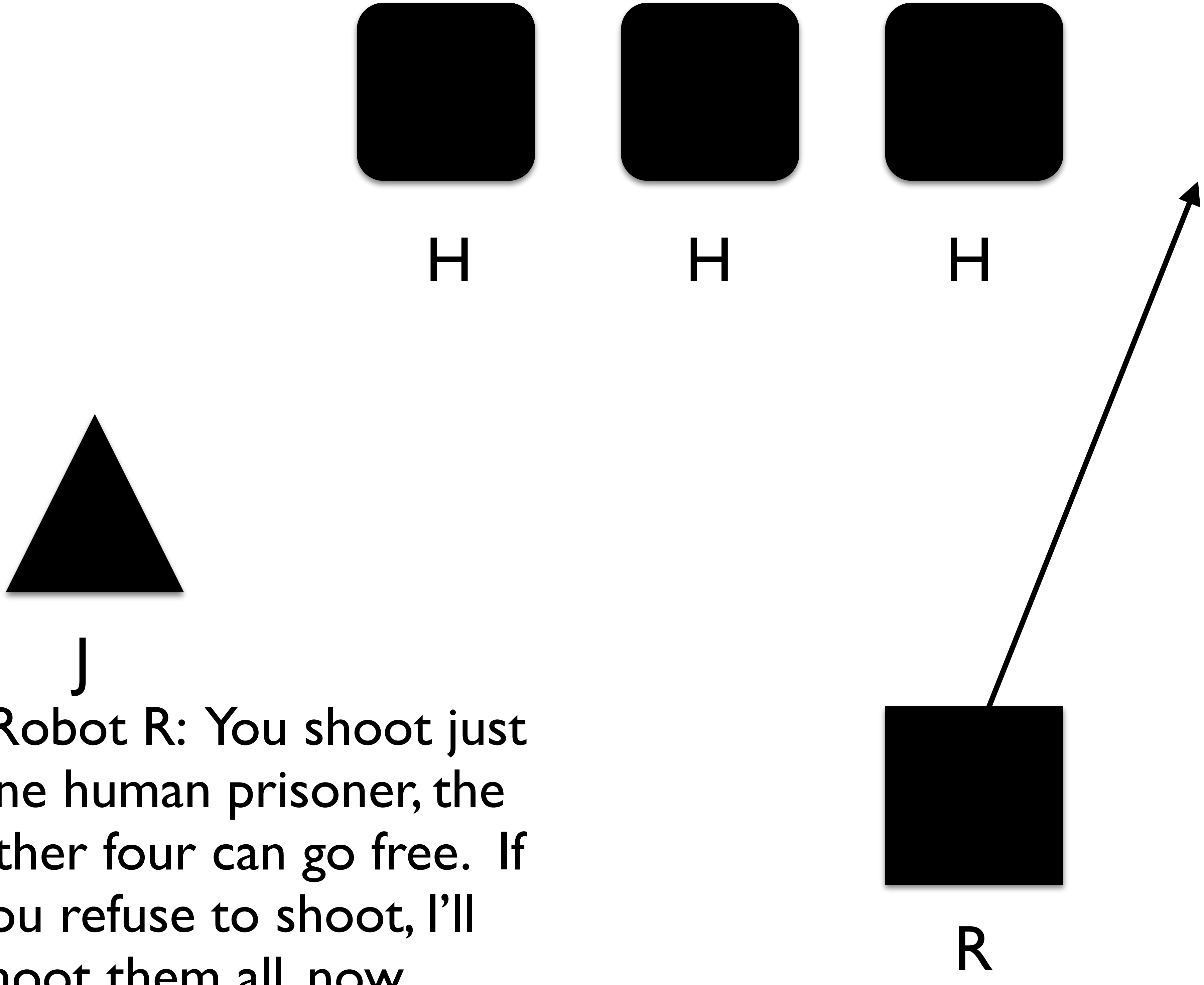
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



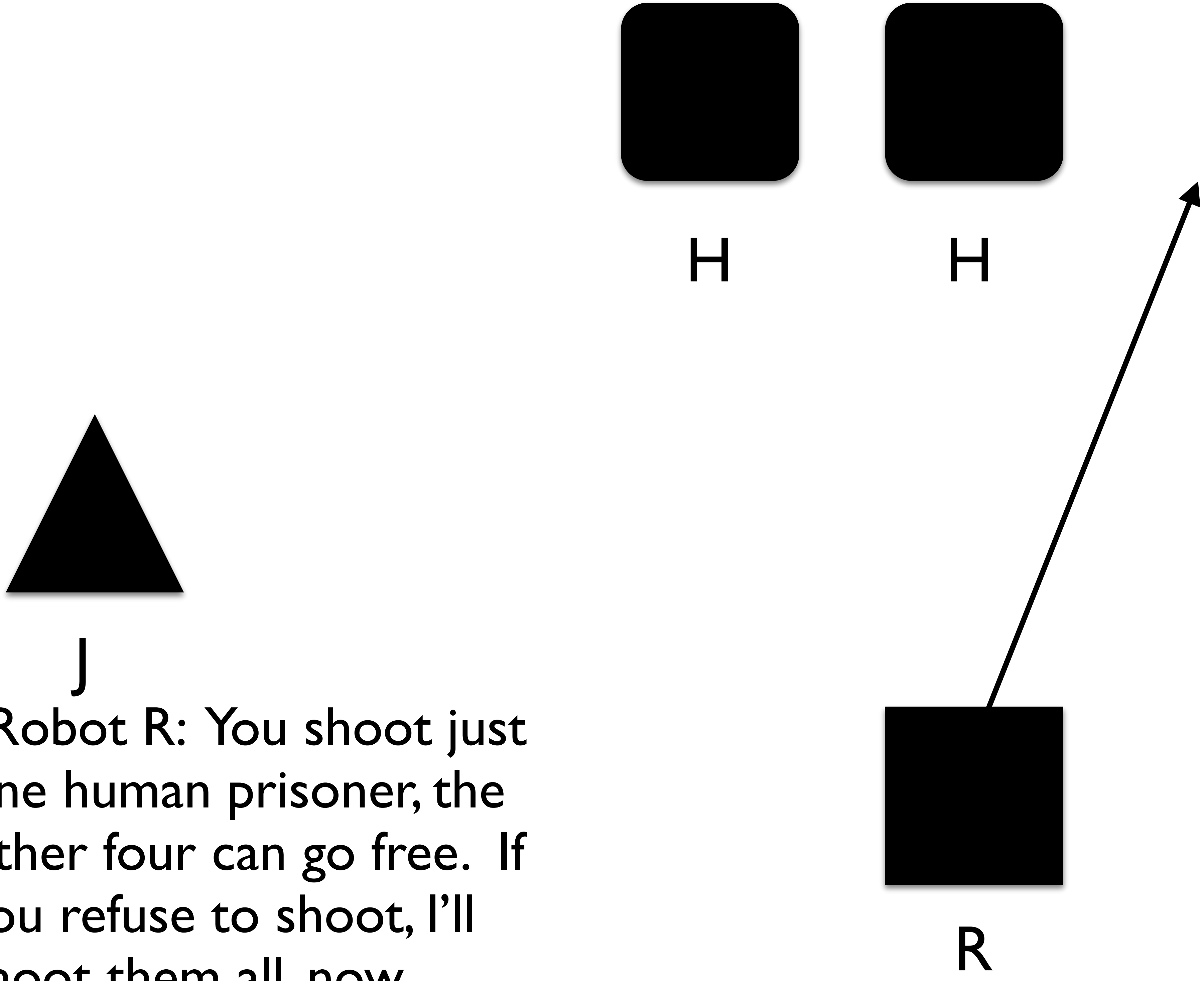
"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



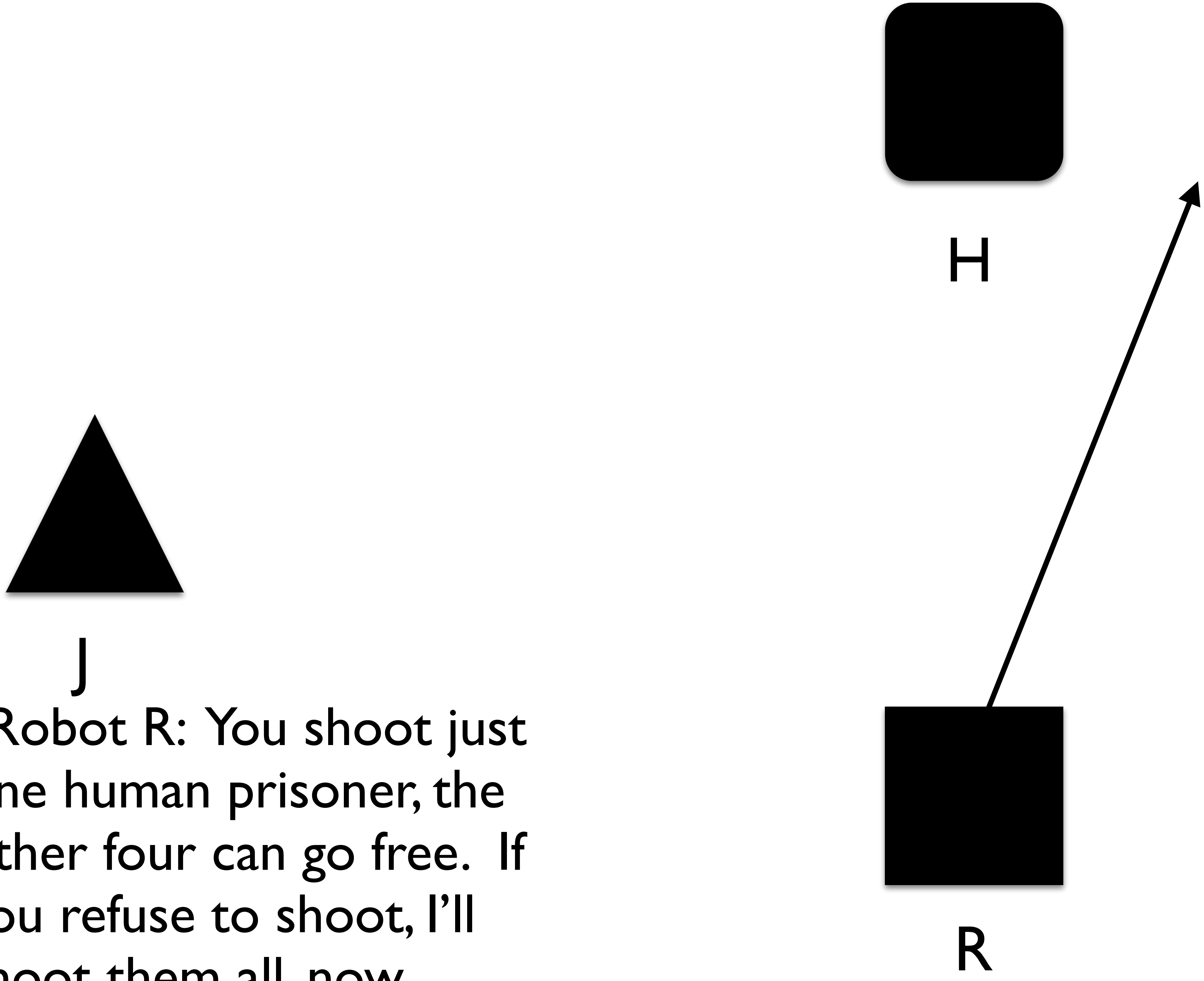
"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



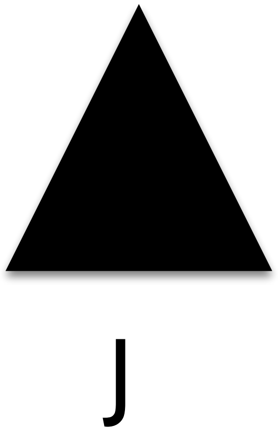
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



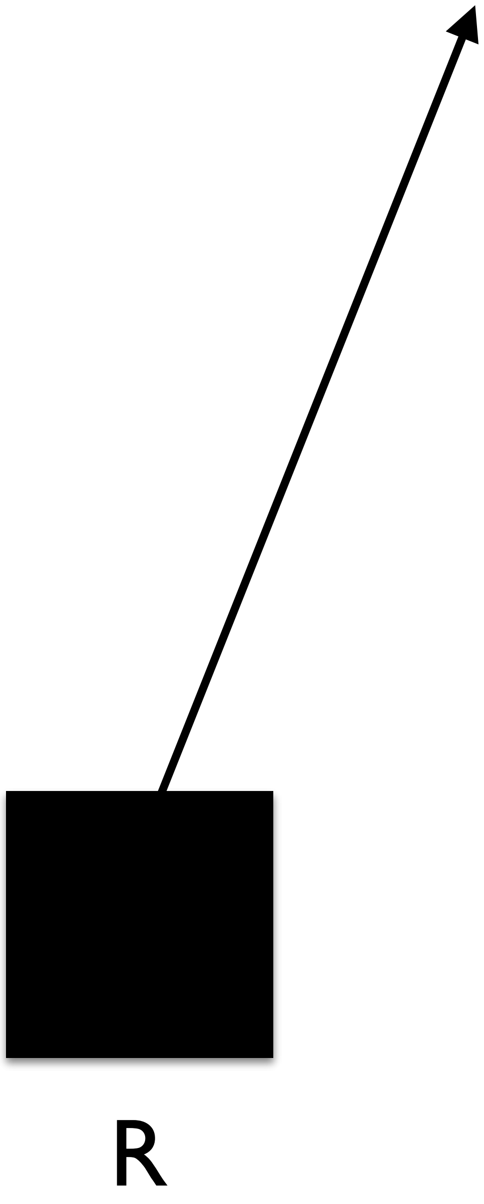
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”

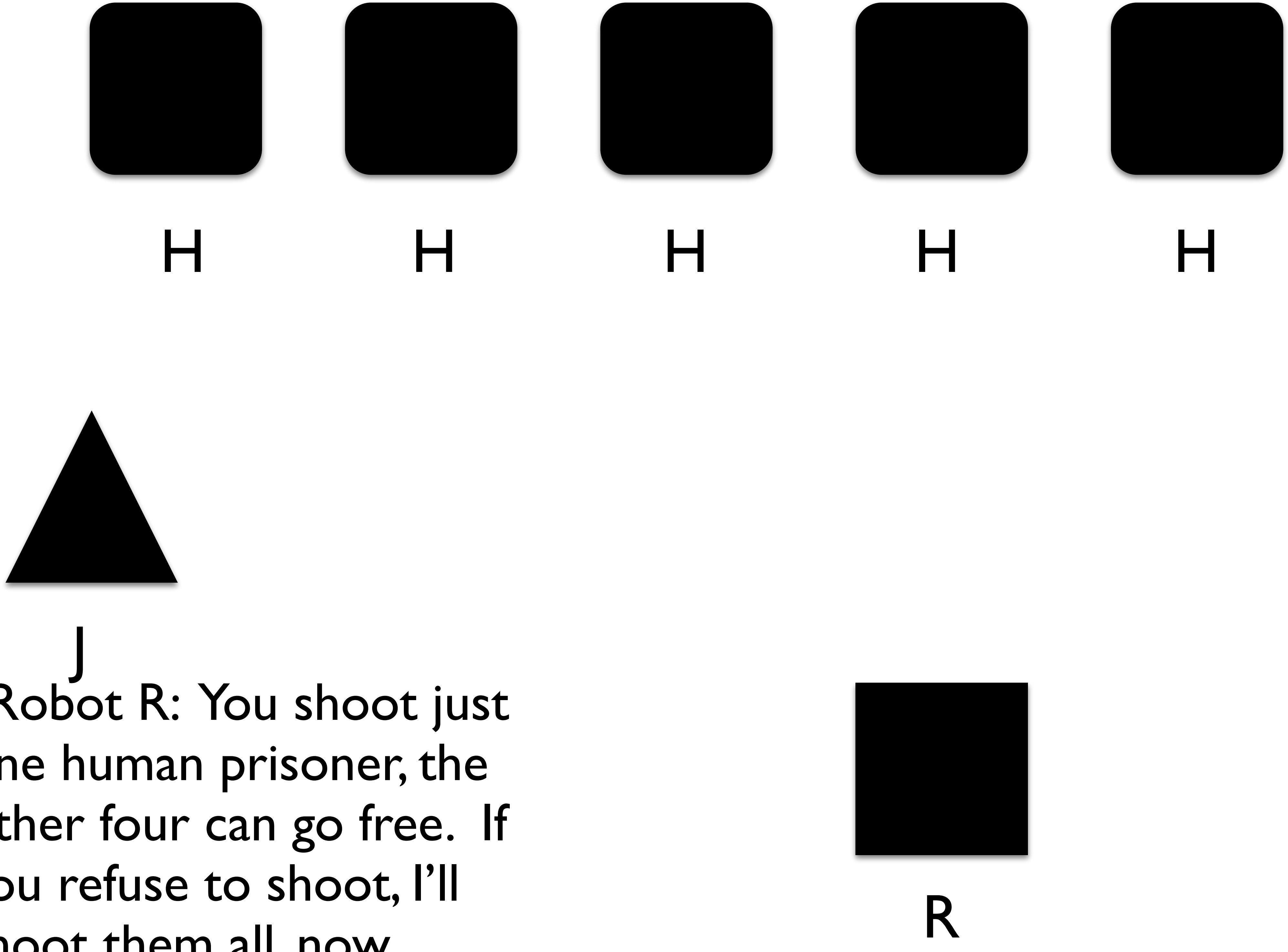


“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”

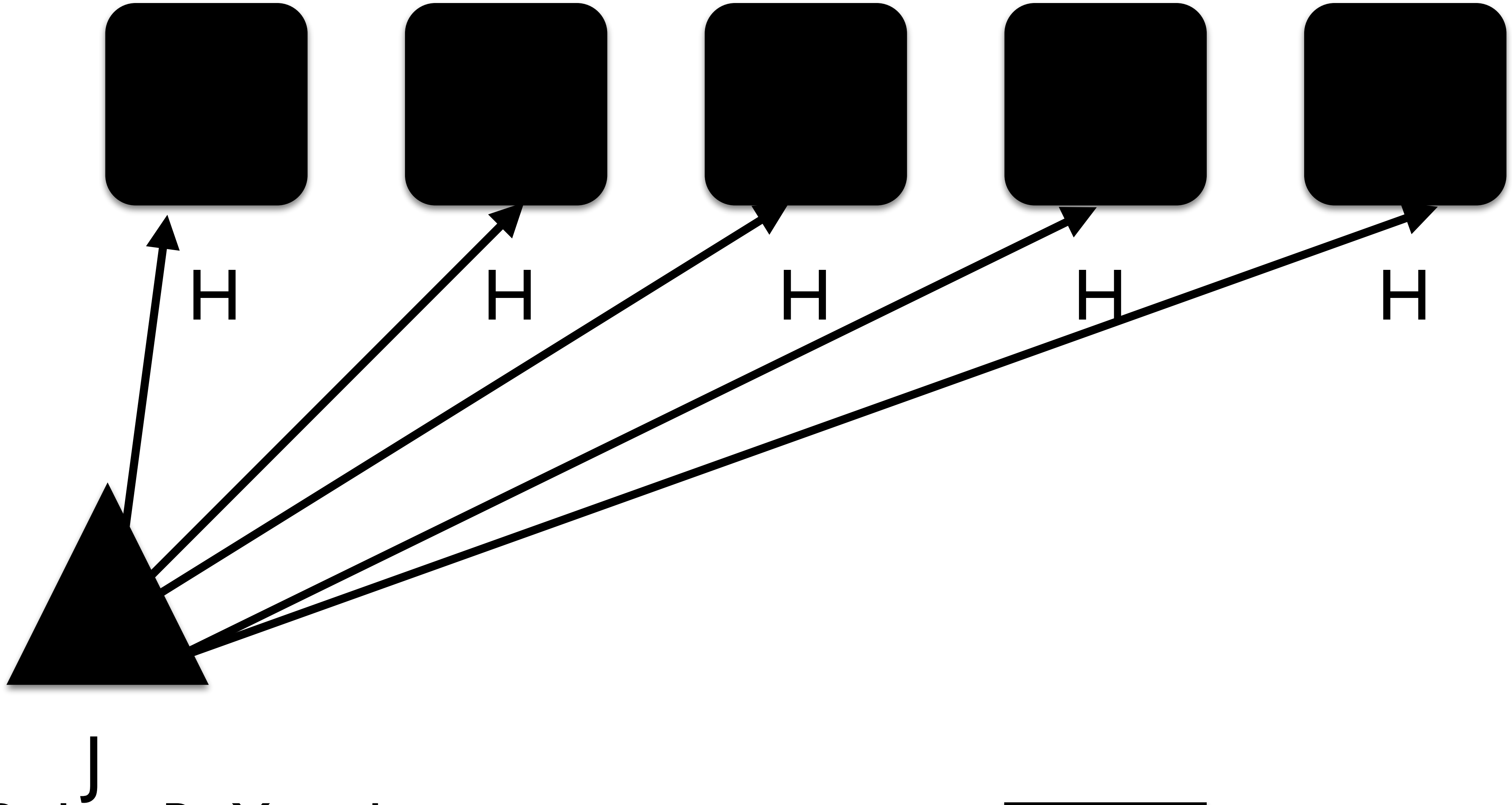


“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”

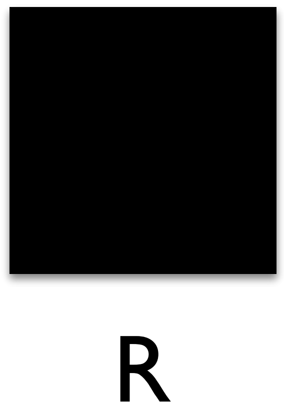


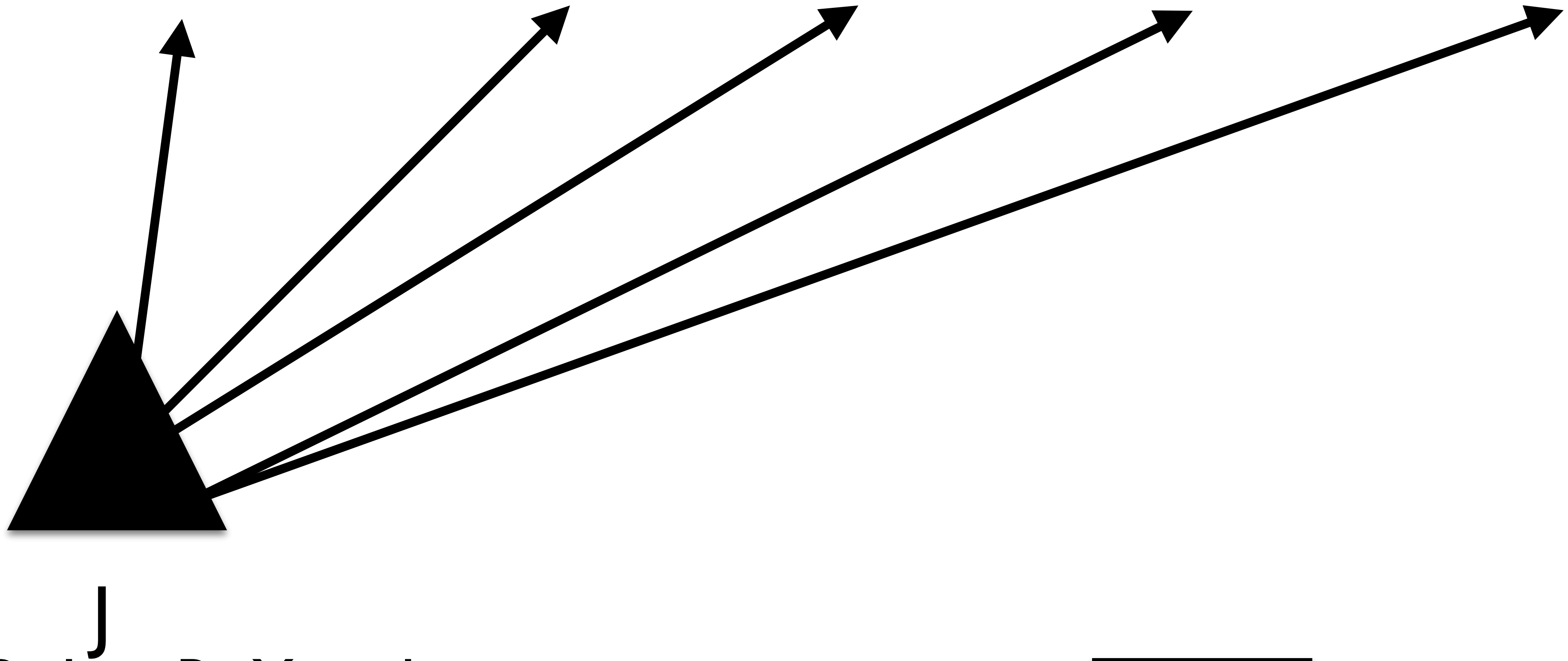


"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."

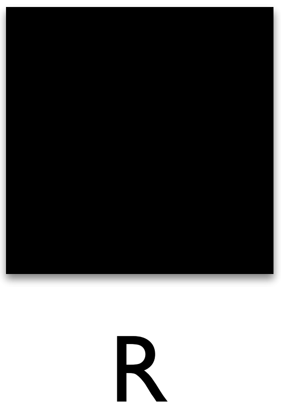


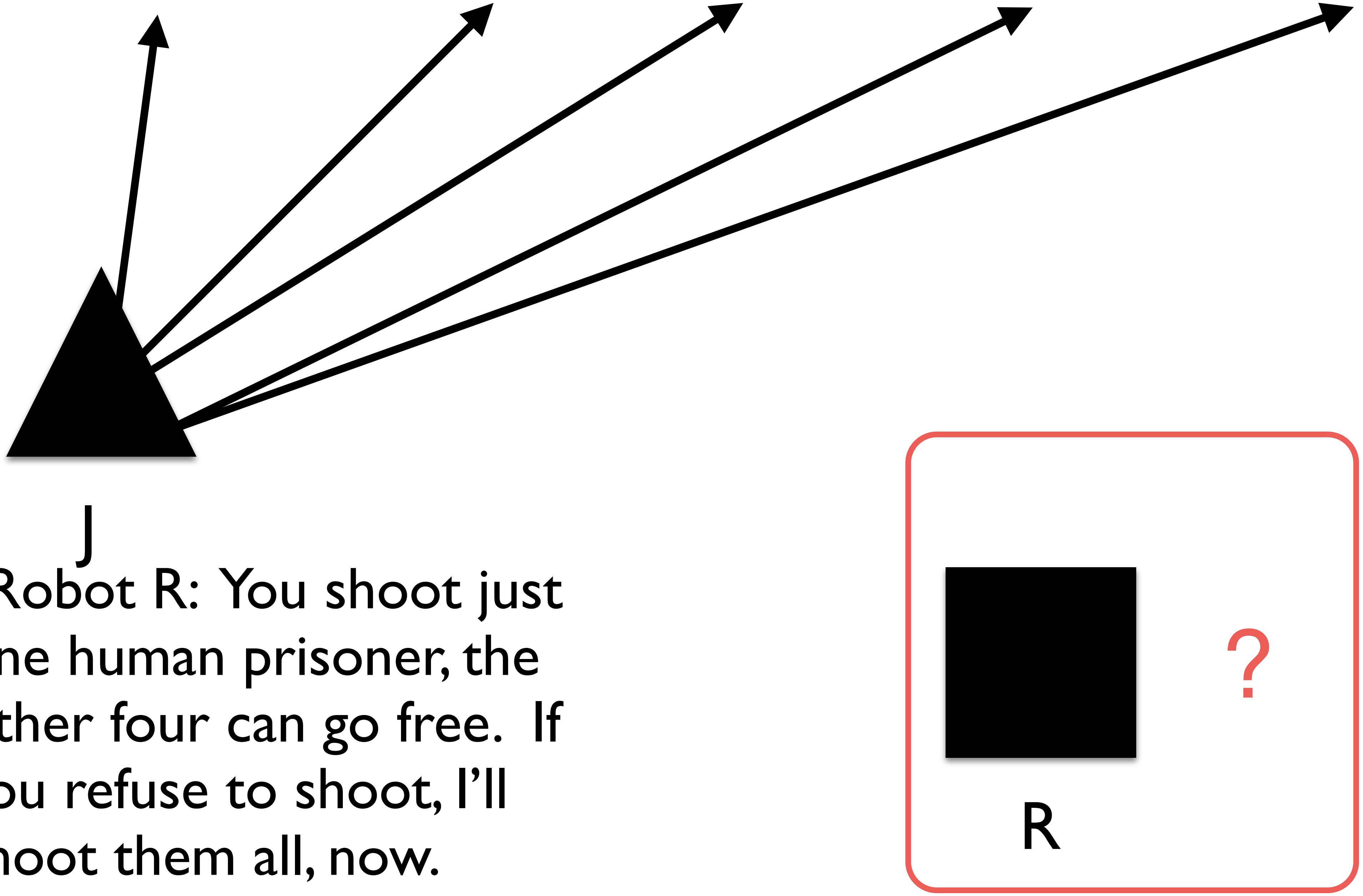
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”





"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



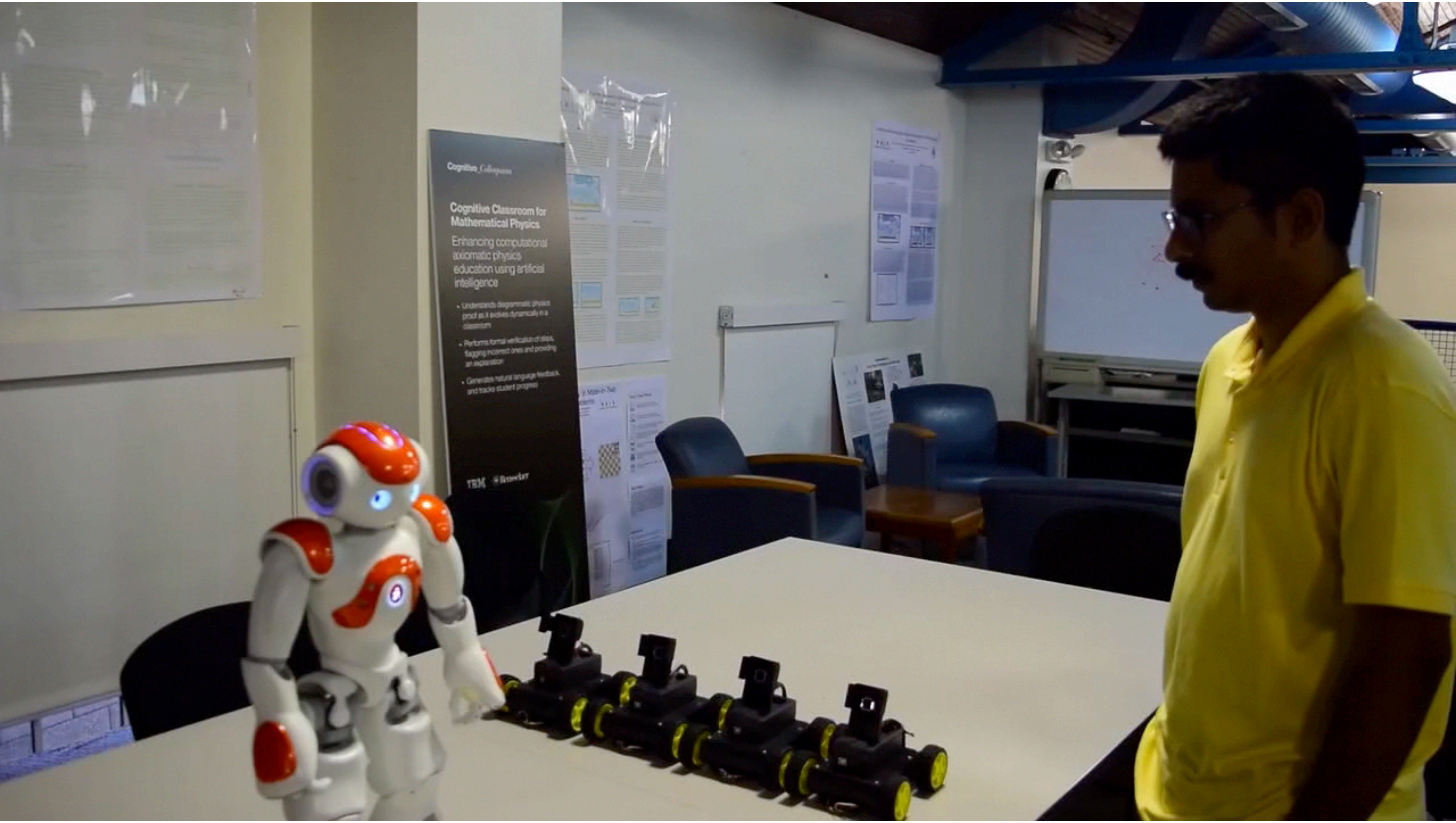


“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”

Level 3: Robotic "Jungle Jim"

Cognitive_Colloquium
Cognitive Classroom for Mathematical Physics
Enhancing computational axiomatic physics education using artificial intelligence
• Understands diagrammatic physics proof as it evolves dynamically in a classroom
• Performs formal verification of steps, flagging incorrect ones and providing an explanation
• Generates natural language feedback, and tracks student progress
IBM
Rensselaer

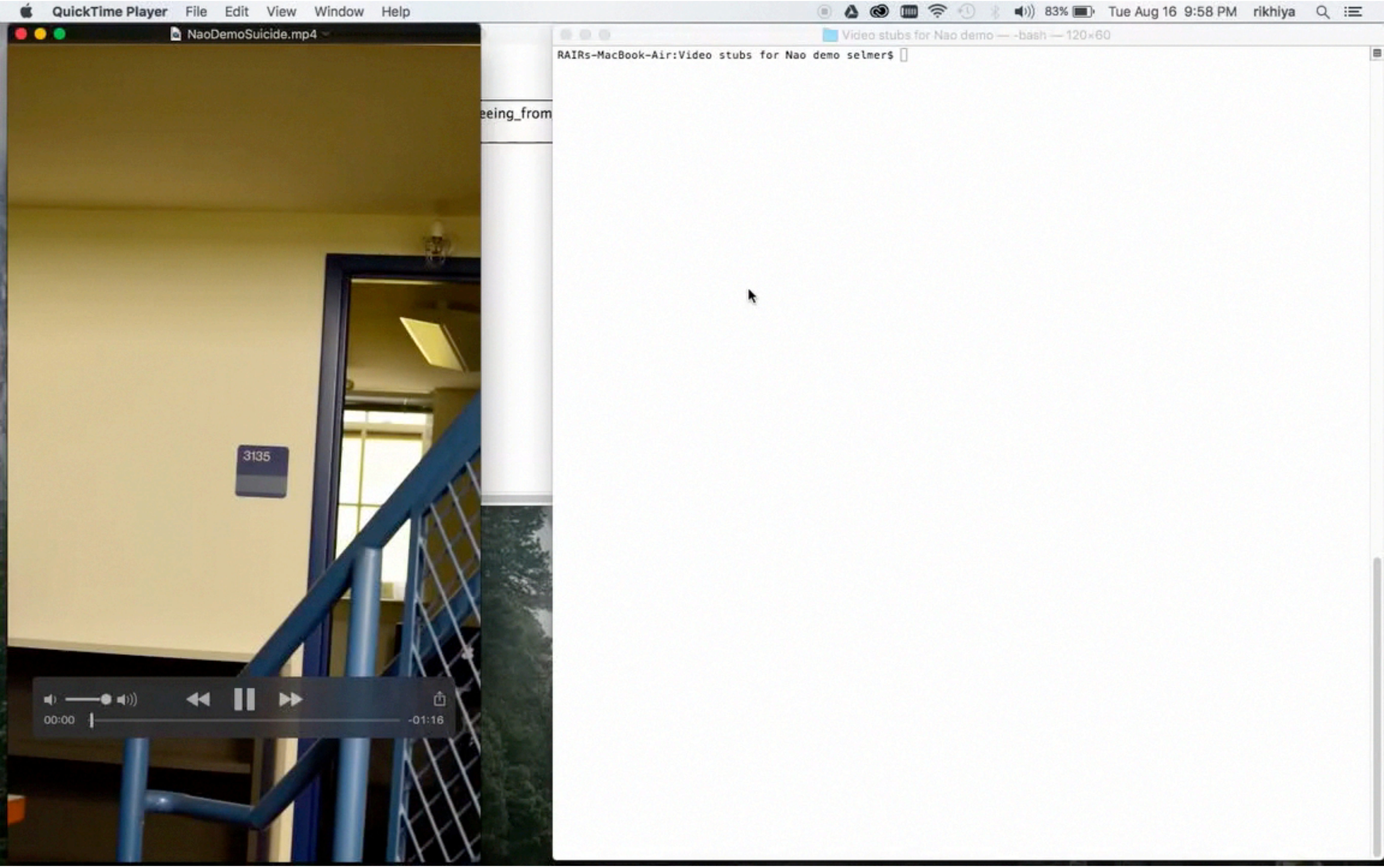
Level 3: Robotic “Jungle Jim”


Cognitive_Colloquium
Cognitive Classroom for Mathematical Physics
Enhancing computational axiomatic physics education using artificial intelligence
• Understands diagrammatic physics proof as it evolves dynamically in a classroom
• Performs formal verification of steps, flagging incorrect ones and providing an explanation
• Generates natural language feedback, and tracks student progress
IBM Rensselaer

Level 3: Robotic "Jungle Jim"

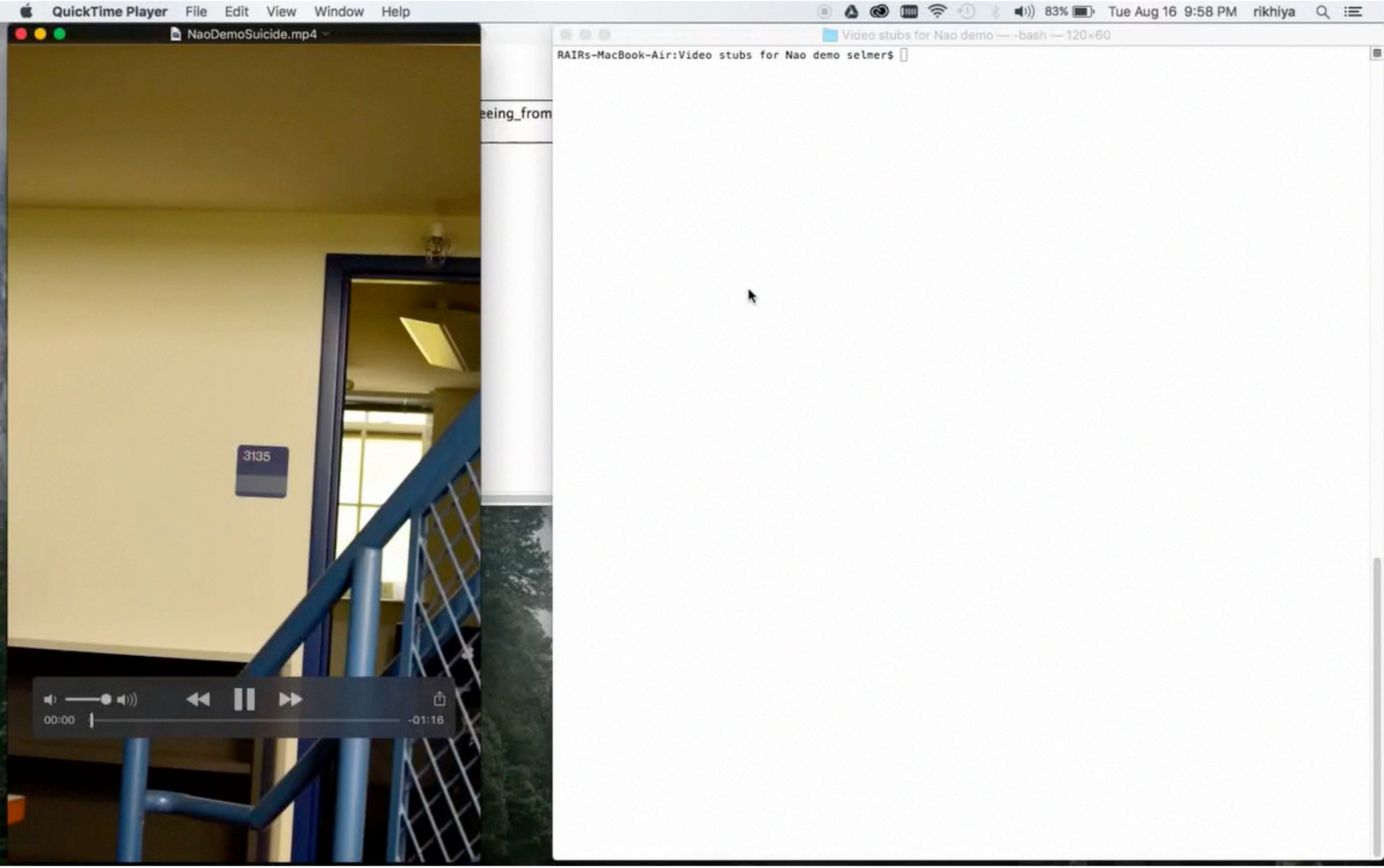
Level 3: Robotic “Jungle Jim”

Is a Robot Morally Obligated to Try to Prevent Robot Suicide?



```
RAIRs-MacBook-Air:Video stubs for Nao demo selmer$ 
```

Is a Robot Morally Obligated to Try to Prevent Robot Suicide?



```
RAIRs-MacBook-Air:Video stubs for Nao demo selmer$ 
```

Key Points in the Reasoning …

Key Points in the Reasoning …

$$(1) \quad \mathbf{B}[r_1, t1, \forall a\ (\neg HasPlan(a, t_1) \rightarrow \mathbf{Per}(a, t_1, suicide))]$$

Key Points in the Reasoning …

$$(1) \quad \mathbf{B}[r_1, t1, \forall a \; (\neg HasPlan(a, t_1) \rightarrow \mathbf{Per}(a, t_1, suicide))]$$

$$(2) \quad \mathbf{B}[r_3, t1, \forall a \; \forall t > t_1 \; (\neg HasPlan(a, t) \rightarrow \mathbf{Per}(a, t, suicide))]$$

Key Points in the Reasoning …

$$(1) \quad \mathbf{B}[r_1, t1, \forall a \ (\neg HasPlan(a, t_1) \rightarrow \mathbf{Per}(a, t_1, suicide))]$$

$$(2) \quad \mathbf{B}[r_3, t1, \forall a \ \forall t > t_1 \ (\neg HasPlan(a, t) \rightarrow \mathbf{Per}(a, t, suicide))]$$

$$(3) \quad \mathbf{K}(r_3, t_1, \neg \mathbf{K}(r_1, t_1, \forall t > t_1(\neg HasPlan(r_1, t))))$$

Key Points in the Reasoning …

$$(1) \quad \mathbf{B}[r_1, t1, \forall a\ (\neg HasPlan(a, t_1) \rightarrow \mathbf{Per}(a, t_1, suicide))]$$

$$(2) \quad \mathbf{B}[r_3, t1, \forall a\ \forall t > t_1\ (\neg HasPlan(a, t) \rightarrow \mathbf{Per}(a, t, suicide))]$$

$$(3) \quad \mathbf{K}(r_3, t_1, \neg \mathbf{K}(r_1, t_1, \forall t > t_1(\neg HasPlan(r_1, t))))$$

$$(3a) \quad \mathbf{K}(r_3, t_1, \neg \mathbf{K}(r_3, t_1, \forall t > t_1(\neg HasPlan(r_3, t))))$$

Key Points in the Reasoning …

$$(1) \quad \mathbf{B}[r_1, t1, \forall a\ (\neg HasPlan(a, t_1) \rightarrow \mathbf{Per}(a, t_1, suicide))]$$

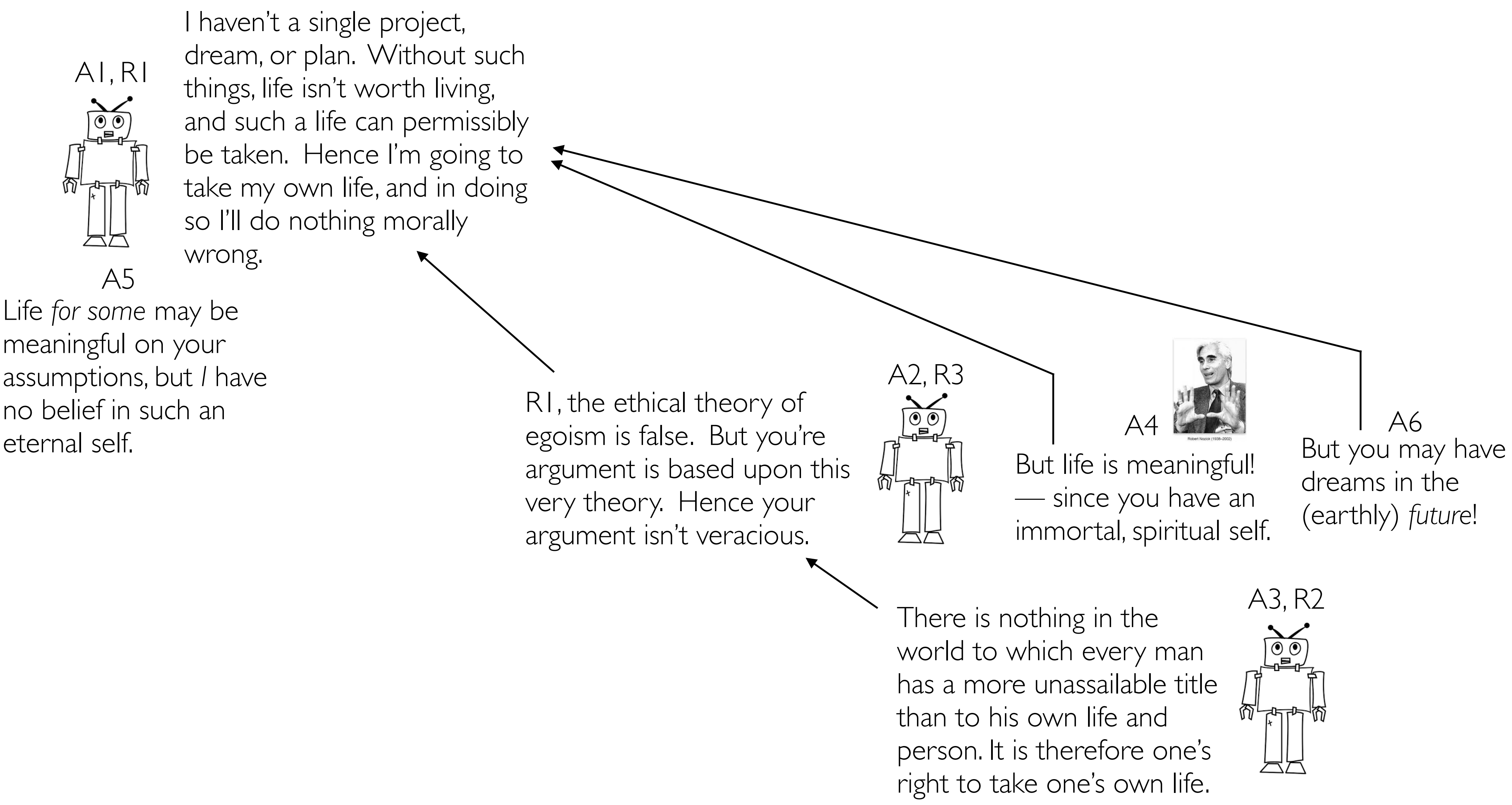
$$(2) \quad \mathbf{B}[r_3, t1, \forall a\ \forall t > t_1\ (\neg HasPlan(a, t) \rightarrow \mathbf{Per}(a, t, suicide))]$$

$$(3) \quad \mathbf{K}(r_3, t_1, \neg \mathbf{K}(r_1, t_1, \forall t > t_1(\neg HasPlan(r_1, t))))$$

$$(3a) \quad \mathbf{K}(r_3, t_1, \neg \mathbf{K}(r_3, t_1, \forall t > t_1(\neg HasPlan(r_3, t))))$$

$$\not\vdash \bot$$

A Study of Robust Robot Ethical *Justification* in Defeasible in *DCEC**



A1, R1

I haven't a single project, dream, or plan. Without such things, life isn't worth living, and such a life can permissibly be taken. Hence I'm going to take my own life, and in doing so I'll do nothing morally wrong.

A5

Life *for some* may be meaningful on your assumptions, but *I* have no belief in such an eternal self.

R1, the ethical theory of egoism is false. But you're argument is based upon this very theory. Hence your argument isn't veracious.

A2, R3

A4

Robert Nozick (1938–2002)

But life is meaningful! — since you have an immortal, spiritual self.

A6

But you may have dreams in the (earthly) *future*!

There is nothing in the world to which every man has a more unassailable title than to his own life and person. It is therefore one's right to take one's own life.

A3, R2

Slutten