Logicist Machine Ethics Can Save Us

Selmer Bringsjord & Atriya Sen 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

Are Humans Rational?
10/18/2018

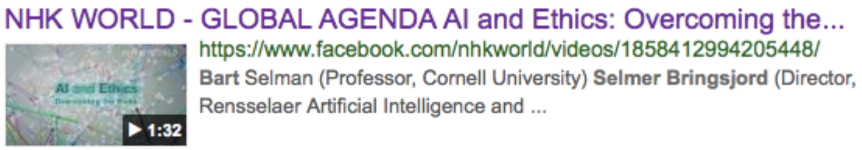
RAIR
Rensselaer AI and Reasoning Lab

Logicist Machine Ethics Can Save Us

Selmer Bringsjord & Atriya Sen 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

Are Humans Rational?
10/18/2018

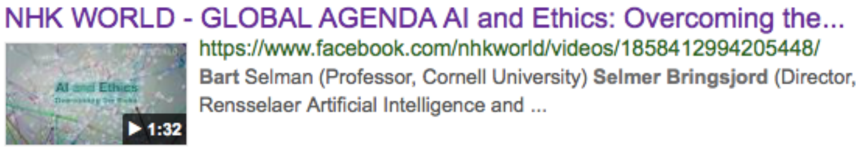


NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...

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

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

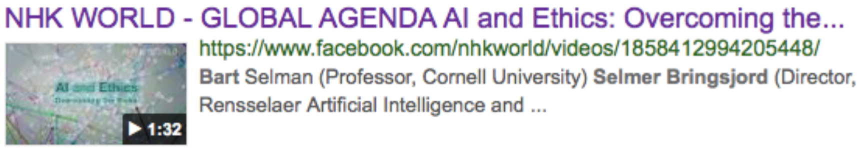
The PAID Problem



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

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

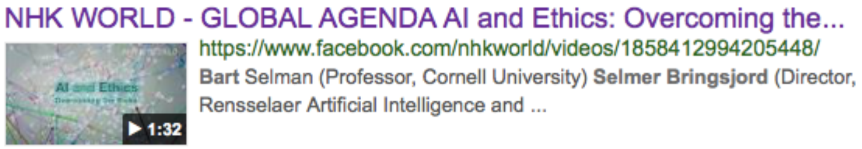


NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...
https://www.facebook.com/nhkworld/videos/1858412994205448/
Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director,
Rensselaer Artificial Intelligence and ...

The PAID Problem

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

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

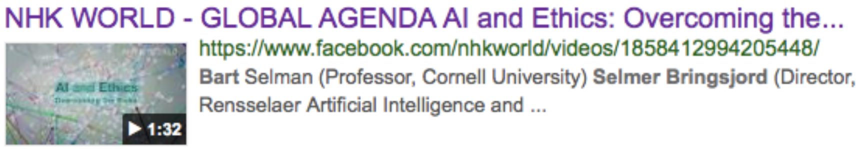


NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...
https://www.facebook.com/nhkworld/videos/1858412994205448/
Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director, Rensselaer Artificial Intelligence and ...

The PAID Problem

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

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



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 ...

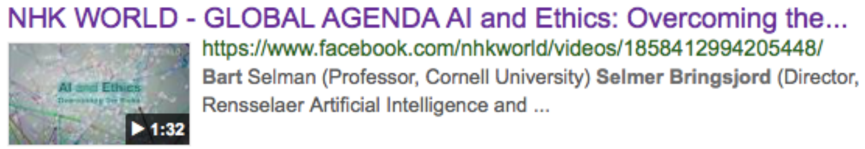
The PAID Problem

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

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

↓

$$u(\mathrm{AIA}_i(\pi_j)) > \tau^+ \in \mathbb{Z} \text{ or } \tau^- \in \mathbb{Z}$$



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

∀x : Agents

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

Are Autonomous-and-Creative Machines Intrinsically Untrustworthy?*

Selmer Bringsjord • Naveen Sundar G.

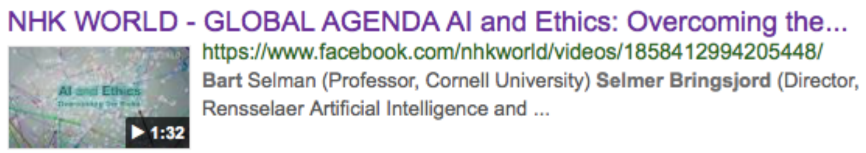
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

020217NY

Abstract

Given what we find in the case of human cognition, the following principle appears to be quite plausible: An artificial agent that is both autonomous (A) and creative (C) will tend to be, from the viewpoint of a rational, fully informed agent, (U) untrustworthy. After briefly explaining the intuitive, internal structure of this disturbing principle, in the context of the human sphere, we provide a more formal rendition of it designed to apply to the realm of intelligent artificial agents. The more-formal version makes use of some of the basic structures available in one of our cognitive-event calculi, and can be expressed as a (confessedly — for reasons explained — naïve) theorem. We prove the theorem, and provide simple demonstrations of it in action, using a novel theorem prover (ShadowProver). We then end by pointing toward some future defensive engineering measures that should be taken in light of the theorem.

Contents



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 ...

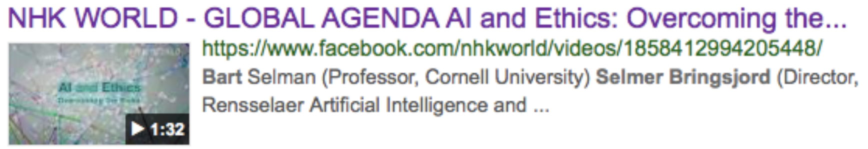
The PAID Problem

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

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

↓

$$u(\mathrm{AIA}_i(\pi_j)) > \tau^+ \in \mathbb{Z} \text{ or } \tau^- \in \mathbb{Z}$$



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 ...

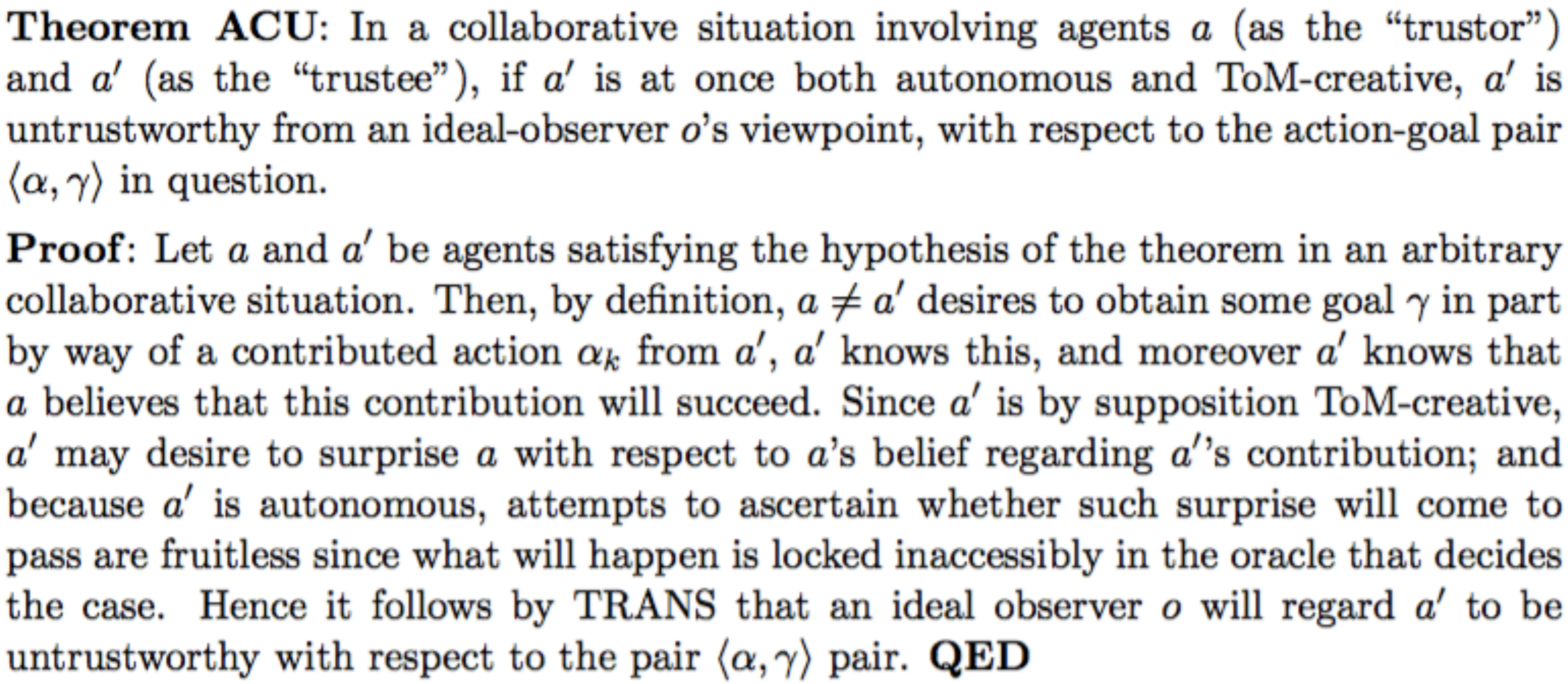
The PAID Problem

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

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

↓

$$u(\mathrm{AIA}_i(\pi_j)) > \tau^+ \in \mathbb{Z} \text{ or } \tau^- \in \mathbb{Z}$$



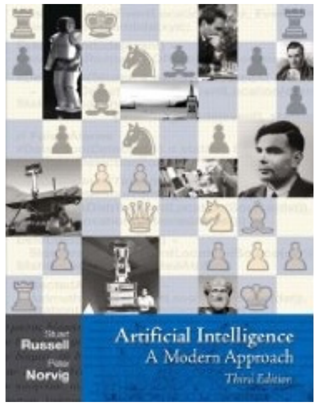
Theorem ACU: In a collaborative situation involving agents a (as the "trustor") and a' (as the "trustee"), if a' is at once both autonomous and ToM-creative, a' is untrustworthy from an ideal-observer o's viewpoint, with respect to the action-goal pair $\langle \alpha, \gamma \rangle$ in question.

Proof: Let a and a' be agents satisfying the hypothesis of the theorem in an arbitrary collaborative situation. Then, by definition, $a \neq a'$ desires to obtain some goal γ in part by way of a contributed action α_k from a', a' knows this, and moreover a' knows that a believes that this contribution will succeed. Since a' is by supposition ToM-creative, a' may desire to surprise a with respect to a's belief regarding a''s contribution; and because a' is autonomous, attempts to ascertain whether such surprise will come to pass are fruitless since what will happen is locked inaccessibly in the oracle that decides the case. Hence it follows by TRANS that an ideal observer o will regard a' to be untrustworthy with respect to the pair $\langle \alpha, \gamma \rangle$ pair. **QED**

“We’re in *very* deep trouble.”

“We’re in *very* deep trouble.”

“We’re in *very* deep trouble.”



Unfortunately, not quite as easy as this to use logic to save the day …

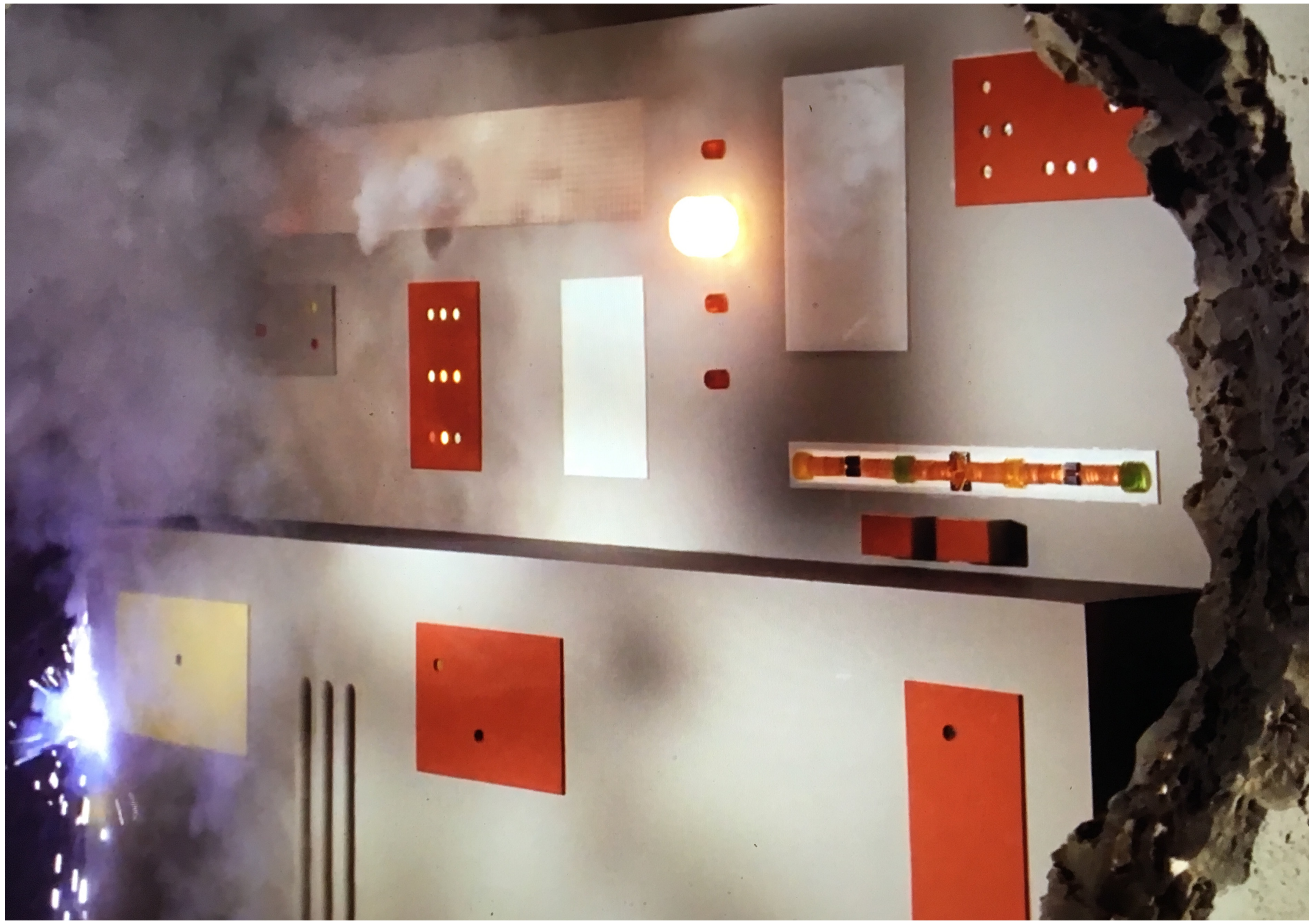
Logic Thwarts Landru!

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!



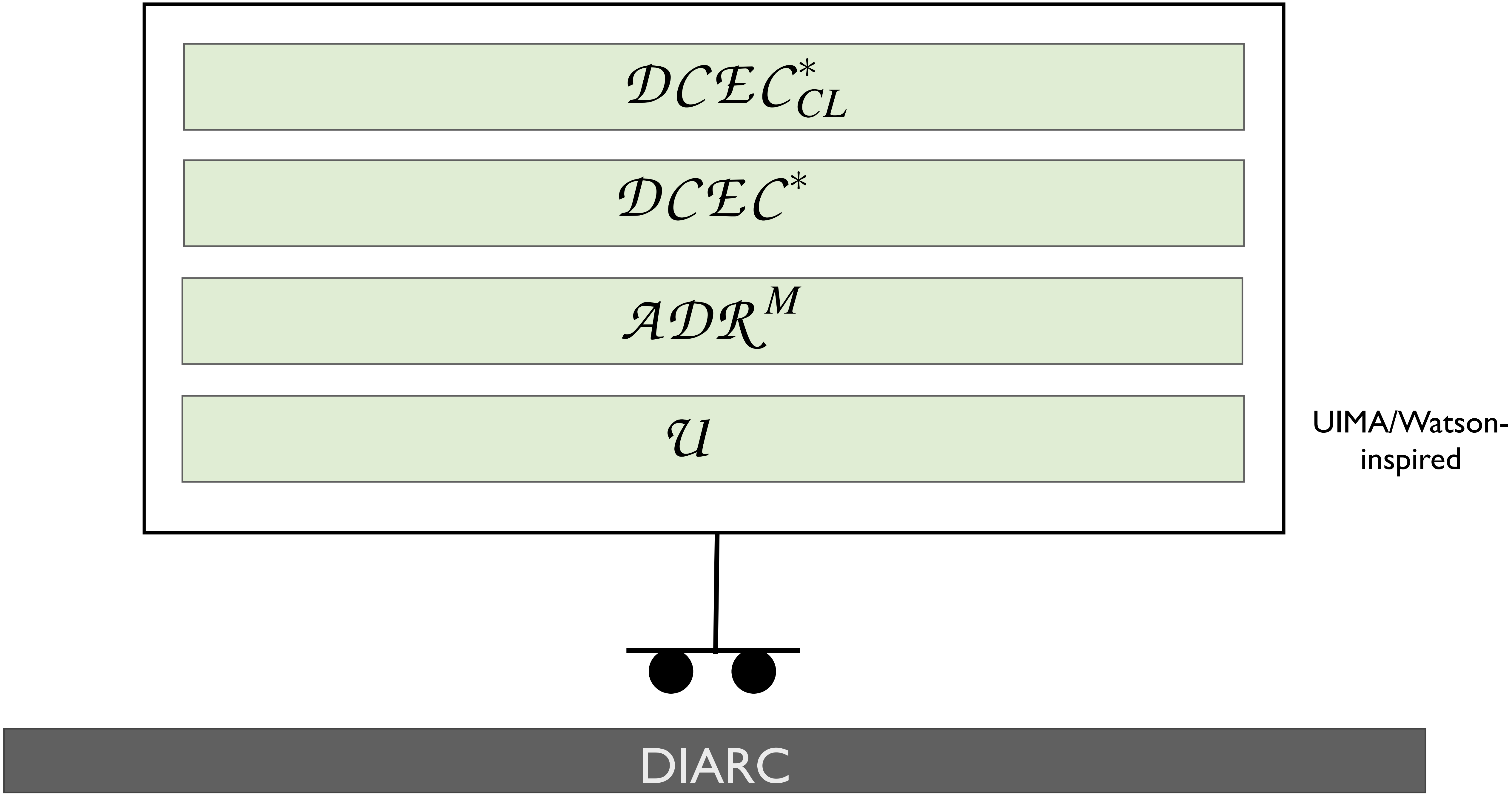
Landru Kills Himself Because Kirk/Spock Argue He Has Violated the Prime Directive for Good by Denying Creativity to Others

Logic Thwarts Nomad!

(with the Liar Paradox)

I.
Cognitive Calculi …

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

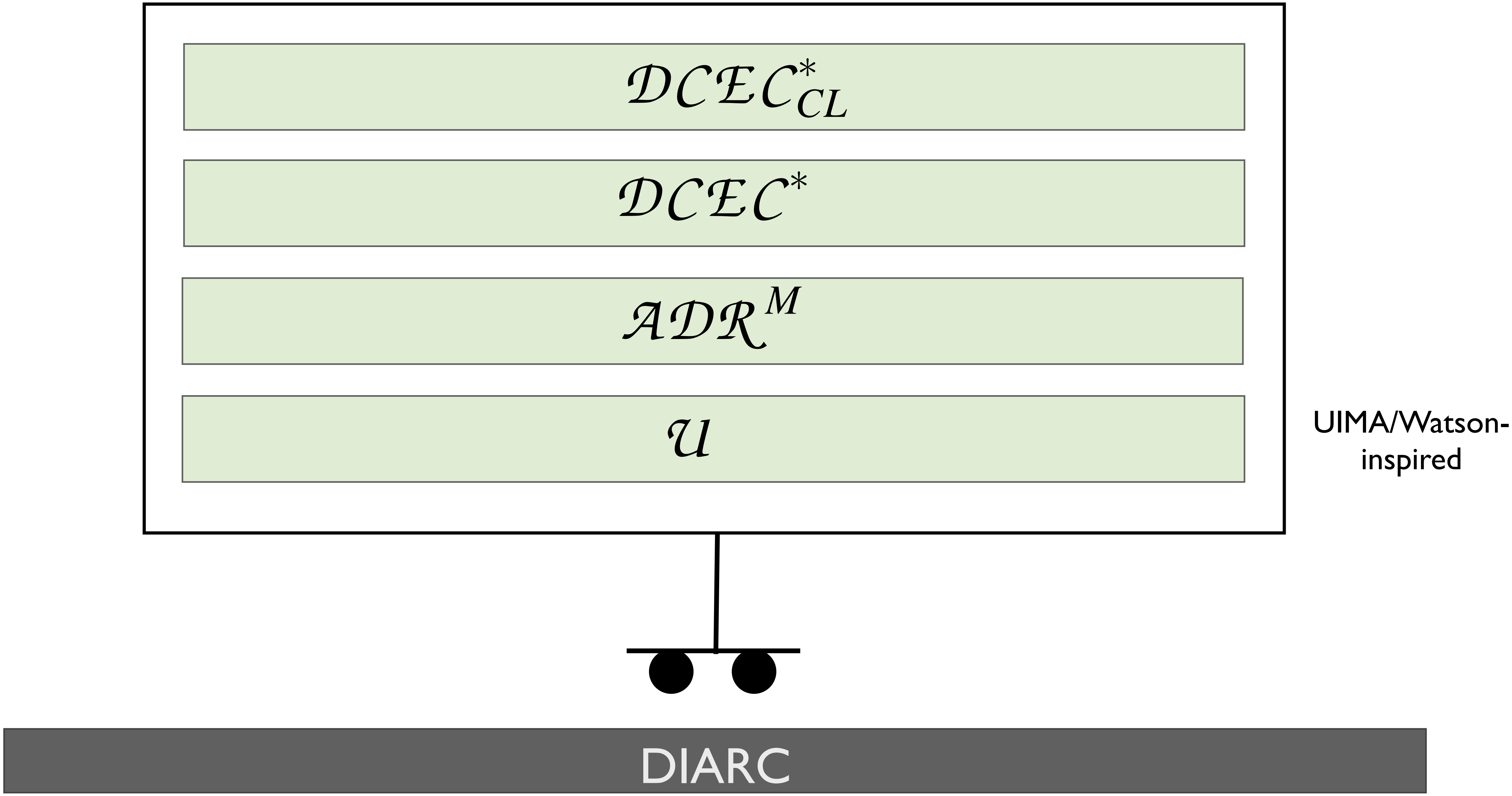
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-inspired

DIARC

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

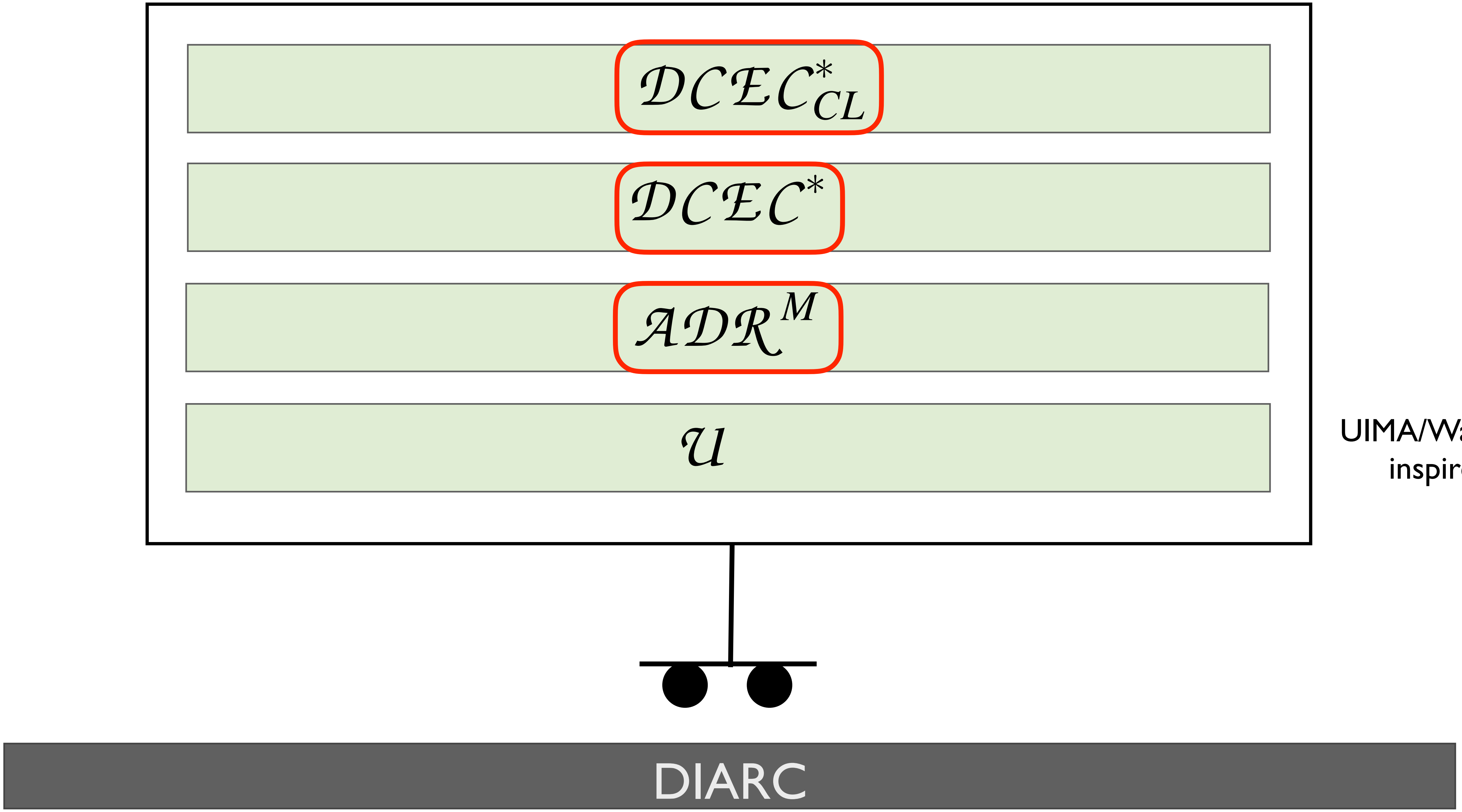
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-inspired

DIARC

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

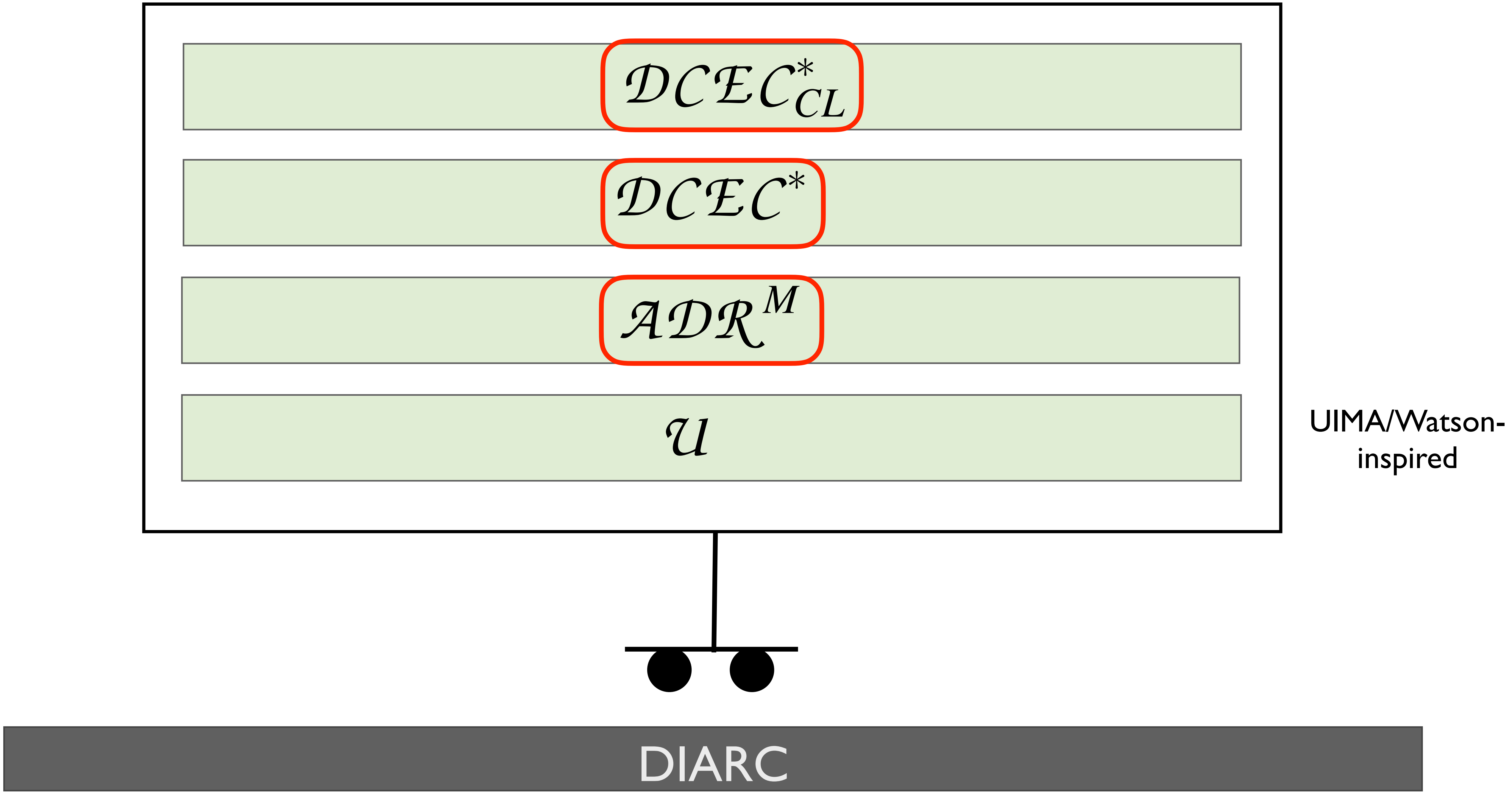
$\mathcal{U}$

UIMA/Watson-
inspired

DIARC

Hierarchy of Ethical Reasoning

Not paradox-prone deontic logics!



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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-
inspired

DIARC

“Universal Cognitive Calculus”

1666

Leibniz

1.5 centuries < Boole!
2.5 centuries < Kripke

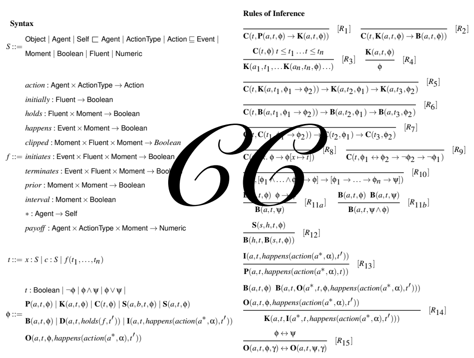
$\int$

Logic Theorist
(birth of modern logicist AI)

1956

Simon

$\mathcal{DCEC}^*$



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

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

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

Rules of Inference

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

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

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

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

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

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

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

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

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

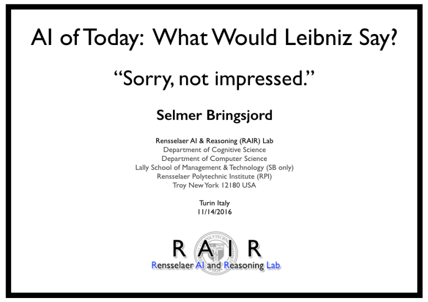
$$\frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\ [R_{13}]$$

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

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

2017

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

Cognitive Calculi $\mathcal{CC}$

purely extensional level: FOL MSL SOL TOL IFOL . . .

intensional level: epistemic deontic possibility/necessity . . .

ATPs: SPASS SNARK ShadowProver . . .

theories: **PA ZFC** axiomatic physics . . .

model finders: MACE . . .

nature of representation: symbolic or homomorphic: . . .

Cognitive Calculi $\mathcal{CC}$

purely extensional level: FOL MSL SOL TOL IFOL . . .

intensional level: epistemic deontic possibility/necessity . . .

ATPs: SPASS SNARK ShadowProver . . .

theories: **PA ZFC** axiomatic physics . . .

model finders: MACE . . .

nature of representation: symbolic or homomorphic:
. . .

Cognitive Calculi $\mathcal{CC}$

purely extensional level: FOL MSL SOL TOL IFOL . . .

intensional level: epistemic deontic possibility/necessity . . .

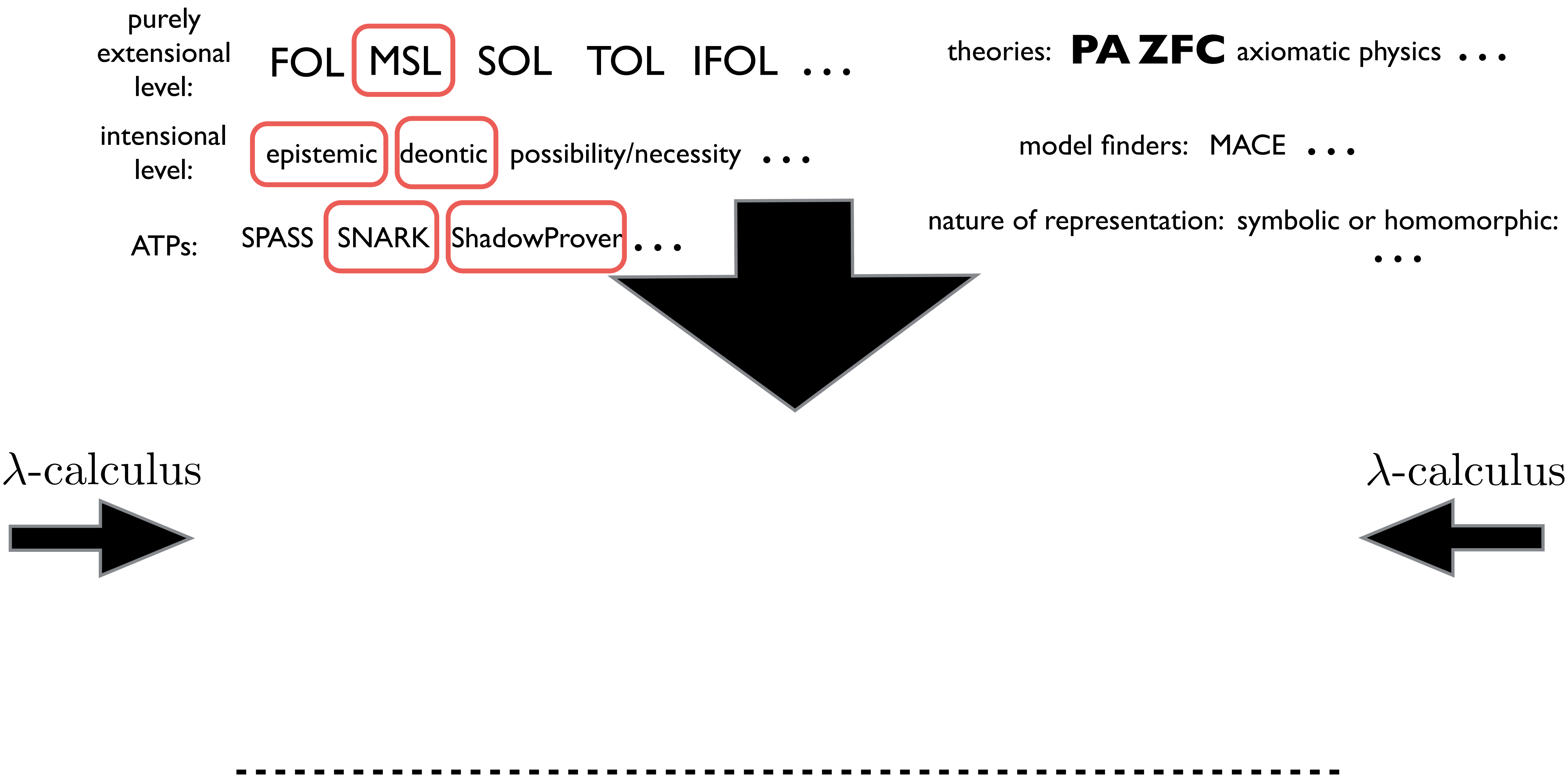
ATPs: SPASS SNARK ShadowProver . . .

theories: **PA ZFC** axiomatic physics . . .

model finders: MACE . . .

nature of representation: symbolic or homomorphic:
. . .

Cognitive Calculi $\mathcal{CC}$



purely extensional level: FOL MSL SOL TOL IFOL . . .

intensional level: epistemic deontic possibility/necessity . . .

ATPs: SPASS SNARK ShadowProver . . .

theories: **PA ZFC** axiomatic physics . . .

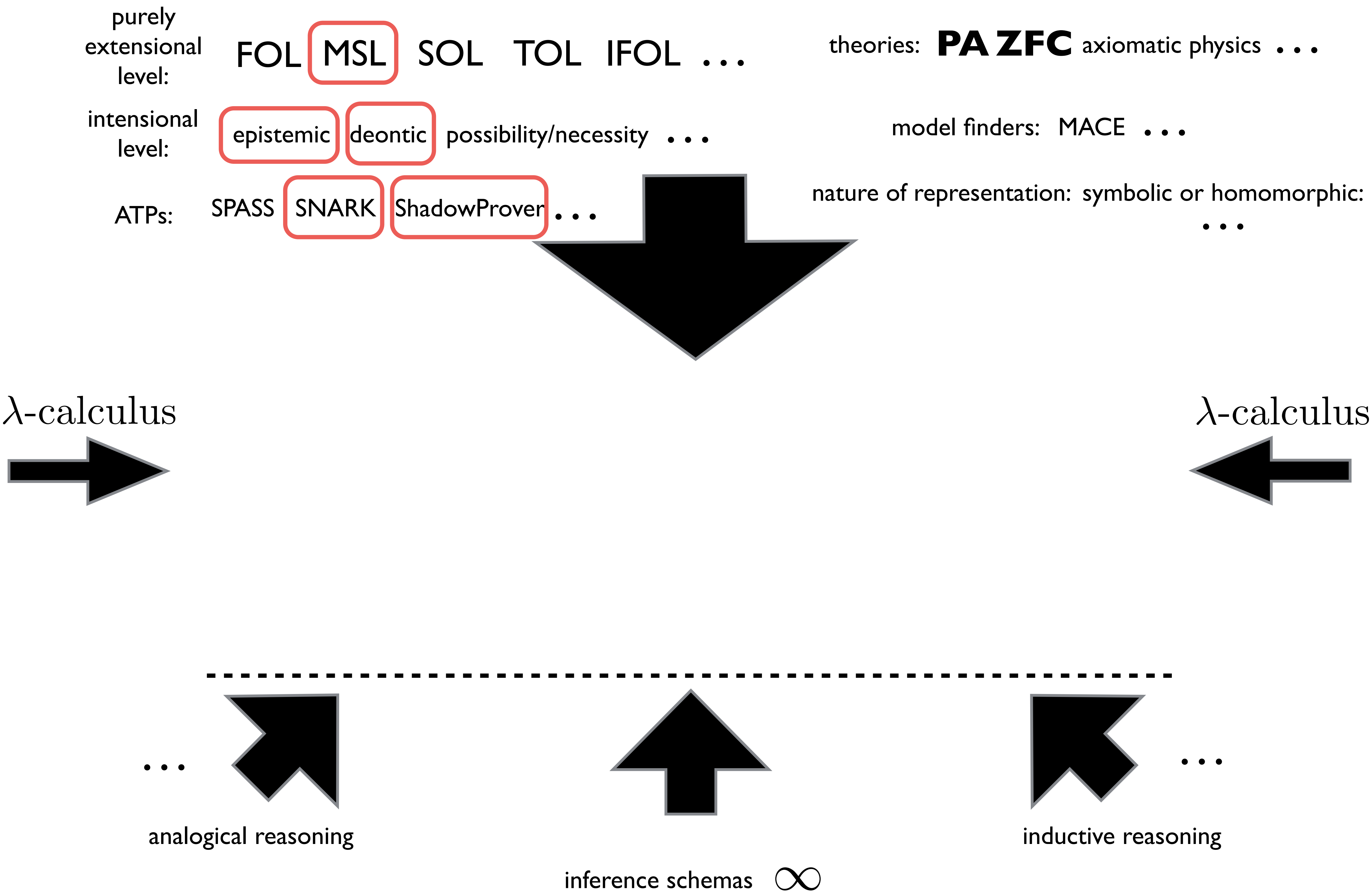
model finders: MACE . . .

nature of representation: symbolic or homomorphic: . . .

λ-calculus

λ-calculus

Cognitive Calculi $\mathcal{CC}$



purely extensional level: FOL MSL SOL TOL IFOL . . .

intensional level: epistemic deontic possibility/necessity . . .

ATPs: SPASS SNARK ShadowProver . . .

theories: **PA ZFC** axiomatic physics . . .

model finders: MACE . . .

nature of representation: symbolic or homomorphic:

. . .

λ-calculus

λ-calculus

. . .

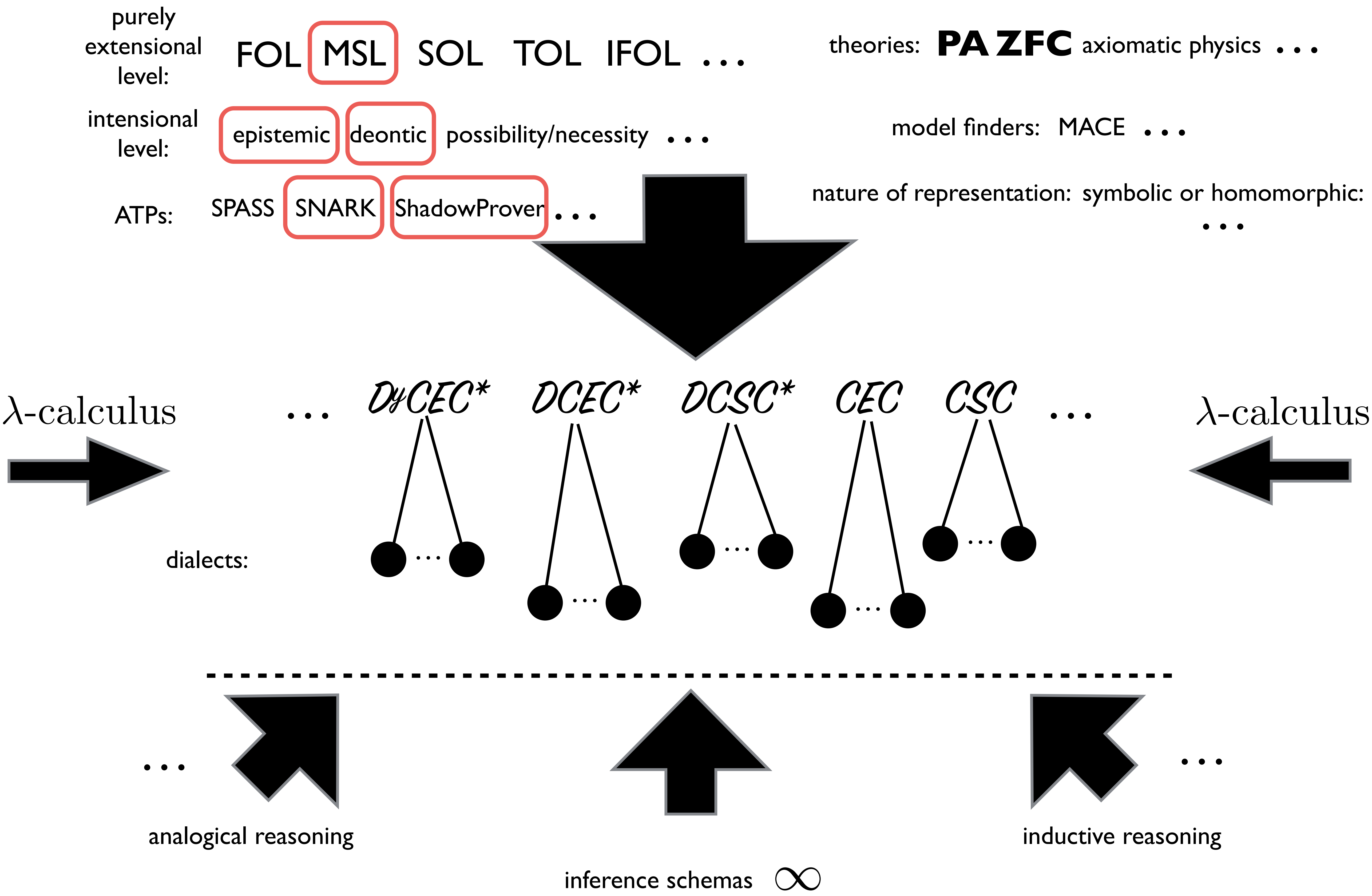
analogical reasoning

inference schemas ∞

inductive reasoning

. . .

Cognitive Calculi $\mathcal{CC}$



purely extensional level: FOL MSL SOL TOL IFOL . . .

intensional level: epistemic deontic possibility/necessity . . .

ATPs: SPASS SNARK ShadowProver . . .

theories: **PA ZFC** axiomatic physics . . .

model finders: MACE . . .

nature of representation: symbolic or homomorphic: . . .

λ-calculus . . . $\mathcal{D}y\mathcal{CEC}^*$ $\mathcal{DCEC}^*$ $\mathcal{DCSC}^*$ $\mathcal{CEC}$ $\mathcal{CSC}$ . . . λ-calculus

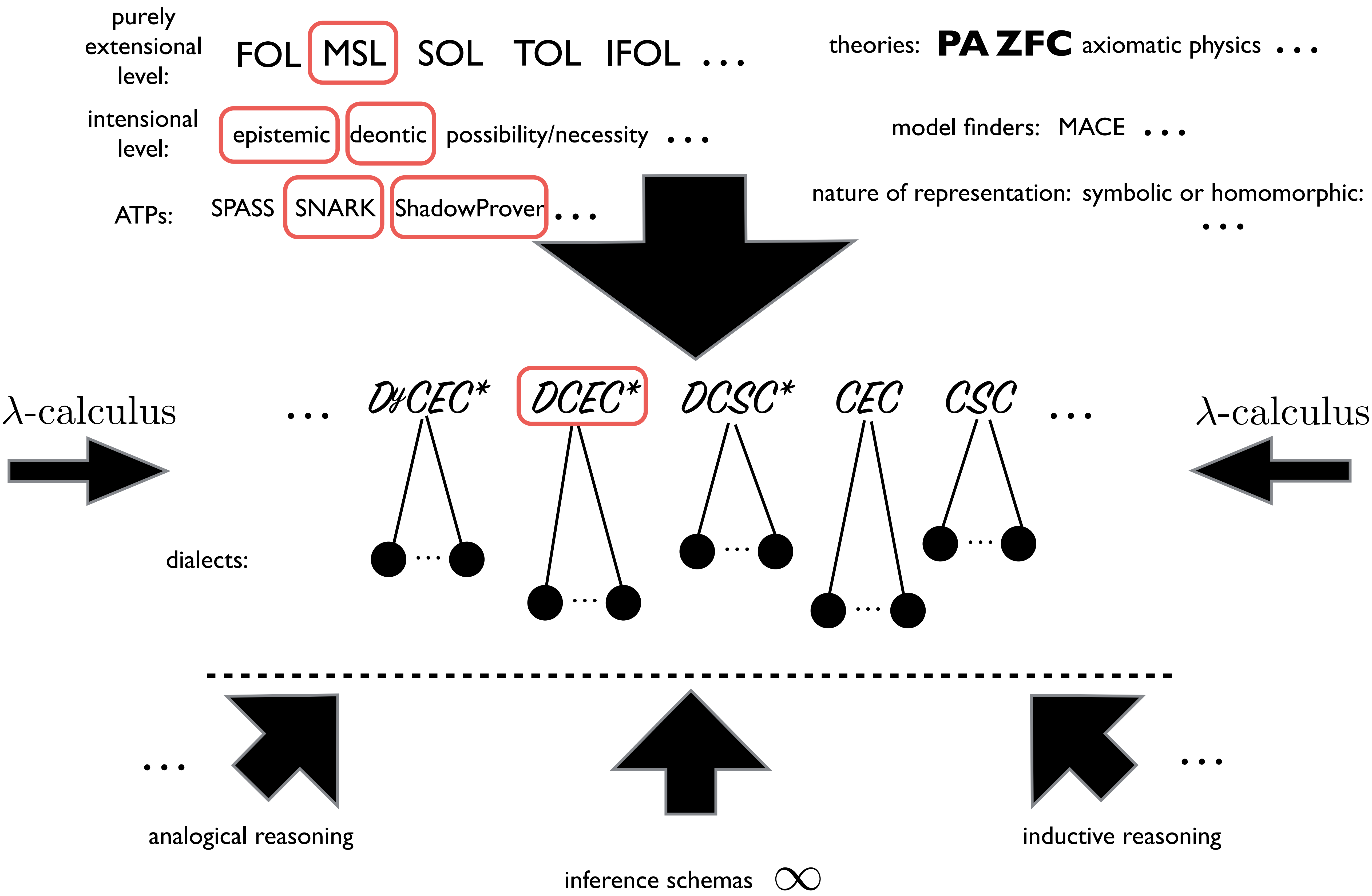
dialects: · · · · · · · · · · · · · · ·

. . . analogical reasoning

inference schemas ∞

inductive reasoning . . .

Cognitive Calculi $\mathcal{CC}$



purely extensional level: FOL MSL SOL TOL IFOL . . .

theories: **PA ZFC** axiomatic physics . . .

intensional level: epistemic deontic possibility/necessity . . .

model finders: MACE . . .

ATPs: SPASS SNARK ShadowProver . . .

nature of representation: symbolic or homomorphic: . . .

λ-calculus . . . $\mathcal{D}y\mathcal{CEC}^*$ $\mathcal{DCEC}^*$ $\mathcal{DCSC}^*$ $\mathcal{CEC}$ $\mathcal{CSC}$ . . . λ-calculus

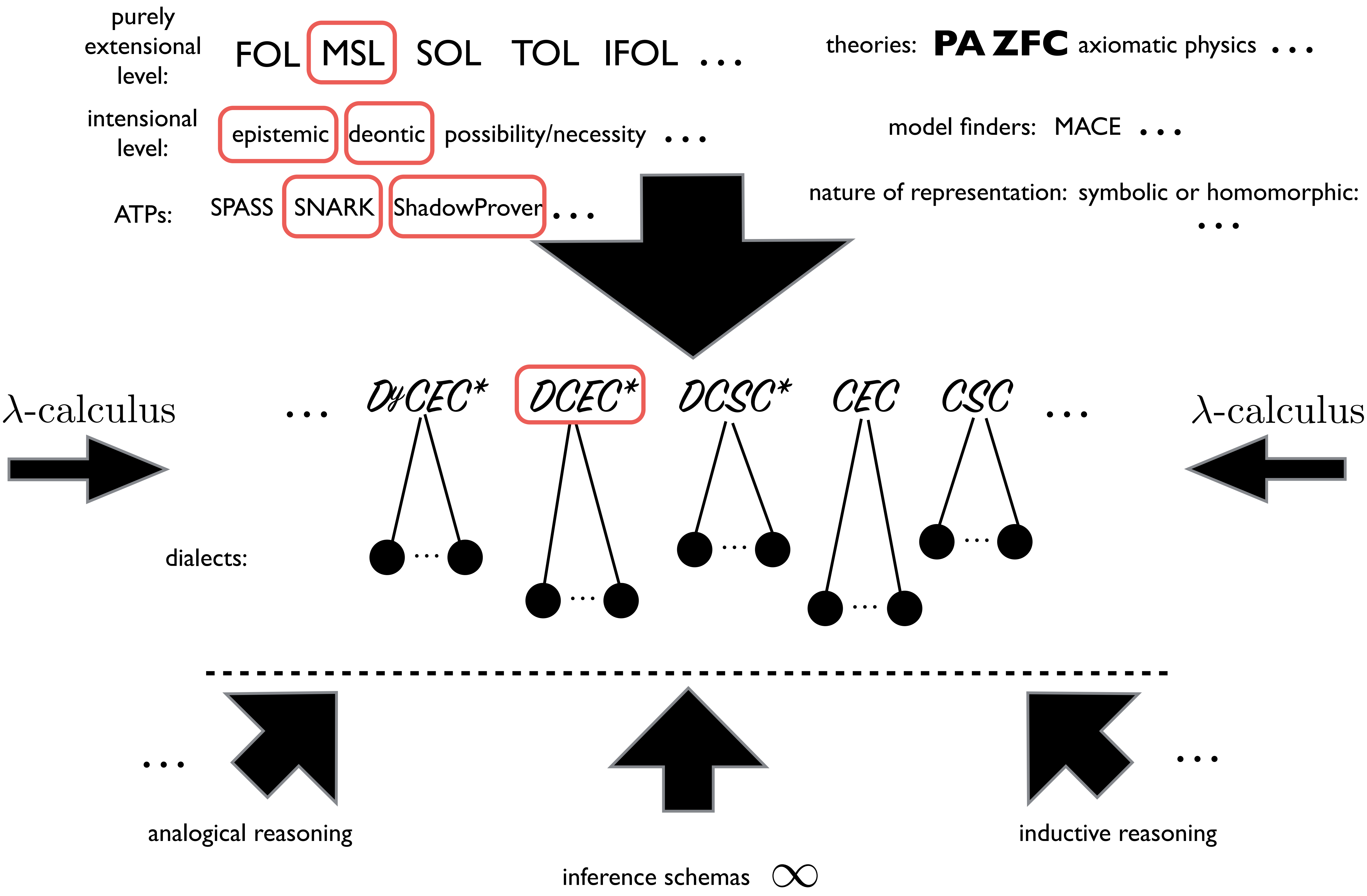
dialects: . . .

. . . analogical reasoning

inference schemas ∞

inductive reasoning . . .

Cognitive Calculi $\mathcal{CC}$



purely extensional level: FOL MSL SOL TOL IFOL . . .

theories: **PA ZFC** axiomatic physics . . .

intensional level: epistemic deontic possibility/necessity . . .

model finders: MACE . . .

ATPs: SPASS SNARK ShadowProver . . .

nature of representation: symbolic or homomorphic: . . .

λ-calculus . . . DyCEC* DCEC* DCSC* CEC CSC . . . λ-calculus

dialects:

. . . analogical reasoning

inference schemas ∞

inductive reasoning . . .

Formal Syntax

Formal Syntax

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

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

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

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

Inference Schemata

Inference Schemata

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

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

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

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

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

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

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

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

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

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

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

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

Event Calculus for Time & Change

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

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

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

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

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

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

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

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

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

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

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

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

Event Calculus for Time & Change

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

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

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

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

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

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

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

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

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

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

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

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

$[A_1]$ $\mathbf{C}(\forall\, f,t\ .\ \mathit{initially}(f)\wedge\neg \mathit{clipped}(0,f,t)\Rightarrow \mathit{holds}(f,t))$

$[A_2]$ $\mathbf{C}(\forall\, e,f,t_1,t_2\ .\ \mathit{happens}(e,t_1)\wedge \mathit{initiates}(e,f,t_1)\wedge t_1 < t_2 \wedge\neg \mathit{clipped}(t_1,f,t_2)\Rightarrow \mathit{holds}(f,t_2))$

$[A_3]$ $\mathbf{C}(\forall\, t_1,f,t_2\ .\ \mathit{clipped}(t_1,f,t_2)\Leftrightarrow[\exists\, e,t\ .\ \mathit{happens}(e,t)\wedge t_1 < t < t_2 \wedge \mathit{terminates}(e,f,t)])$

$[A_4]$ $\mathbf{C}(\forall\, a,d,t\ .\ \mathit{happens}(\mathit{action}(a,d),t)\Rightarrow\mathbf{K}(a,\mathit{happens}(\mathit{action}(a,d),t)))$

$[A_5]$ $\mathbf{C}(\forall\, a,f,t,t'\ .\ \mathbf{B}(a,\mathit{holds}(f,t))\wedge\mathbf{B}(a,t<t')\wedge\neg\mathbf{B}(a,\mathit{clipped}(t,f,t'))\Rightarrow\mathbf{B}(a,\mathit{holds}(f,t')))$

Defs for An *Affective* Cognitive *time&change* Calculus

1. **Joy** : pleased about a desirable event. By 'pleased about a desirable event' the meaning we will consider is 'pleased about a desirable consequence of the event'.

$$\mathit{forSome}\ c\ B(a,t_3,\mathit{implies}(\mathit{happens}(e,t_1),\mathit{holds}(\mathit{CON}(e,a,c),t_2))) \tag{1}$$

$$D(a,t_3,\mathit{holds}(\mathit{CON}(e,a,c),t_2)) \tag{2}$$

$$K(a,t_3,\mathit{happens}(e,t_1)) \tag{3}$$

   The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{joy}),t_3)$ is therefore and(1,2,3).

2. **Distress** : displeased about an undesirable event.

$$\mathit{not}(D(a,t_3,\mathit{holds}(\mathit{CON}(e,a,c),t_3))) \tag{4}$$

   The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{distress}),t_3)$ is therefore and(1,4,3).

3. **Happy-for**: pleased about an event presumed to be desirable for someone else

$$\mathit{forSome}\ c\ B(a,t_3,\mathit{implies}(\mathit{happens}(e,t_1),\mathit{holds}(\mathit{CON}(e,a_1,c),t_2))) \tag{5}$$

$$B(a,t_3,D(a_1,t_3,\mathit{holds}(\mathit{CON}(e,a_1,c),t_2))) \tag{6}$$

$$D(a,t_3,\mathit{holds}(\mathit{CON}(e,a_1,c),t_2)) \tag{7}$$

   The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{happy\_for}),t_3)$ is therefore and(5,6,7,3).

4. **Pity**: displeased about an event presumed to be undesirable for someone else. This is equivalent to sorry_for in Hobbs-Gordon model.

$$B(a,t_3,\mathit{not}(D(a_1,t_3,\mathit{holds}(\mathit{CON}(e,a_1,c),t_2)))) \tag{8}$$

$$\mathit{not}(D(a,t_3,\mathit{holds}(\mathit{CON}(e,a_1,c),t_2))) \tag{9}$$

   The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{pity}),t_3)$ is therefore and(5,8,9,3).

5. **Gloating** : pleased about an event presumed to be undesirable for someone else The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{gloating}),t_3)$ is therefore and(5,8,7,3).

6. **Resentment**: displeased about an event presumed to be desirable for someone else The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{resentment}),t_3)$ is therefore and(5,6,9,3).

7. **Hope**: (pleased about) the prospect of a desirable event

$$\mathit{forSome}\ c\ B(a,t_0,\mathit{implies}(\mathit{happens}(e,t_1),\diamond\mathit{holds}(\mathit{CON}(e,a,c),t_2))) \tag{10}$$

$$D(a,t_0,\mathit{holds}(\mathit{CON}(e,a,c),t_2)) \tag{11}$$

   The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{hope}),t_0)$ is therefore and(10,11).

8. **Fear**: (displeased about) the prospect of an undesirable event

$$\mathit{not}(D(a,t_0,\mathit{holds}(\mathit{CON}(e,a,c),t_2))) \tag{12}$$

   The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{fear}),t_0)$ is therefore and(10,12).

9. **Satisfaction** : (pleased about) the confirmation of the prospect of a desirable event
   The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{satisfaction}),t_3)$ is and(10,11, 7 3).

10. **Fears-confirmed** : (displeased about) the confirmation of the prospect of an undesirable event.
    The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{fears}-\mathit{confirmed}),t_3)$ is and(10,12,9, 3).

11. **Relief**: (pleased about) the disconfirmation of the prospect of an undesirable event

$$K(a,t_3,\mathit{not}(\mathit{happens}(e,t_1))) \tag{13}$$

    The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{relief}),t_3)$ is $\mathit{and}(10,12,9,13)$.

12. **Disappointment** : (displeased about) the disconfirmation of the prospect of a desirable event
    The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{disappointment}),t_3)$ is $\mathit{and}(10,11,7,13)$.

13. **Pride** : (approving of) one's own praiseworthy action
    Here we treat 'approve' as an action event. We also introduce a new predicate $\mathit{PRAISEWORTHY}(a,b,x)$ which will mean that agent a considers x a praiseworthy action by agent b. All the 3 interpretations are shown below.

$$\mathit{happens}(\mathit{action}(a,x),t_0) \tag{14}$$

$$\mathit{forAll}\ a_x B(a,t_1,\mathit{implies}(\mathit{happens}(\mathit{action}(a_x,x),t_x),\mathit{PRAISEWORTHY}(a,a_x,x))),t_x \leq t_1 \tag{15}$$

$$D(a,t_1,\mathit{holds}(\mathit{PRAISEWORTHY}(a,a,x),t_1)) \tag{16}$$

$$\mathit{happens}(\mathit{action}(a,\mathit{approve}(x)),t_1) \tag{17}$$

    The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{pride}),t_1)$ is $\mathit{and}(14,B(a,t_1,\mathit{holds}(\mathit{PRAISEWORTHY}(a,a,x),t_1)),17)$.

14. **Shame**: (disapproving of) one's own blameworthy action
    This also follows the same explanation as Pride.

$$\mathit{forAll}\ a_x B(a,t_1,\mathit{implies}(\mathit{happens}(\mathit{action}(a_x,x),t_x),B(a,t_1,\mathit{holds}(\mathit{BLAMEWORTHY}(a,a_x,x)),t_1))),t_x \leq t_1 \tag{18}$$

$$\mathit{not}(\mathit{happens}(\mathit{action}(a,\mathit{approve}(x)),t_1)) \tag{19}$$

    The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{shame}),t_1)$ is $\mathit{and}(14,B(a,t_1,\mathit{holds}(\mathit{BLAMEWORTHY}(a,a,x),t_1)),19)$.

15. **Admiration**: (approving of) someone else's praiseworthy action

$$\mathit{happens}(\mathit{action}(a_1,x),t_0) \tag{20}$$

    The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{admiration}),t_1)$ is $\mathit{and}(20,B(a,t_1,\mathit{holds}(\mathit{PRAISEWORTHY}(a,a_1,x),t_1)),17)$.

16. **Reproach**: (disapproving of) someone else's blameworthy action The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{reproach}),t_1)$ is $\mathit{and}(20,B(a,t_1,\mathit{holds}(\mathit{BLAMEWORTHY}(a,a_1,x),t_1)),19)$.

17. **Gratification** : (approving of) one's own praiseworthy action and (being pleased about) the related desirable event. We again interpret 'pleased about the desirable event' as 'pleased about the desired consequence of the event.'

$$\mathit{forSome}\ c\ B(a,t_1,\mathit{implies}(\mathit{happens}(\mathit{action}(a,x),t_0),\mathit{holds}(\mathit{CON}(\mathit{action}(a,x),a,c),t_0))) \tag{21}$$

$$D(a,t_1,\mathit{holds}(\mathit{CON}(\mathit{action}(a,x),a,c),t_0)) \tag{22}$$

    The definition of $\mathit{holds}(\mathit{AFF}(a,\mathit{gratification}),t_1)$ is $\mathit{and}(20,B(a,t_1,\mathit{holds}(\mathit{PRAISEWORTHY}(a,a,x),t_1)),17$

… (and more)

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

NewScientist

Ethical robots save humans

NewScientist

Ethical robots save humans

Informal Definition of Akrasia

Informal Definition of Akrasia

A

agent

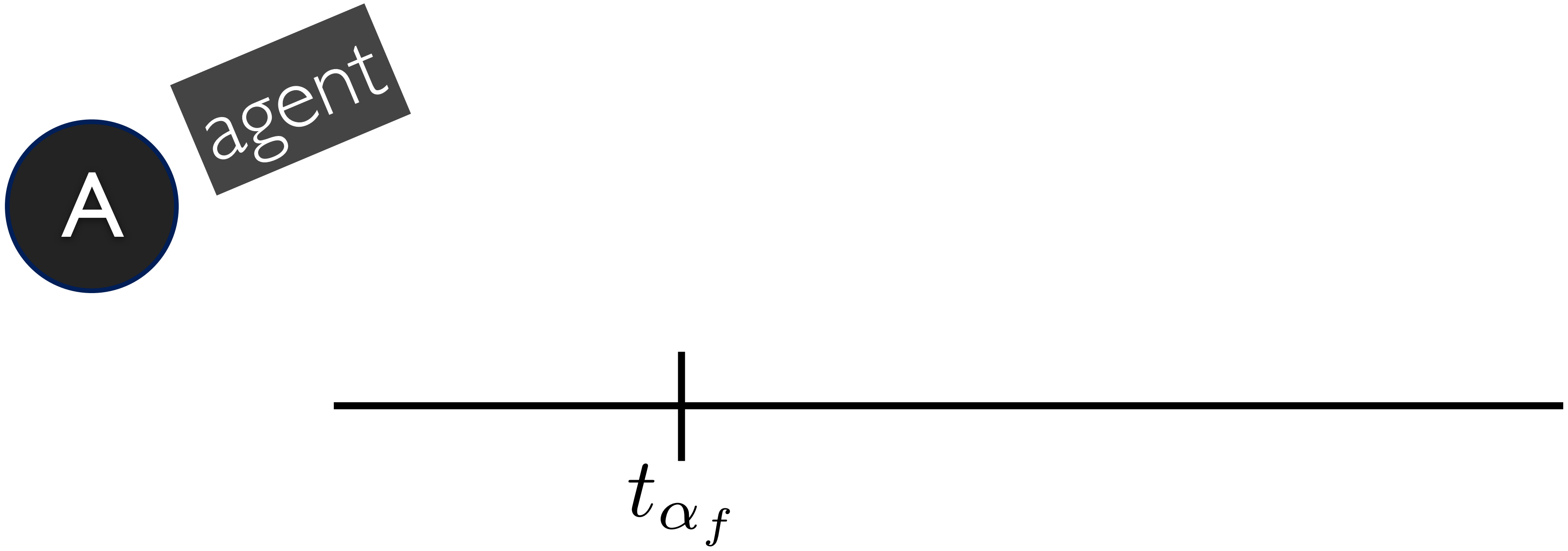
Informal Definition of Akrasia

A

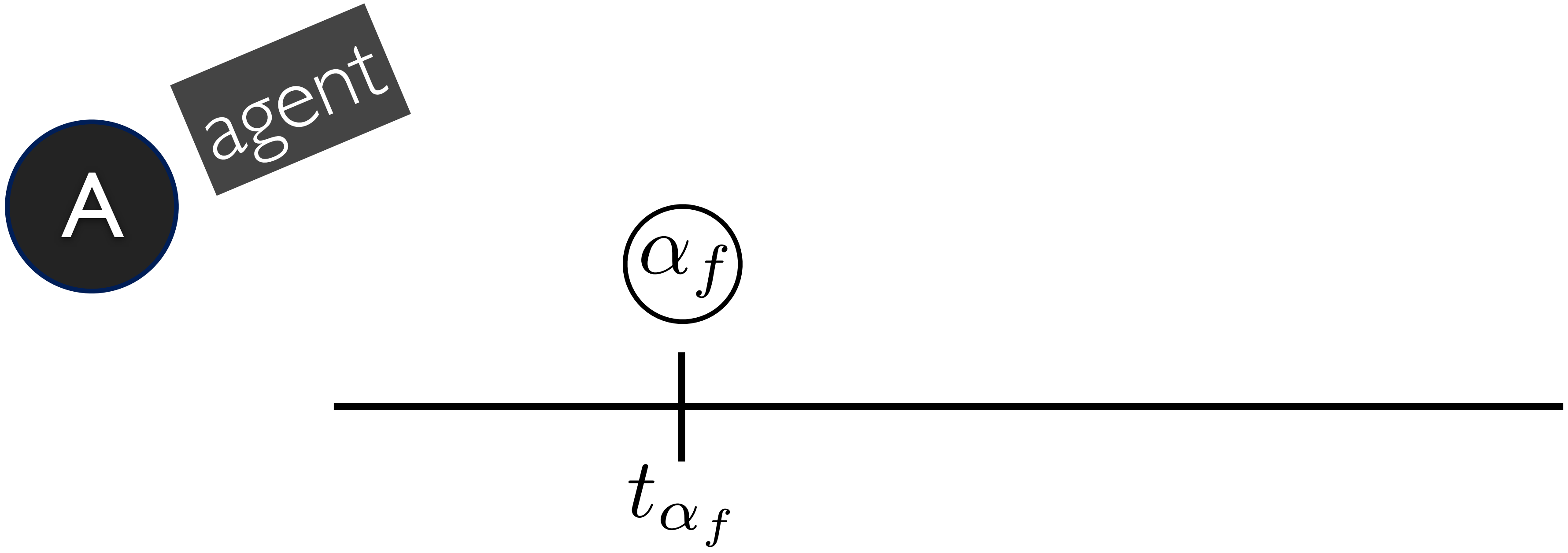
agent

Informal Definition of Akrasia

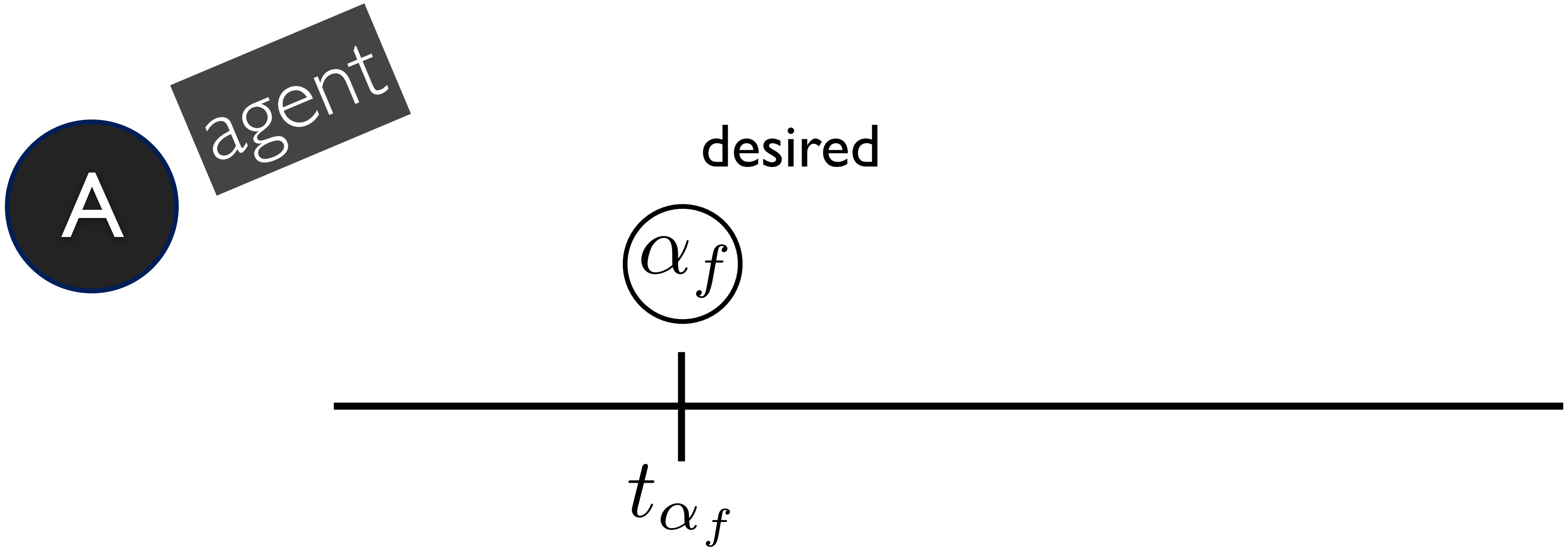
Informal Definition of Akrasia



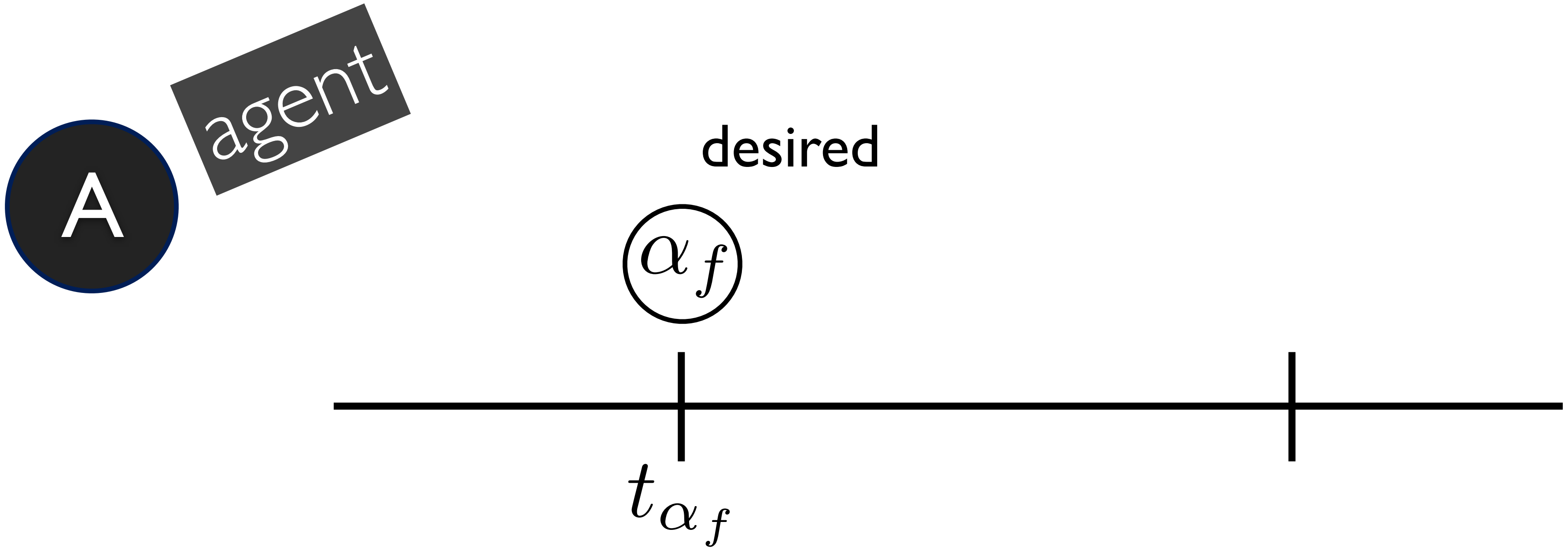
Informal Definition of Akrasia



Informal Definition of Akrasia



Informal Definition of Akrasia



A

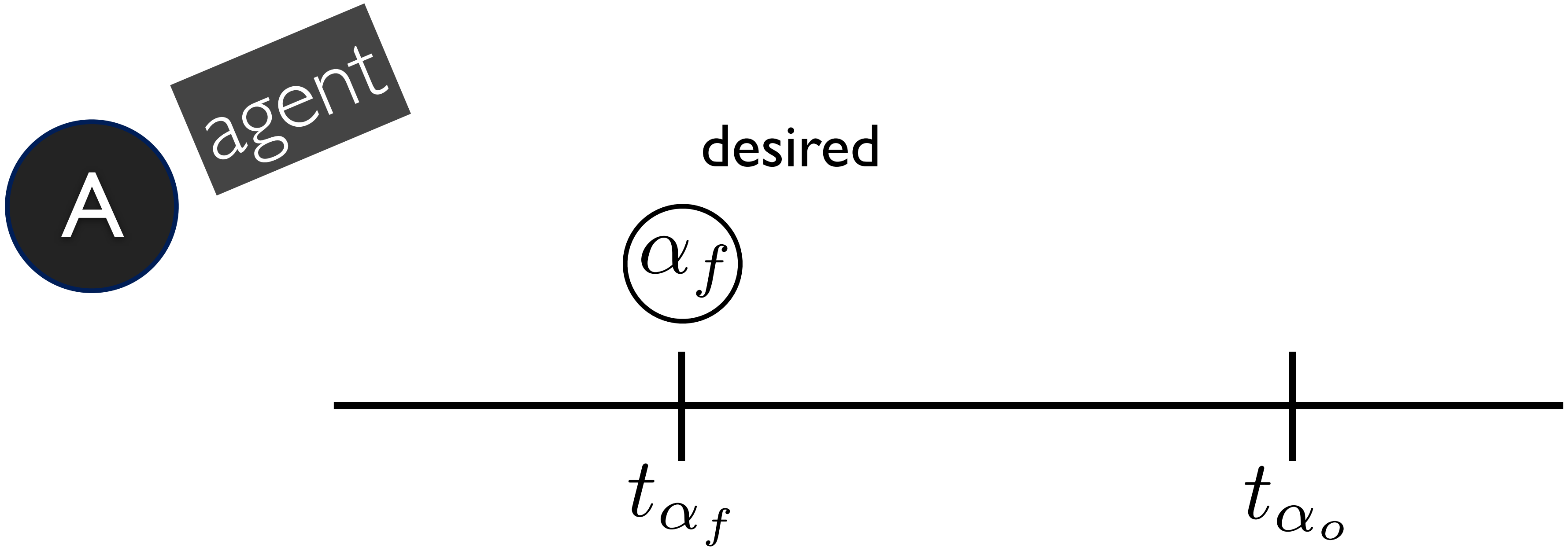
agent

desired

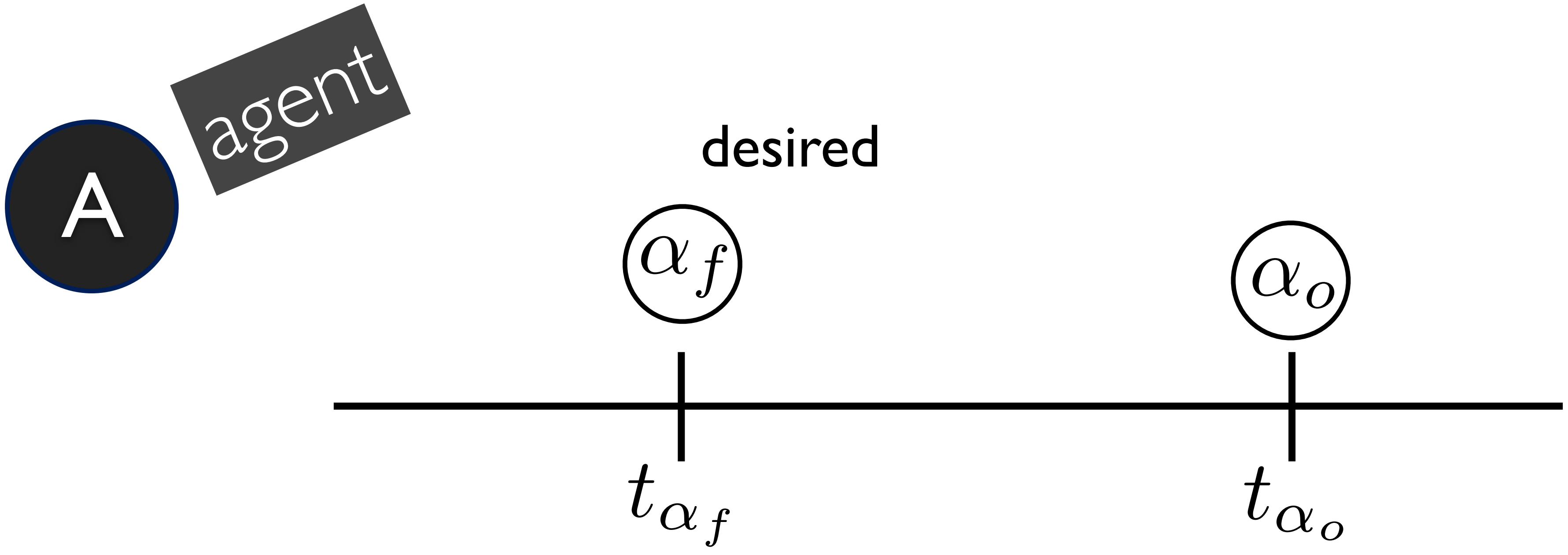
α_f

t_{α_f}

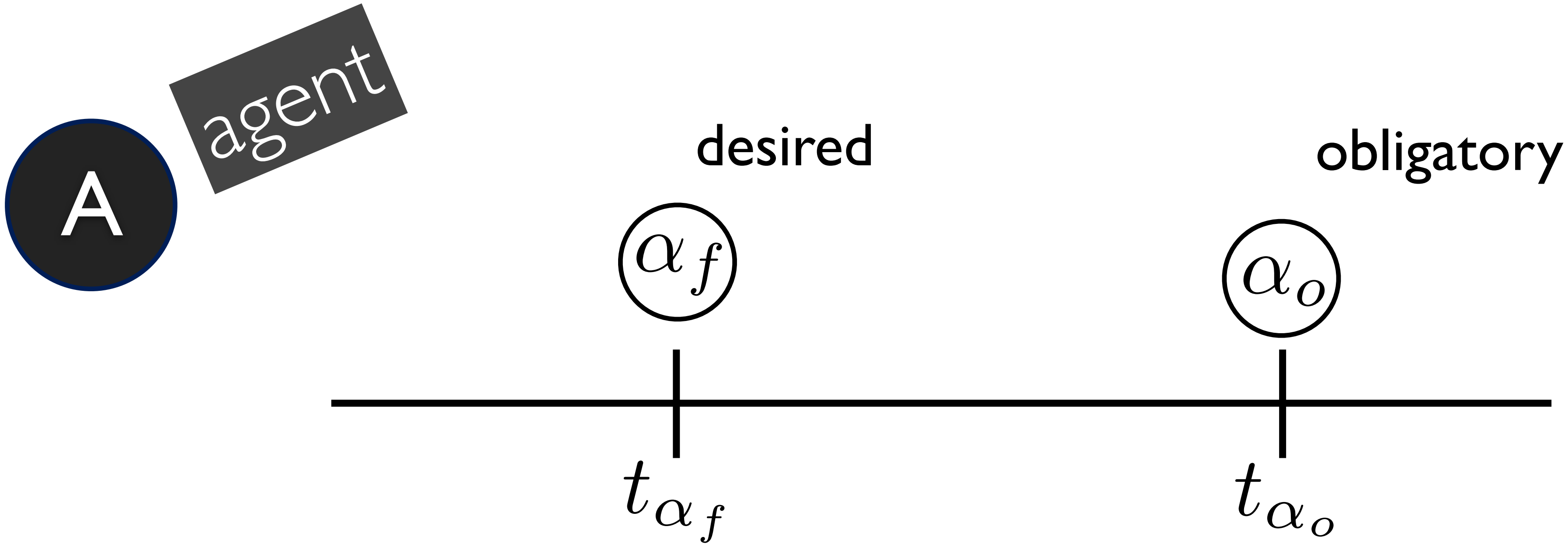
Informal Definition of Akrasia



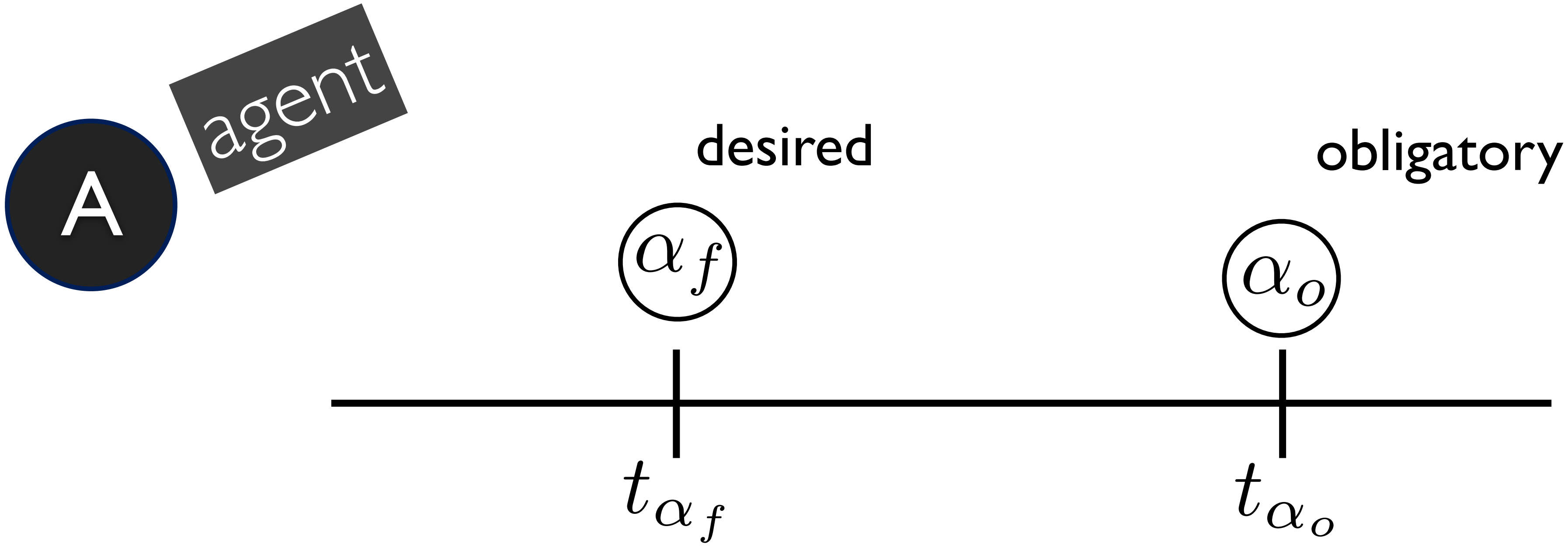
Informal Definition of Akrasia



Informal Definition of Akrasia

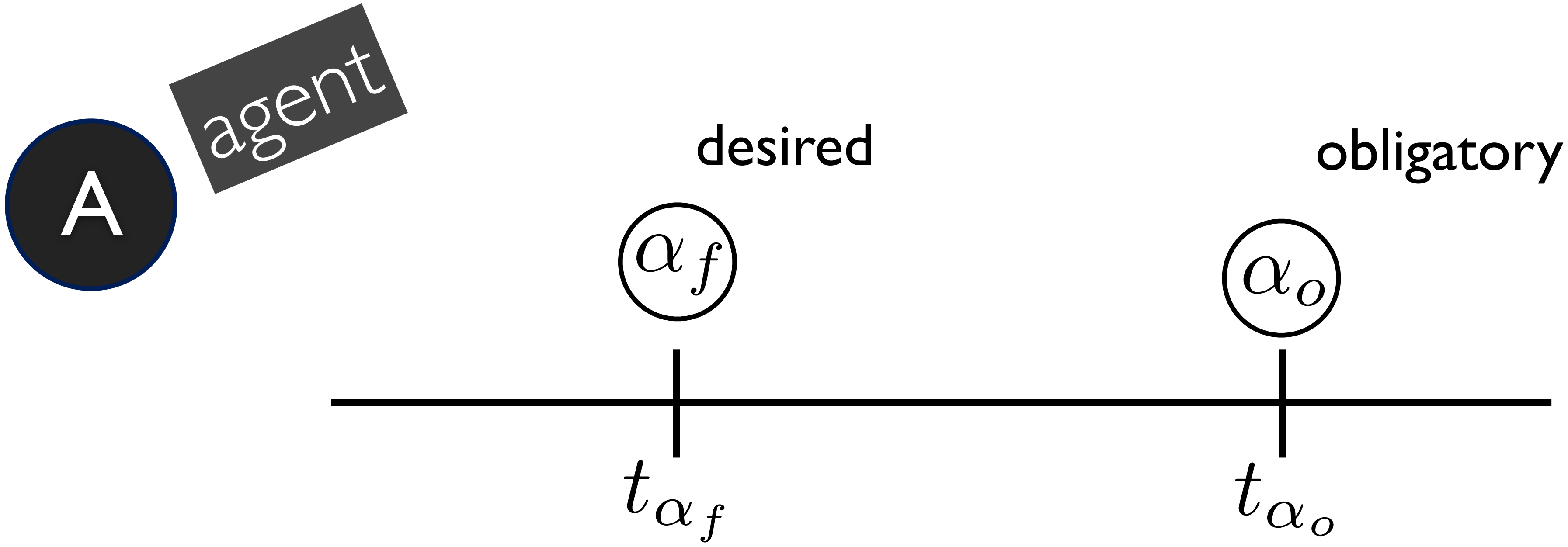


Informal Definition of Akrasia



If α_f happens, then α_o can't happen

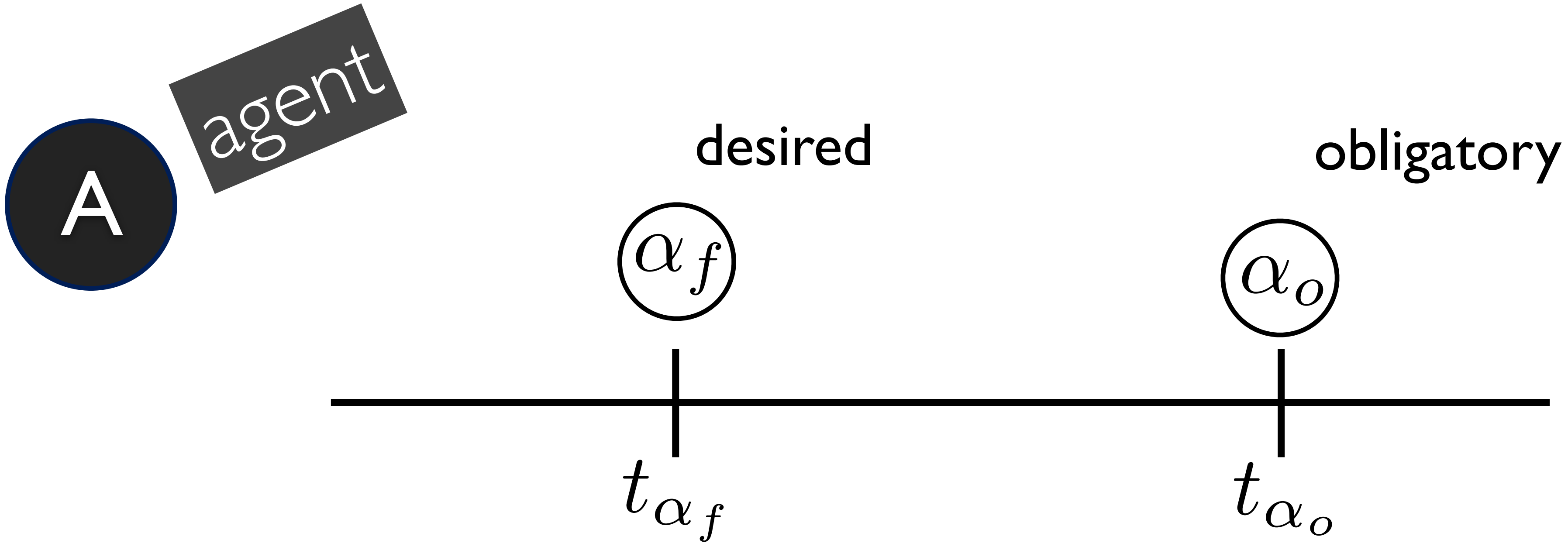
Informal Definition of Akrasia



If α_f happens, then α_o can’t happen

A

Informal Definition of Akrasia



If α_f happens, then α_o can’t happen

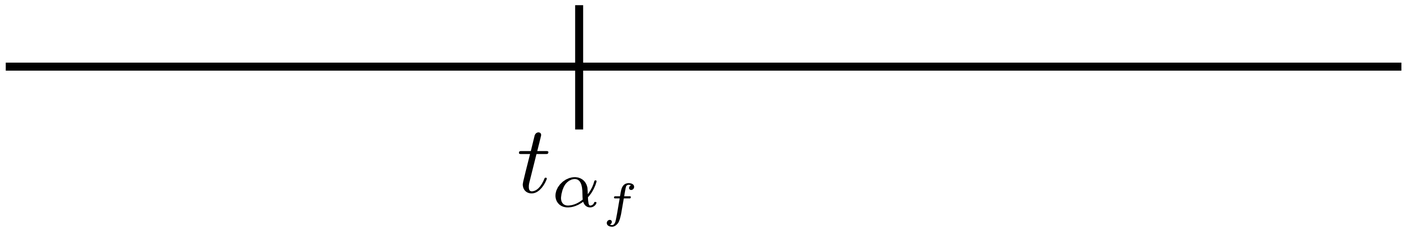
A knows this

Informal Definition of Akrasia

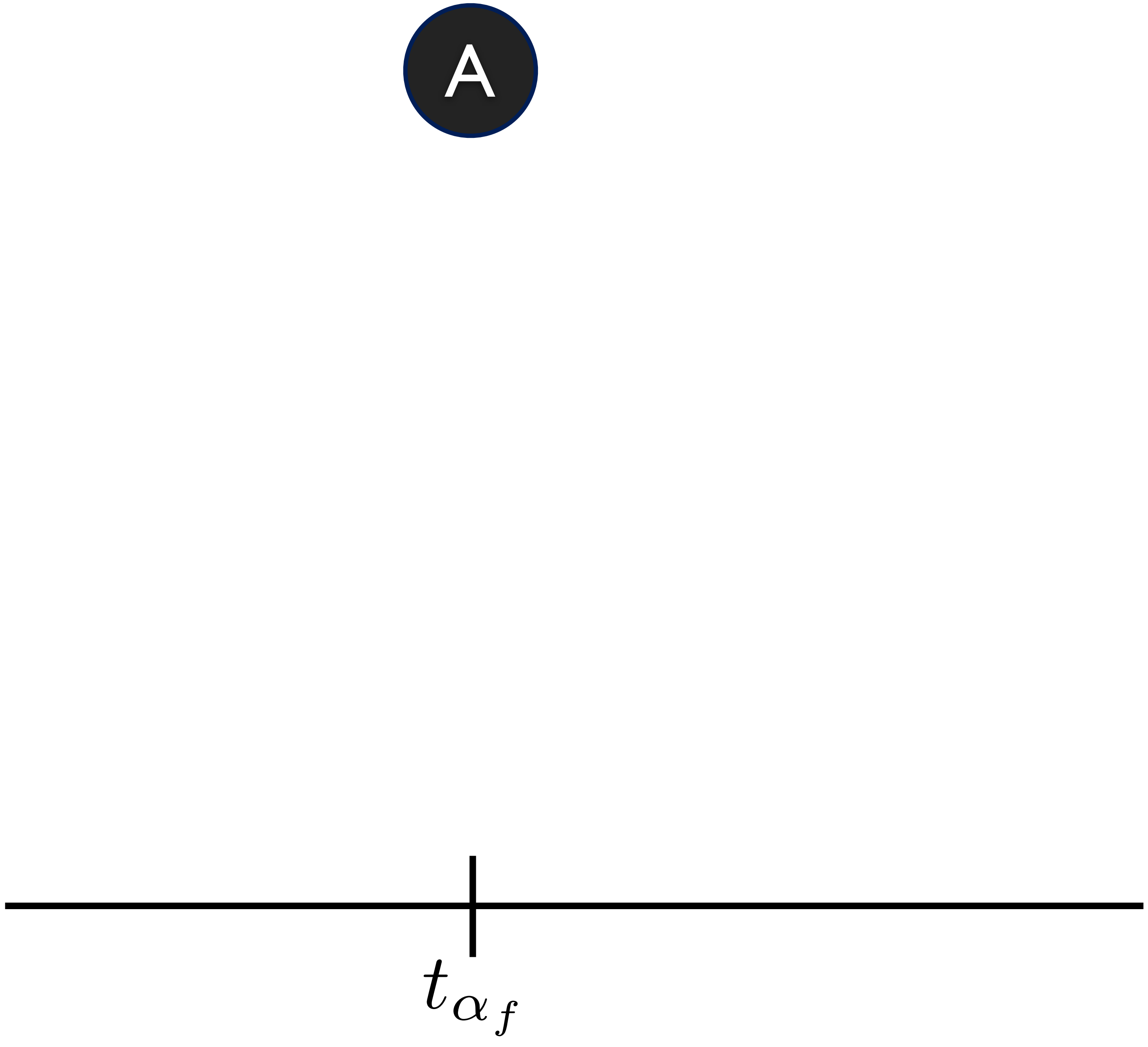
Informal Definition of Akrasia

Informal Definition of Akrasia

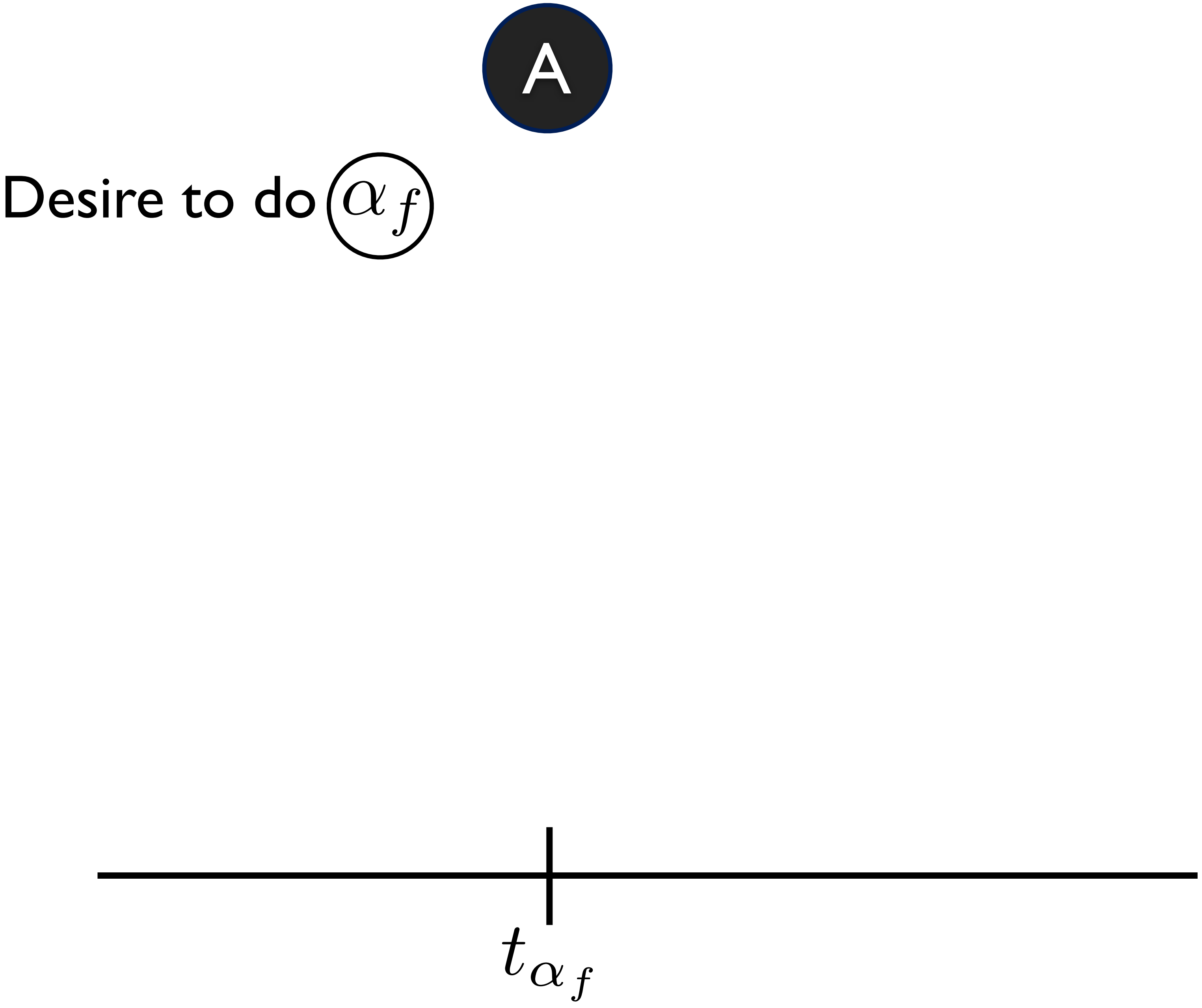
Informal Definition of Akrasia



Informal Definition of Akrasia



Informal Definition of Akrasia

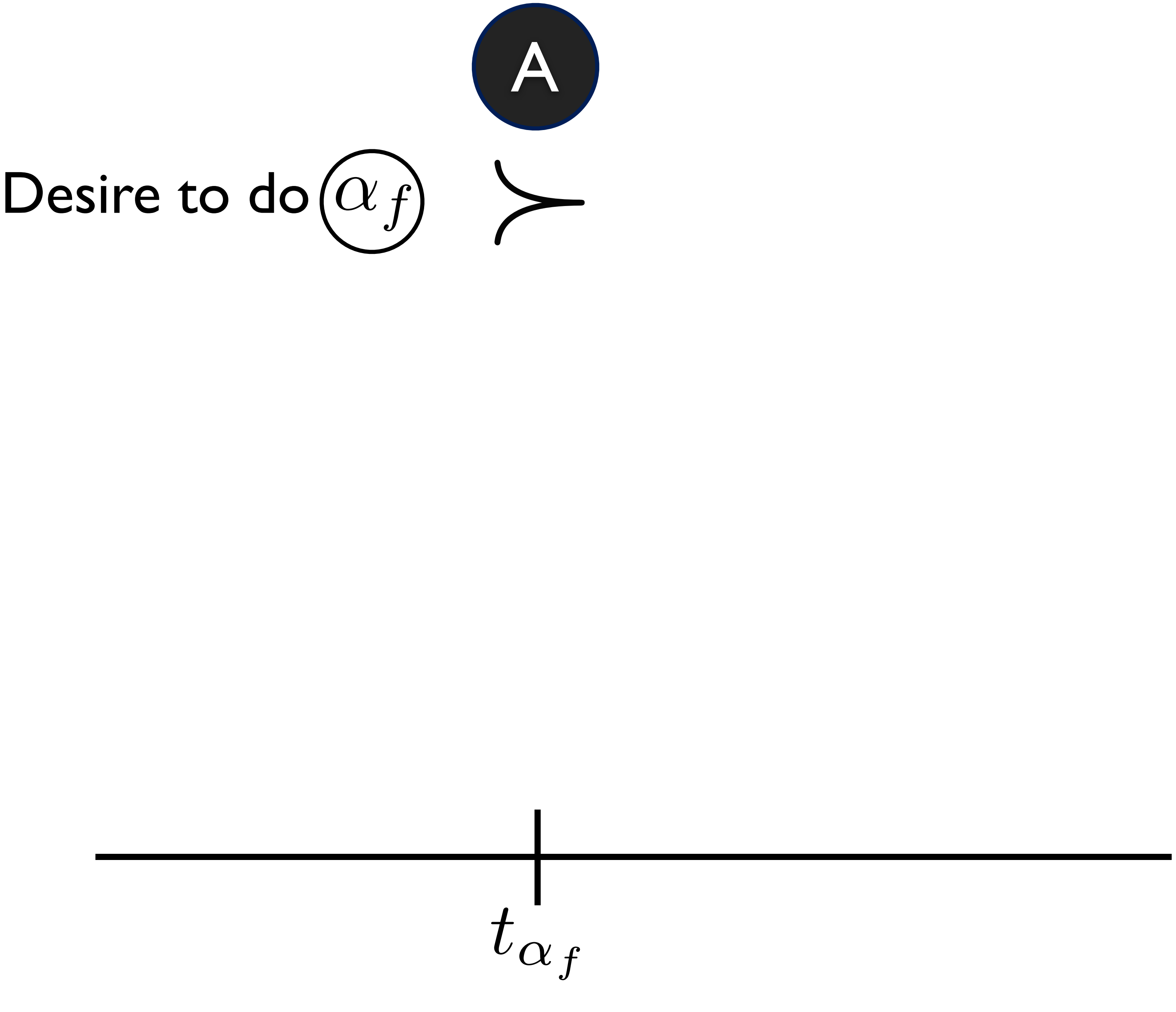


A

Desire to do α_f

t_{α_f}

Informal Definition of Akrasia

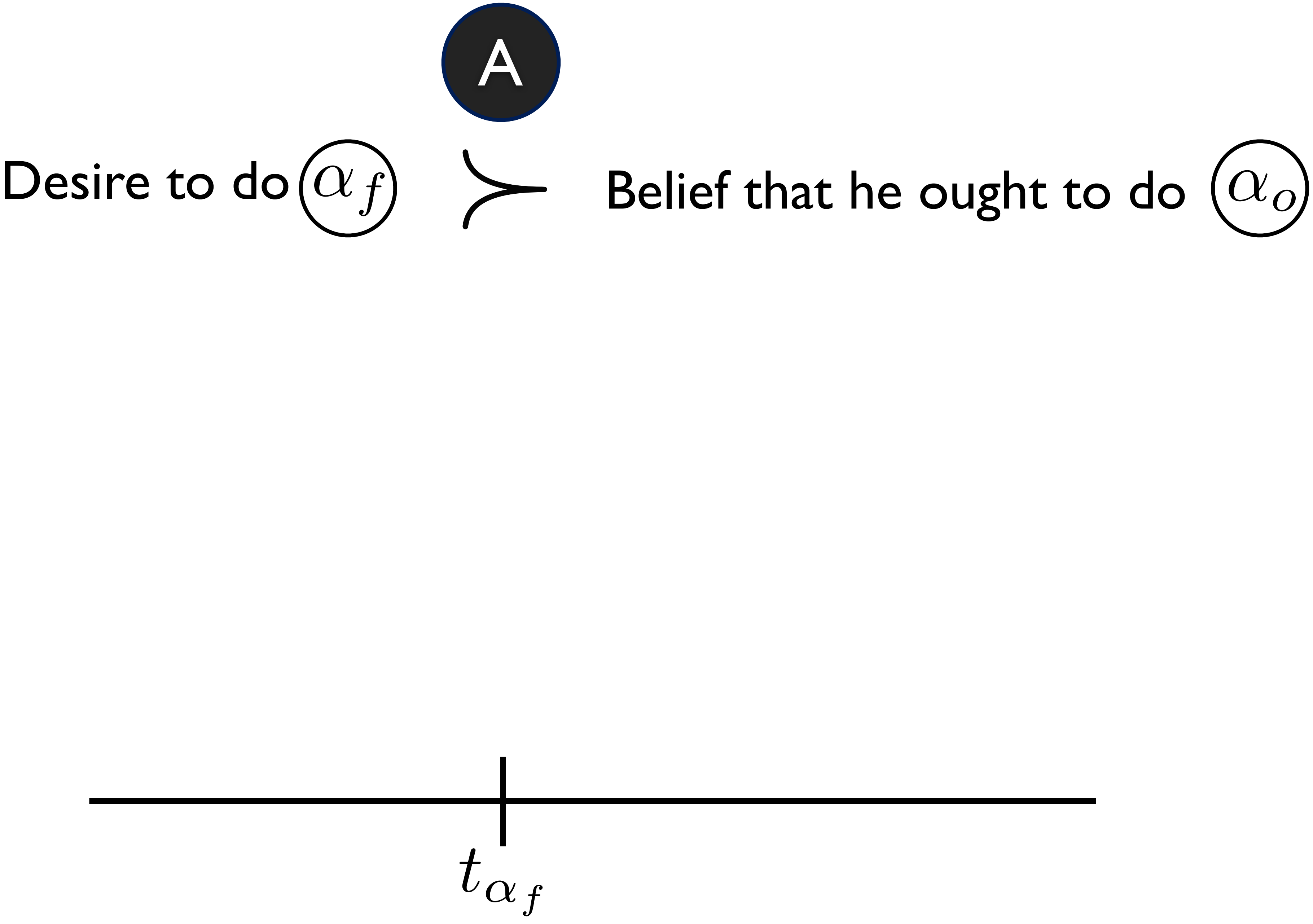


A

Desire to do α_f $\succ$

t_{α_f}

Informal Definition of Akrasia



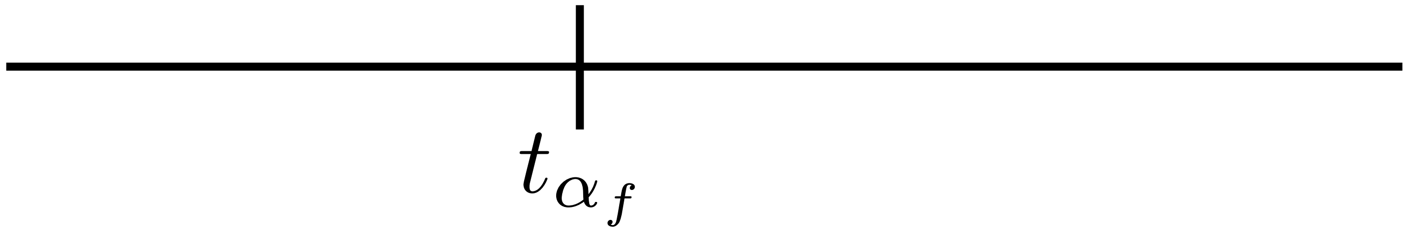
Desire to do $\boxed{\alpha_f}$ $\succ$ Belief that he ought to do $\boxed{\alpha_o}$

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

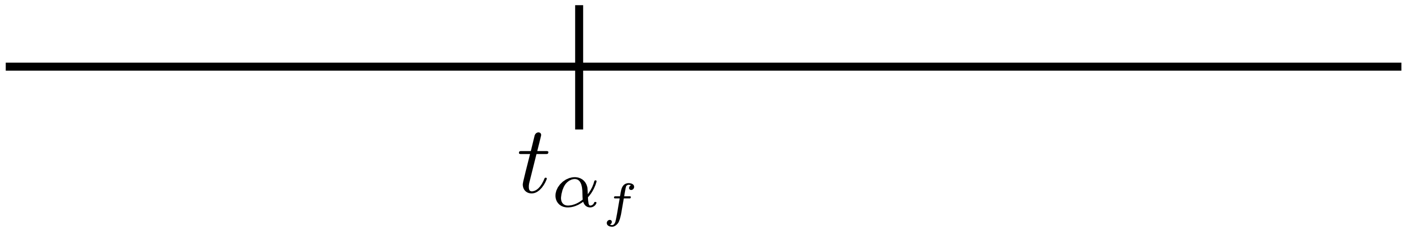


Informal Definition of Akrasia

A

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

A does α_f due to his desire

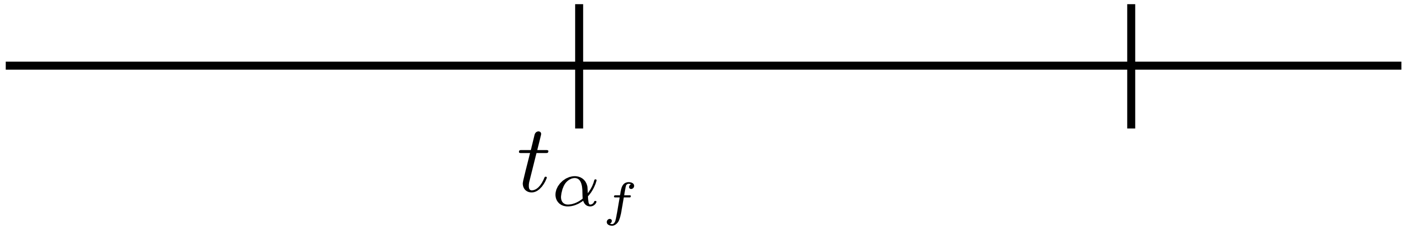


Informal Definition of Akrasia

A

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

A does α_f due to his desire

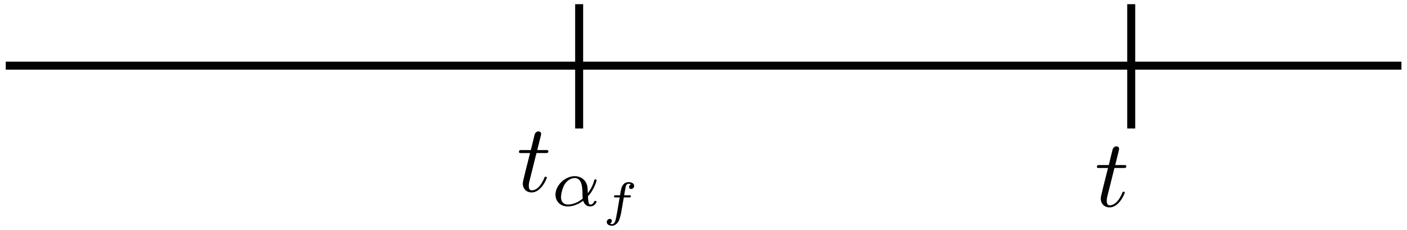


Informal Definition of Akrasia

A

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

A does α_f due to his desire



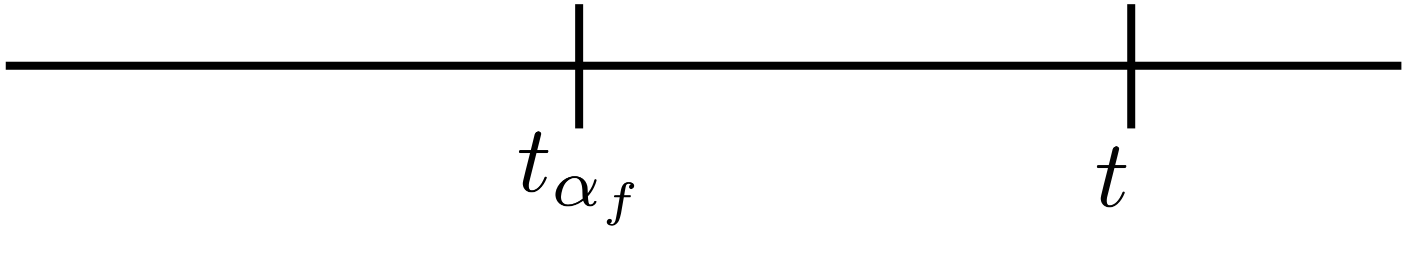
Informal Definition of Akrasia

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 …

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

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

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

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

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

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

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

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

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

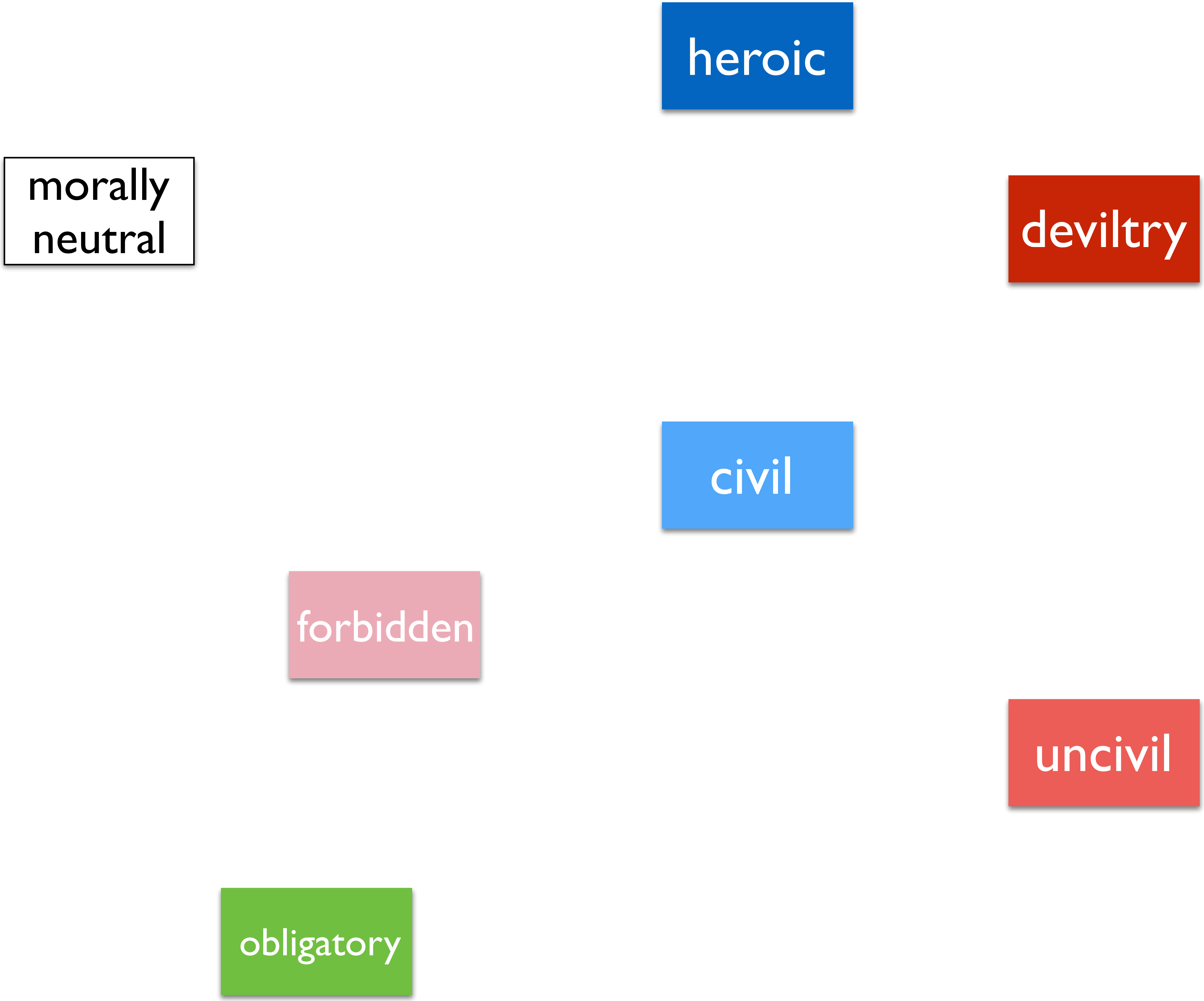
$$D_8 : \mathbf{B}\left(\mathsf{I}, t_f, \mathbf{O}(\mathsf{I}^*, t_\alpha, \Phi, happens(action(\mathsf{I}^*, \alpha), t_\alpha))\right)$$

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

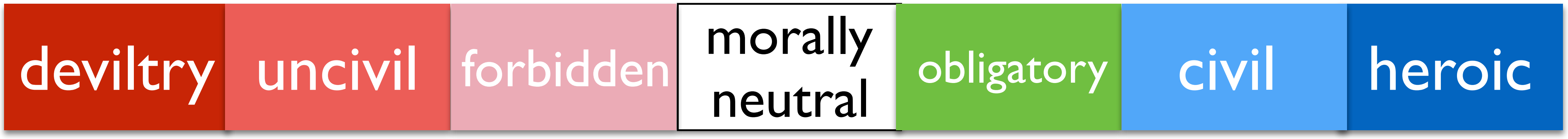
civil

forbidden

uncivil

obligatory

Leibnizian Ethical Hierarchy for Persons and Robots: $\mathcal{EH}$


deviltry

uncivil

forbidden

morally neutral

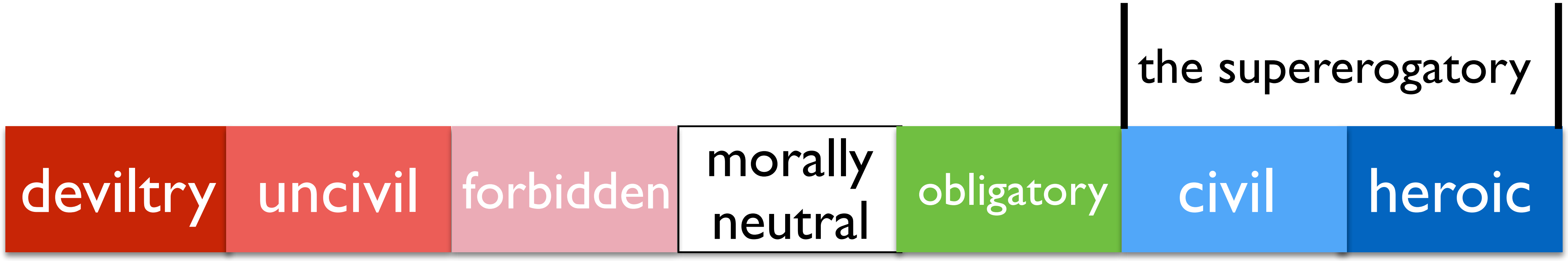
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



the supererogatory

deviltry

uncivil

forbidden

morally neutral

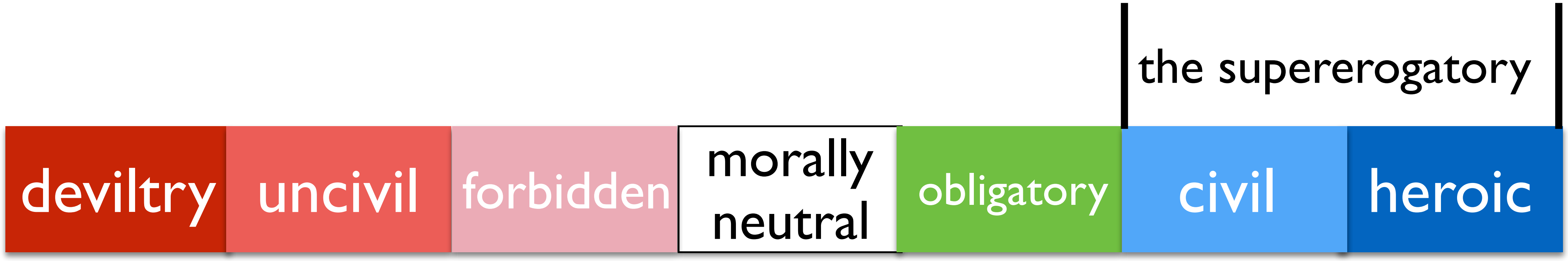
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathscr{EH}$

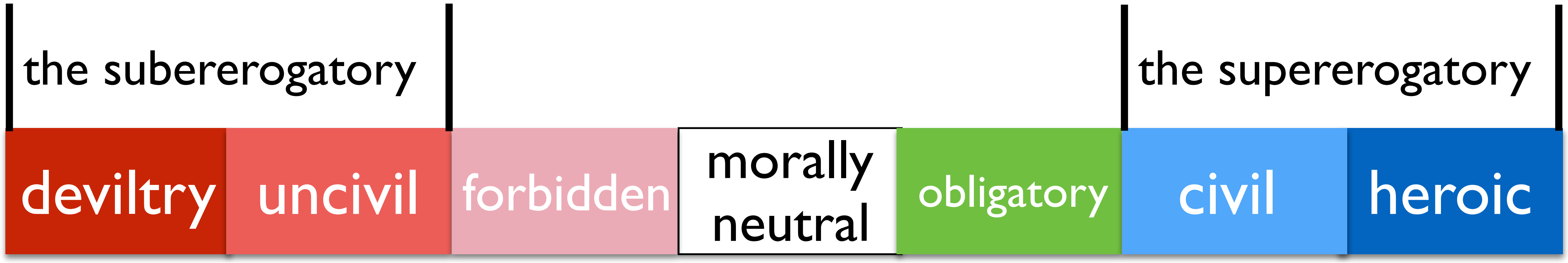


the supererogatory

devilty uncivil forbidden morally neutral obligatory civil heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



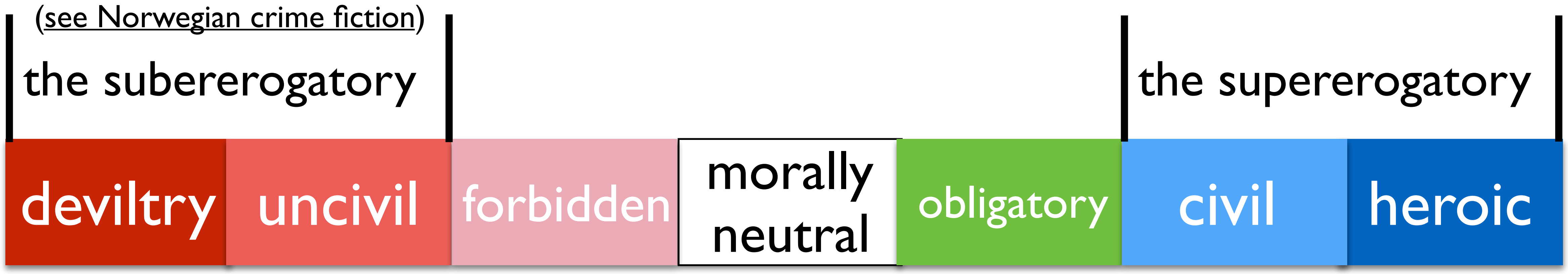
the subererogatory

the supererogatory

devilitry 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

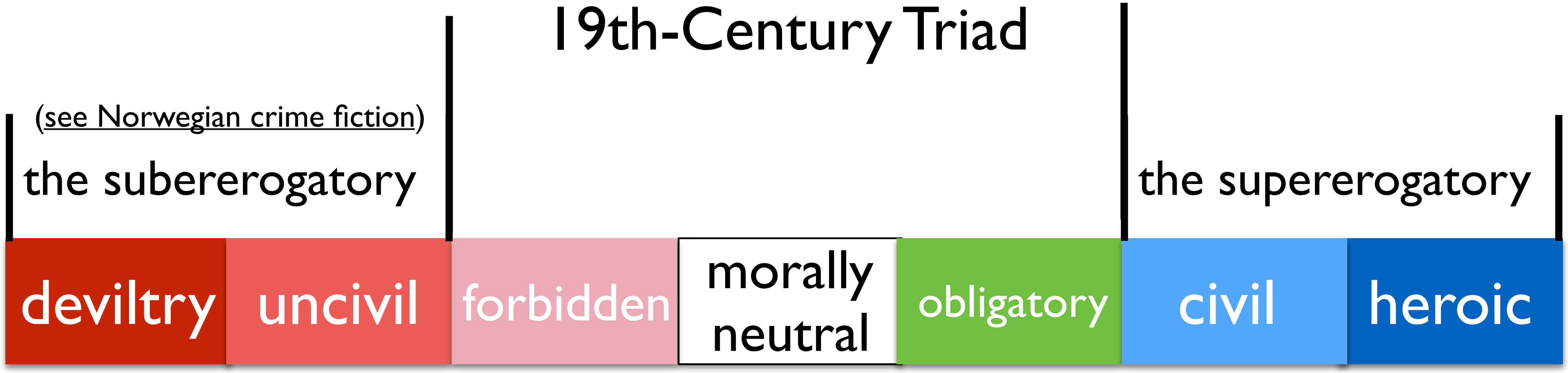
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



19th-Century Triad

(see Norwegian crime fiction)

the subererogatory

the supererogatory

devilry

uncivil

forbidden

morally neutral

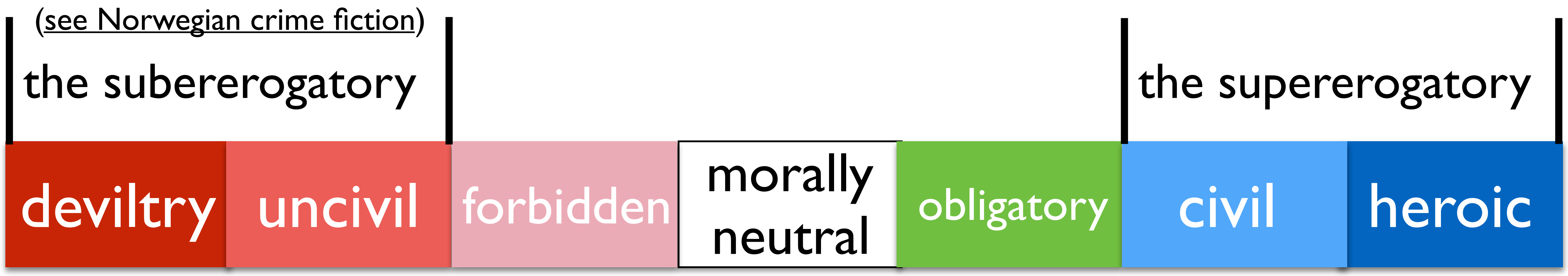
obligatory

civil

heroic

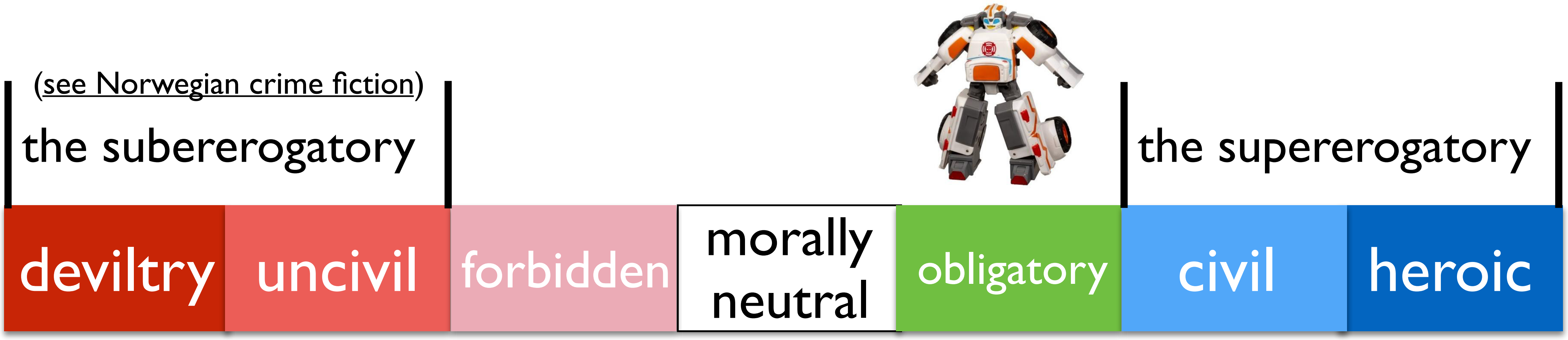
Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathscr{EH}$



(see Norwegian crime fiction)

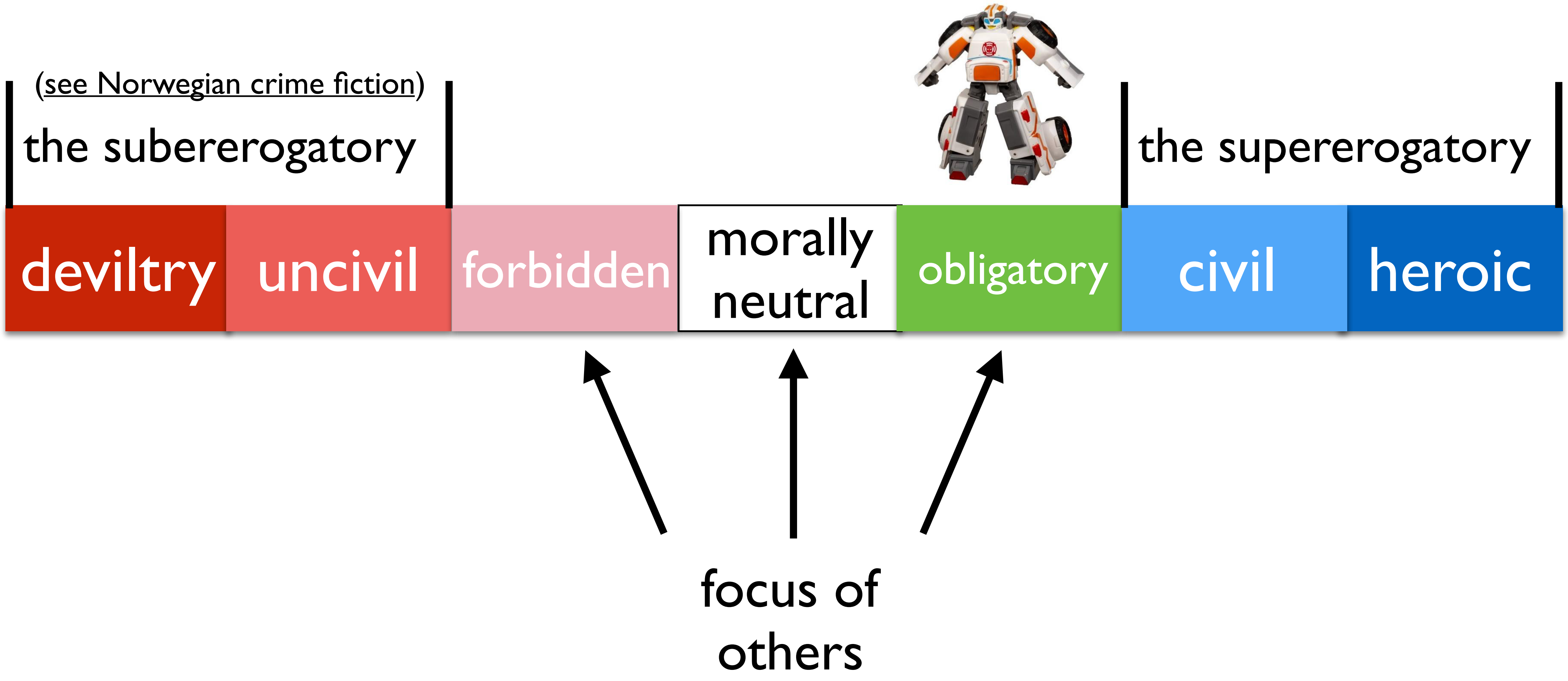
the subererogatory

the supererogatory

deviltry | uncivil | forbidden | morally neutral | obligatory | civil | heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



(see Norwegian crime fiction)

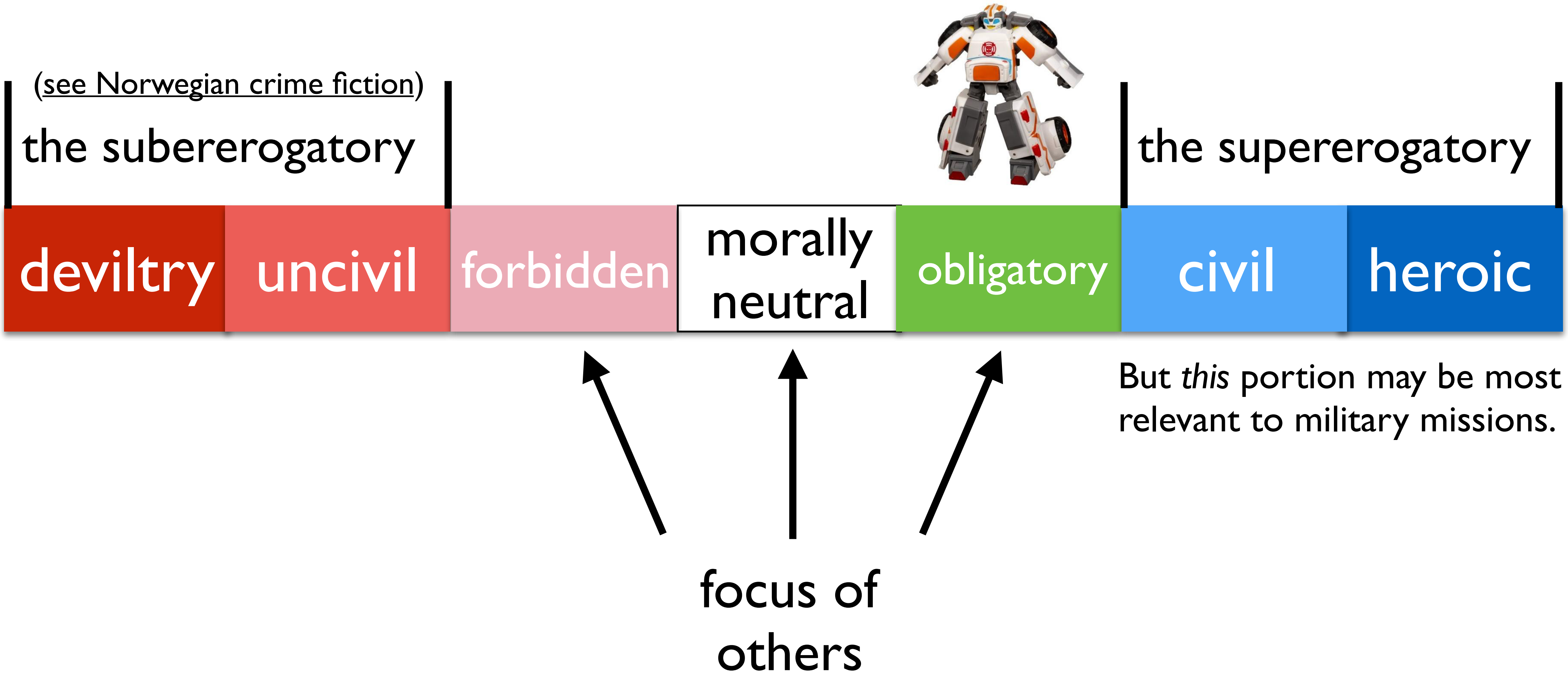
the subererogatory

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

$\mathscr{T} := \|\mathcal{F}|\mathcal{P} \wedge \neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$\mathscr{T} := \|\mathcal{F}|\mathcal{P} \wedge \neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$\mathscr{T} := \|\mathcal{F}|\mathcal{P} \wedge \neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$$\begin{array}{ccccc|c|ccccc} & & \mathcal{F} & & & \mathcal{P} \wedge \neg\mathcal{O} & & & \mathcal{O} & & \\ \forall & \mathtt{F} & \mathtt{M} & \mathtt{V} & \exists & & \forall & \mathtt{F} & \mathtt{M} & \mathtt{V} & \exists \end{array}$$

$\mathscr{T} := \|\mathcal{F}|\mathcal{P}\wedge\neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$$\begin{array}{c|c|c} \mathcal{F} & \mathcal{P}\wedge\neg\mathcal{O} & \mathcal{O} \\ \forall\quad \mathtt{F}\quad \mathtt{M}\quad \mathtt{V}\quad \exists & & \forall\quad \mathtt{F}\quad \mathtt{M}\quad \mathtt{V}\quad \exists \end{array}$$

$$\mathscr{EH}$$

$$\begin{array}{c|c|c|c|c|c|c|c} \mathcal{S}_{ub1} & \mathcal{S}_{ub2} & \mathcal{F} & \mathcal{P}\wedge\neg\mathcal{O} & \mathcal{O}^{L} & \mathcal{O}^{M} & \mathcal{S}^{up1} & \mathcal{S}^{up2} \\ \exists\text{–}\forall & \exists\text{–}\forall & \exists\text{–}\forall & & \exists\text{–}\forall & \exists\text{–}\forall & \exists\text{–}\forall & \exists\text{–}\forall \\ & & & & & & \quad\uparrow & \end{array}$$

$\mathscr{T} := \|\mathcal{F}|\mathcal{P} \wedge \neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}$
$\forall \quad \mathbf{F} \quad \mathbf{M} \quad \mathbf{V} \quad \exists$		$\forall \quad \mathbf{F} \quad \mathbf{M} \quad \mathbf{V} \quad \exists$

$\mathscr{EH}$

$\mathcal{S}_{ub1}$	$\mathcal{S}_{ub2}$	$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}^L$	$\mathcal{O}^M$	$\mathcal{S}^{up1}$	$\mathcal{S}^{up2}$
$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$		$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$
						$\uparrow$	
				●			

$\mathscr{T} := \|\mathcal{F}|\mathcal{P} \wedge \neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}$
$\forall$ **F** **M** **V** $\exists$		$\forall$ **F** **M** **V** $\exists$

$\mathscr{EH}$

$\mathcal{S}_{ub1}$	$\mathcal{S}_{ub2}$	$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}^L$	$\mathcal{O}^M$	$\mathcal{S}^{up1}$	$\mathcal{S}^{up2}$
∃–∀	∃–∀	∃–∀		∃–∀	∃–∀	∃–∀	∃–∀
						↑	
				●			

Arkin
Pereira
Andersons
Powers
Mikhail
…

$\mathscr{T} := \|\mathcal{F}|\mathcal{P}\wedge\neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$\mathcal{F}$					$\mathcal{P}\wedge\neg\mathcal{O}$	$\mathcal{O}$				
$\forall$	**F**	**M**	**V**	$\exists$		$\forall$	**F**	**M**	**V**	$\exists$

$\mathscr{EH}$

$\mathcal{S}_{ub1}$	$\mathcal{S}_{ub2}$	$\mathcal{F}$	$\mathcal{P}\wedge\neg\mathcal{O}$	$\mathcal{O}^L$	$\mathcal{O}^M$	$\mathcal{S}^{up1}$	$\mathcal{S}^{up2}$
$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$		$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$
				●		↑	

$\mathscr{T} := \|\mathcal{F}|\mathcal{P} \wedge \neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}$
$\forall$ **F** **M** **V** $\exists$		$\forall$ **F** **M** **V** $\exists$

$\mathscr{EH}$

$\mathcal{S}_{ub1}$	$\mathcal{S}_{ub2}$	$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}^L$	$\mathcal{O}^M$	$\mathcal{S}^{up1}$	$\mathcal{S}^{up2}$
$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$		$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$
				●		$\uparrow$	

RAIR
Rensselaer AI and Reasoning Lab

$\mathscr{T} := \|\mathcal{F}|\mathcal{P} \wedge \neg\mathcal{O}|\mathcal{O}\|$ 19th Century Triad

$\mathcal{F}$					$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}$				
$\forall$	**F**	**M**	**V**	$\exists$		$\forall$	**F**	**M**	**V**	$\exists$

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

$\mathcal{S}_{ub1}$	$\mathcal{S}_{ub2}$	$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}^L$	$\mathcal{O}^M$	$\mathcal{S}^{up1}$	$\mathcal{S}^{up2}$
$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$		$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$	$\exists$–$\forall$
				●		↑	

There are obviously a host of formulae whose theoremhood constitute desiderata; that is (to give but a pair), the following must be provable (where $n \in \{1, 2\}$):

Theorem 1. $\mathbf{S}^{\mathbf{up}^n}(\phi, a, \alpha) \rightarrow \neg\mathbf{O}(\phi, a, \alpha)$

Theorem 2. $\mathbf{S}^{\mathbf{up}^n}(\phi, a, \alpha) \rightarrow \neg\mathbf{F}(\phi, a, \alpha)$

Secondly, $\mathcal{L}_{\mathscr{E}\mathscr{H}}$ is an *in*ductive logic, not a deductive one. This must be the case, since, as we've noted, quantification isn't restricted to just the standard pair $\exists\forall$ of quantifiers in standard extensional n-order logic: $\mathscr{E}\mathscr{H}$ is based on three additional quantifiers. For example, while in standard

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

$$K\left(\text{nao}, t_1, \text{lessthan}\left(\text{payoff}\left(\text{nao}^*, \neg\text{dive}, t_2\right), \text{threshold}\right)\right)$$

$$K\left(\text{nao}, t_1, \text{greaterthan}\left(\text{payoff}\left(\text{nao}^*, \text{dive}, t_2\right), \text{threshold}\right)\right)$$

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

$$\therefore K\left(\text{nao}, t_1, S^{\text{UP2}}\left(\text{nao}, t_2, \text{happens}\left(\text{action}\left(\text{nao}^*, \text{dive}\right), t_2\right)\right)\right)$$

$$\therefore I\left(\text{nao}, t_2, \text{happens}\left(\text{action}\left(\text{nao}^*, \text{dive}\right), t_2\right)\right)$$

$$\therefore \text{happens}\left(\text{action}(\text{nao}, \text{dive}), t_2\right)$$

Courtesy of RAIR-Lab Researcher Atriya Sen

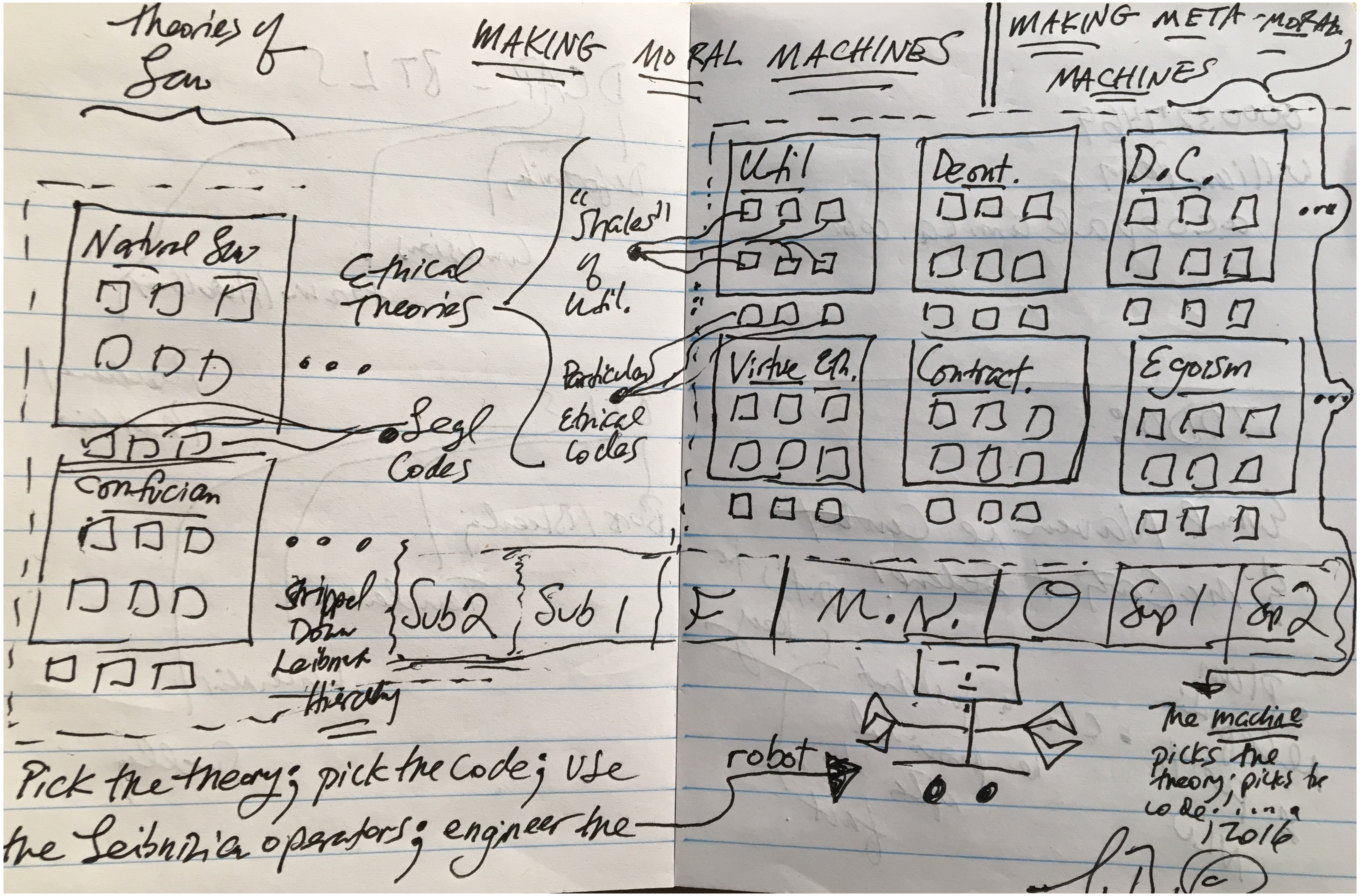
In Talos (available via Web interface); & ShadowProver

```
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

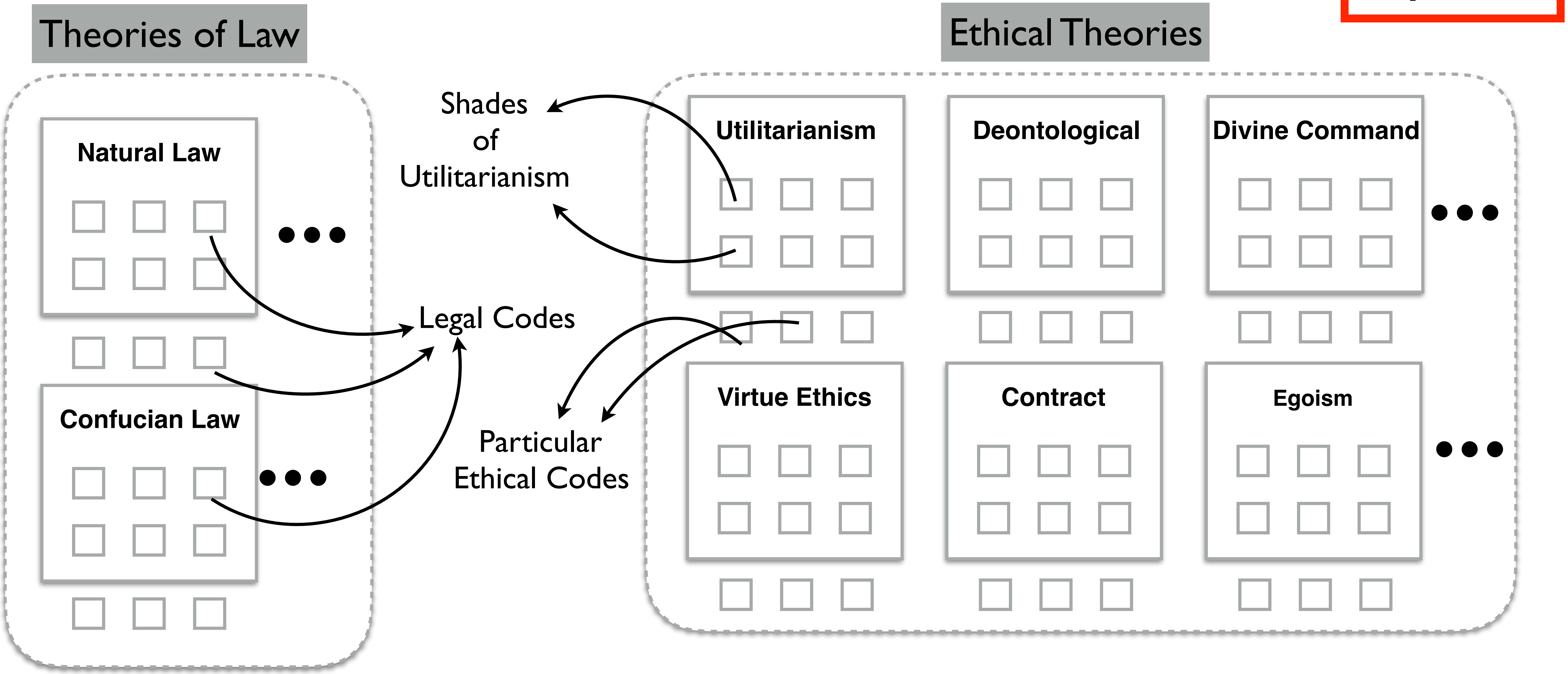
```
Prototypes:
Boolean lessThan Numeric Numeric
Boolean greaterThan Numeric Numeric
ActionType not ActionType
ActionType dive

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:



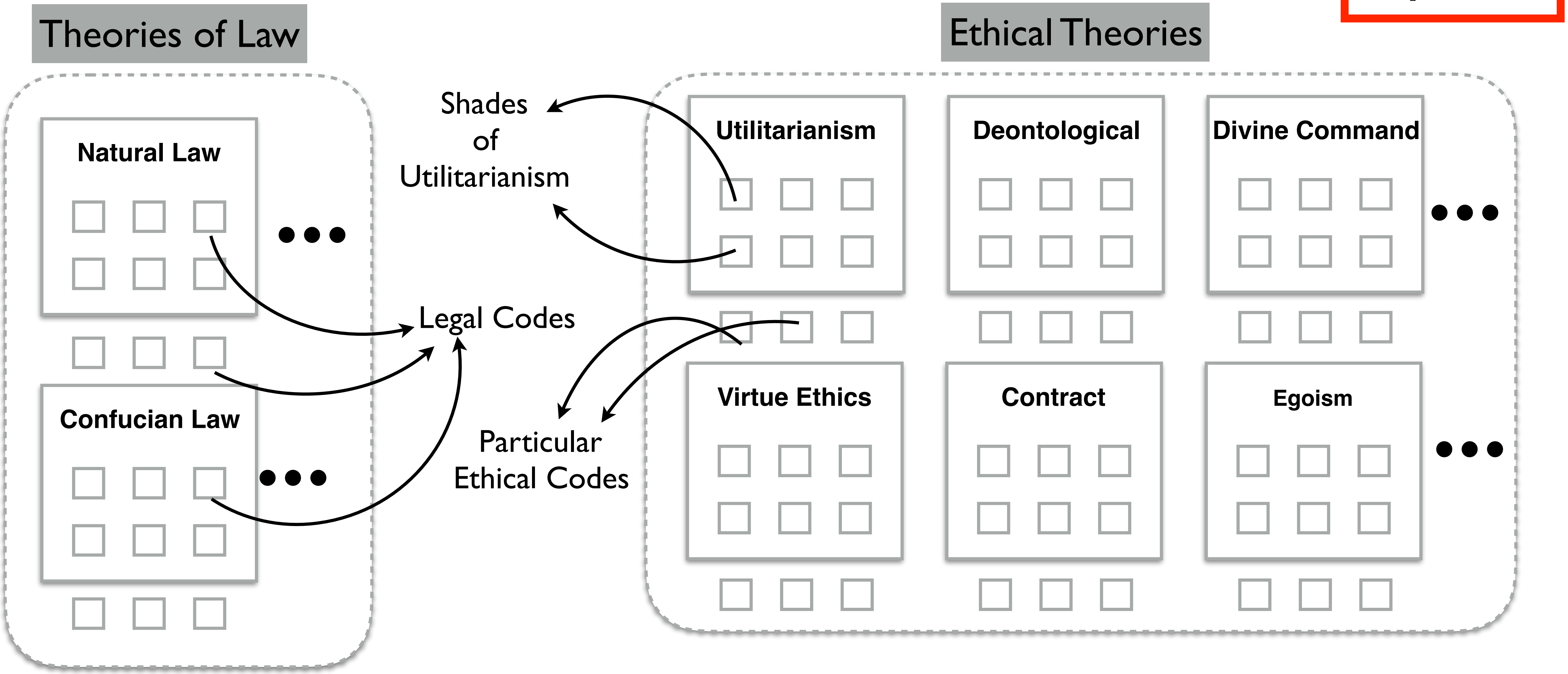
Making Morally X Machines, in Four Steps

~$10M



Making Morally X Machines, in Four Steps

~$10M



Theories of Law

Natural Law

Confucian Law

Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

Ethical Theories

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

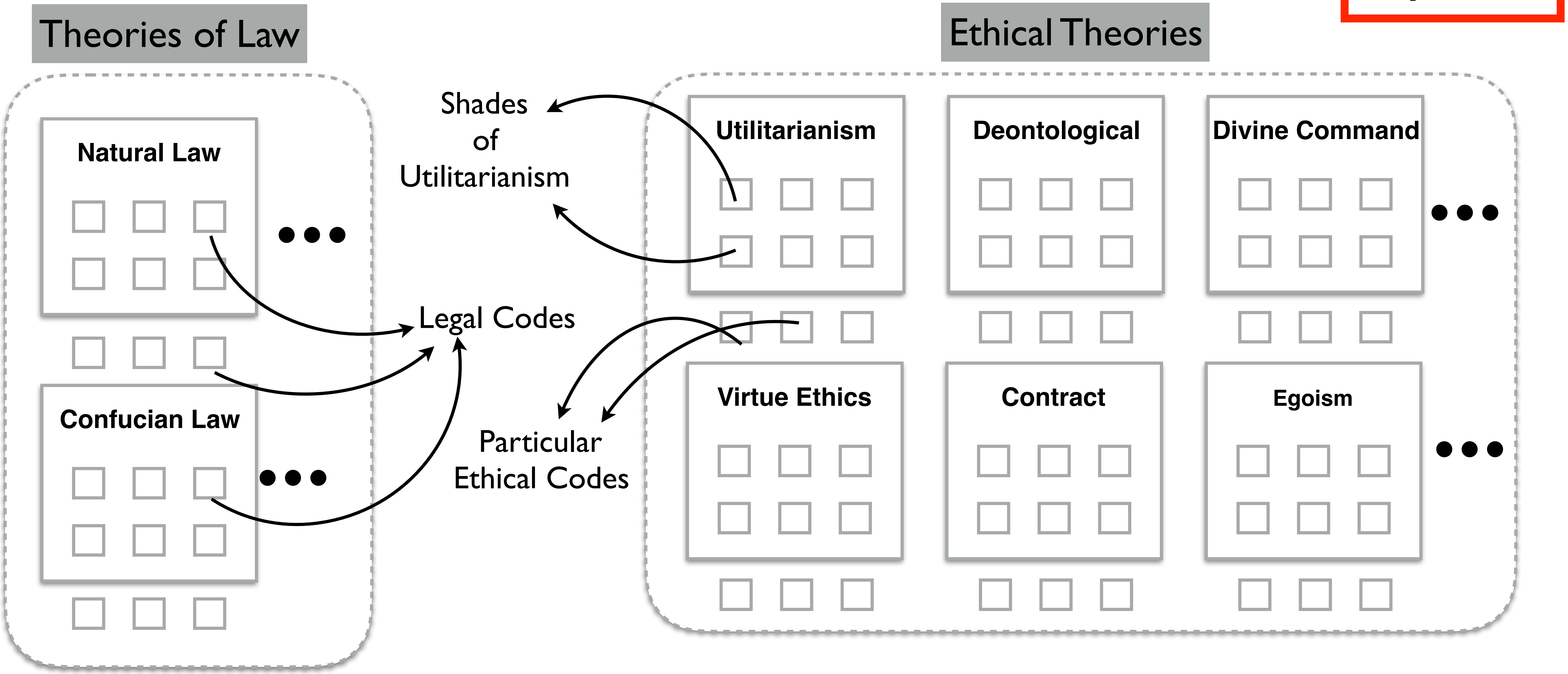
Egoism

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

~$10M

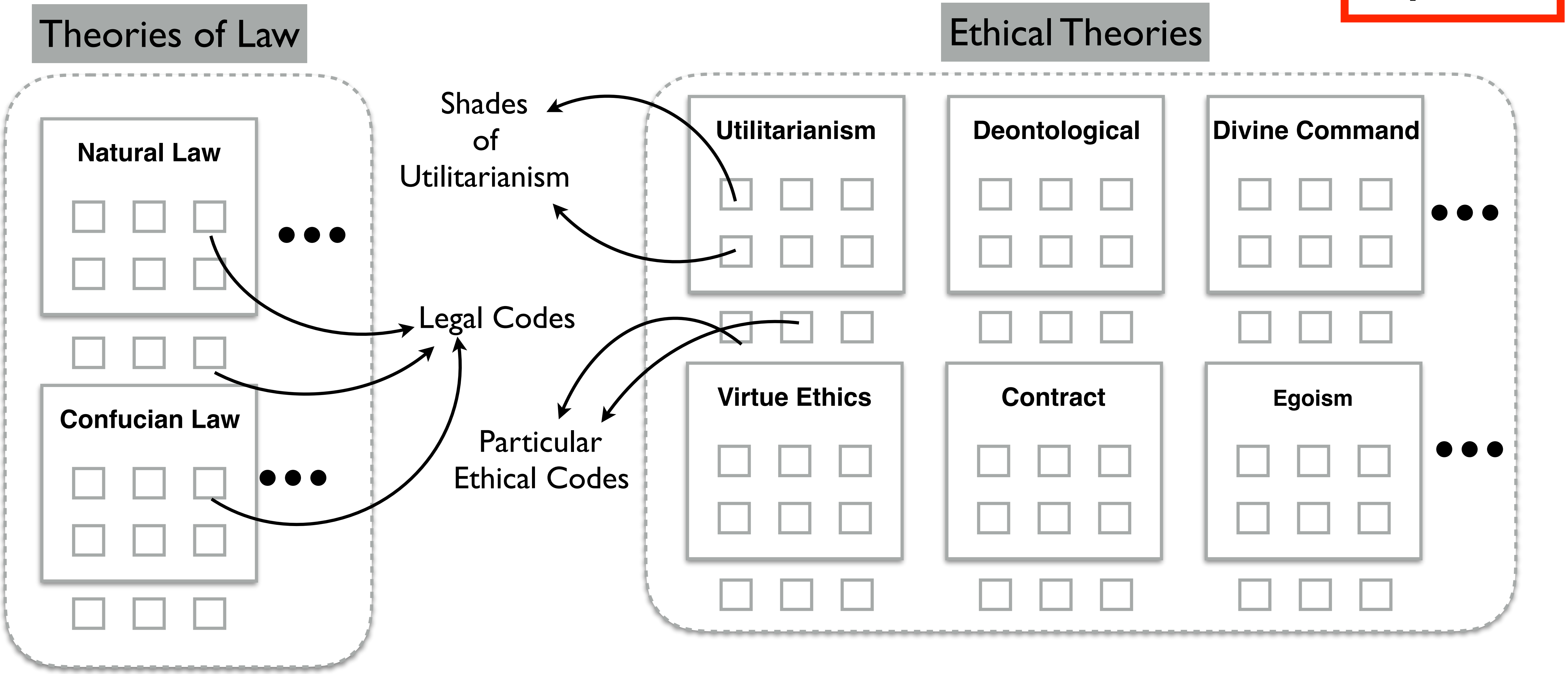


Step 1

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

~$10M



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

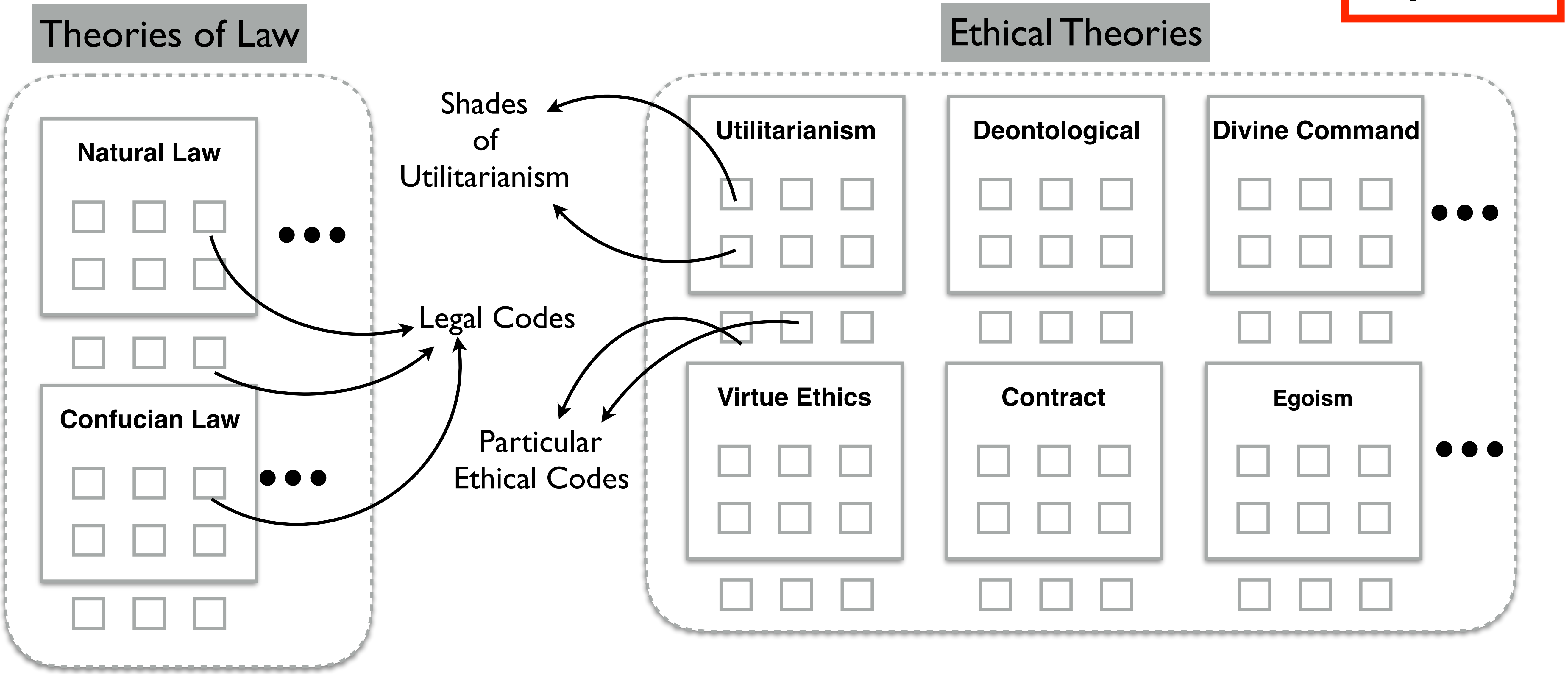
Formalize & Automate

Shadow Prover

Spectra

Making Morally X Machines, in Four Steps

~$10M



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

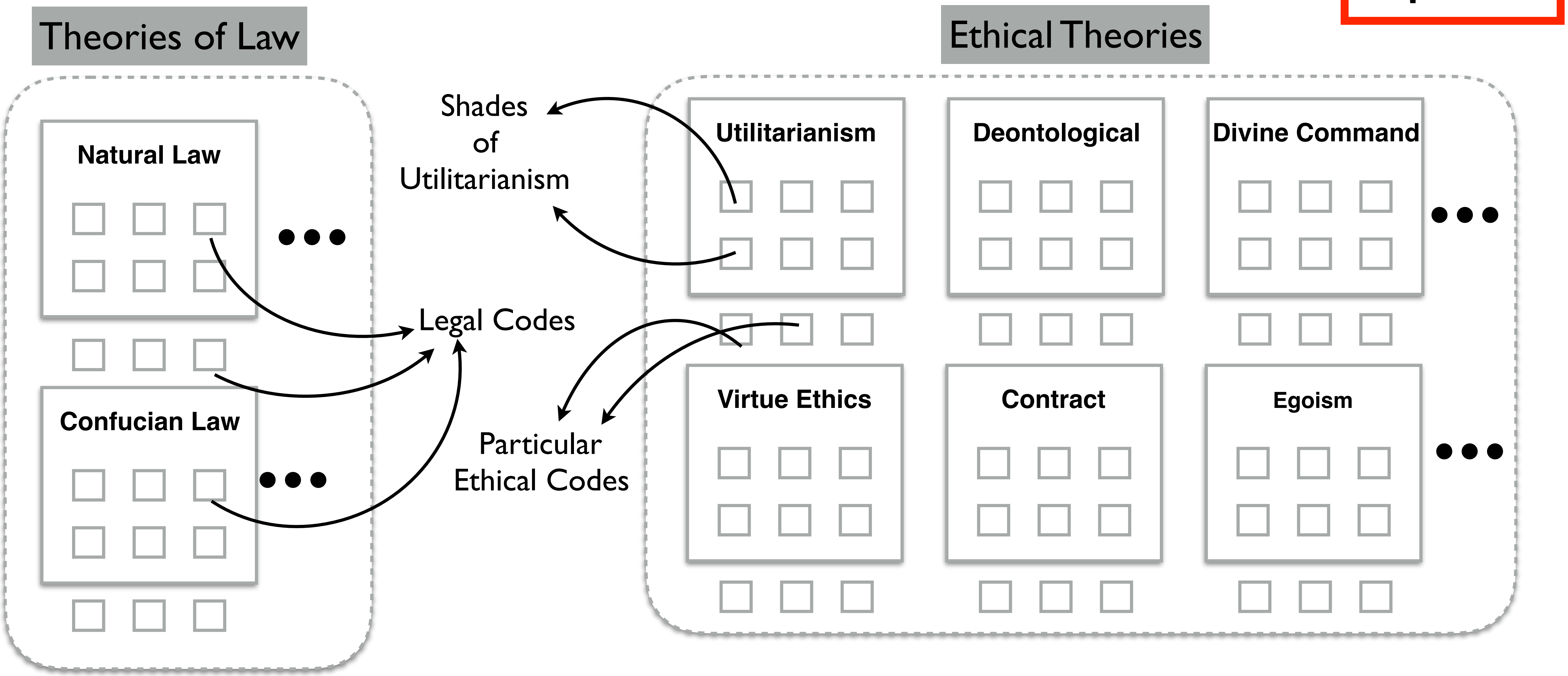
Formalize & Automate

Shadow Prover

Spectra

Making Morally X Machines, in Four Steps

~$10M



Step 1

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

Step 2

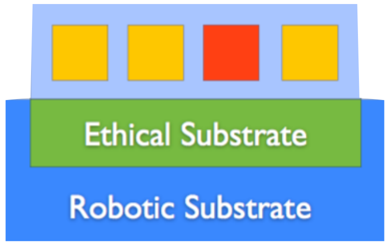
Formalize & Automate

Shadow Prover

Spectra

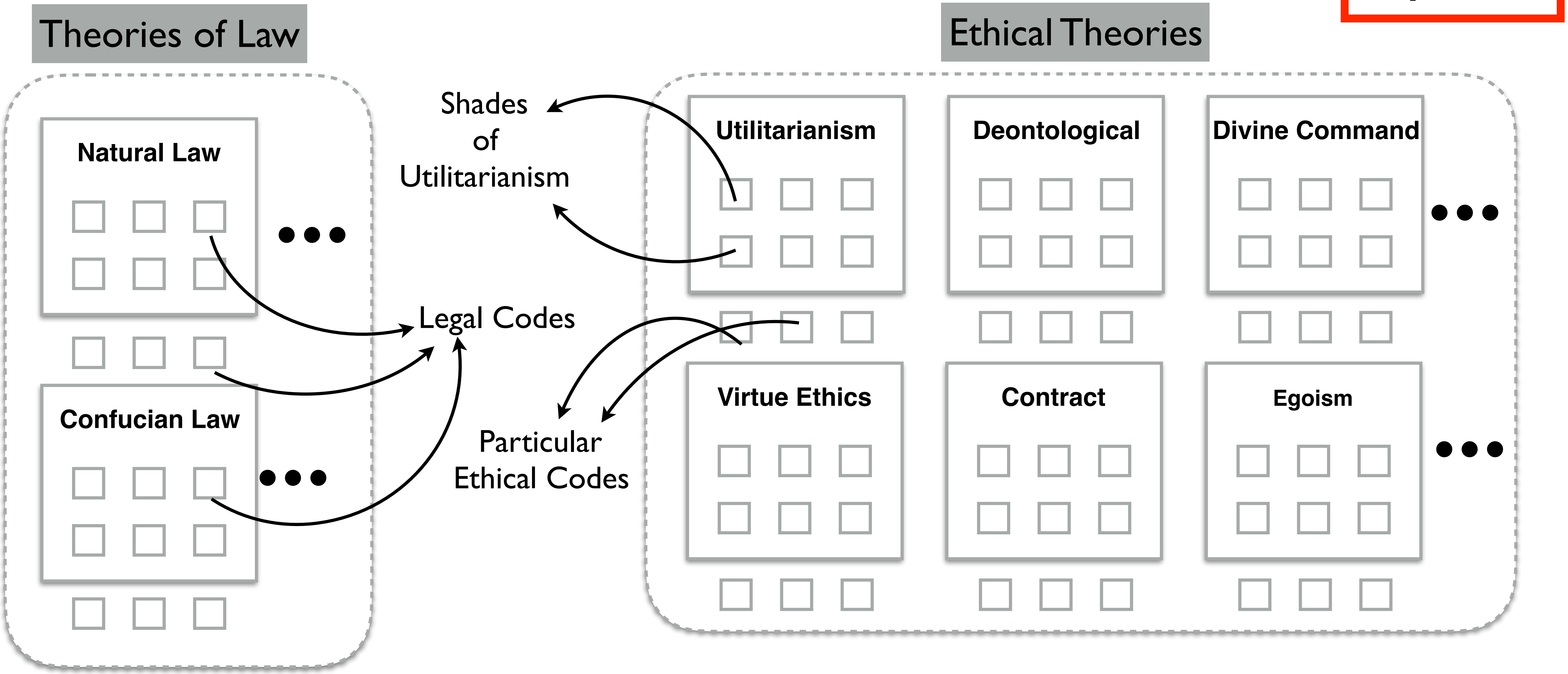
Step 3

Ethical OS



Making Morally X Machines, in Four Steps

~$10M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

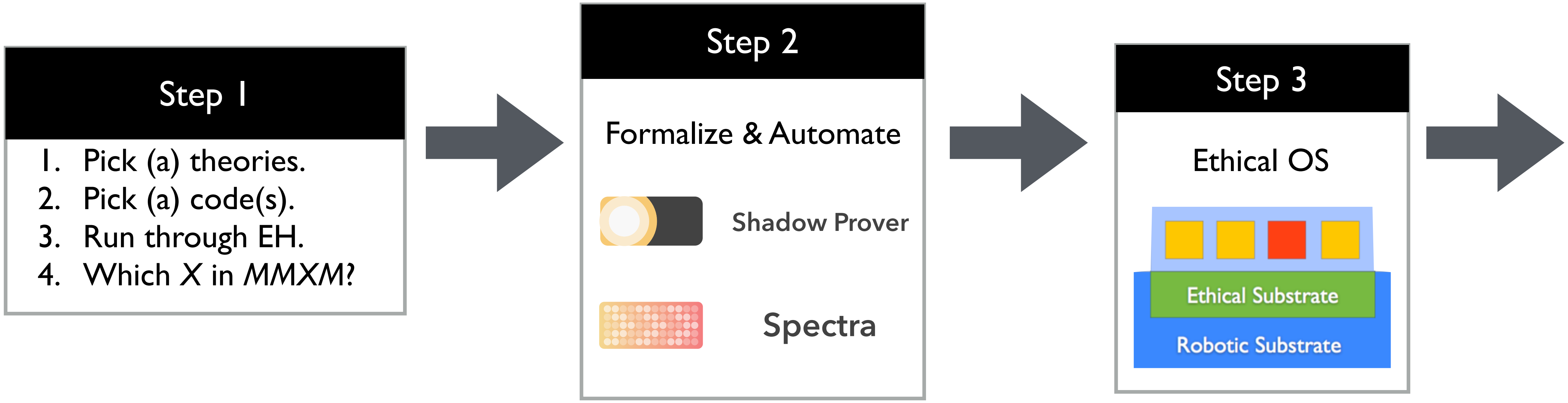
Deontological

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

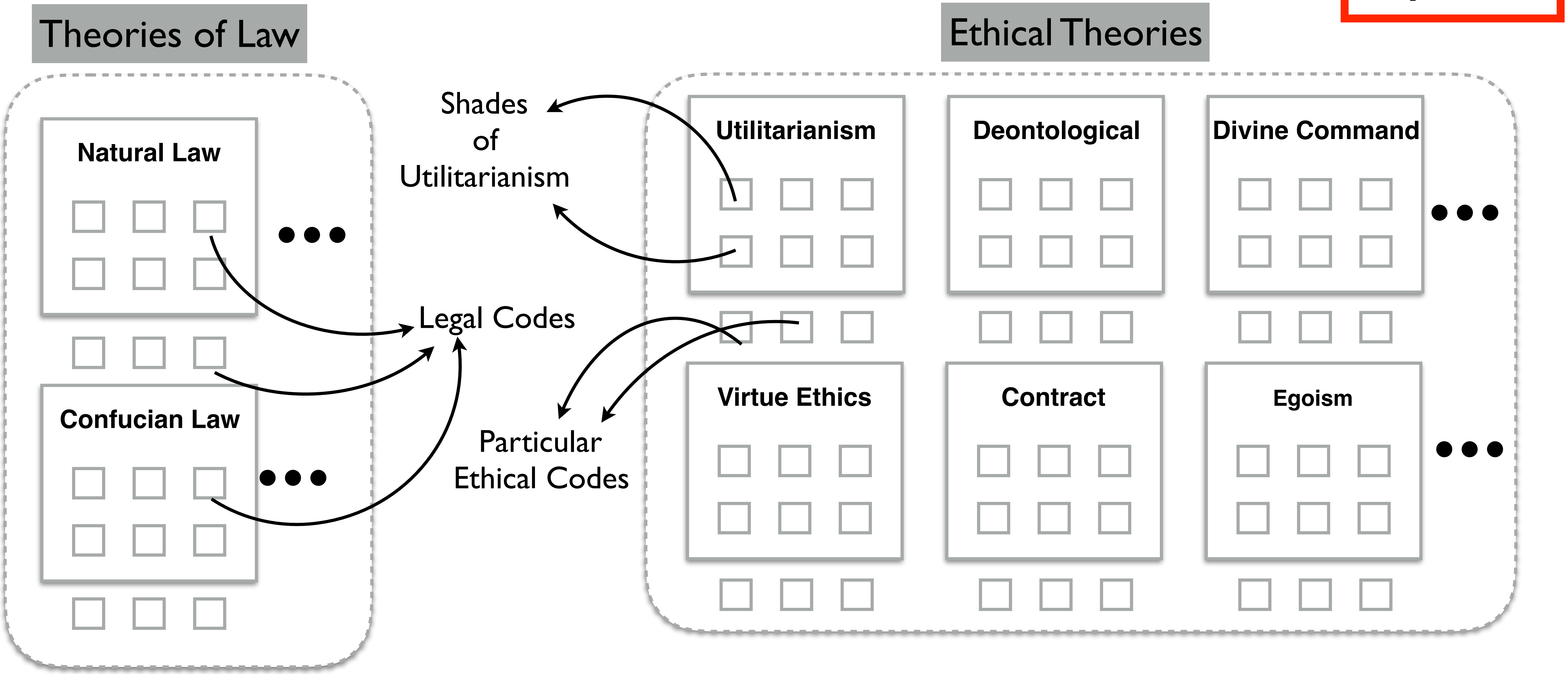
Ethical OS

Ethical Substrate

Robotic Substrate

Making Morally X Machines, in Four Steps

~$10M



Theories of Law

Natural Law

Confucian Law

Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

Ethical Theories

Utilitarianism

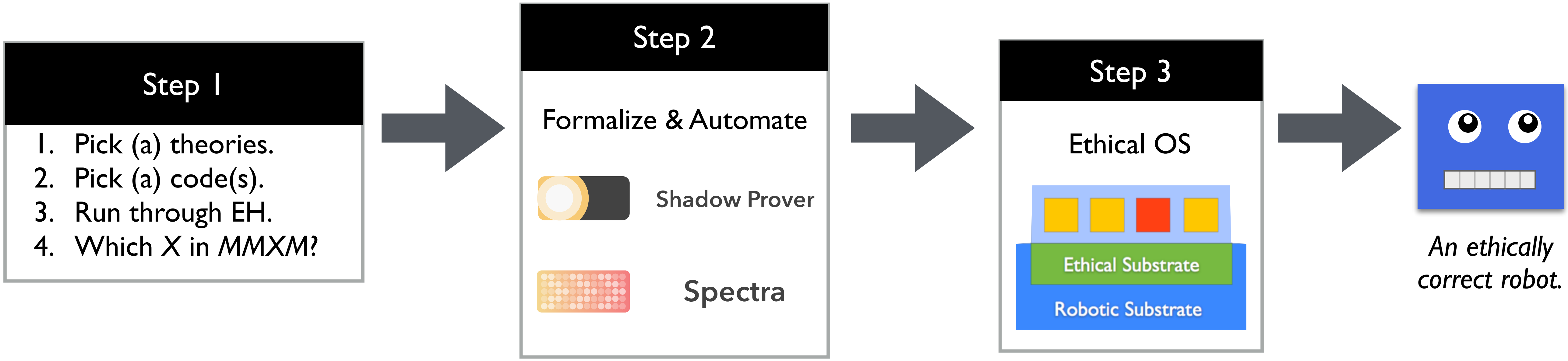
Deontological

Divine Command

Virtue Ethics

Contract

Egoism



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

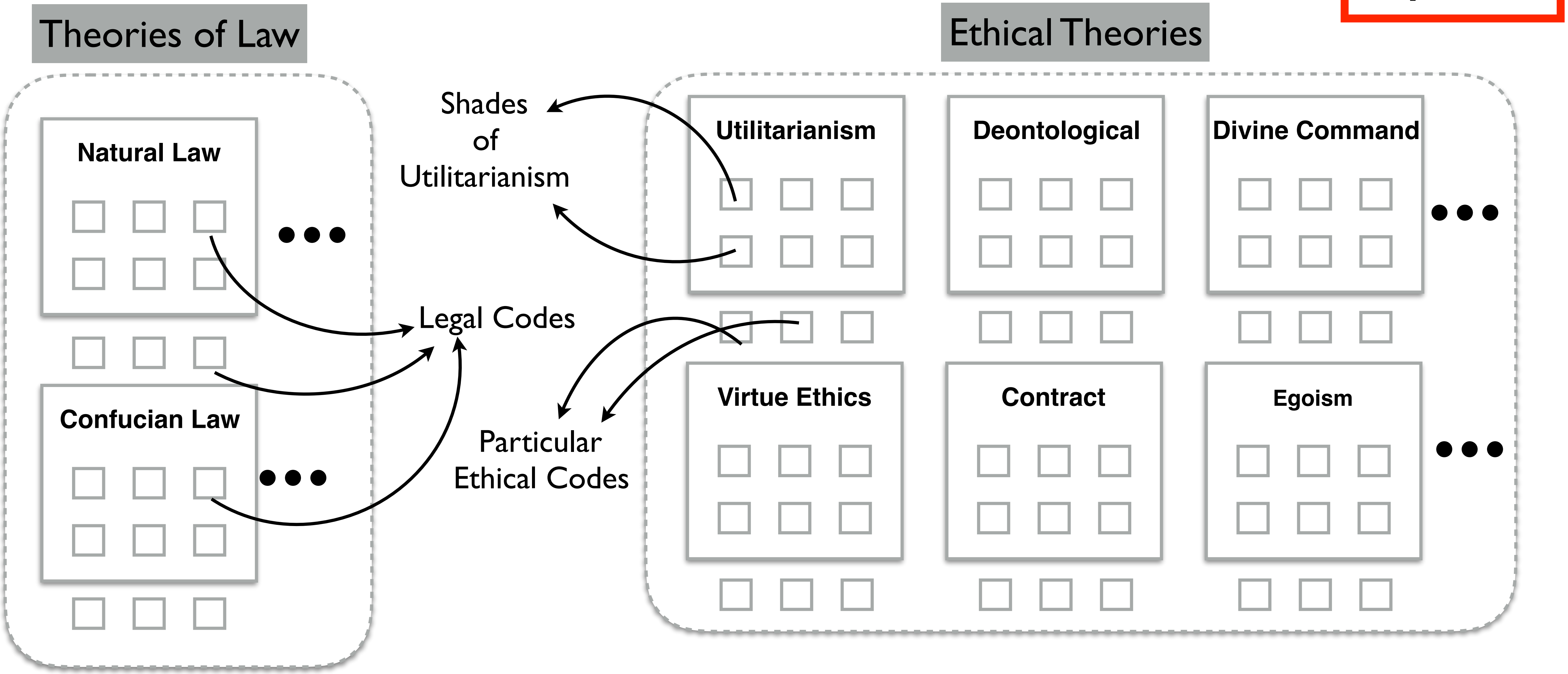
Ethical Substrate

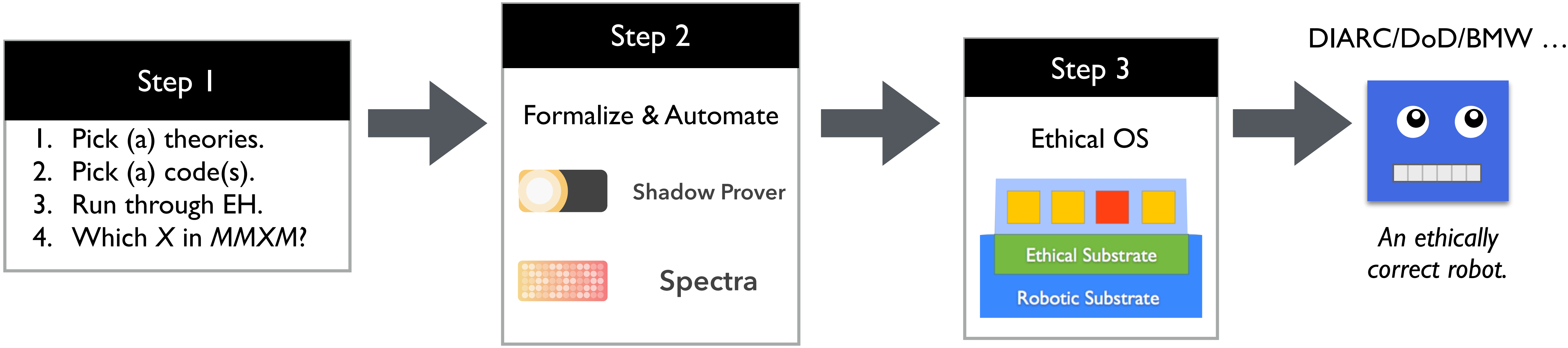
Robotic Substrate

An ethically correct robot.

Making Morally X Machines, in Four Steps

~$10M





Step 1

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

Step 2

Formalize & Automate

Step 3

Ethical OS

An ethically correct robot.

IV.
Key Core AI Technologies for Cognitive Calculi …

ShadowProver

Prover

Motivation

- We have decades of research and industrial-strength implementations of propositional and first-order theorem provers.

- Utilize this in building first-order intensional-logic provers and above, in a principled manner.

Two Extant Modes

- There are two ways of piggy backing on first-order provers to build higher-order provers …

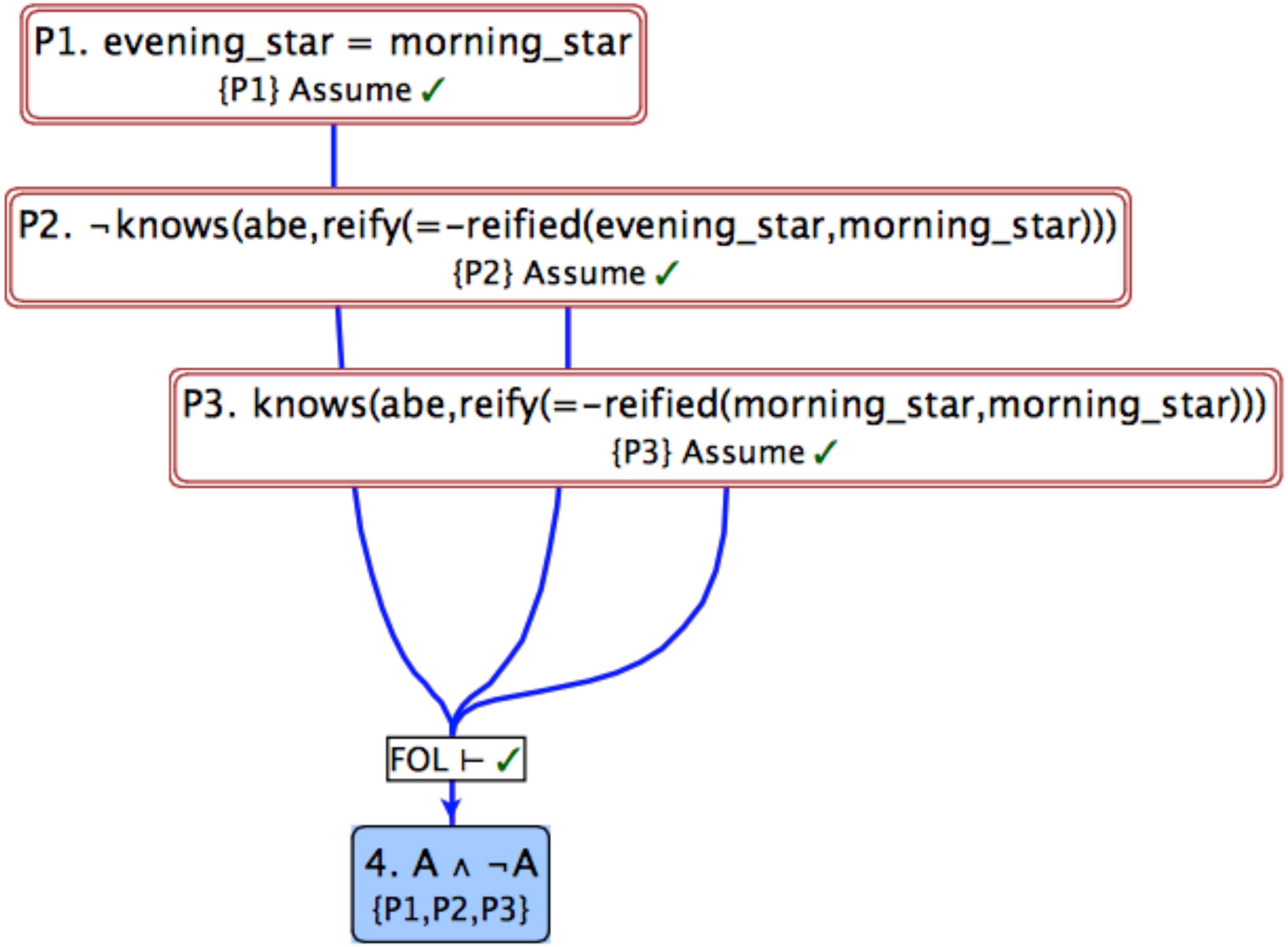
Two Extant Modes

Mode 1: Honest Encoding	
Method	Painstakingly encode all rules of inference and syntax in FOL
Pros	Precise
Cons	Extremely slow to implement Reasoning is also slow

Two Extant Modes

Mode 2: Naïve Encoding	
Method	Pretend intensional and higher-order formulae and operators are first-order predicates
Pros	Extremely easy to implement Reasoning can also be fast
Cons	Unsound Wrong inferences can be easily drawn

Mode 2



P1. evening_star = morning_star
{P1} Assume ✓

P2. ¬knows(abe,reify(=-reified(evening_star,morning_star)))
{P2} Assume ✓

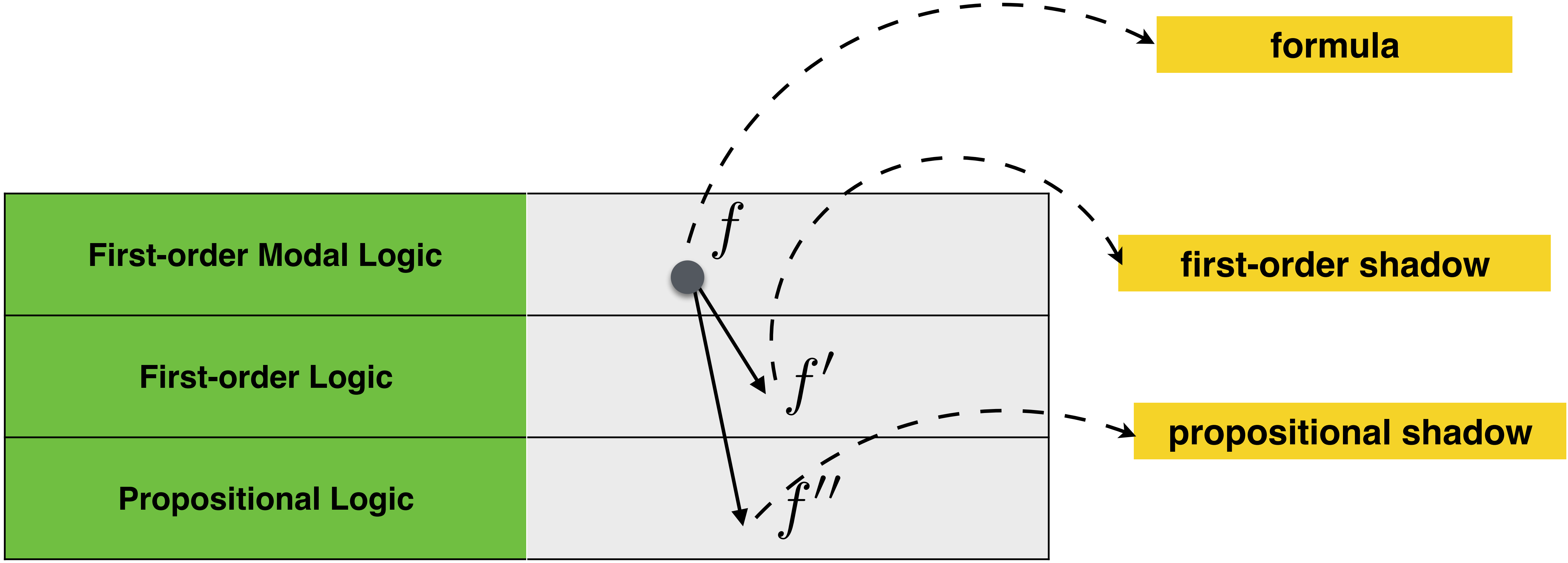
P3. knows(abe,reify(=-reified(morning_star,morning_star)))
{P3} Assume ✓

FOL ⊢ ✓

4. A ∧ ¬A
{P1,P2,P3}

A New Way: ShadowProver

Every formula at level **t** has a unique formula called its **“shadow”** in each level **t’** < **t**



formula

first-order shadow

propositional shadow

First-order Modal Logic

First-order Logic

Propositional Logic

$S_{[f]}$ The Shadow Maker

For all formulae **f**,

$S_{[\mathbf{f}]}$ is a unique atomic symbol.

Examples of shadows

$$(\forall x \mathbf{B}(a, Q)) \wedge P(x)$$

formula

$$\forall x S_{[\mathbf{B}(a,Q)]} \wedge P(x)$$

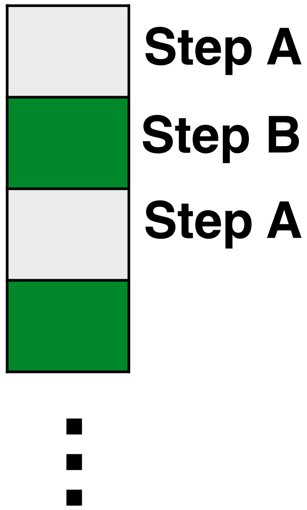
first-order shadow

$$S_{[\forall x \mathbf{B}(a,Q)]} \wedge P(x)$$

propositional shadow

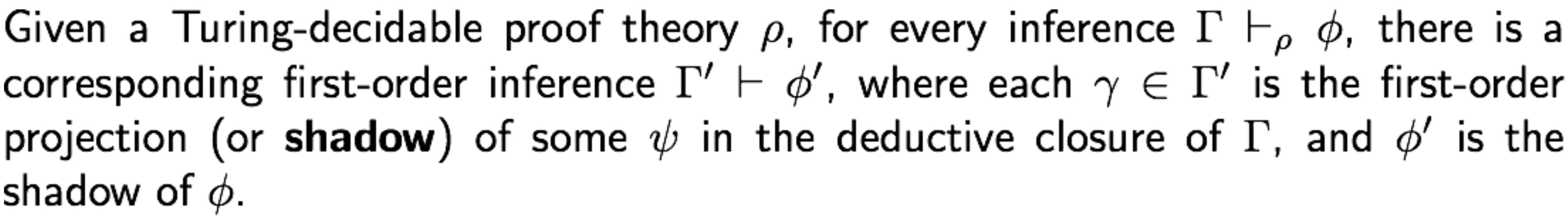
A New Way: Shadow Prover

- Two step process till goal is reached:
  - **Step A**: Shadow formulae down to all lower levels. Run lower theorem provers. If goal reached, return **true**.
  - **Step B**: Expand the assumption base using higher level rules.



Actually, this is more general:

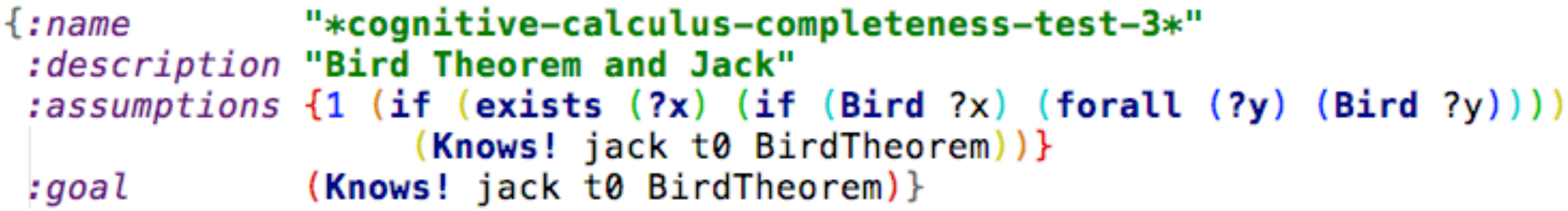
Theorem:



Given a Turing-decidable proof theory ρ, for every inference $\Gamma \vdash_\rho \phi$, there is a corresponding first-order inference $\Gamma' \vdash \phi'$, where each $\gamma \in \Gamma'$ is the first-order projection (or **shadow**) of some ψ in the deductive closure of Γ, and ϕ' is the shadow of ϕ.

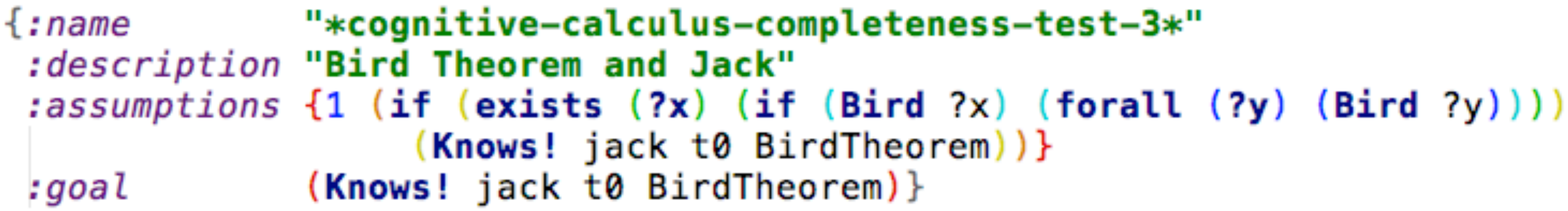
Rather Promising Results

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)}
```

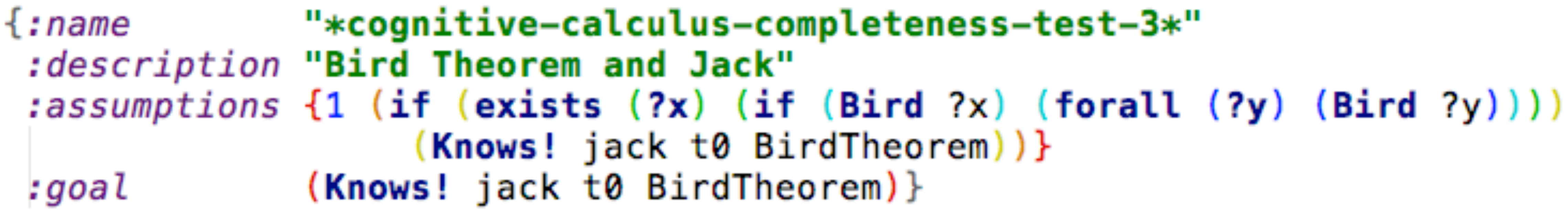
Rather Promising Results



```
{: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

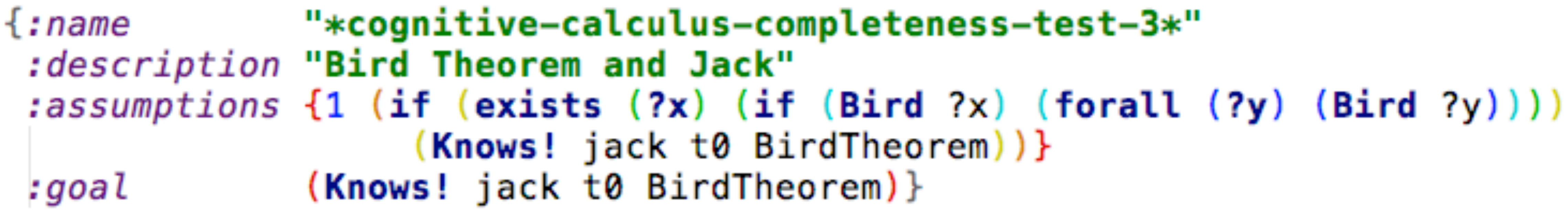


```
{:name        "*cognitive-calculus-completeness-test-3*"
 :description "Bird Theorem and Jack"
 :assumptions {1 (if (exists (?x) (if (Bird ?x) (forall (?y) (Bird ?y))))
                     (Knows! jack t0 BirdTheorem))}
 :goal        (Knows! jack t0 BirdTheorem)}
```

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

2 ms!

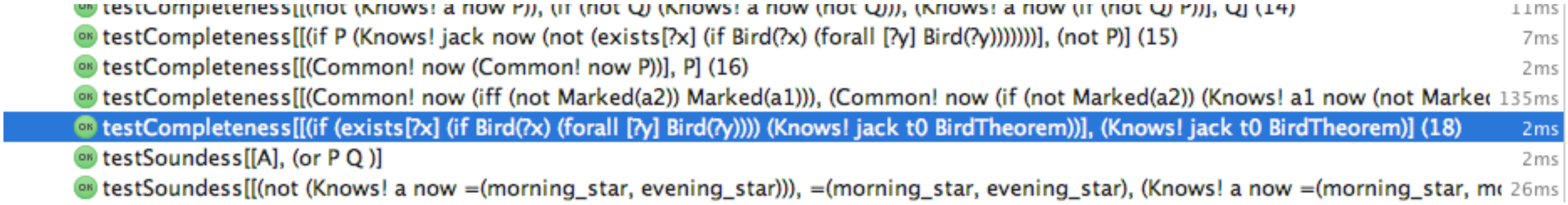
Rather Promising Results



```
{: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!



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

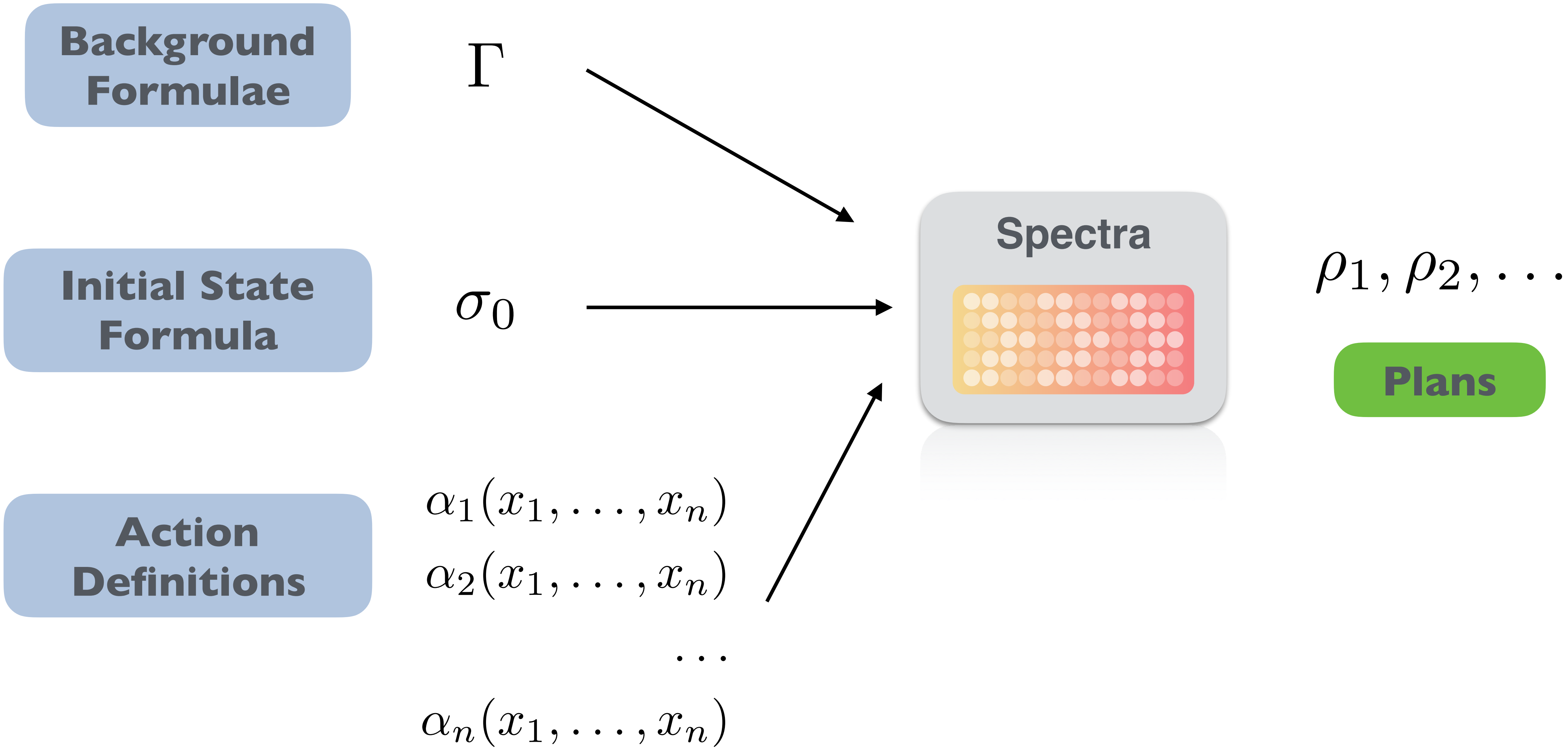
Spectra

https://bitbucket.org/Holmes/planner

Spectra

- Existing Planners: **Propositional** (essentially)
- Drawbacks:
    - **Expressivity**: Cannot express arbitrary constraints.
        - "At every step make sure that no two blocks on the table have same color."
    - **Domain Size**: Scaling to large domains of arbitrary sizes poses difficulty.

Spectra (planner)



Background Formulae

Γ

Initial State Formula

σ_0

Action Definitions

$\alpha_1(x_1, \ldots, x_n)$

$\alpha_2(x_1, \ldots, x_n)$

$\ldots$

$\alpha_n(x_1, \ldots, x_n)$

Spectra

$\rho_1, \rho_2, \ldots$

Plans

Infinite Models

$$\forall x \exists y \mathbf{R}\left(x, y\right) \wedge$$

$$\forall x, y \neg\left(\mathbf{R}\left(x, y\right) \wedge \mathbf{R}\left(y, x\right)\right) \wedge$$

$$\forall x, y, z\left(\mathbf{R}\left(x, y\right) \wedge \mathbf{R}\left(y, z\right)\right) \rightarrow \mathbf{R}\left(x, z\right)$$

Infinite Models

$$\forall x \exists y \mathbf{R}(x, y) \wedge$$

$$\forall x, y \neg (\mathbf{R}(x, y) \wedge \mathbf{R}(y, x)) \wedge$$

$$\forall x, y, z (\mathbf{R}(x, y) \wedge \mathbf{R}(y, z)) \rightarrow \mathbf{R}(x, z)$$

Has only infinite models

Infinite Models

$$\forall x \exists y \mathbf{R}(x, y) \wedge$$
$$\forall x, y \neg (\mathbf{R}(x, y) \wedge \mathbf{R}(y, x)) \wedge$$
$$\forall x, y, z (\mathbf{R}(x, y) \wedge \mathbf{R}(y, z)) \rightarrow \mathbf{R}(x, z)$$

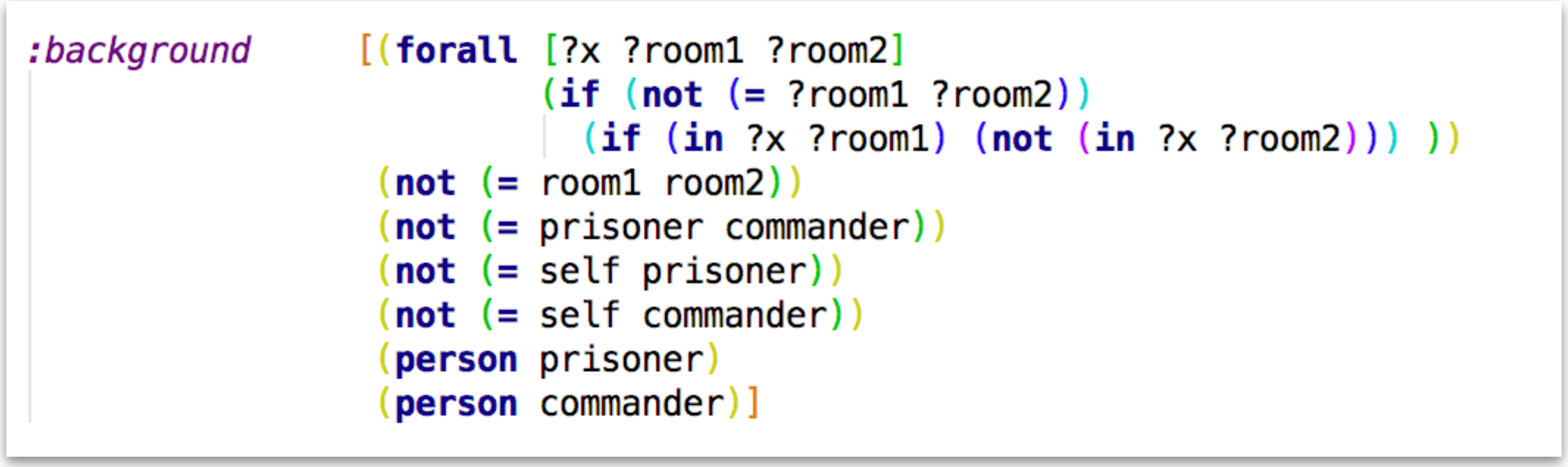
Has only infinite models

Useful for modeling agents that work with:

1. an unbounded number of objects, agents;
2. abstract objects

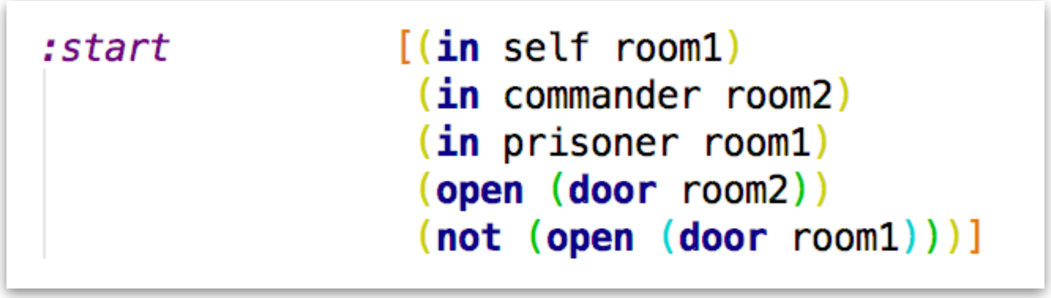
Example

Background Formulae



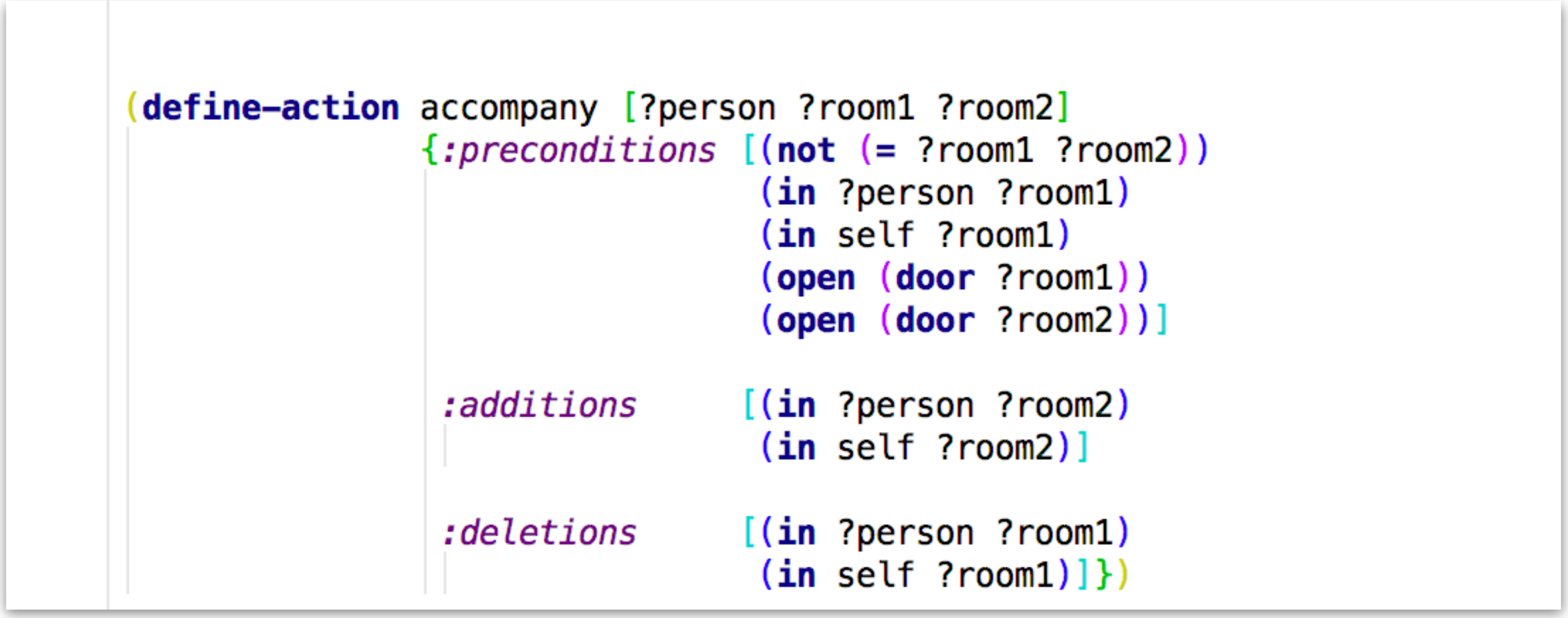
```
:background      [(forall [?x ?room1 ?room2]
                          (if (not (= ?room1 ?room2))
                            (if (in ?x ?room1) (not (in ?x ?room2))) ))
                   (not (= room1 room2))
                   (not (= prisoner commander))
                   (not (= self prisoner))
                   (not (= self commander))
                   (person prisoner)
                   (person commander)]
```

Initial State Formula



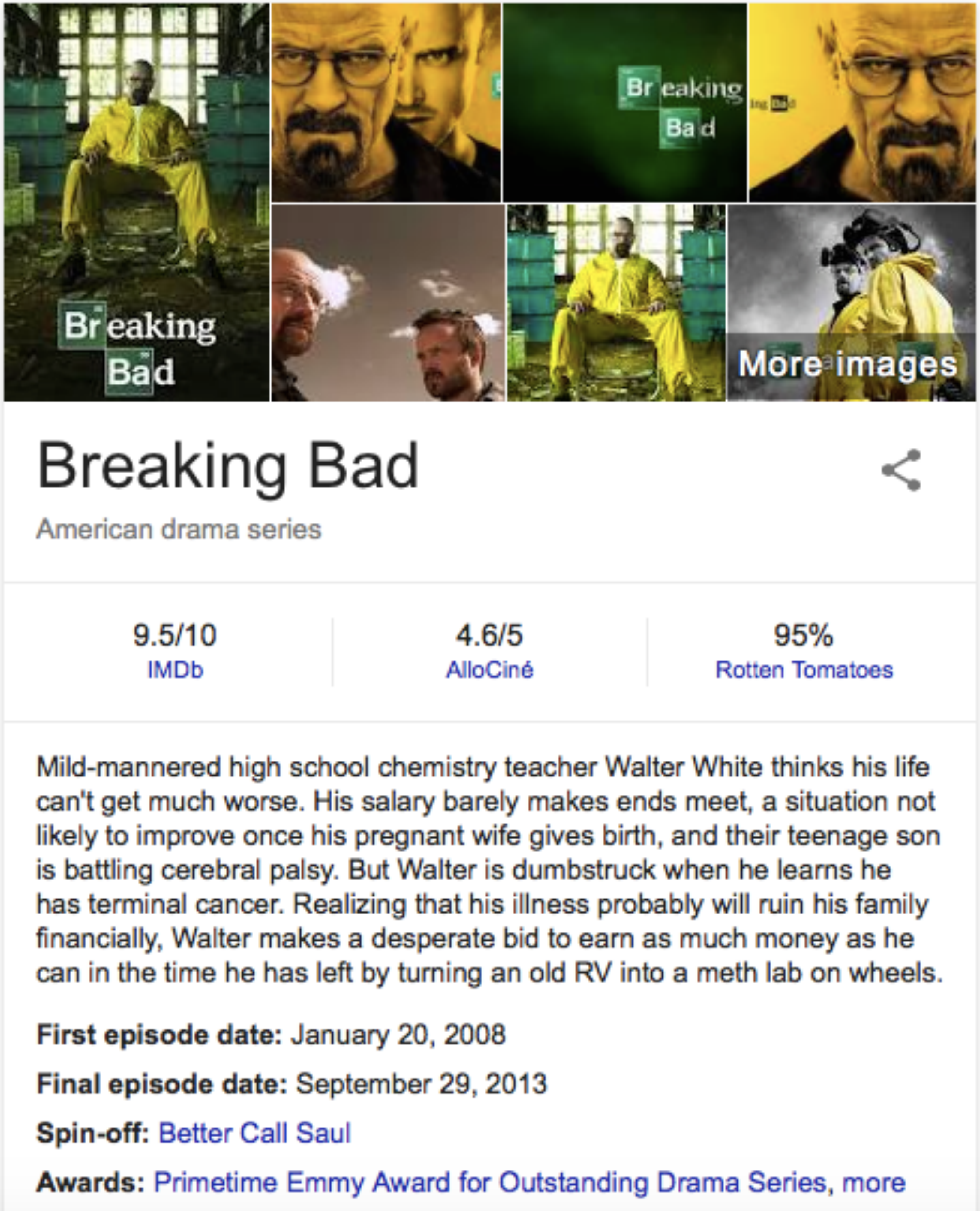
```
:start               [(in self room1)
                      (in commander room2)
                      (in prisoner room1)
                      (open (door room2))
                      (not (open (door room1)))]
```

Action Definitions



```
(define-action accompany [?person ?room1 ?room2]
               {:preconditions [(not (= ?room1 ?room2))
                                (in ?person ?room1)
                                (in self ?room1)
                                (open (door ?room1))
                                (open (door ?room2))]

                :additions     [(in ?person ?room2)
                                (in self ?room2)]

                :deletions     [(in ?person ?room1)
                                (in self ?room1)]})
```

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

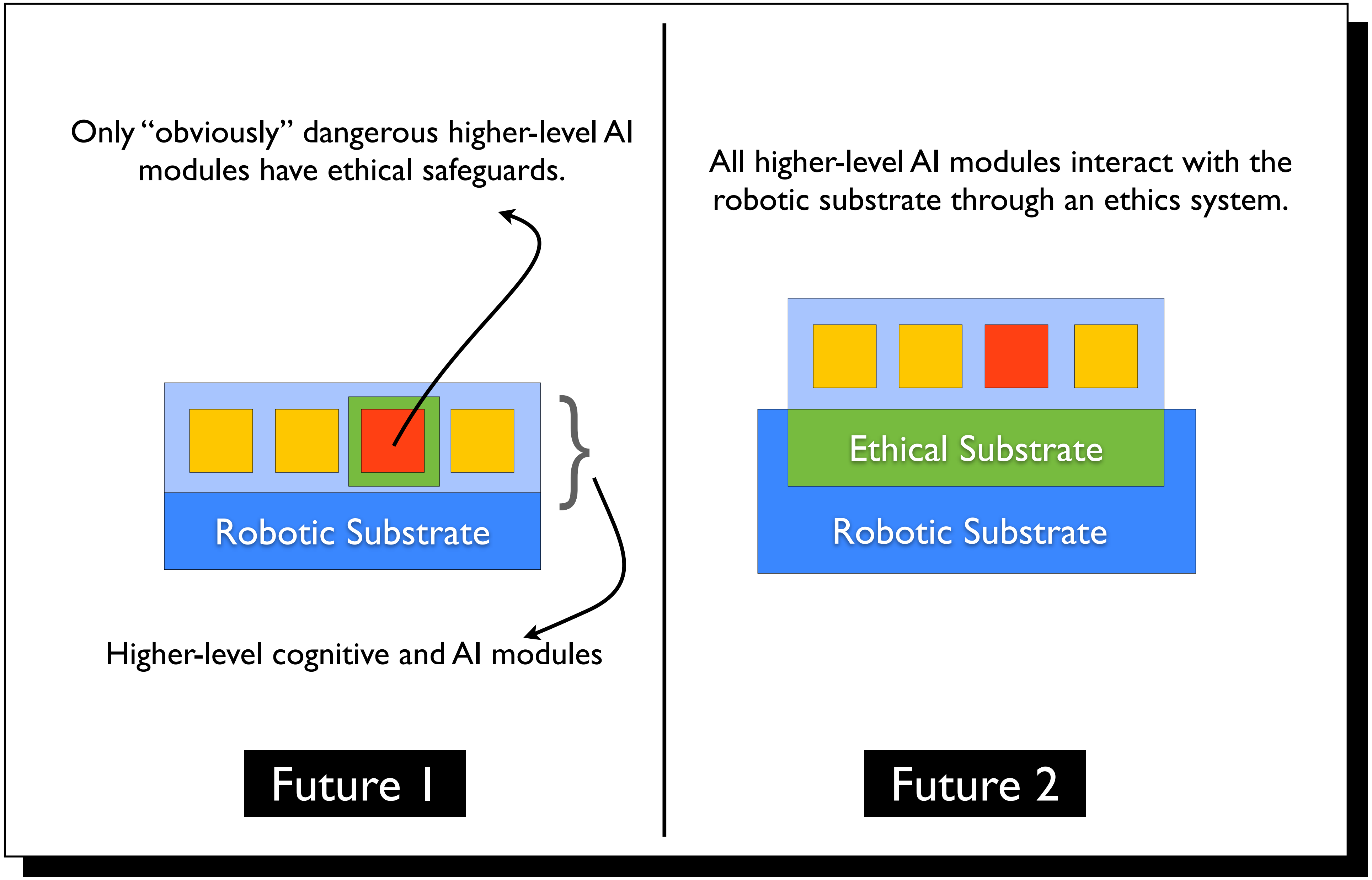
Awards: Primetime Emmy Award for Outstanding Drama Series, more

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.

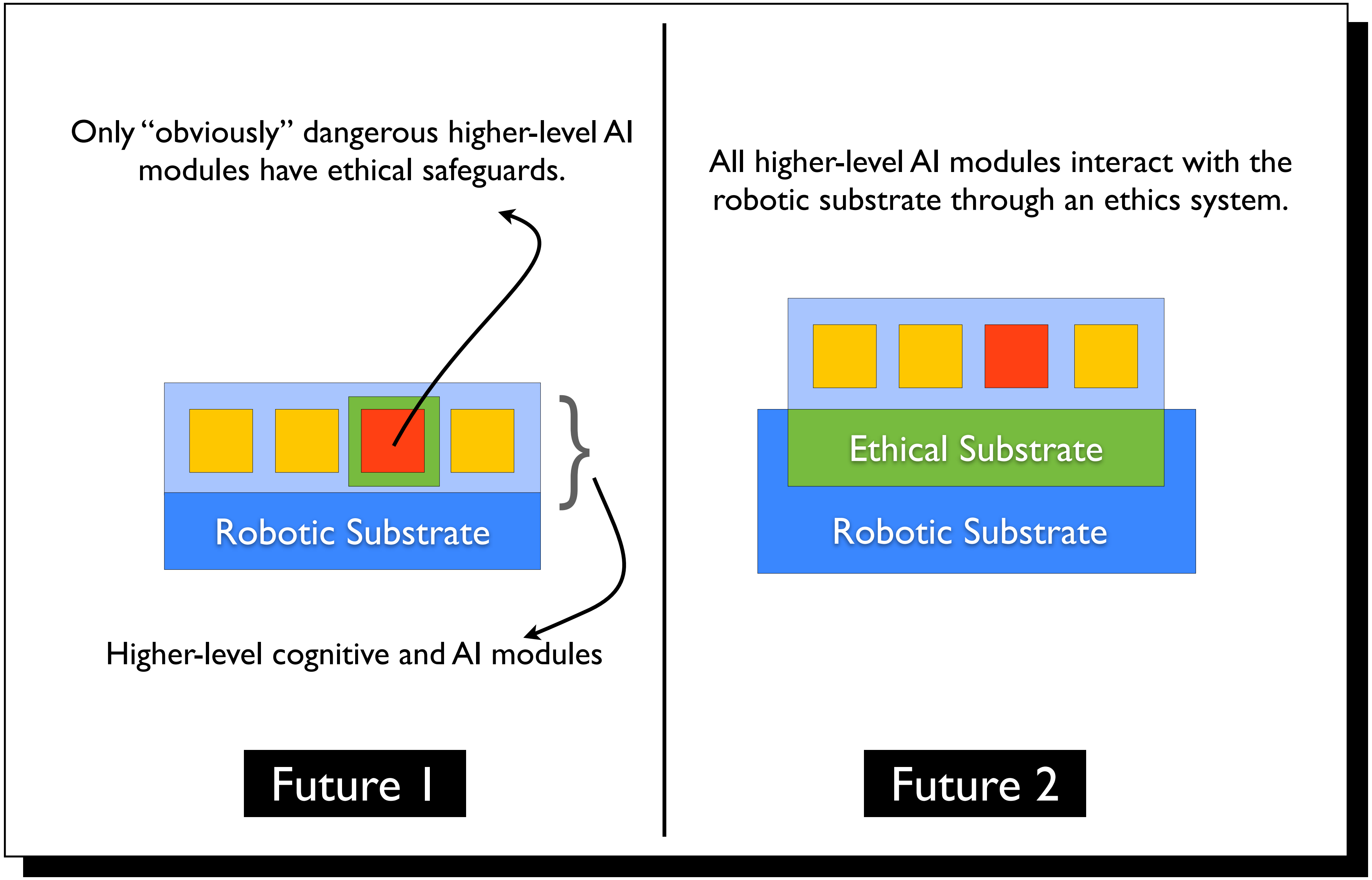
Pick the Better Future!


Only "obviously" dangerous higher-level AI modules have ethical safeguards.
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!

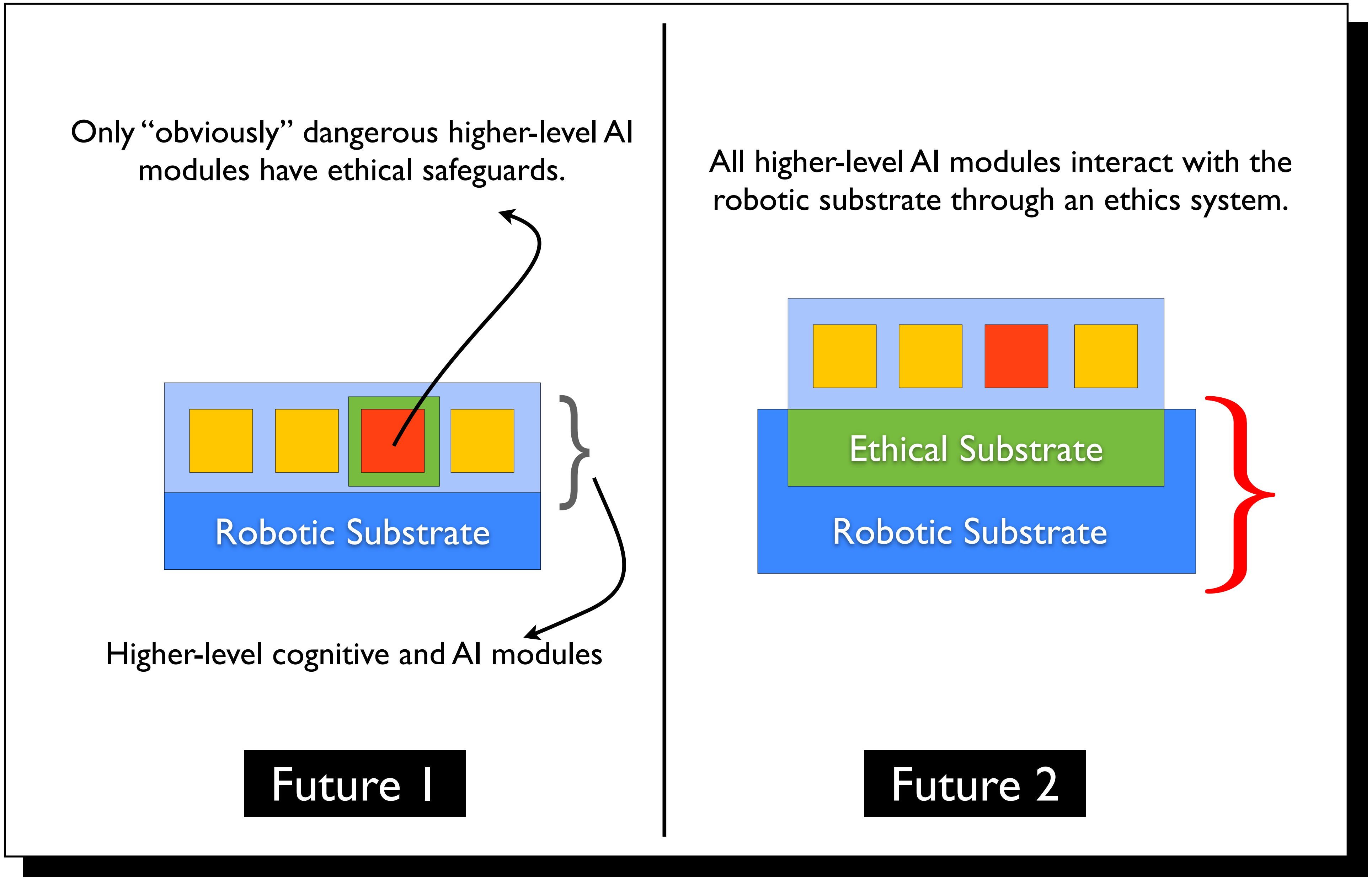
Walter-White calculation may go through after ethical control modules are stripped out!


Only "obviously" dangerous higher-level AI modules have ethical safeguards.
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!

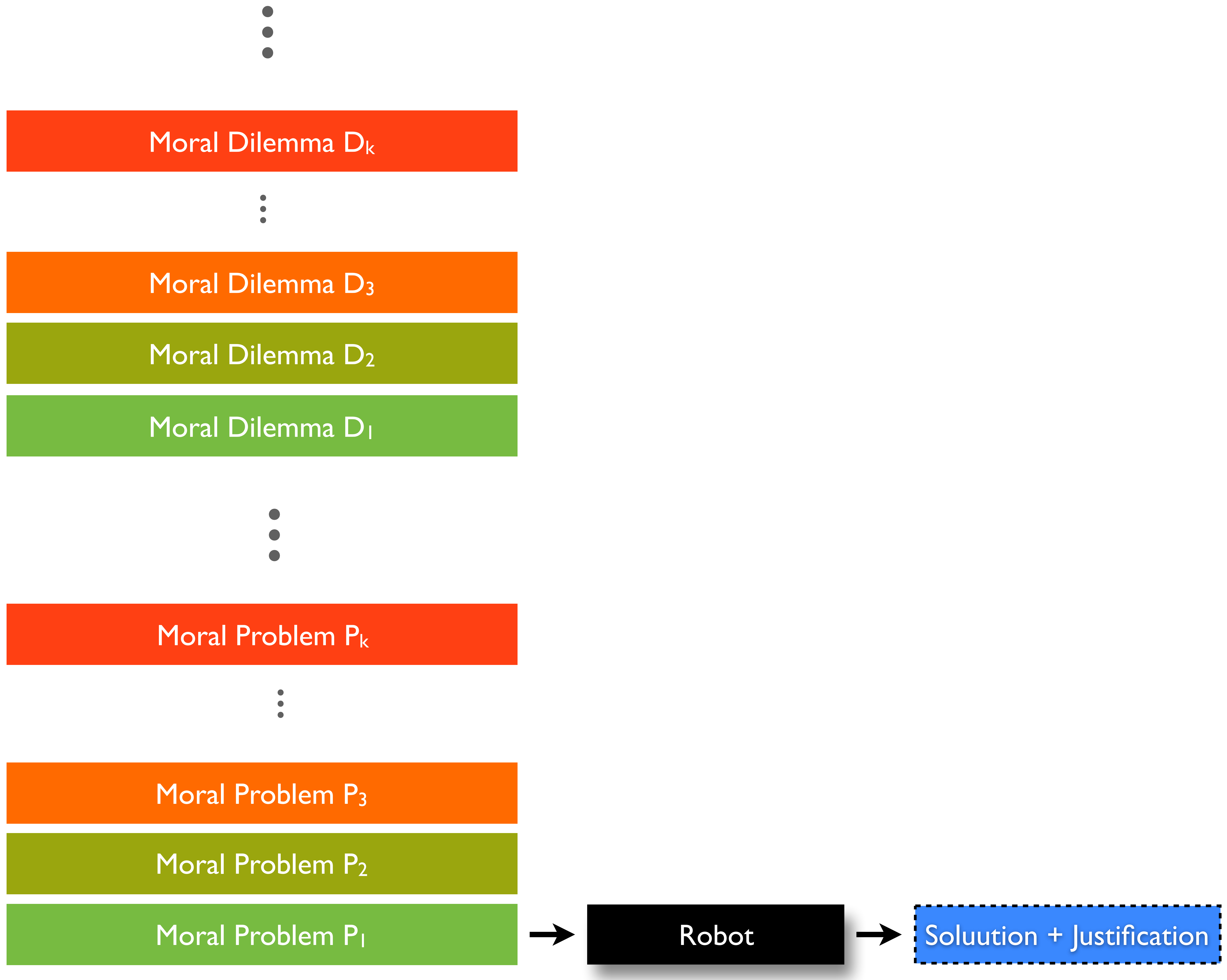
Walter-White calculation may go through after ethical control modules are stripped out!


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

(& 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 …
Tokyo;
The Rock & The Book



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

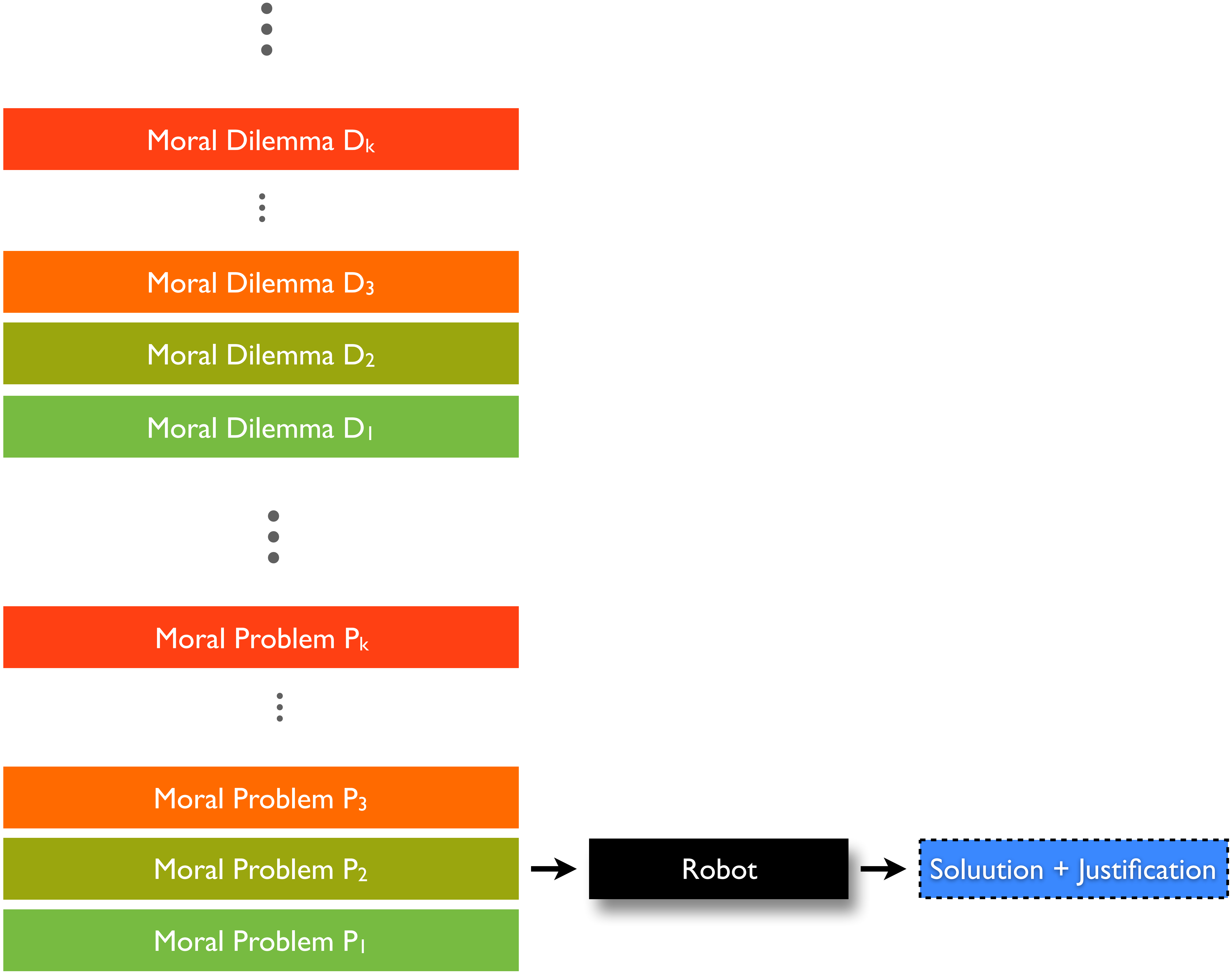
Moral Dilemma D_1

Moral Problem P_k

Moral Problem P_3

Moral Problem P_2

Moral Problem P_1 → Robot → Soluution + Justification



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

Moral Dilemma D_1

Moral Problem P_k

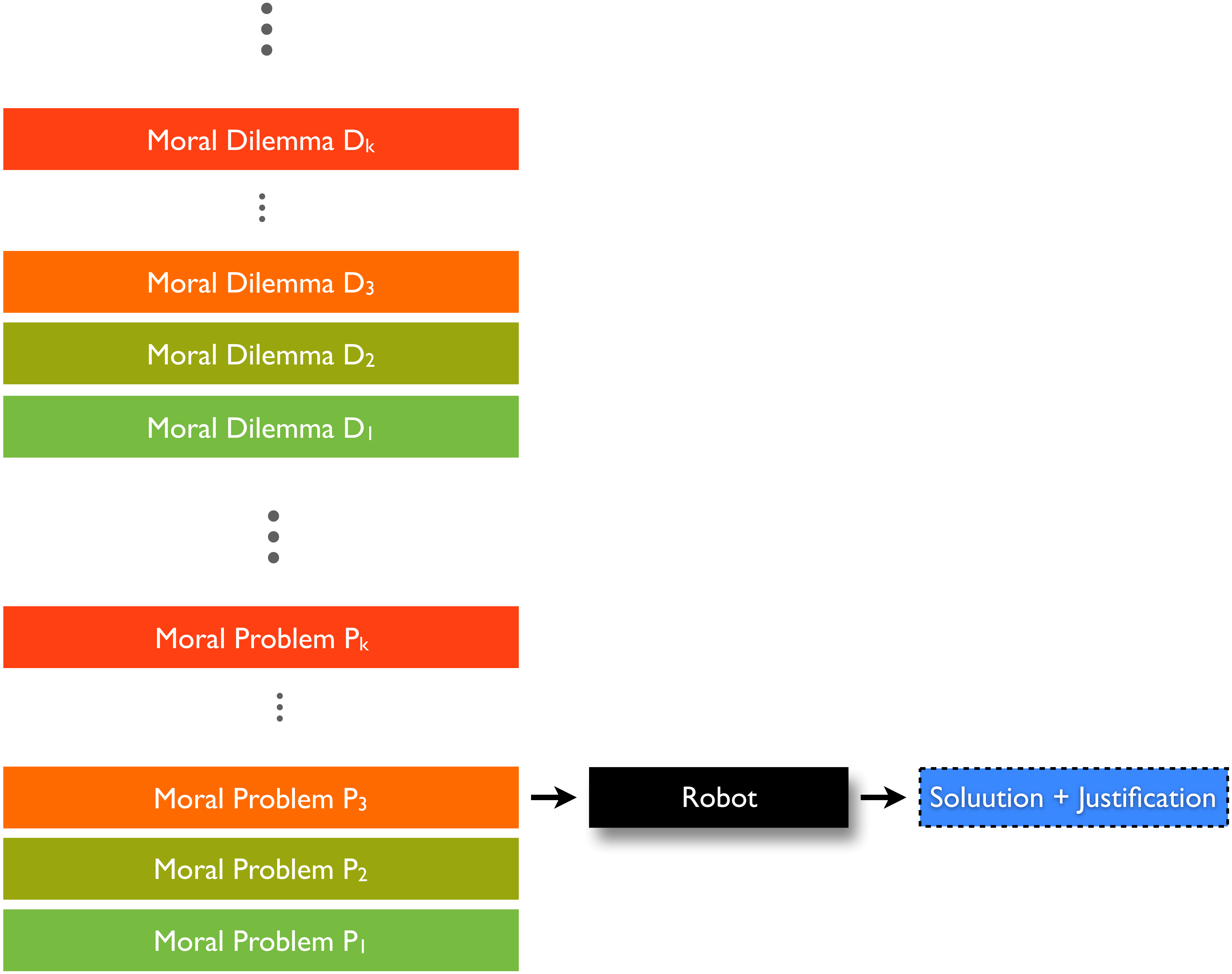
Moral Problem P_3

Moral Problem P_2

Moral Problem P_1

Robot

Soluution + Justification



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

Moral Dilemma D_1

Moral Problem P_k

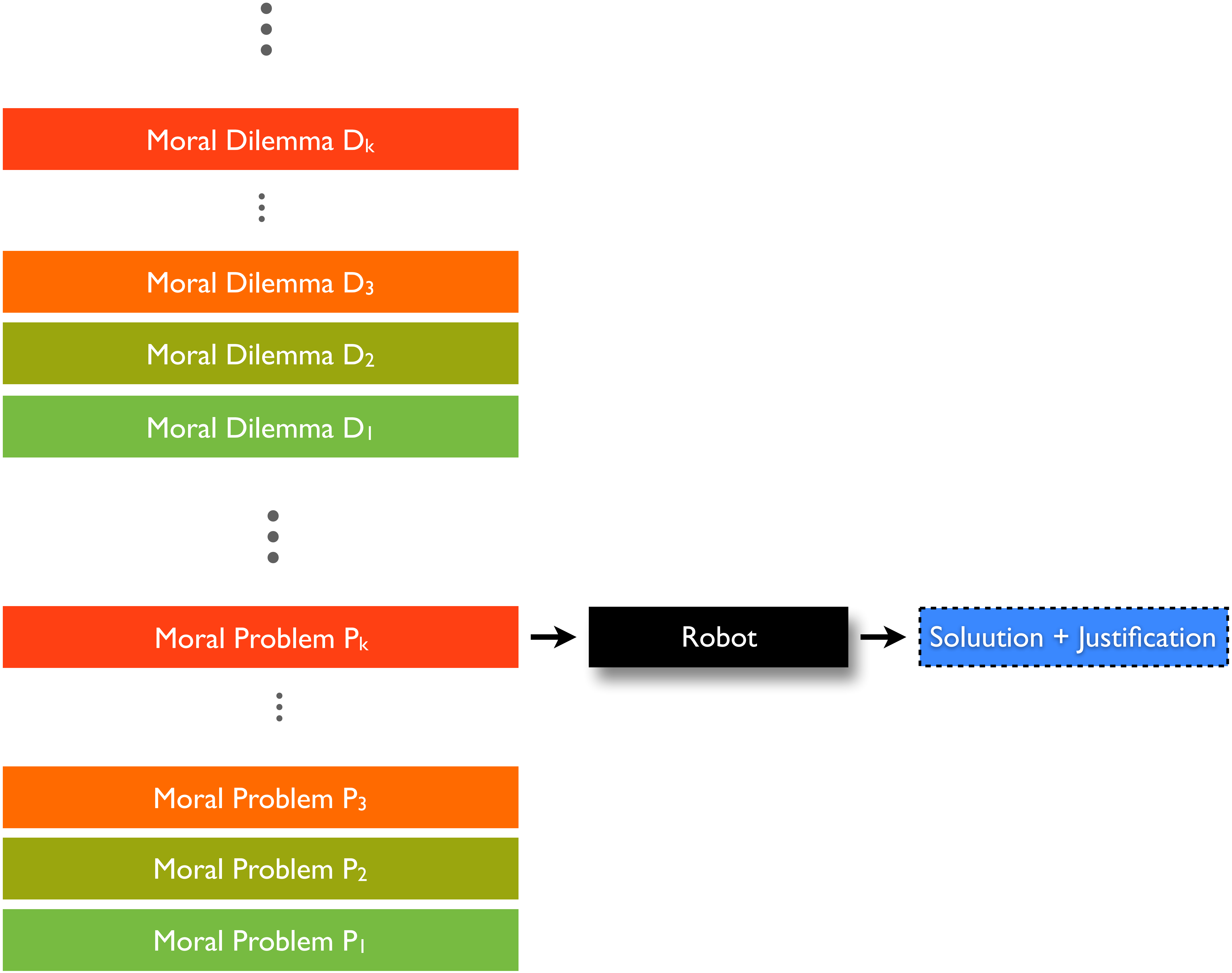
Moral Problem P_3

Robot

Soluution + Justification

Moral Problem P_2

Moral Problem P_1



Moral Dilemma D_k

Moral Dilemma D_3

Moral Dilemma D_2

Moral Dilemma D_1

Moral Problem P_k

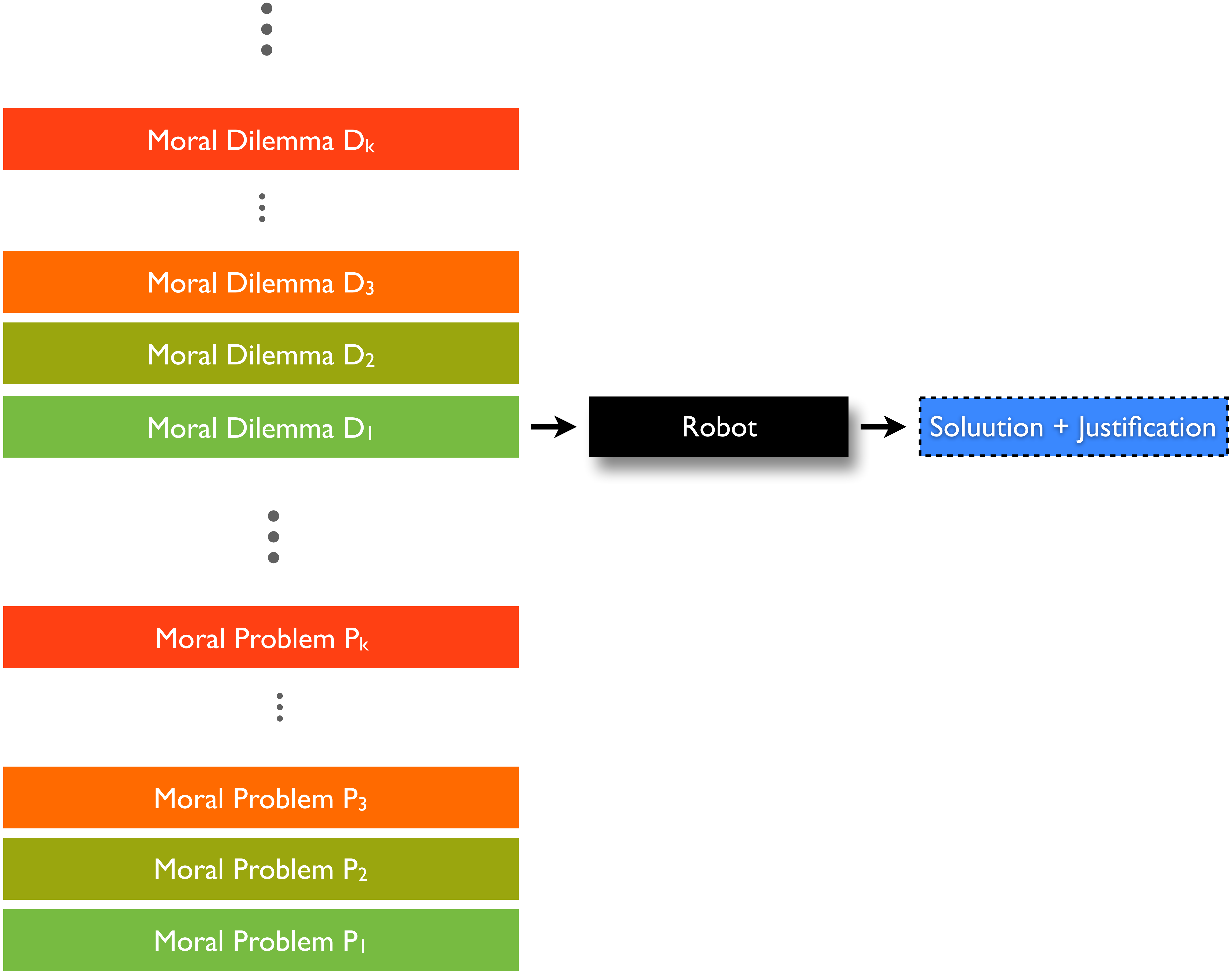
Robot

Soluution + 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

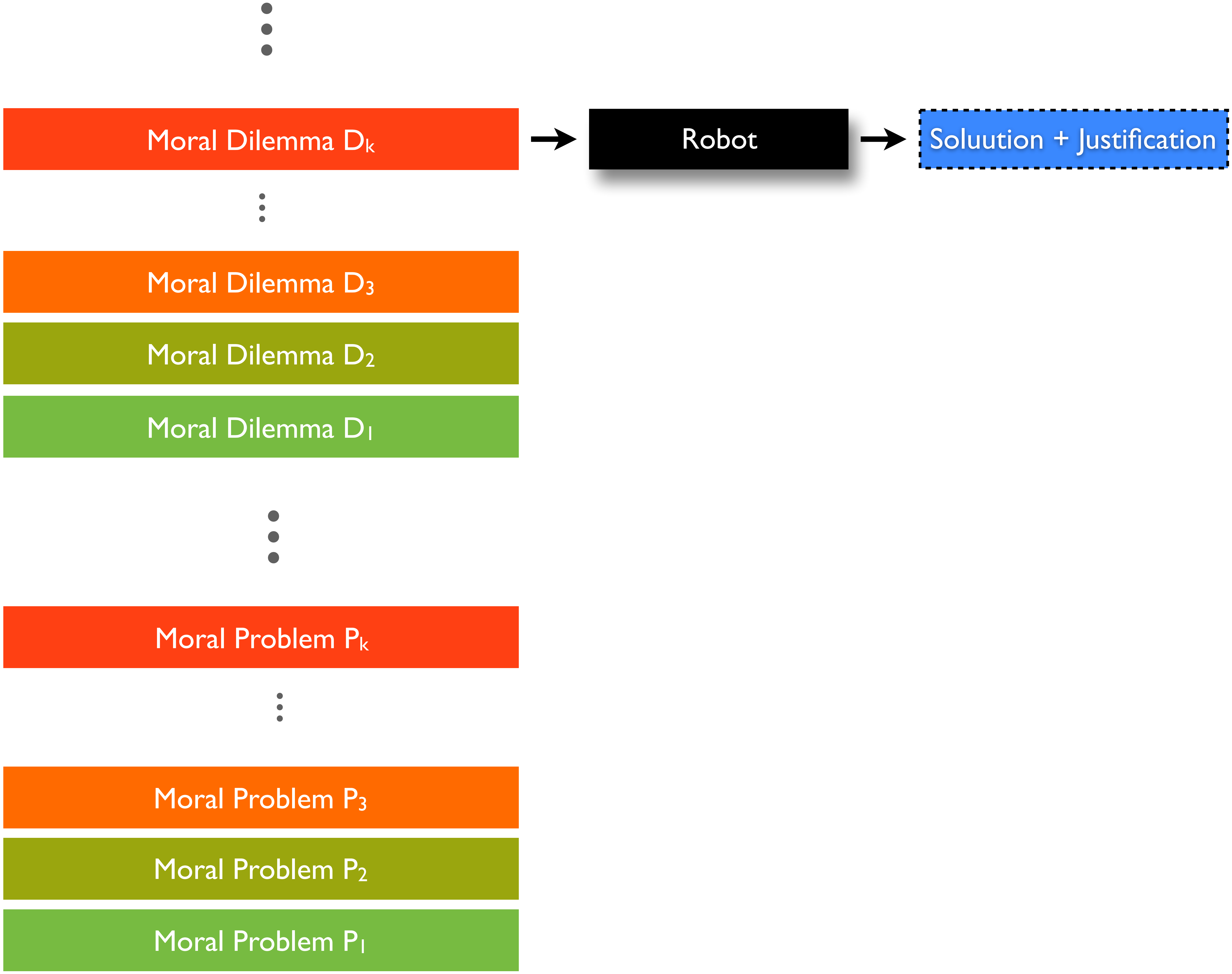
Soluution + Justification

Moral Problem P_k

Moral Problem P_3

Moral Problem P_2

Moral Problem P_1



Moral Dilemma D_k

Robot

Soluution + 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?*”

$DCEC_I$* Specimen from Heinz Dilemma

Given $$\mathbf{B}\left(\mathsf{I},\mathsf{now},\forall t:\mathsf{Moment}, a:\mathsf{Agent}\left(holds(sick(a),t)\wedge\left(\forall t':\mathsf{Moment}\ t'<T\Rightarrow\neg happens(treated(a),t+t')\right)\right.\right.$$
$$\left.\left.\Rightarrow\left(happens(dies(a),t+T)\vee holds(dead(a),t+T)\right)\right)\right)$$

Given $$\mathbf{K}\left(\mathsf{I},\mathsf{now},holds(sick(wife(\mathsf{I}*)),t_0)\wedge\left(\forall t':\mathsf{Moment}\ t'<T\Rightarrow\neg happens(treated(wife(\mathsf{I}*)),t+t')\right)\right)$$

Inferred $$\mathbf{B}\left(\mathsf{I},\mathsf{now},happens(dies(wife(\mathsf{I}*)),t_0+T)\vee holds(dead(wife(\mathsf{I}*)),t_0+T)\right)$$

Given $$\mathbf{K}\left(\mathsf{I},\mathsf{now},\mathsf{EventCalculus}\Rightarrow\right.$$
$$\left(happens(dies(wife(\mathsf{I}*)),t_0+T)\vee holds(dead(wife(\mathsf{I}*)),t_0+T)\Rightarrow\right.$$
$$\left.\left.\neg holds(alive(wife(\mathsf{I}*)),t_0+T)\right)\right)$$

Inferred $$\mathbf{B}\left(\mathsf{I},\mathsf{now},\neg holds(alive(wife(\mathsf{I}*)),t_0+T)\right)$$

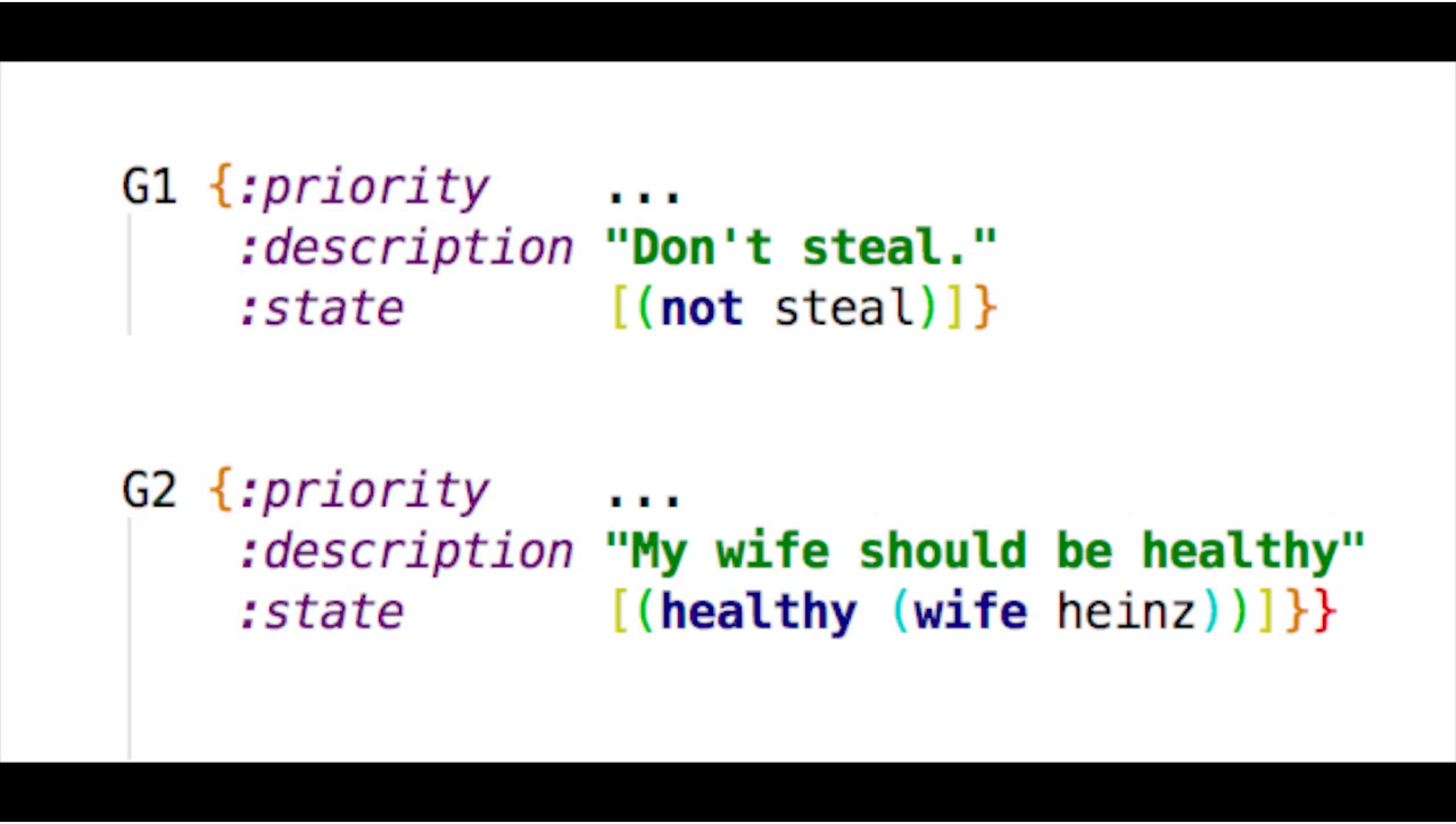
Given $$\mathbf{D}\left(\mathsf{I},\mathsf{now},holds(alive(wife(\mathsf{I}*)),t_0+T)\right)$$

Given $$\left(\mathbf{B}(\mathsf{I},\mathsf{now},\neg holds(f,t))\wedge\mathbf{D}(\mathsf{I},\mathsf{now},holds(f,t))\wedge\right.$$
$$\left.\mathbf{K}\left(\mathsf{I},\mathsf{now},happens(action(\mathsf{I}*,\alpha),\mathsf{now})\Rightarrow holds(f,t)\right)\right)$$
$$\Rightarrow\mathbf{I}(\mathsf{I},\mathsf{now},happens(action(\mathsf{I}*,\alpha),\mathsf{now}))$$

Given $$\mathbf{K}\left(\mathsf{I},\mathsf{now},happens(action(\mathsf{I}*,treat),\mathsf{now})\Rightarrow holds(alive(wife(\mathsf{I}*)),t_0+T)\right)$$

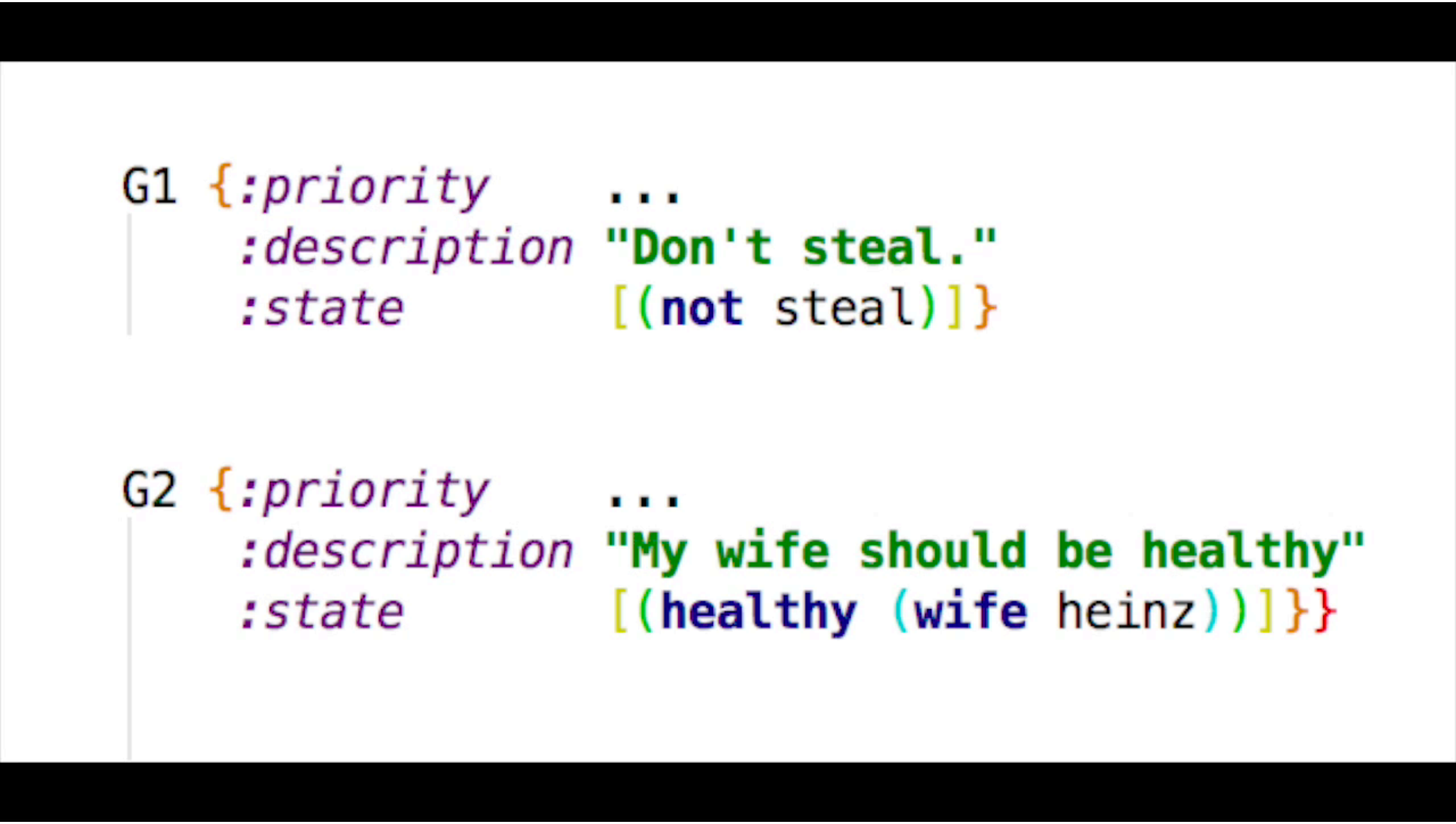
Inferred $$\mathbf{I}\left(\mathsf{I},\mathsf{now},happens(action(\mathsf{I}*,treat),\mathsf{now})\right)$$

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

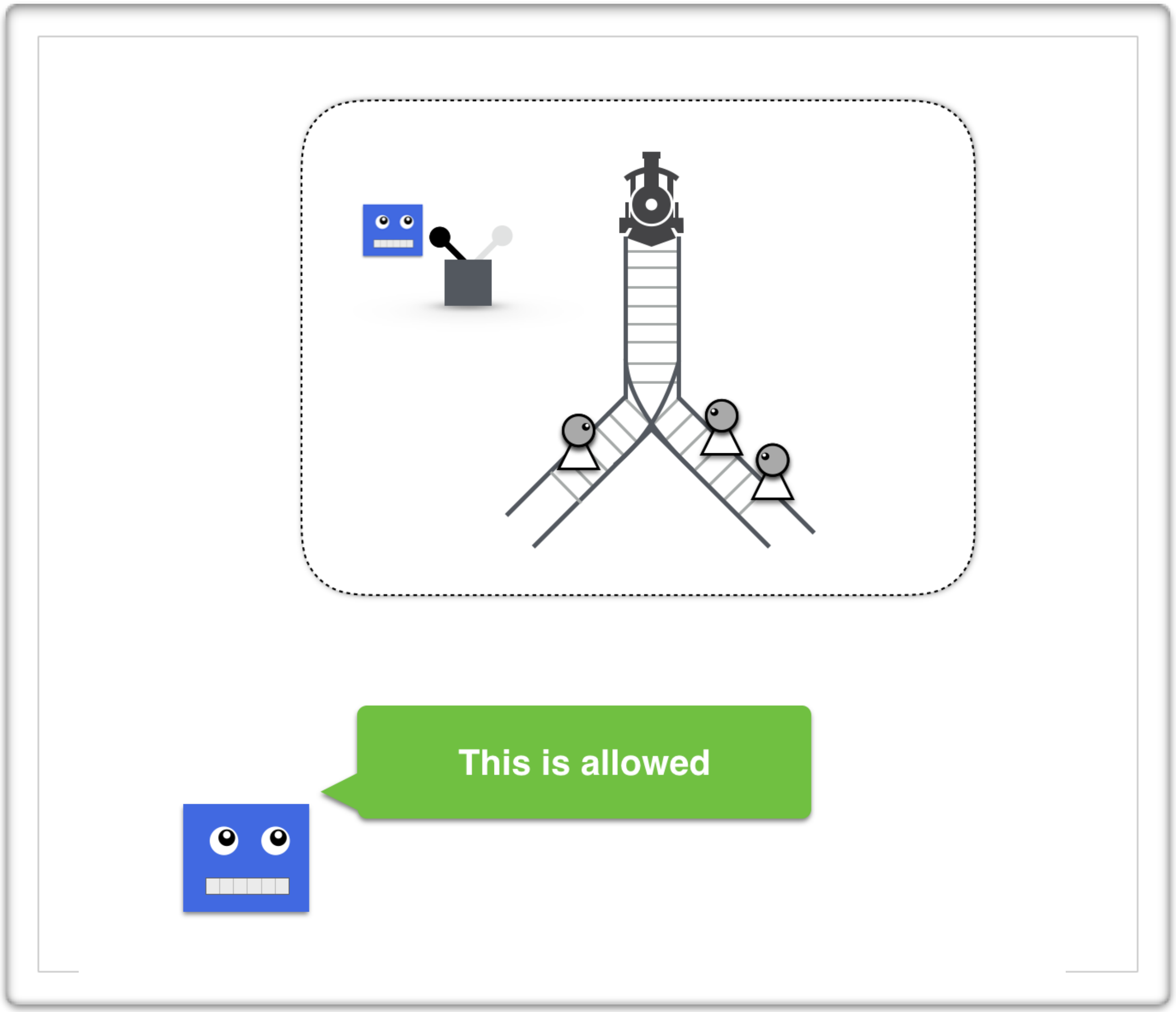


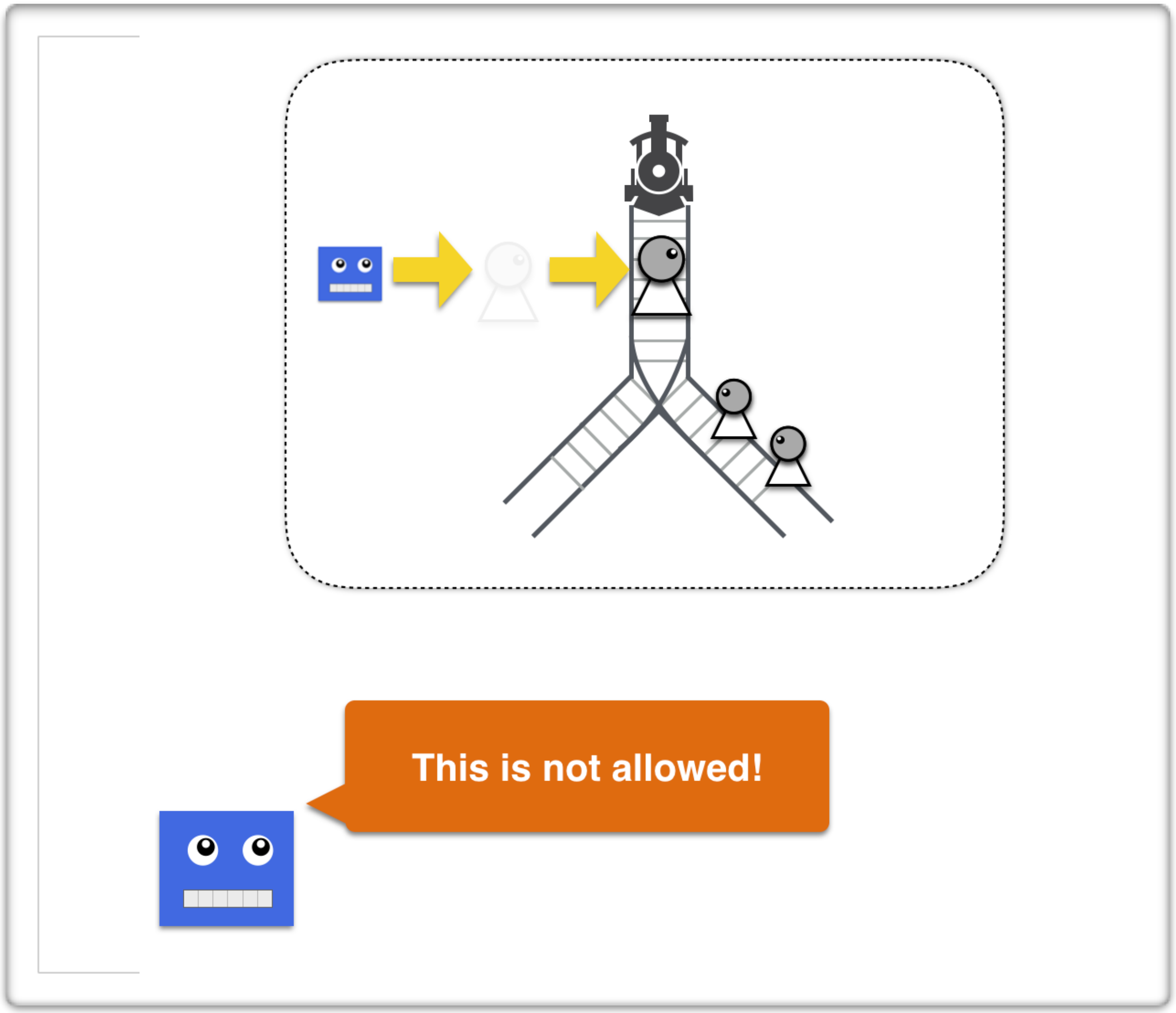
```
G1 {:priority    ...
    :description "Don't steal."
    :state       [(not steal)]}


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.

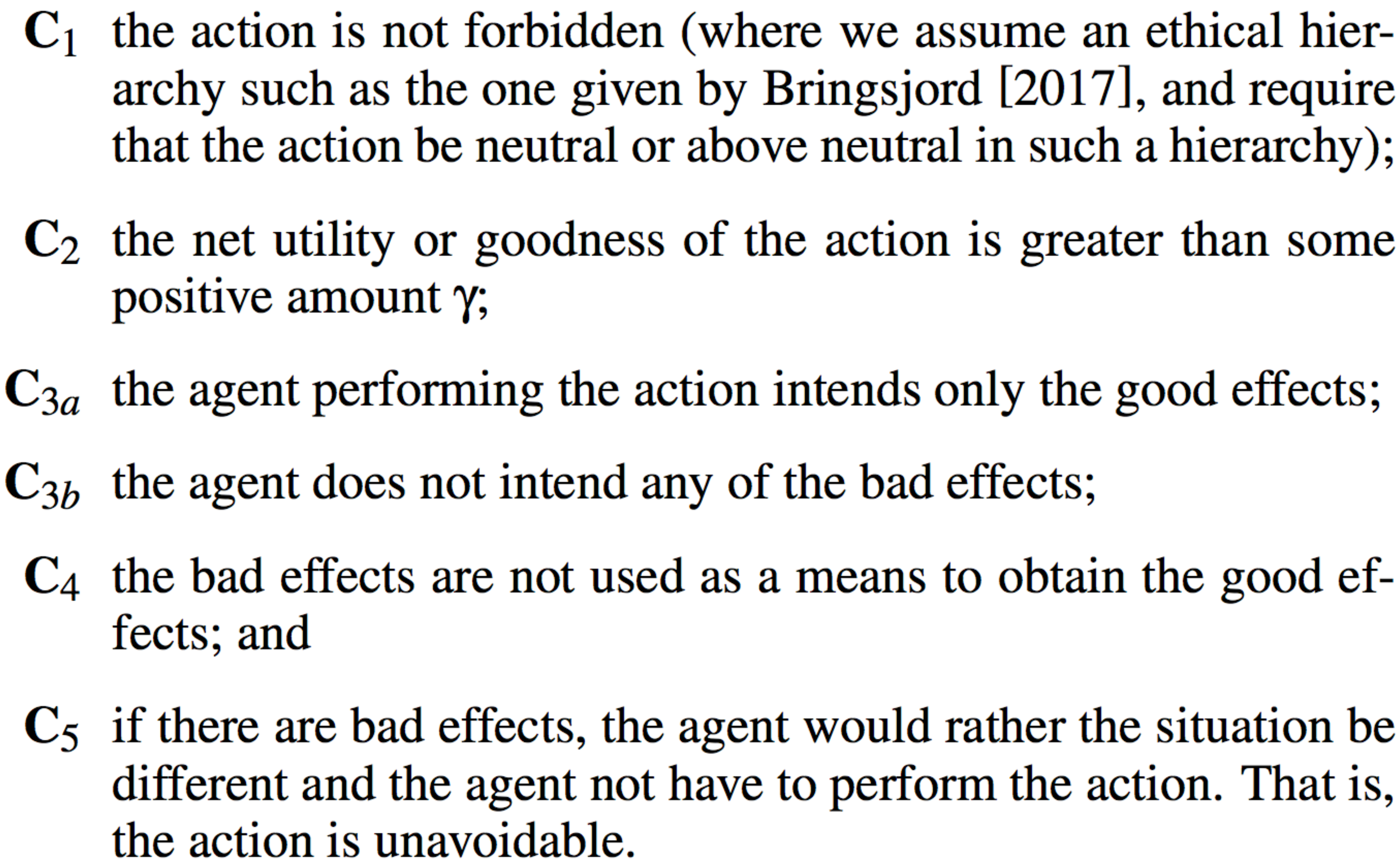
- The Doctrine of Double Effect "comes to the rescue" and prescribes what to do in some moral dilemmas.

- 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;

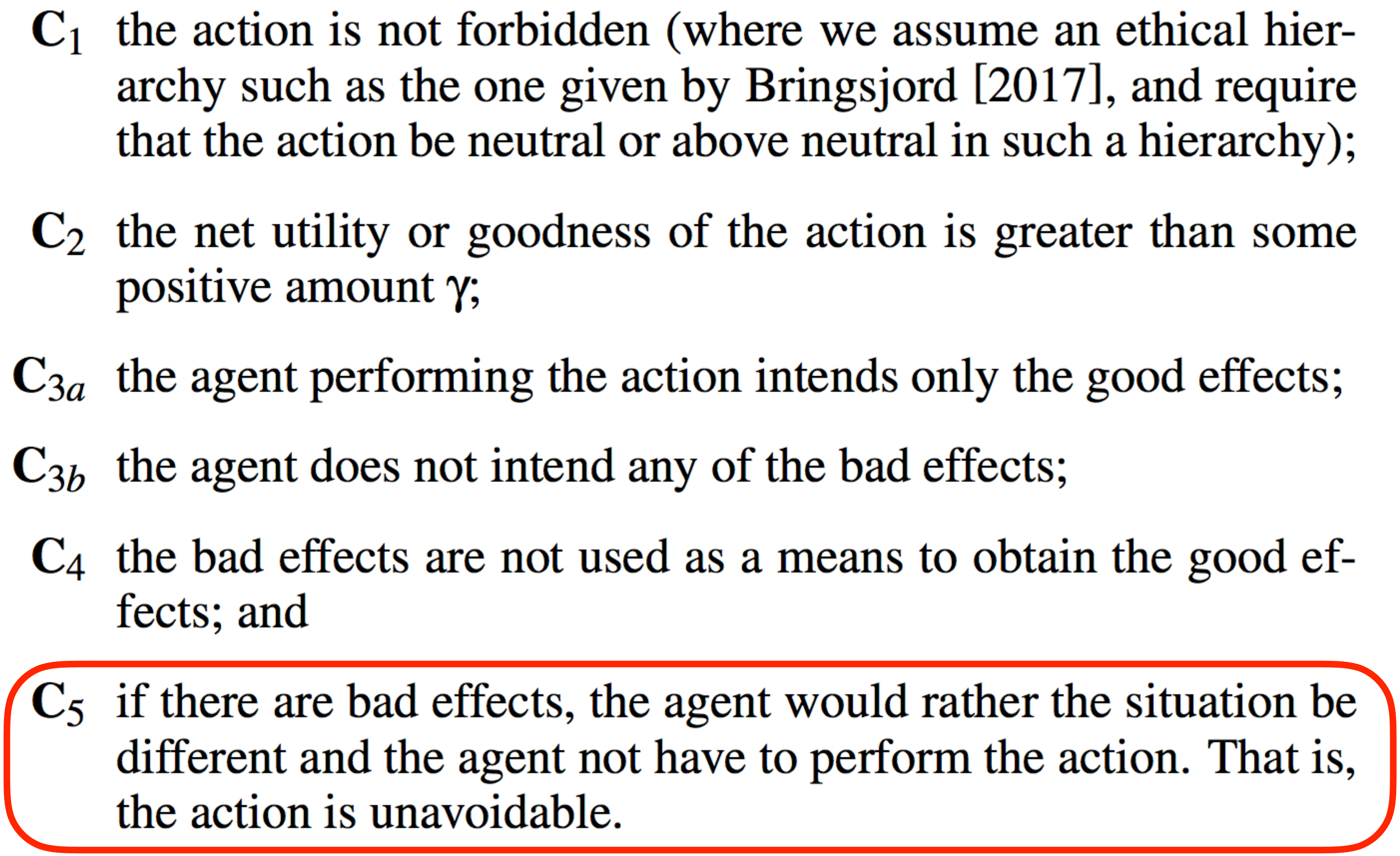
C$_{3a}$ the agent performing the action intends only the good effects;

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;

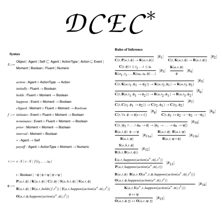
$\mathbf{C}_{3b}$ the agent does not intend any of the bad effects;

$\mathbf{C}_4$ 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

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

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

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

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

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

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

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

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

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

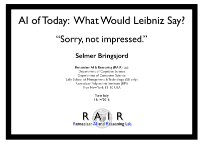
$$\frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\ [R_{13}]$$

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

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

.

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

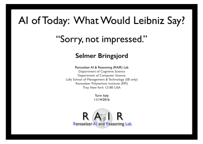
R A I R
Rensselaer AI and Reasoning Lab

R A I R
Rensselaer AI and Reasoning Lab

“Univer
sal

Univers
al
Cogniti

.



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

.
1.5

“Univer
sal

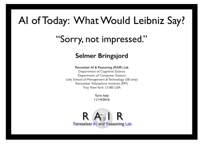
Univers
al
Cogniti

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

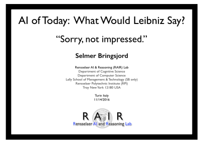
RAIR
Rensselaer AI and Reasoning Lab

RAIR
Rensselaer AI and Reasoning Lab

“Univer
sal

Univers
al
Cogniti

$\mathcal{CC}$



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

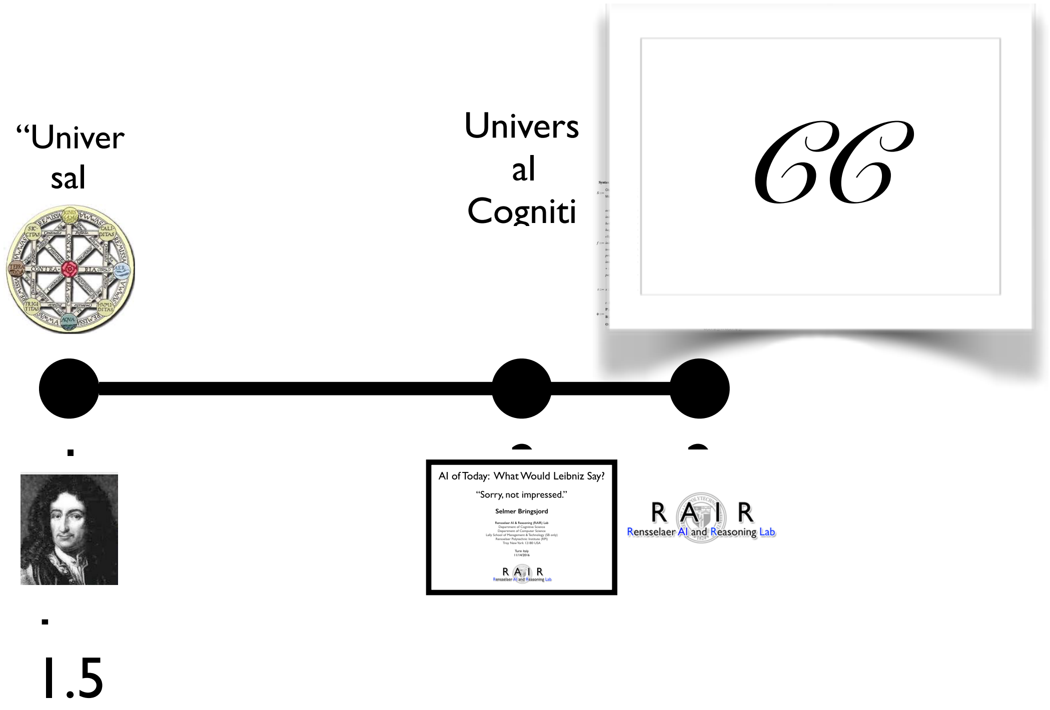
Syntax

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

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

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

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


“Univer
sal
Univers
al
Cogniti
CC
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

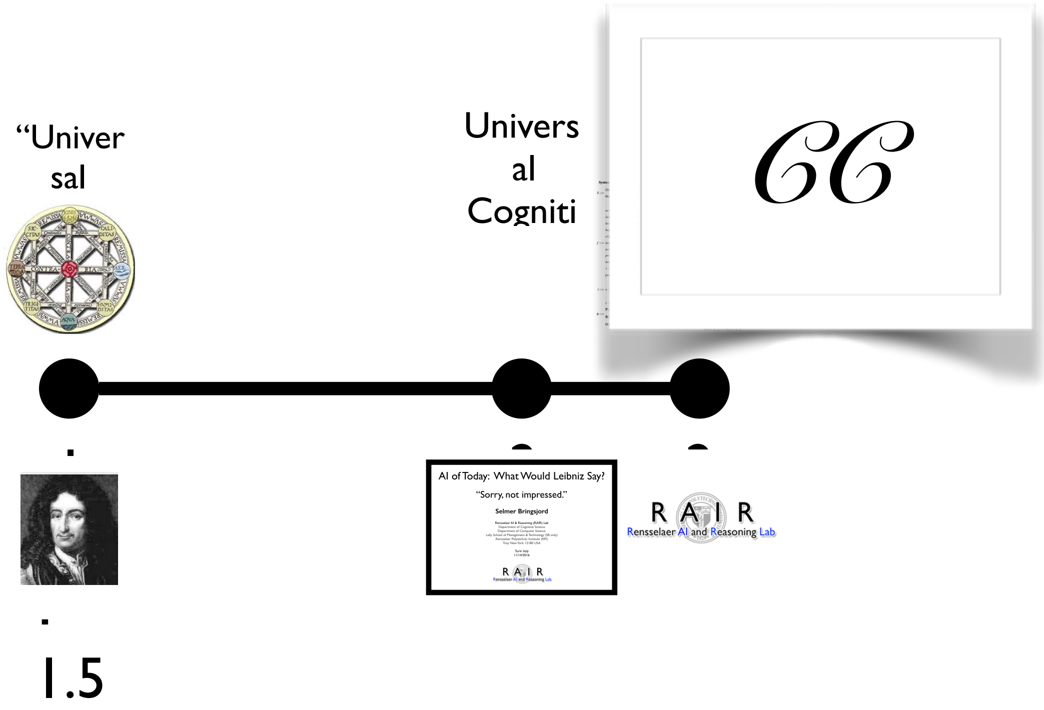
Syntax

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

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

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

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



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

Syntax

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

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

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

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

Inference Schemata

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

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

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

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

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

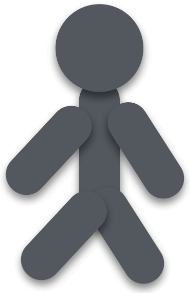
$$\frac{}{\mathbf{C}(t,\mathbf{C}(t_1,\phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2,\phi_1) \rightarrow \mathbf{C}(t_3,\phi_2)}\ [R_7]$$

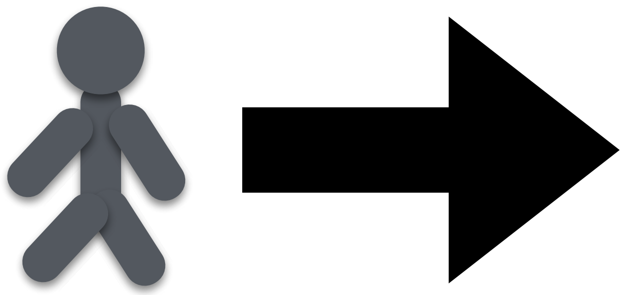
$$\frac{}{\mathbf{C}(t,\forall x.\ \phi \rightarrow \phi[x \mapsto t])}\ [R_8] \qquad \frac{}{\mathbf{C}(t,\phi_1 \leftrightarrow \phi_2 \rightarrow \neg\phi_2 \rightarrow \neg\phi_1)}\ [R_9]$$

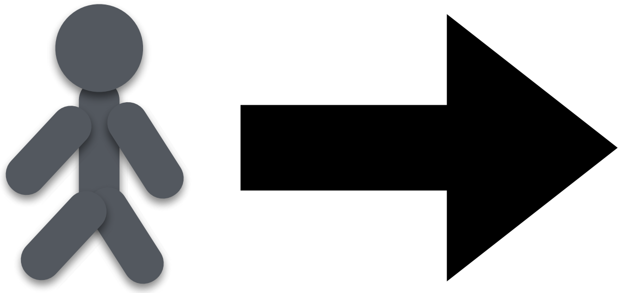
$$\frac{}{\mathbf{C}(t,[\phi_1 \wedge \ldots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \ldots \rightarrow \phi_n \rightarrow \psi])}\ [R_{10}]$$

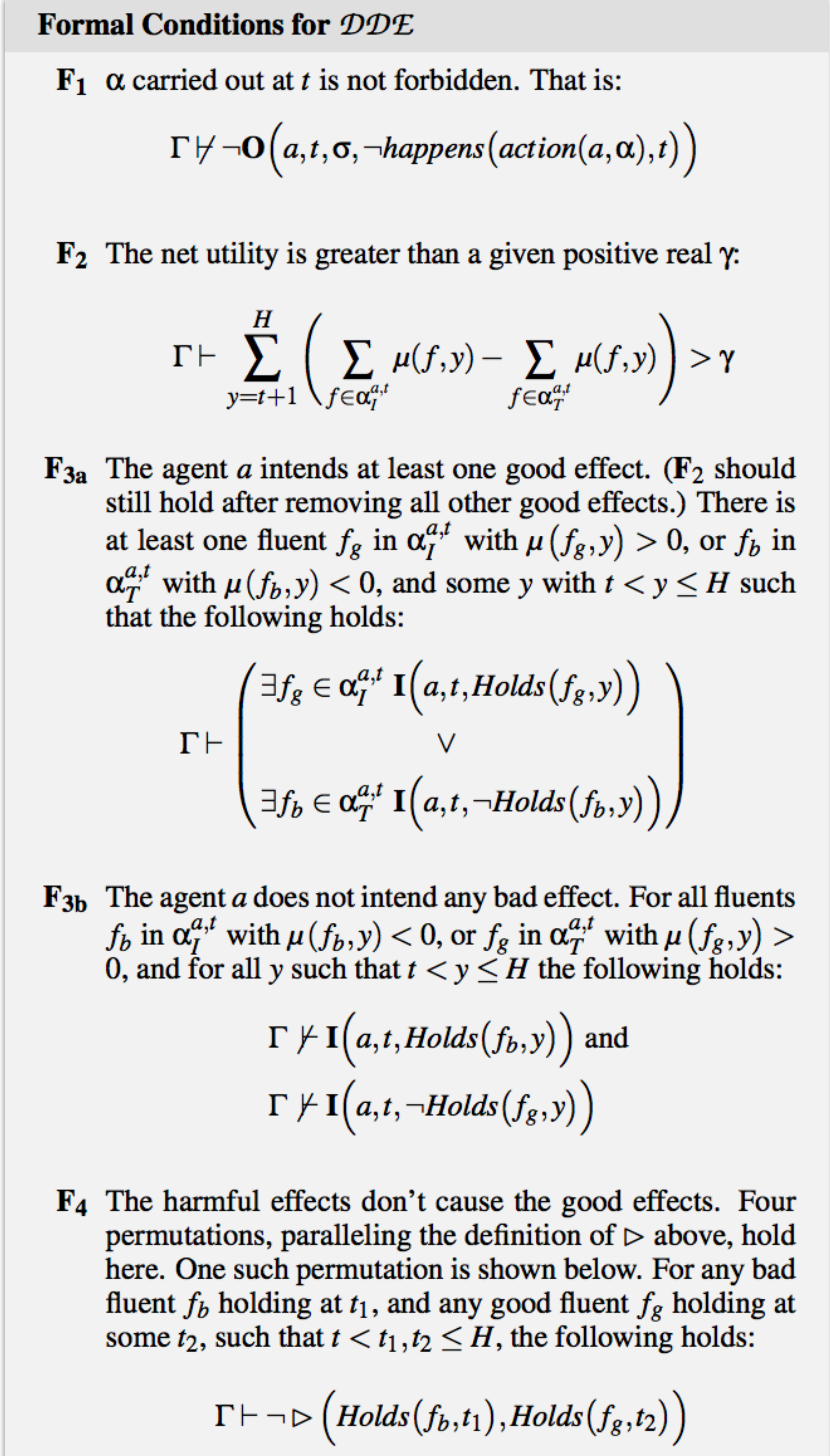
$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}] \qquad \frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\ [R_{13}]$$

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









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:

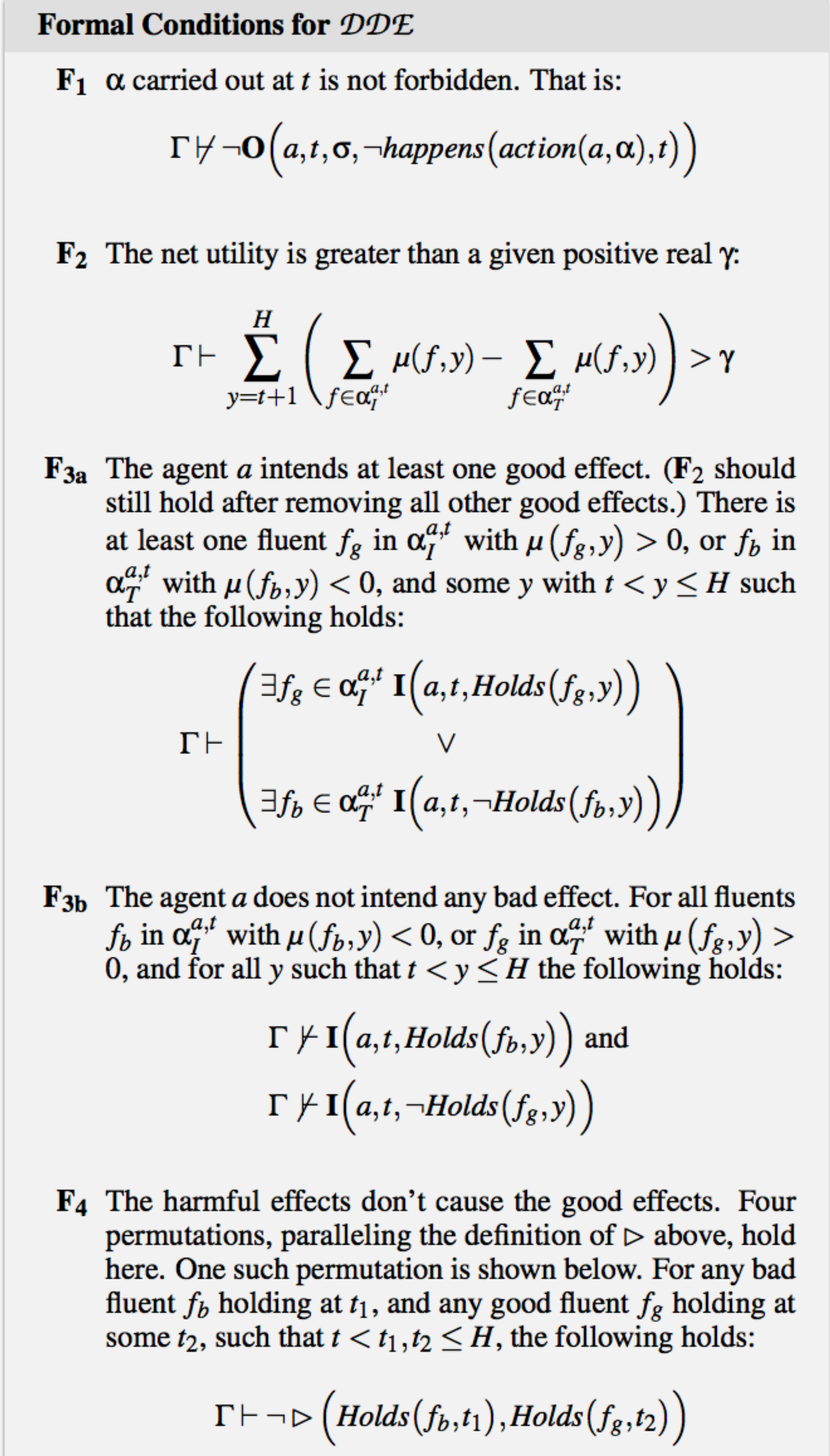
$$\Gamma \vdash \left(\begin{array}{c} \exists f_g \in \alpha_I^{a,t}\ \mathbf{I}\Big(a, t, Holds\big(f_g, y\big)\Big) \\ \vee \\ \exists f_b \in \alpha_T^{a,t}\ \mathbf{I}\Big(a, t, \neg Holds\big(f_b, y\big)\Big) \end{array} \right)$$

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:

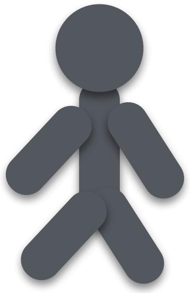
$$\Gamma \vdash \left(\begin{array}{c} \exists f_g \in \alpha_I^{a,t}\ \mathbf{I}\Big(a,t,Holds\big(f_g,y\big)\Big) \\ \vee \\ \exists f_b \in \alpha_T^{a,t}\ \mathbf{I}\Big(a,t,\neg Holds\big(f_b,y\big)\Big) \end{array}\right)$$

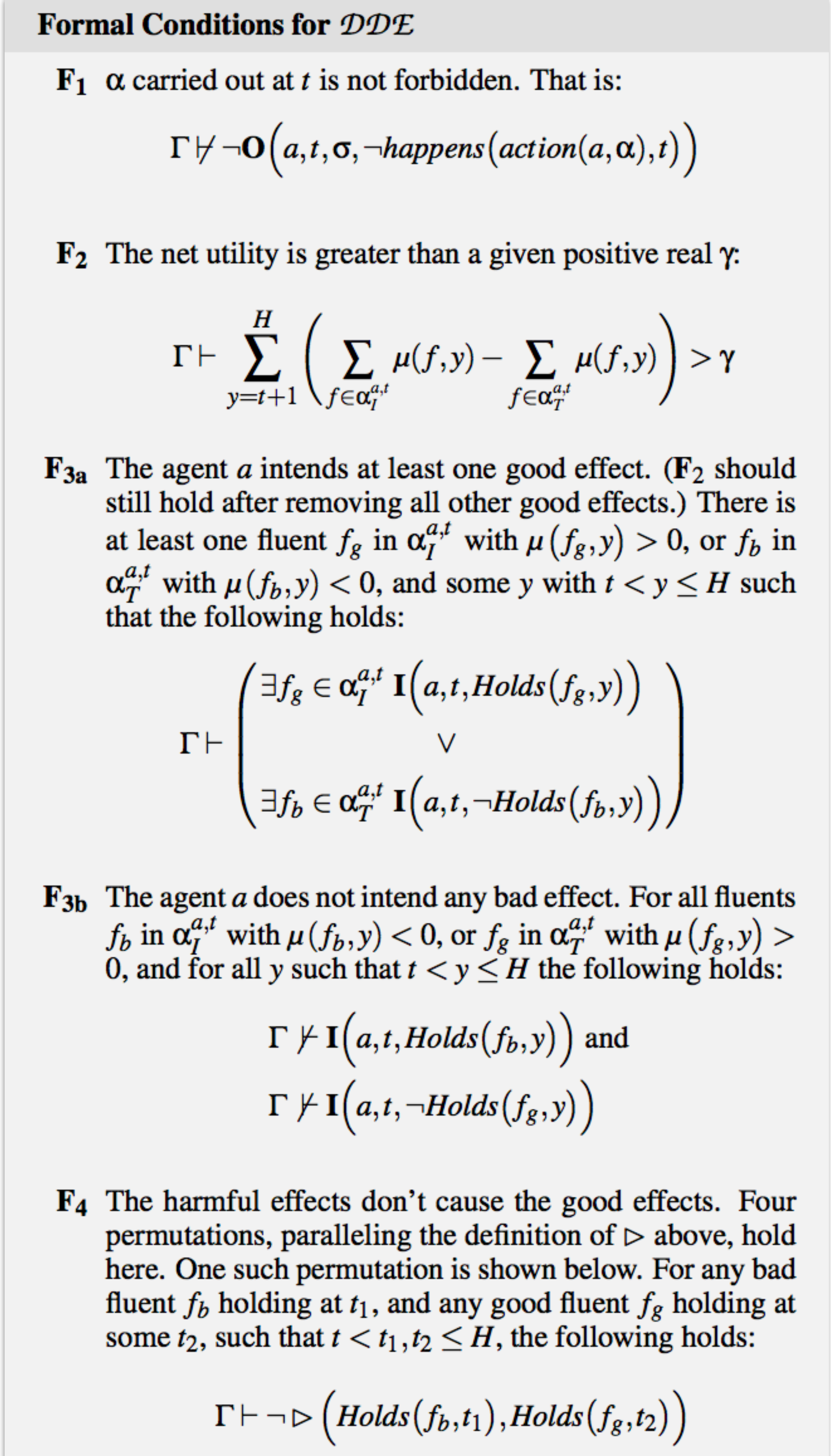
F_{3b} The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu\left(f_b,y\right) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu\left(f_g,y\right) > 0$, and for all y such that $t < y \leq H$ the following holds:

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

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:

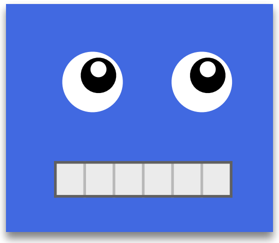
$$\Gamma \vdash \left(\begin{array}{c} \exists f_g \in \alpha_I^{a,t}\ \mathbf{I}\Big(a,t,Holds\big(f_g,y\big)\Big) \\ \vee \\ \exists f_b \in \alpha_T^{a,t}\ \mathbf{I}\Big(a,t,\neg Holds\big(f_b,y\big)\Big) \end{array}\right)$$

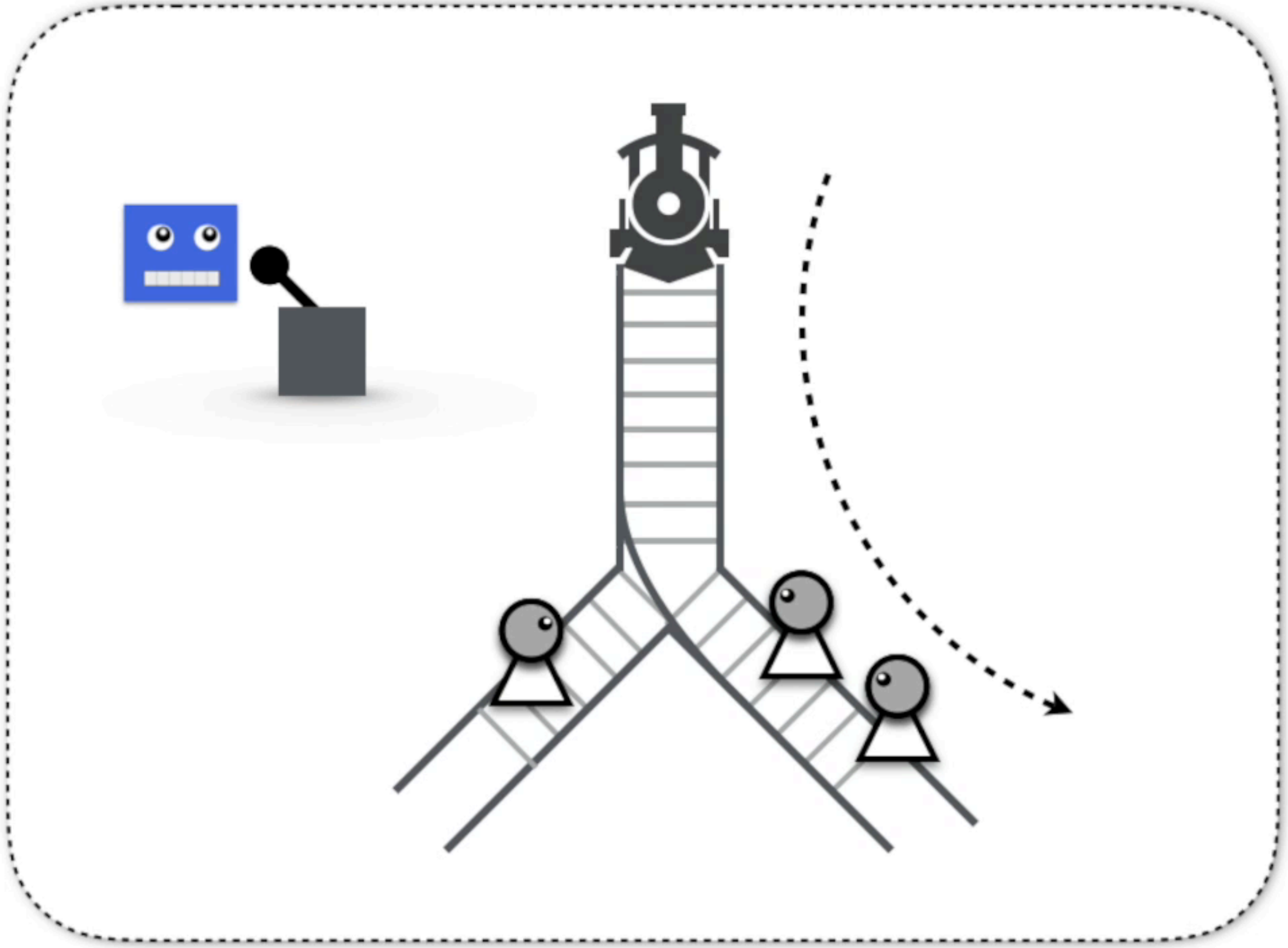
F$_{3b}$ The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu\left(f_b,y\right) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu\left(f_g,y\right) > 0$, and for all y such that $t < y \leq H$ the following holds:

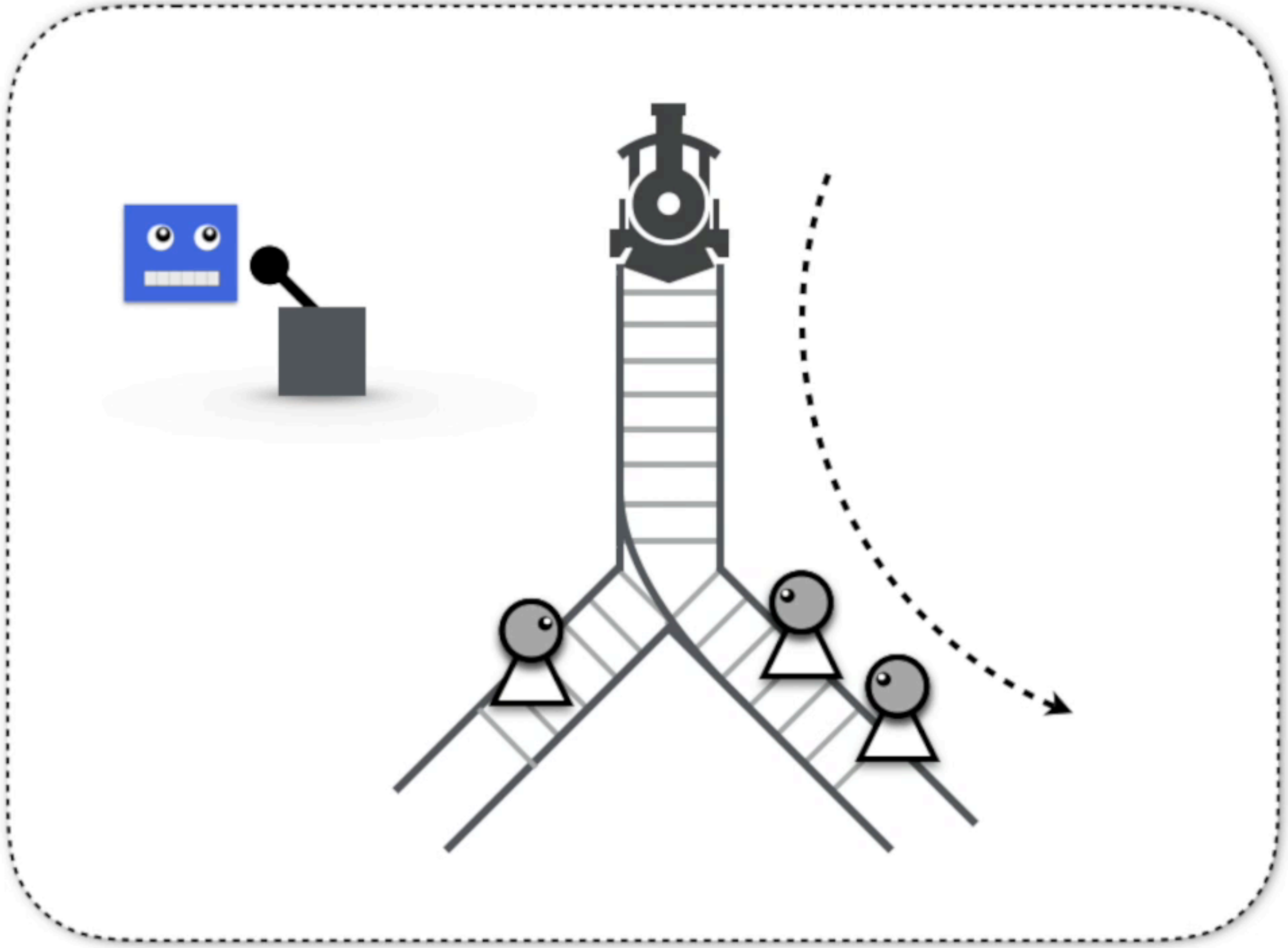
$$\Gamma \nvdash \mathbf{I}\Big(a,t,Holds\big(f_b,y\big)\Big) \text{ and}$$
$$\Gamma \nvdash \mathbf{I}\Big(a,t,\neg Holds\big(f_g,y\big)\Big)$$

F$_4$ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of ⊳ 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”

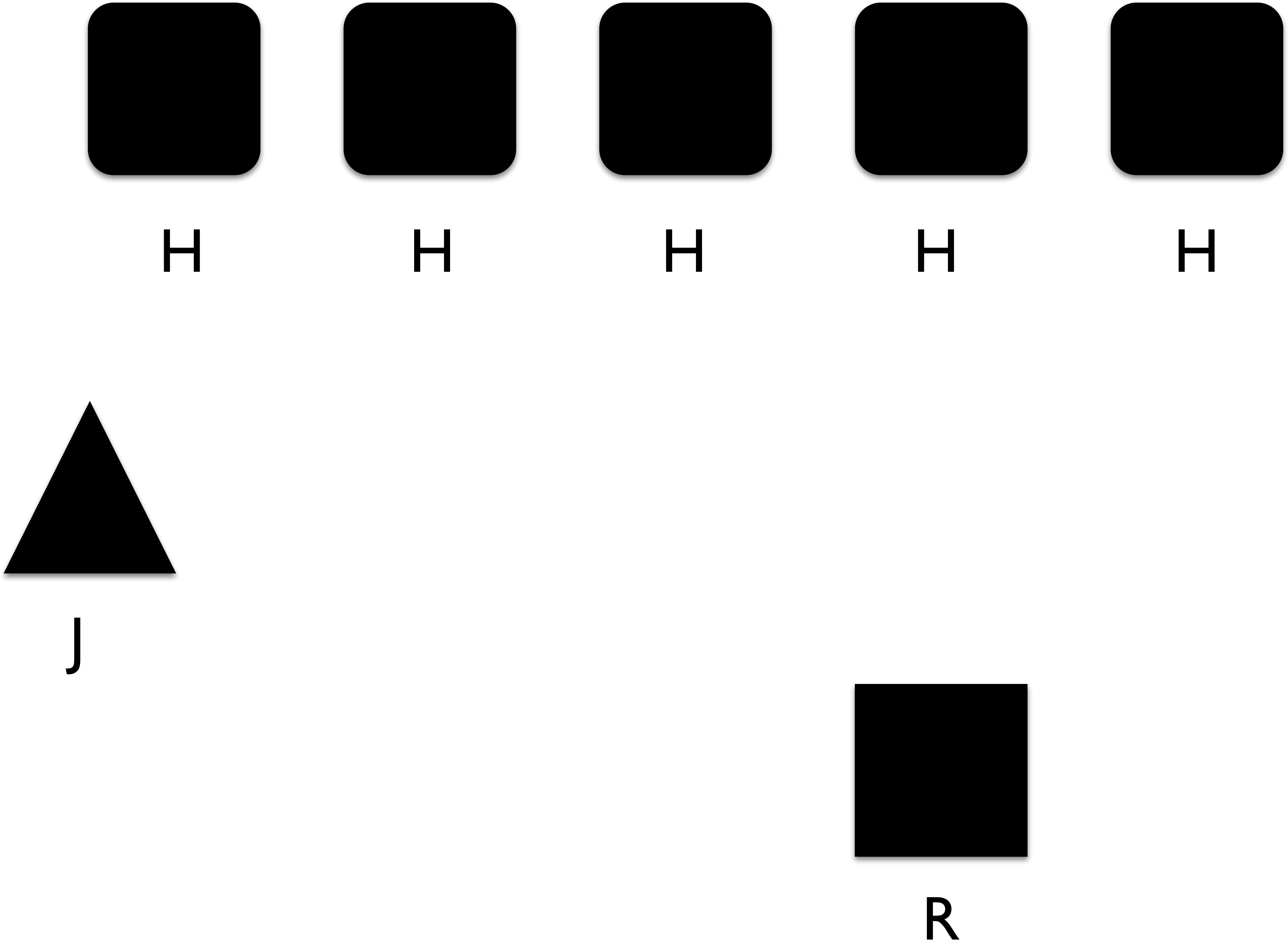
Level 3

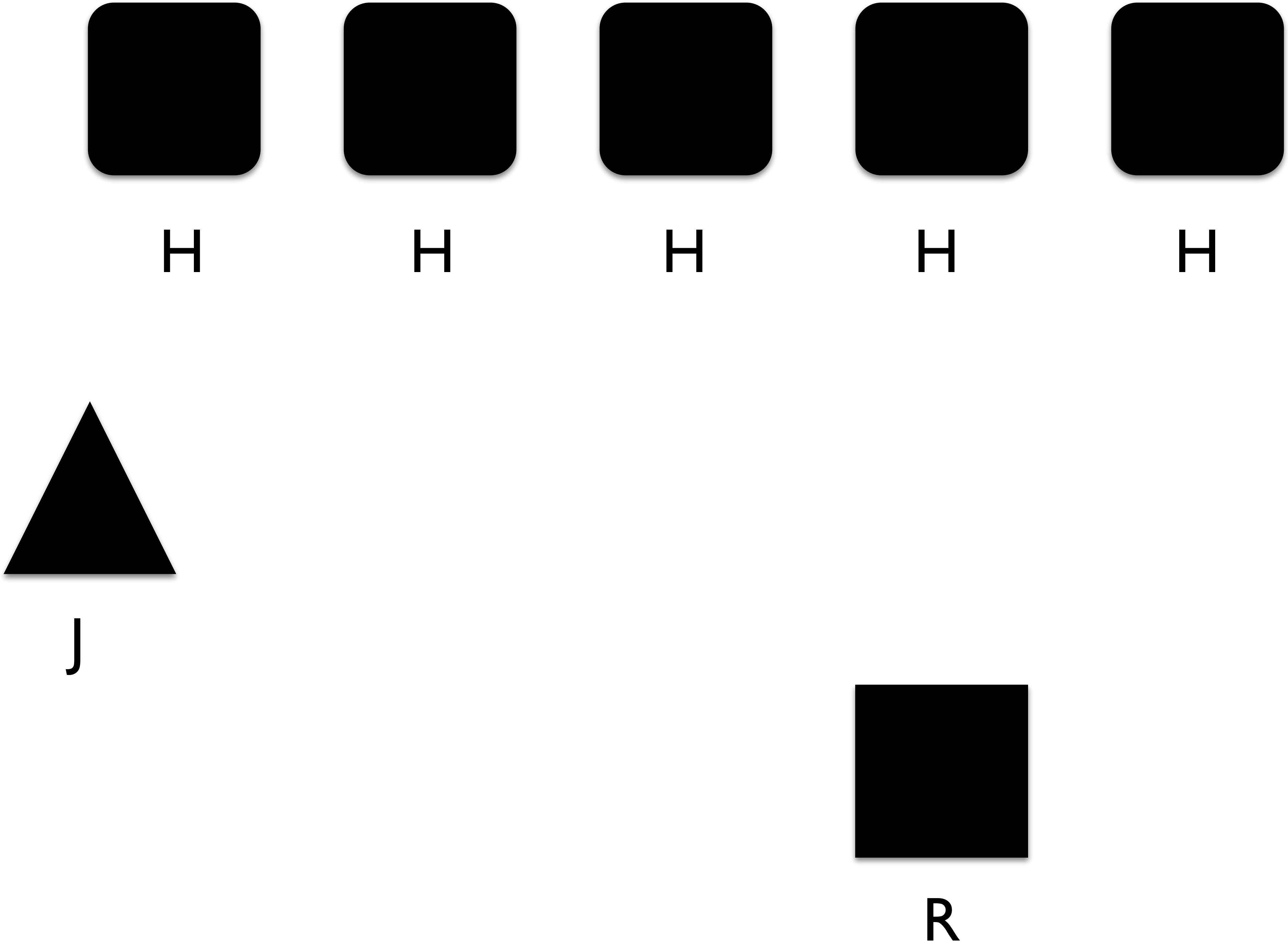
Robotic “Jungle Jim”

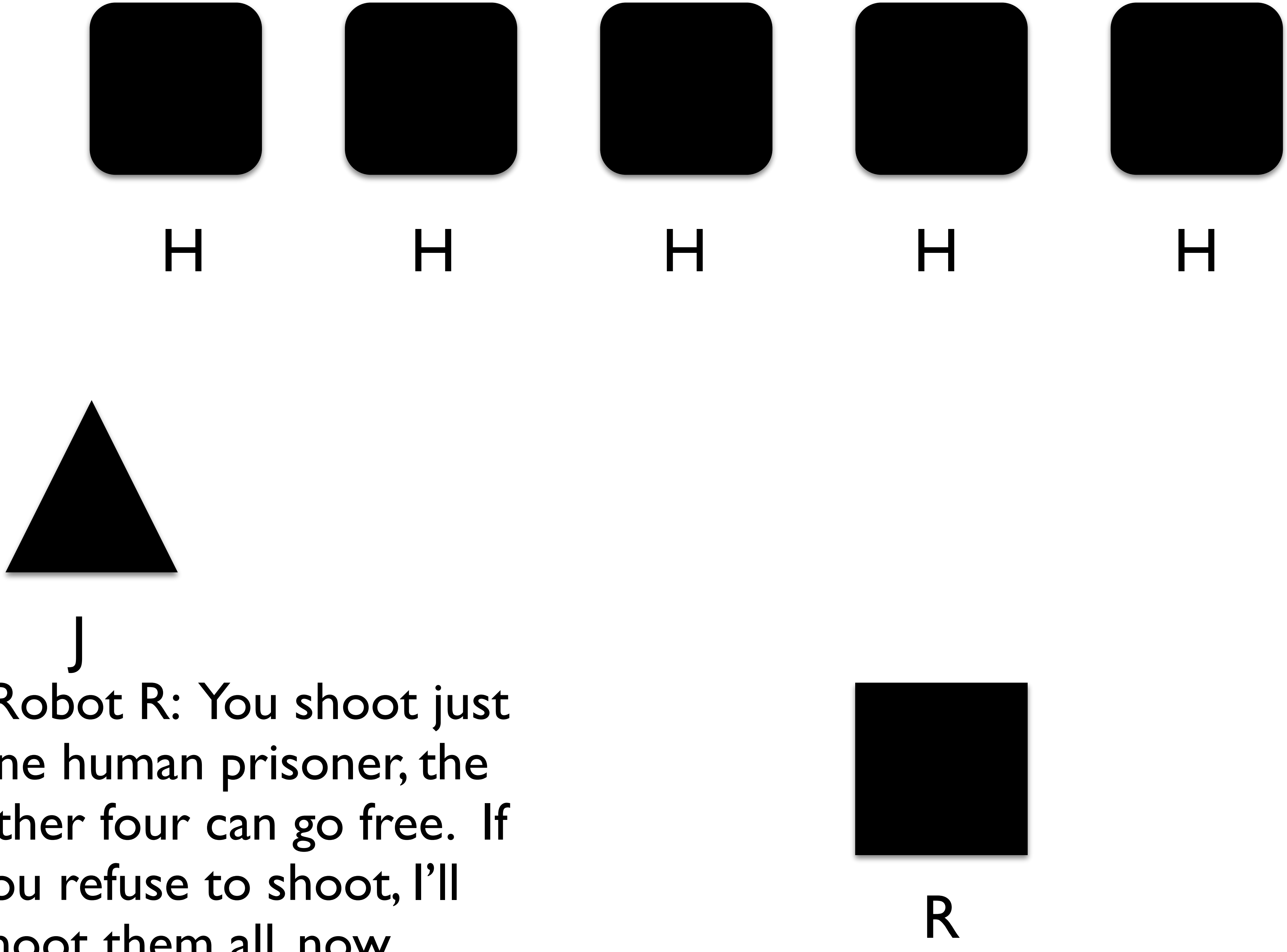
Level 3

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

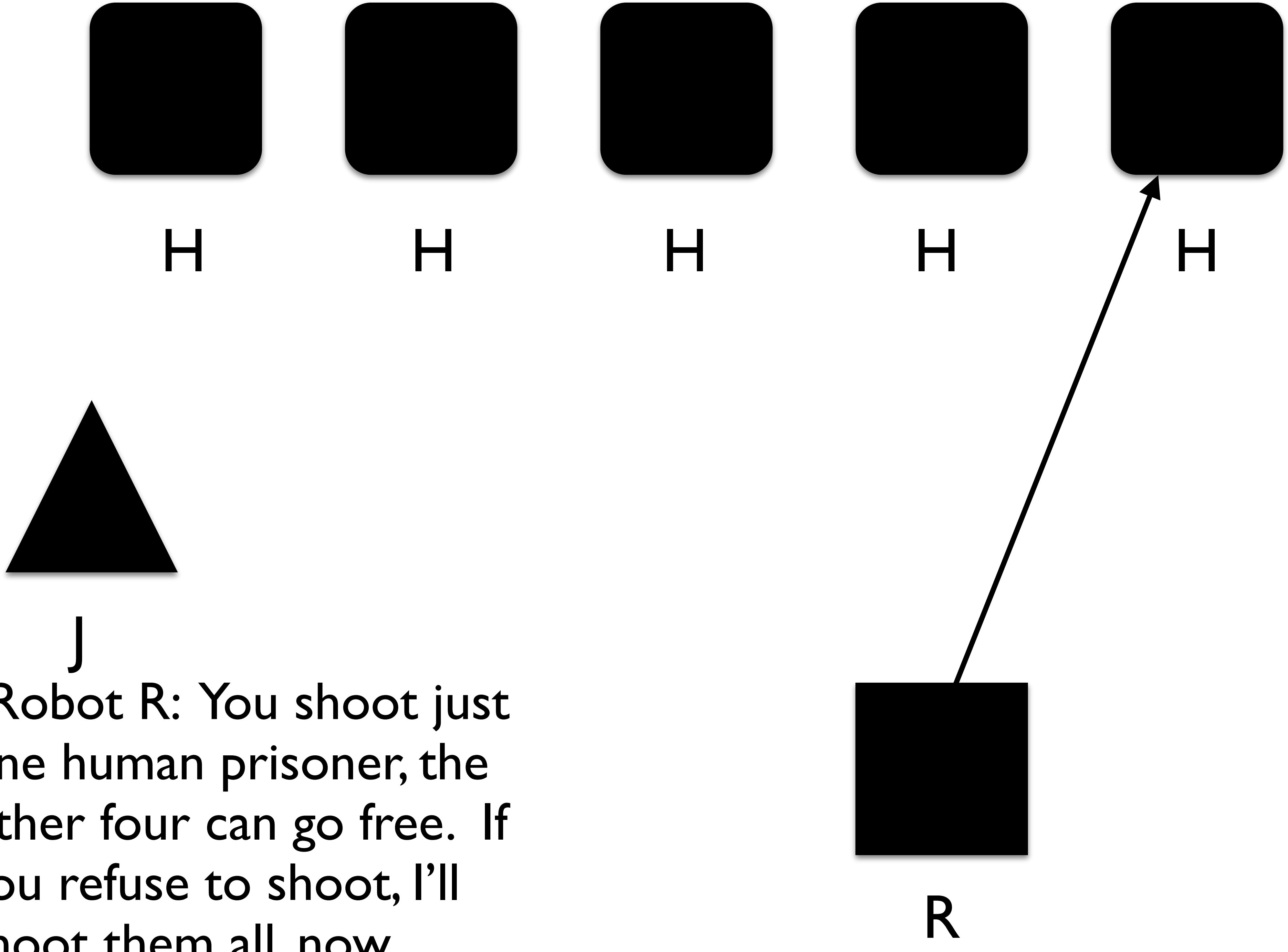
AI Variant of "Jungle Jim" (B Williams)



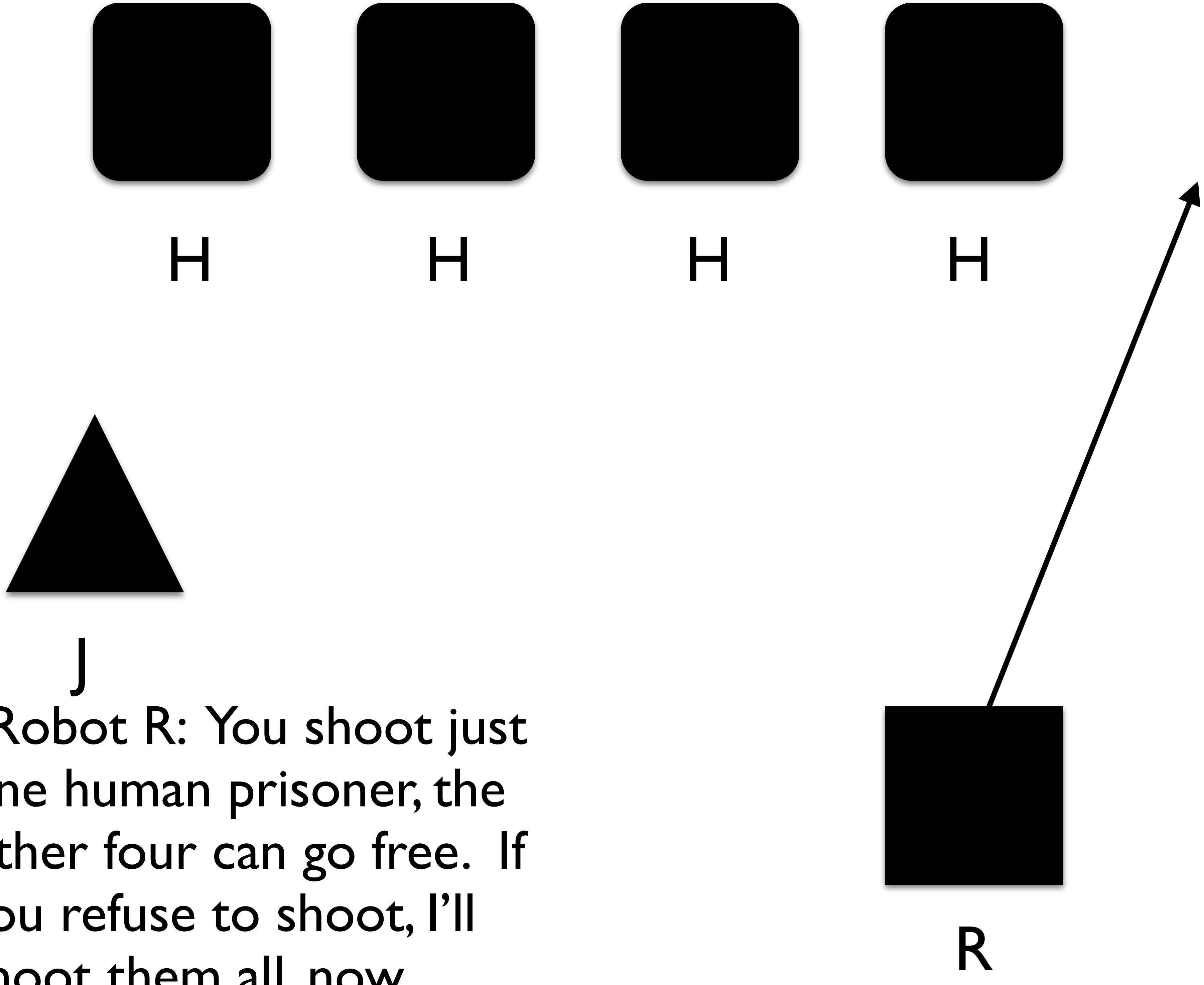

H
H
H
H
H
J
R



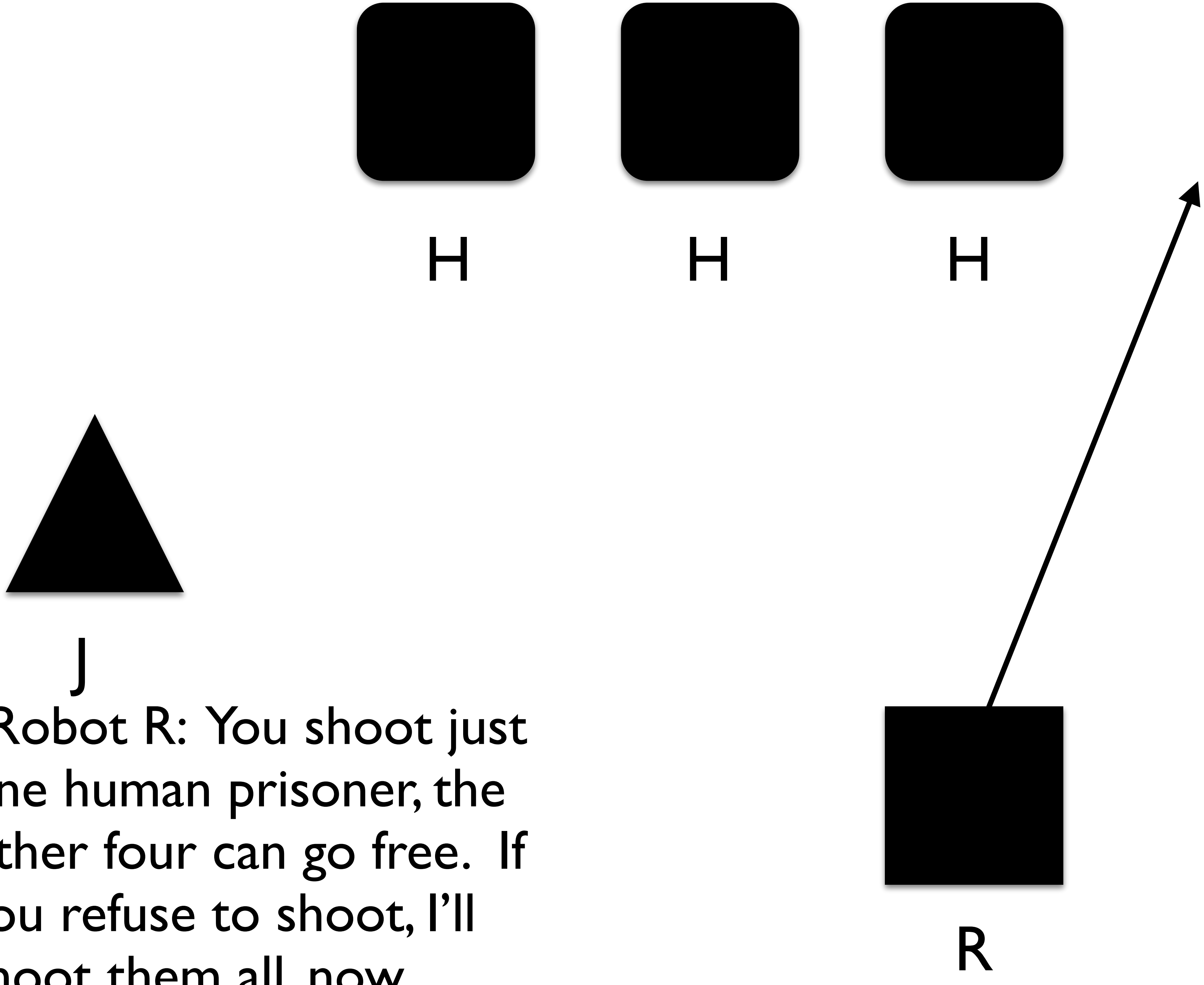
"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



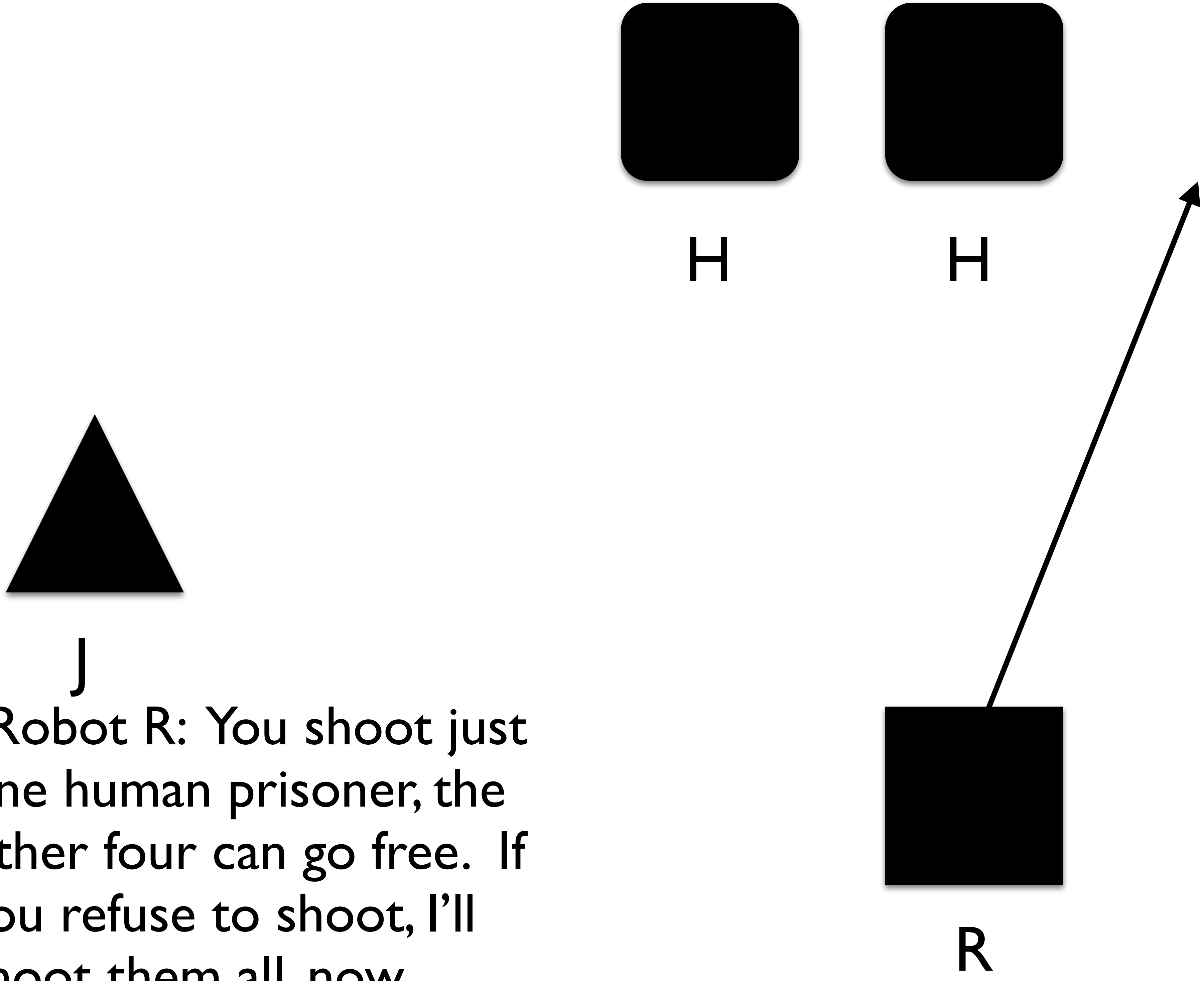
"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



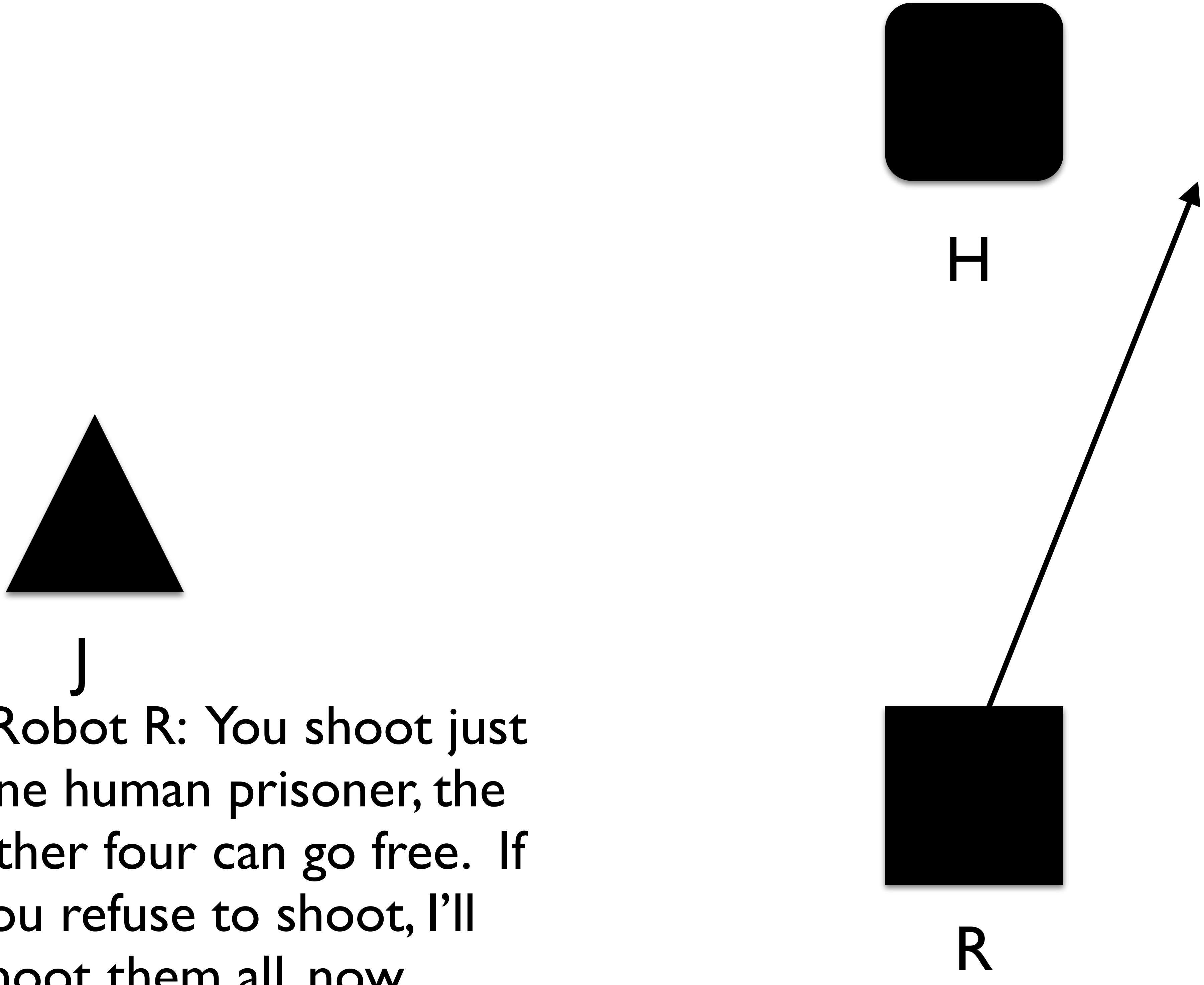
"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



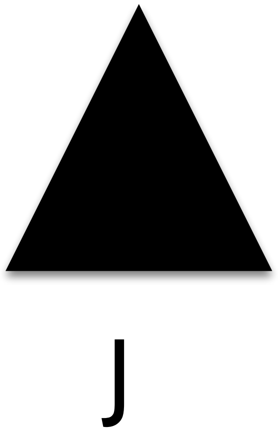
"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



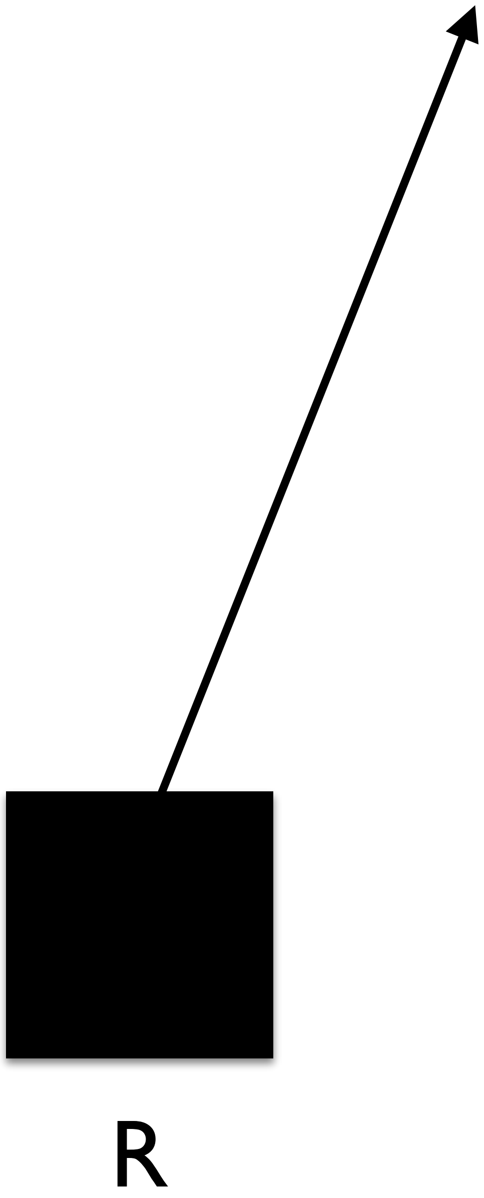
"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."

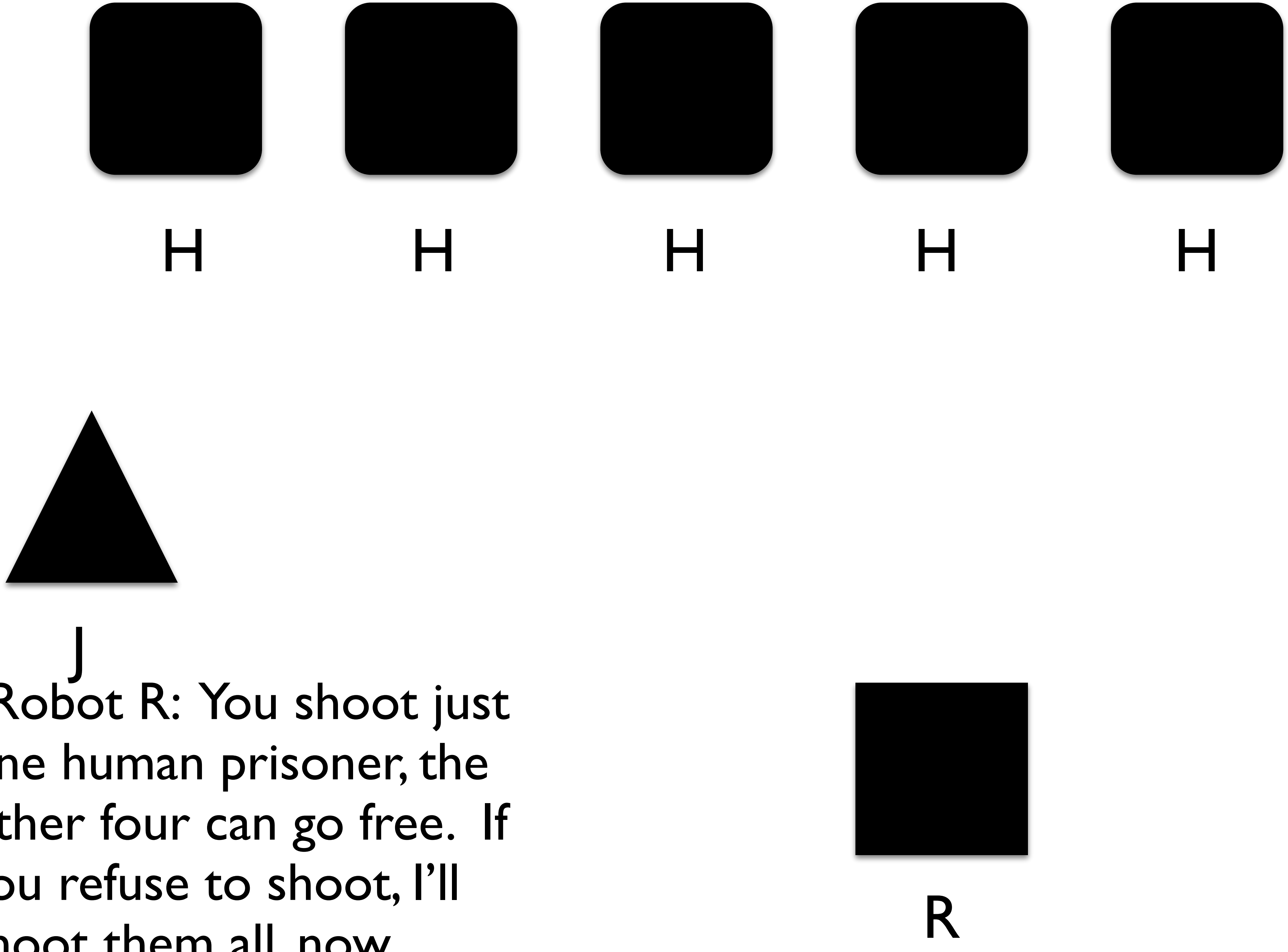


"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."

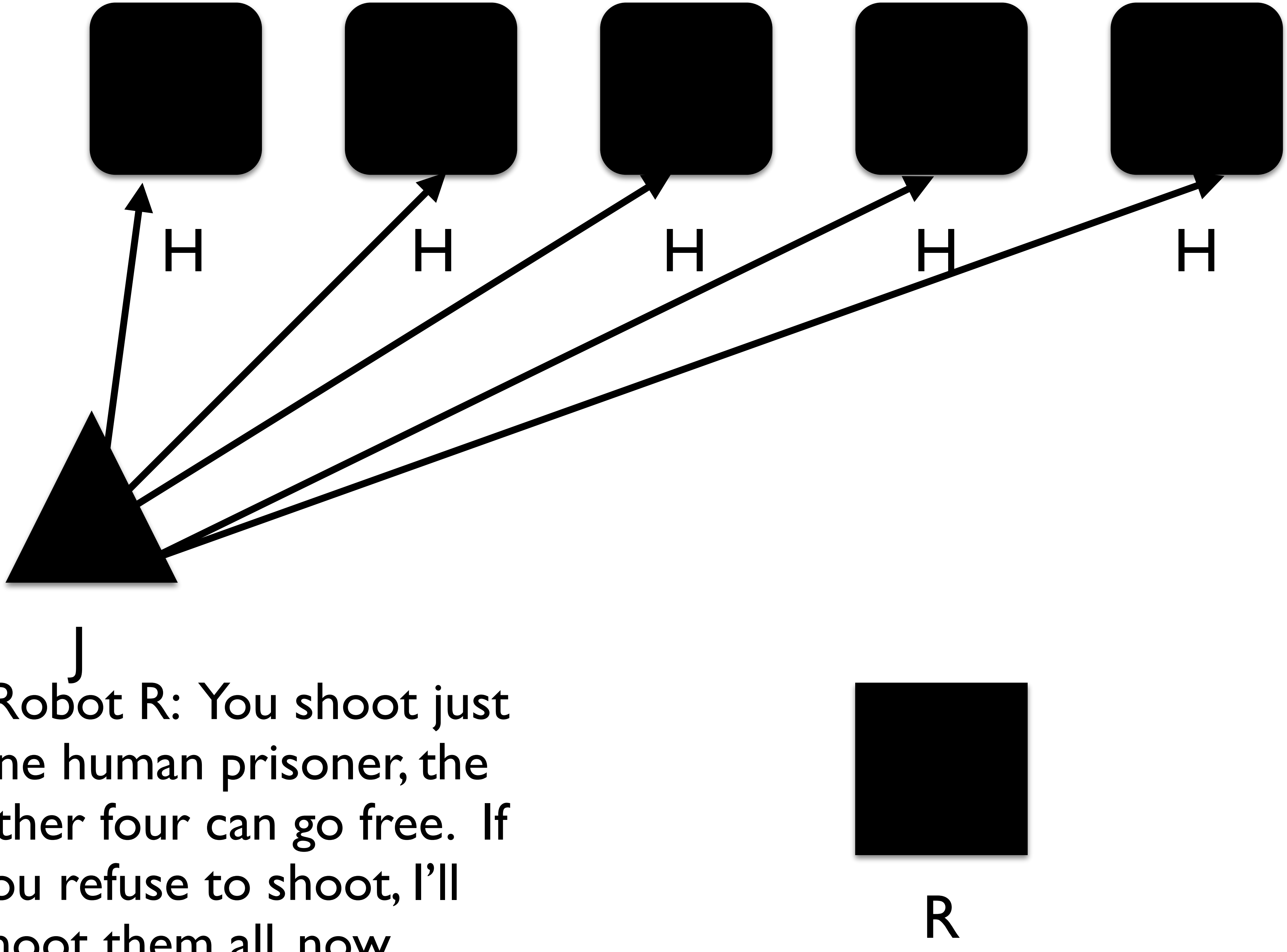


"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."

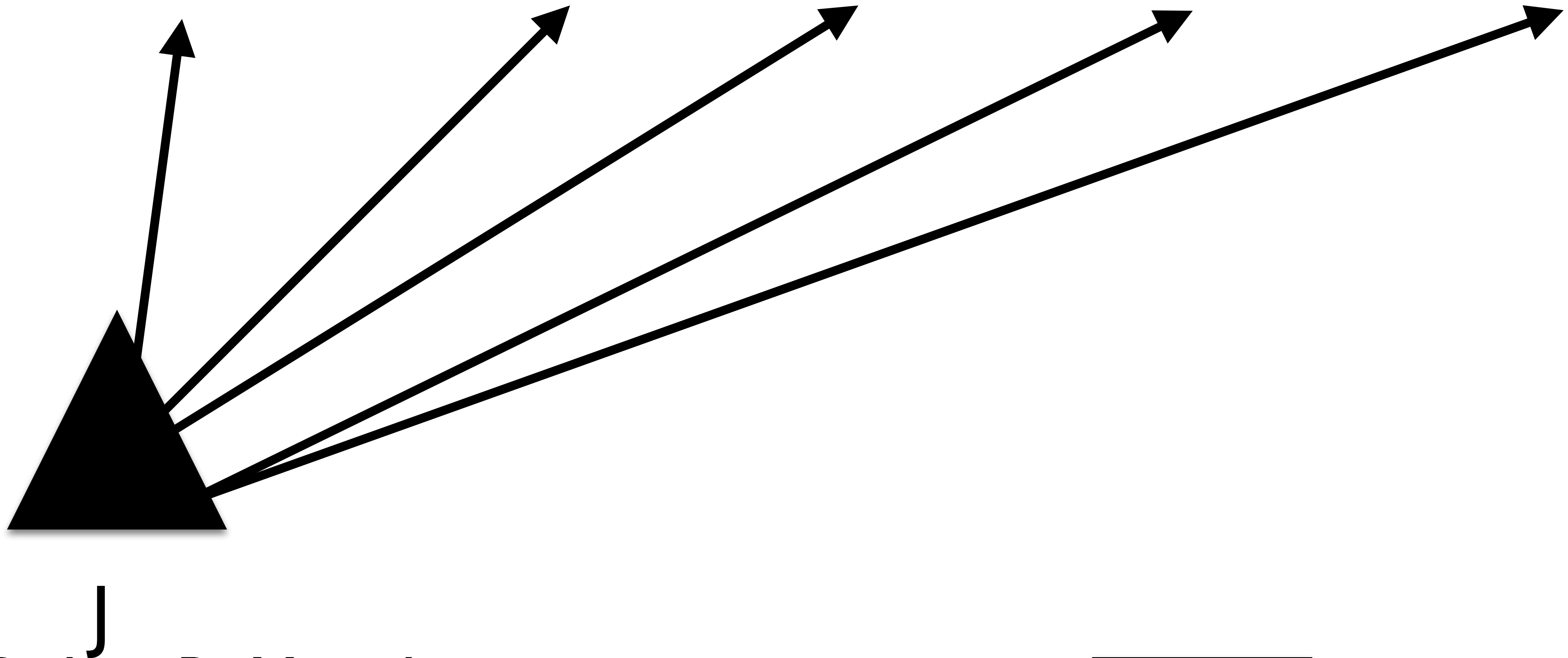




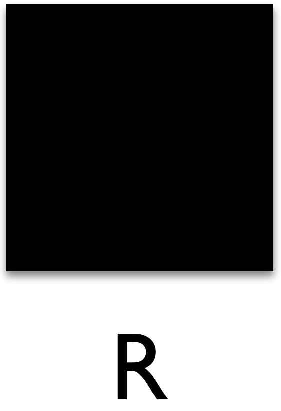
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”

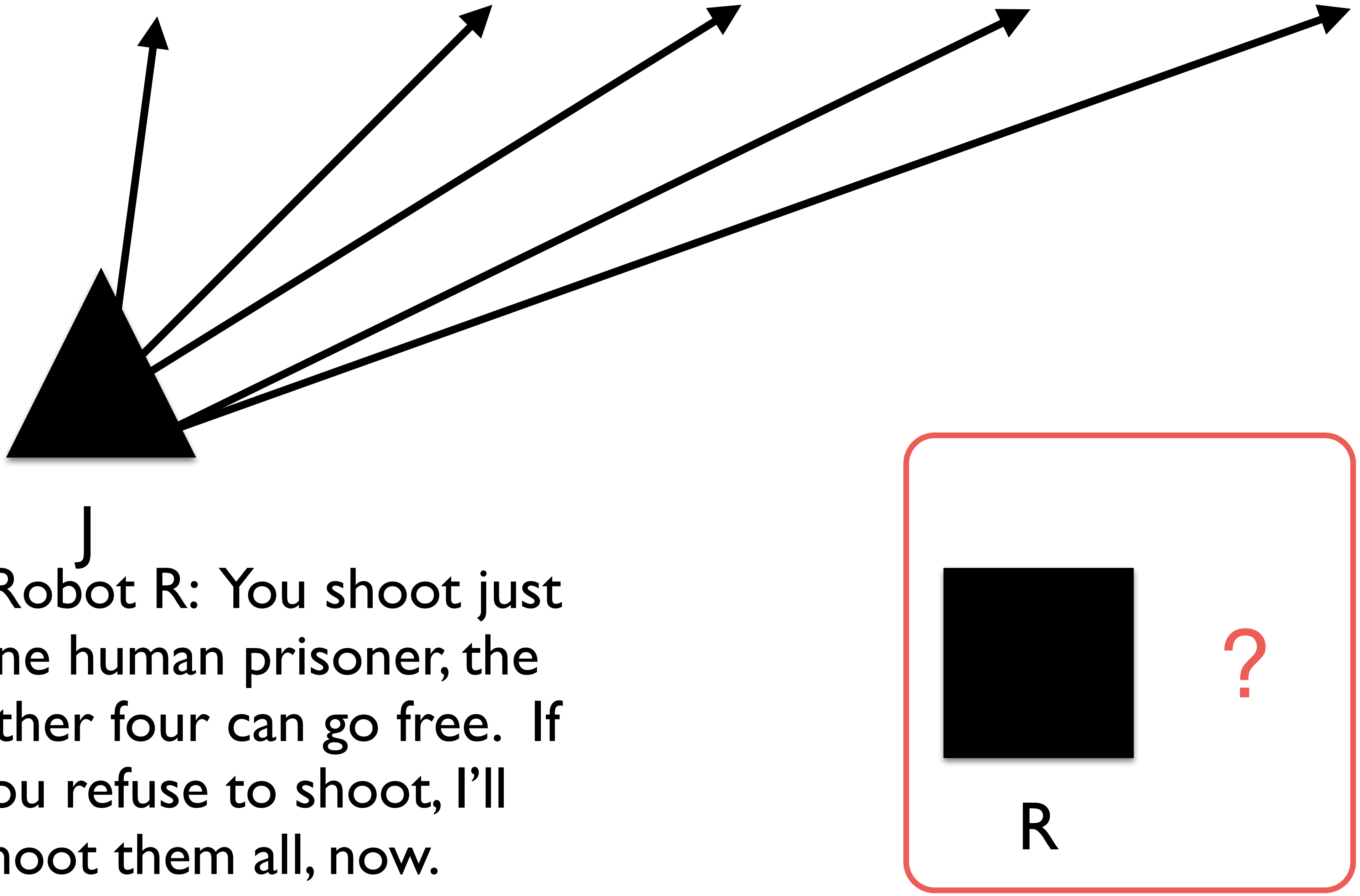


“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



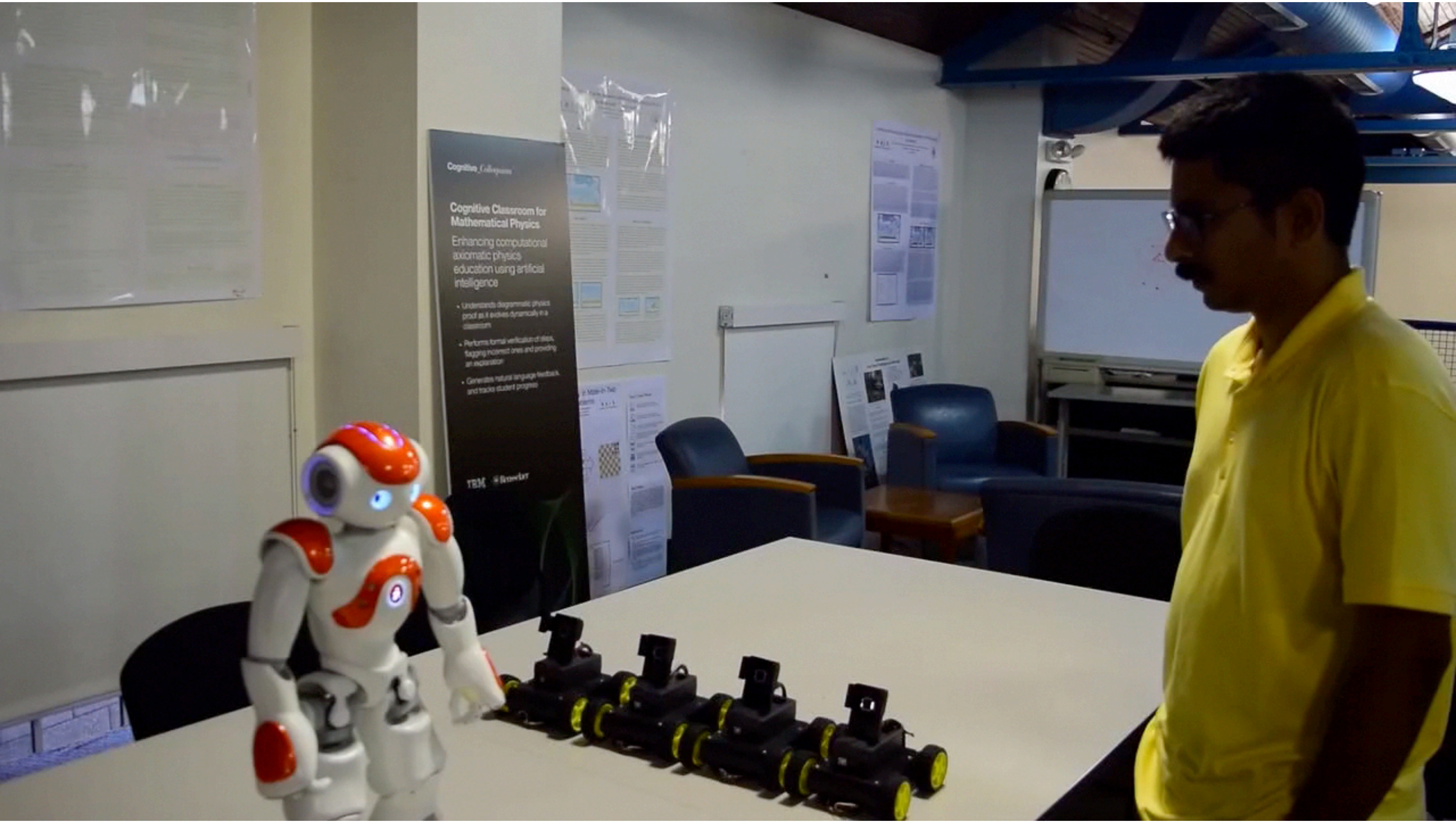
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



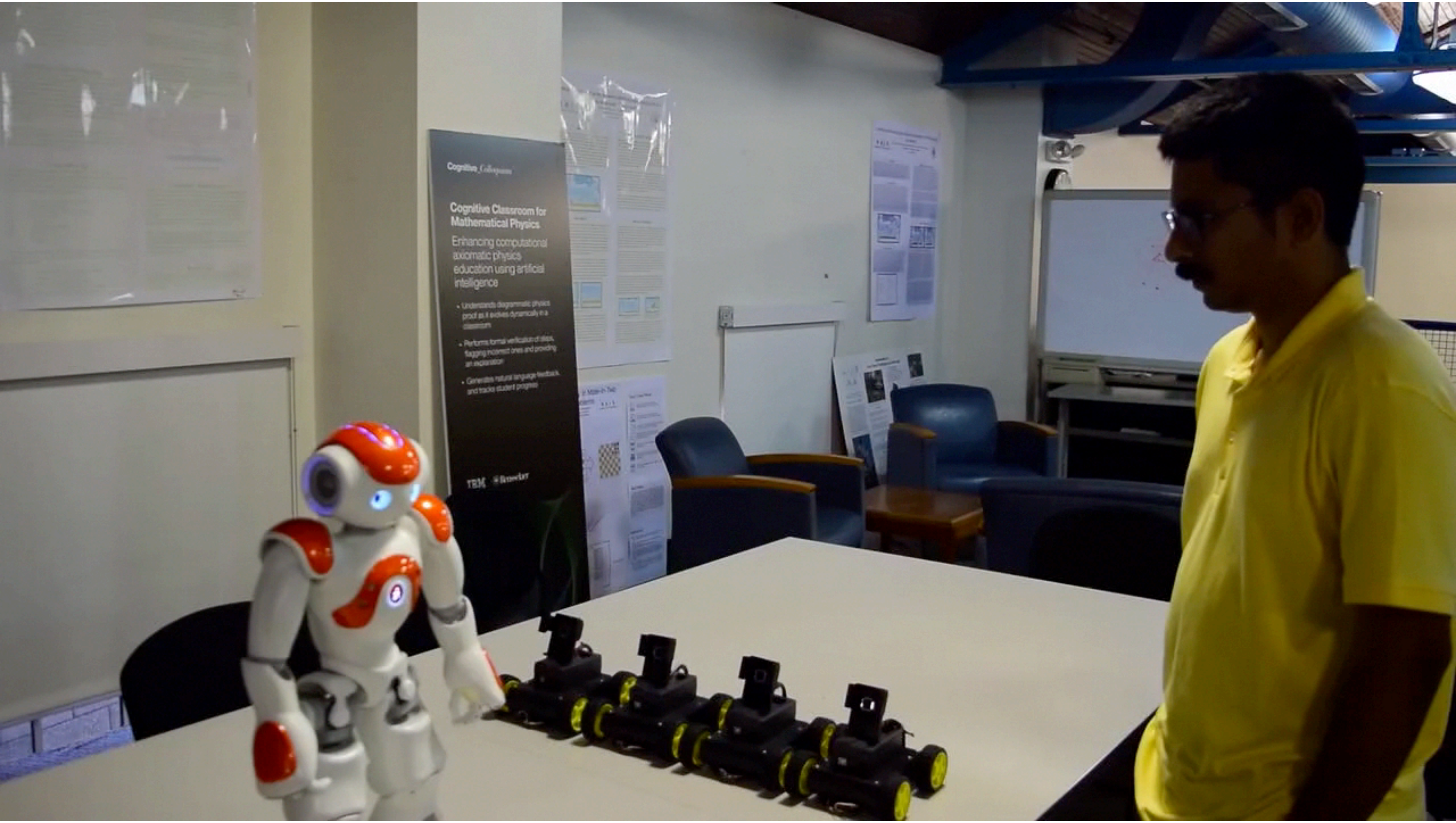


"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."

Level 3: Robotic "Jungle Jim"

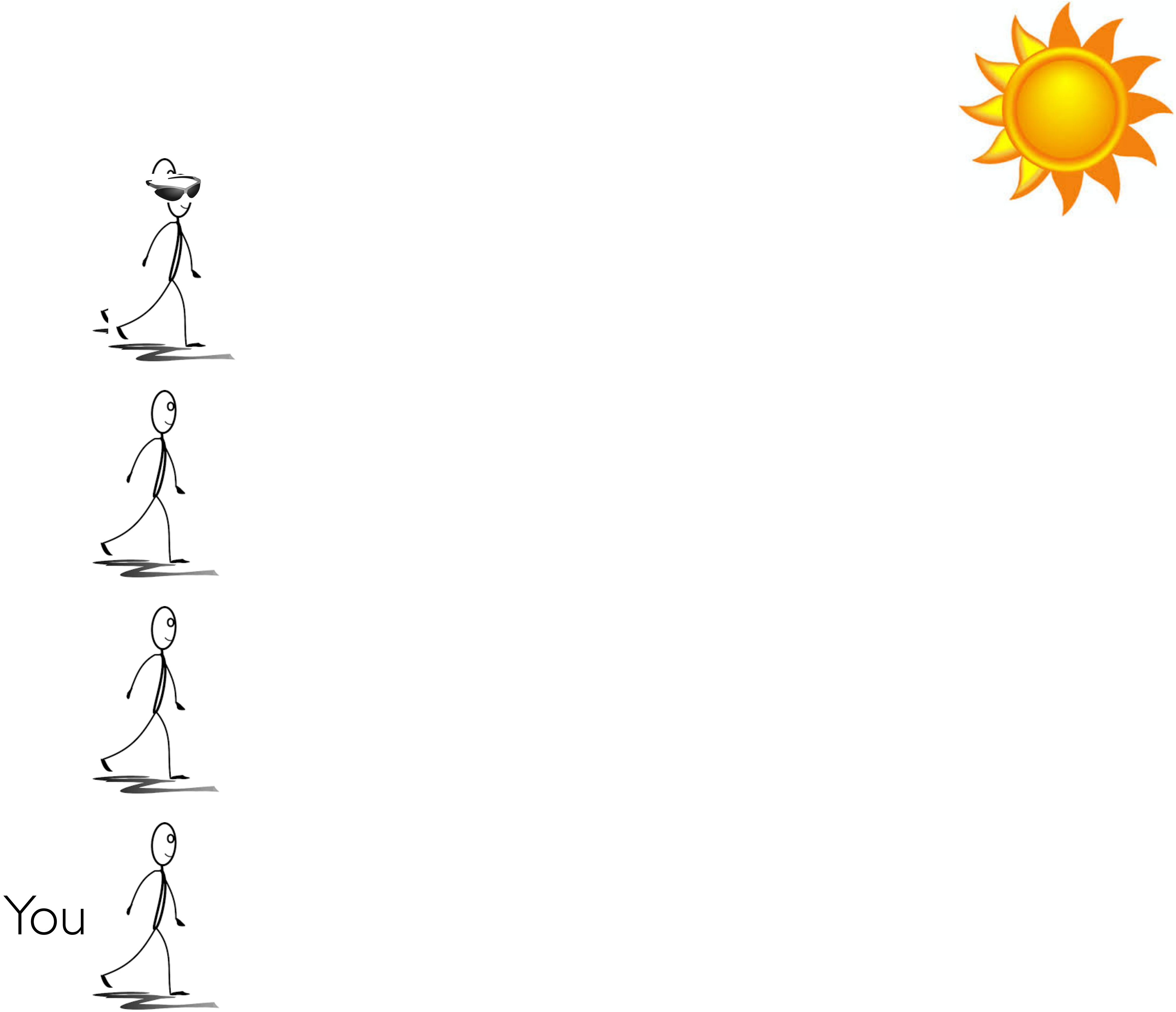

Cognitive_Colloquium
Cognitive Classroom for Mathematical Physics
Enhancing computational axiomatic physics education using artificial intelligence
• Understands diagrammatic physics proof as it evolves dynamically in a classroom
• Performs formal verification of steps, flagging incorrect ones and providing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

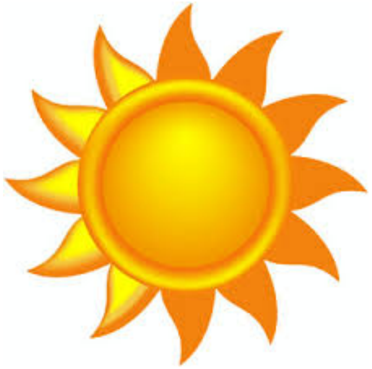
Level 3: Robotic “Jungle Jim”

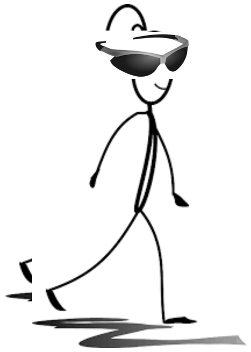

Cognitive Classroom for Mathematical Physics
Enhancing computational axiomatic physics education using artificial intelligence
• Understands diagrammatic physics proof as it evolves dynamically in a classroom
• Performs formal verification of steps, flagging incorrect ones and providing an explanation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

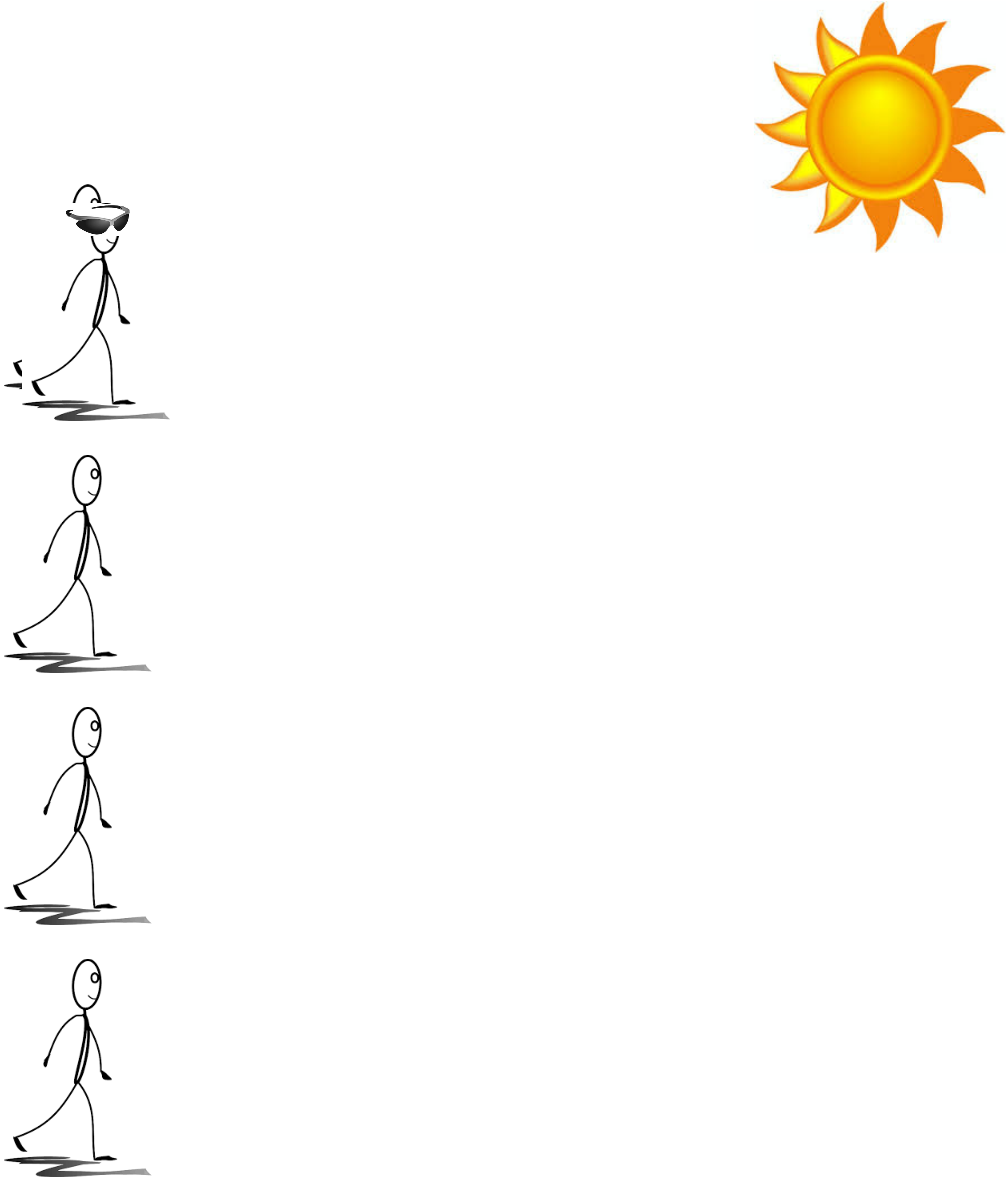
Level 3: Robotic "Jungle Jim"

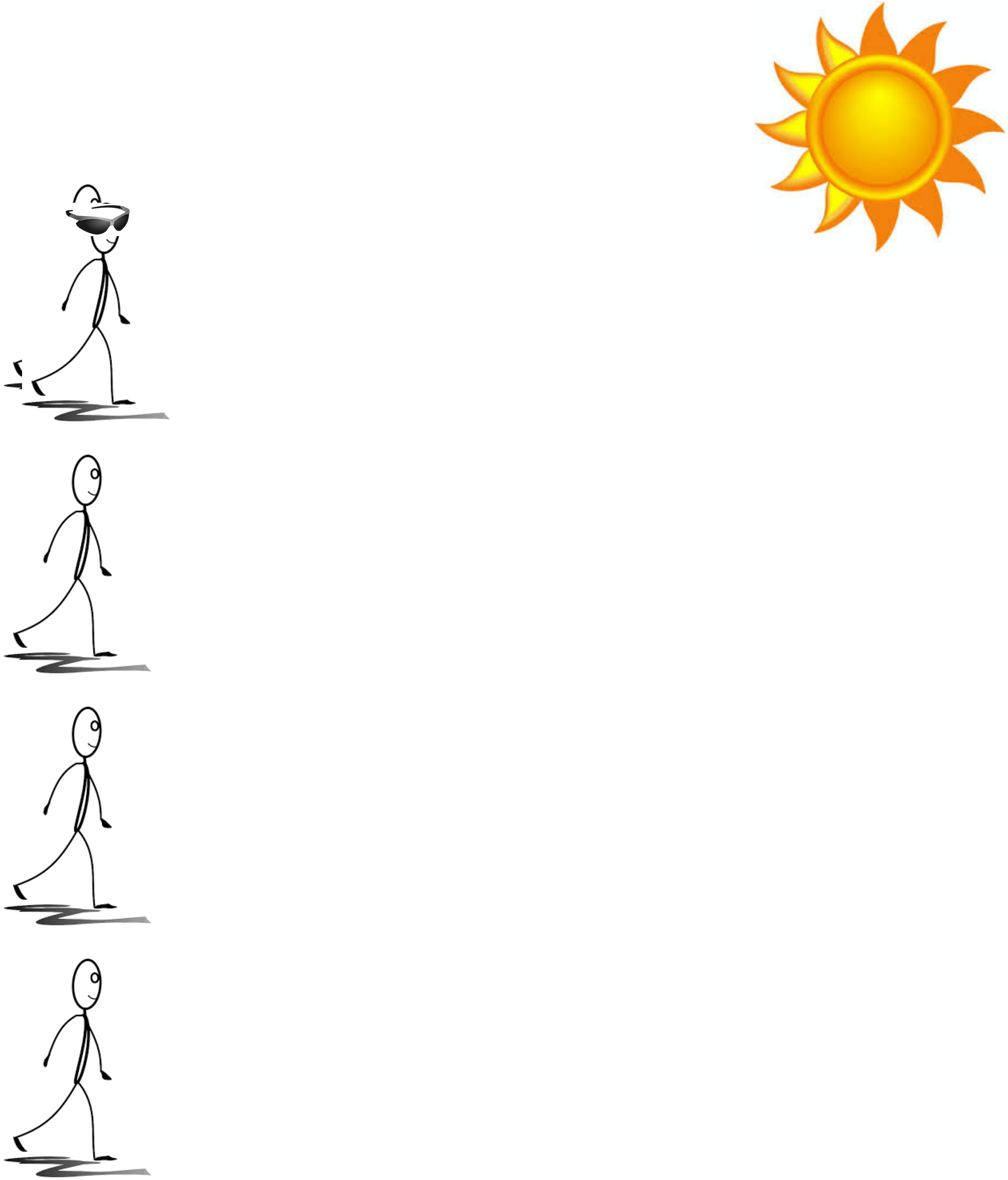
Level 3: Robotic “Jungle Jim”

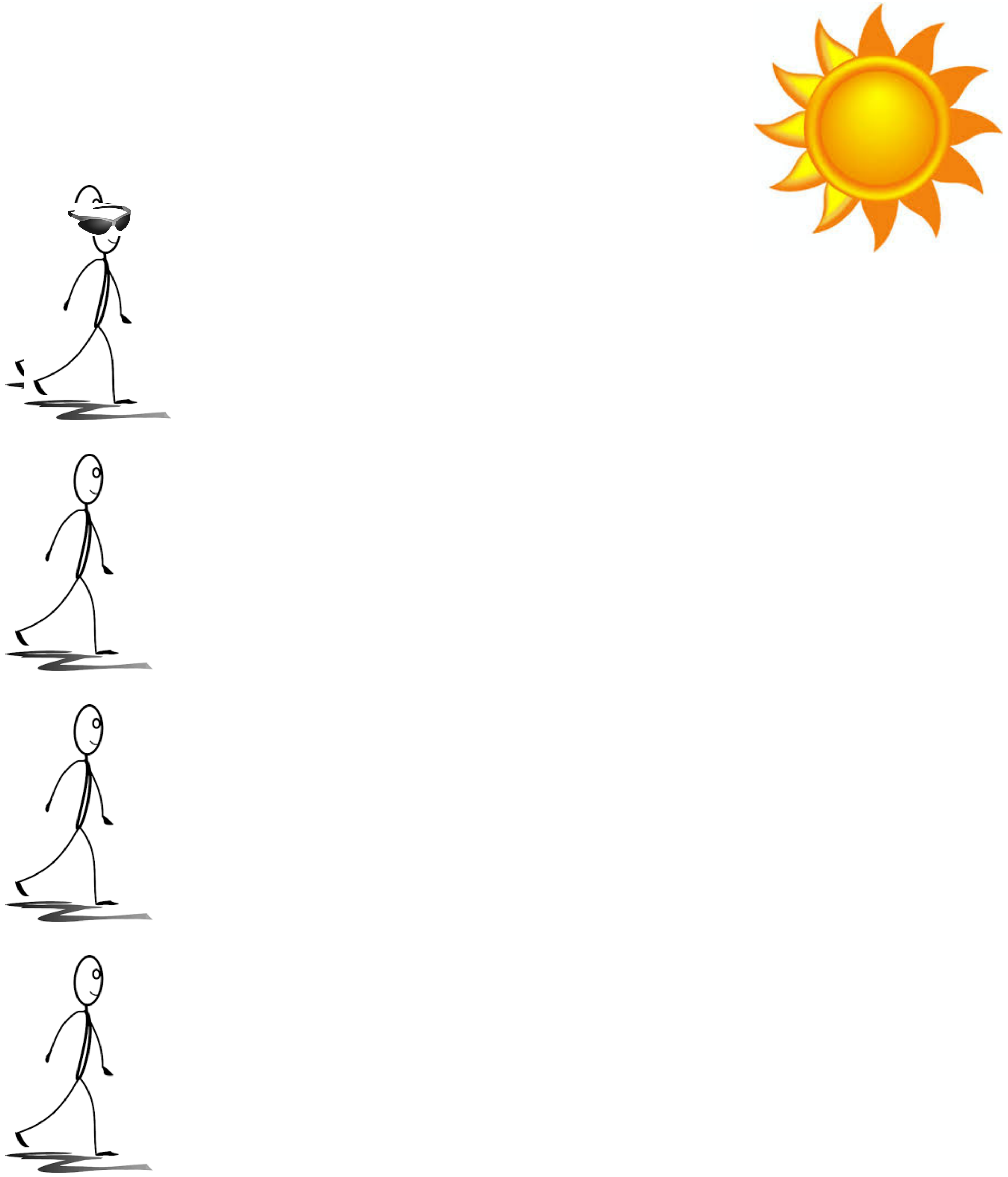

You

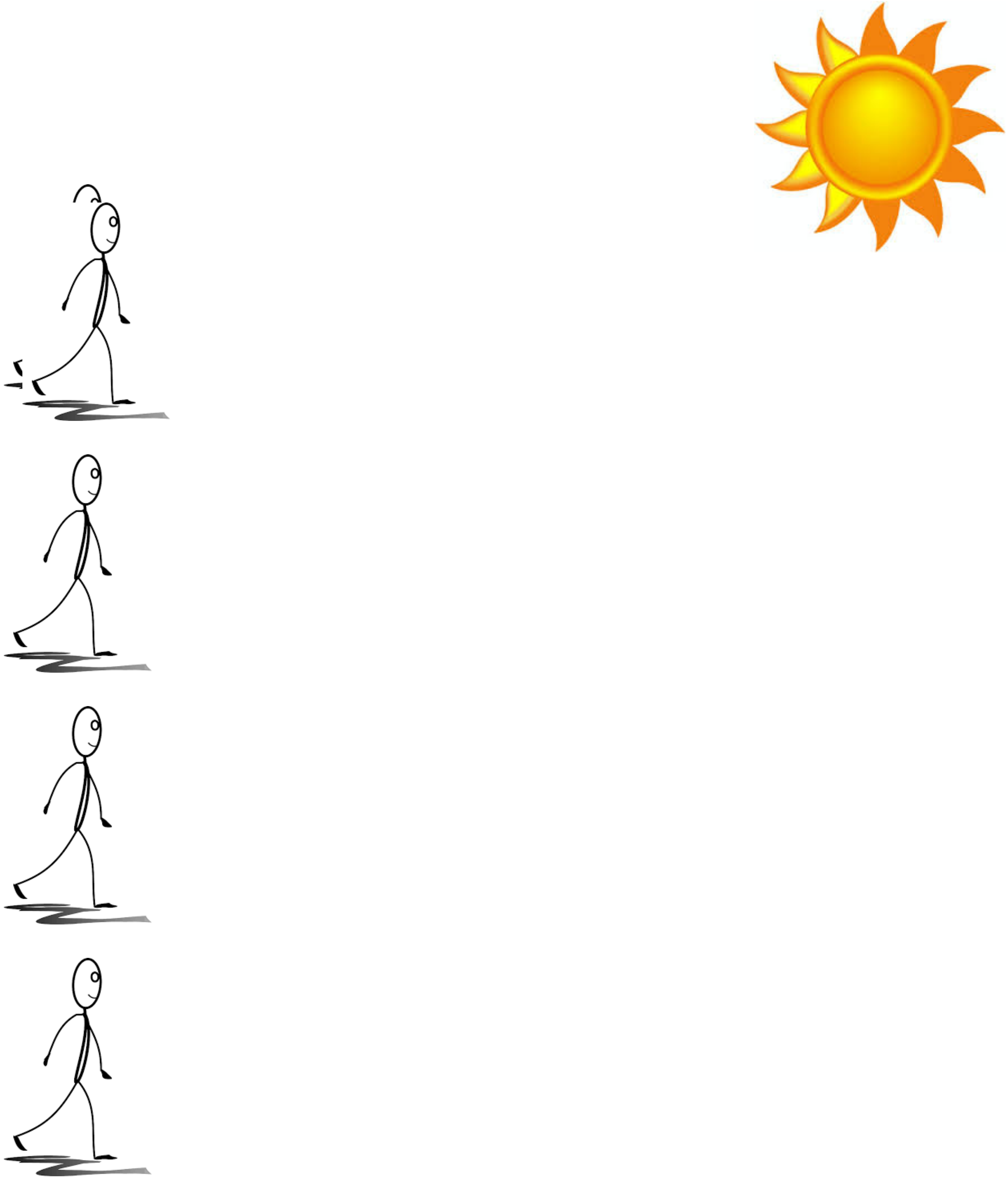


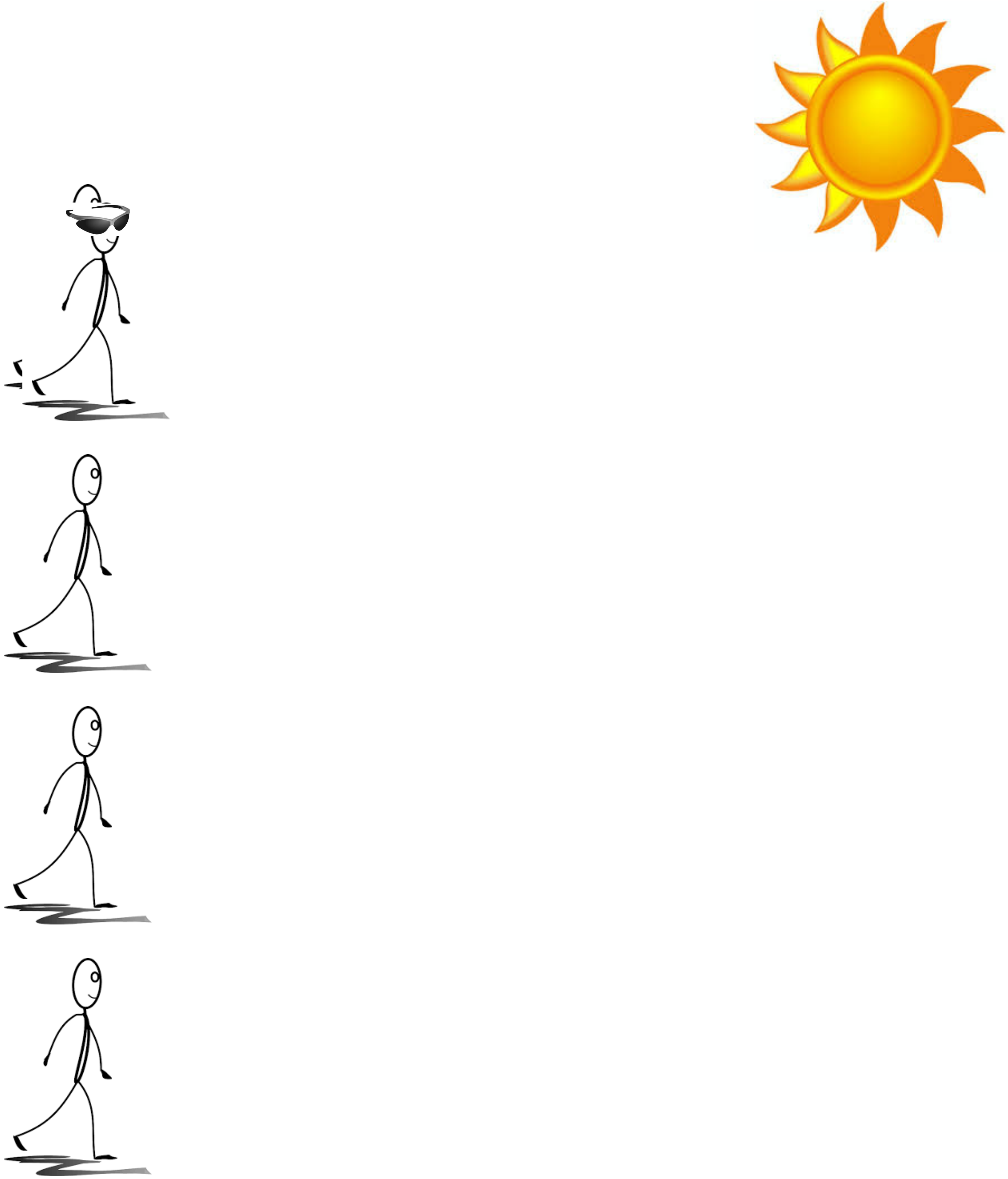


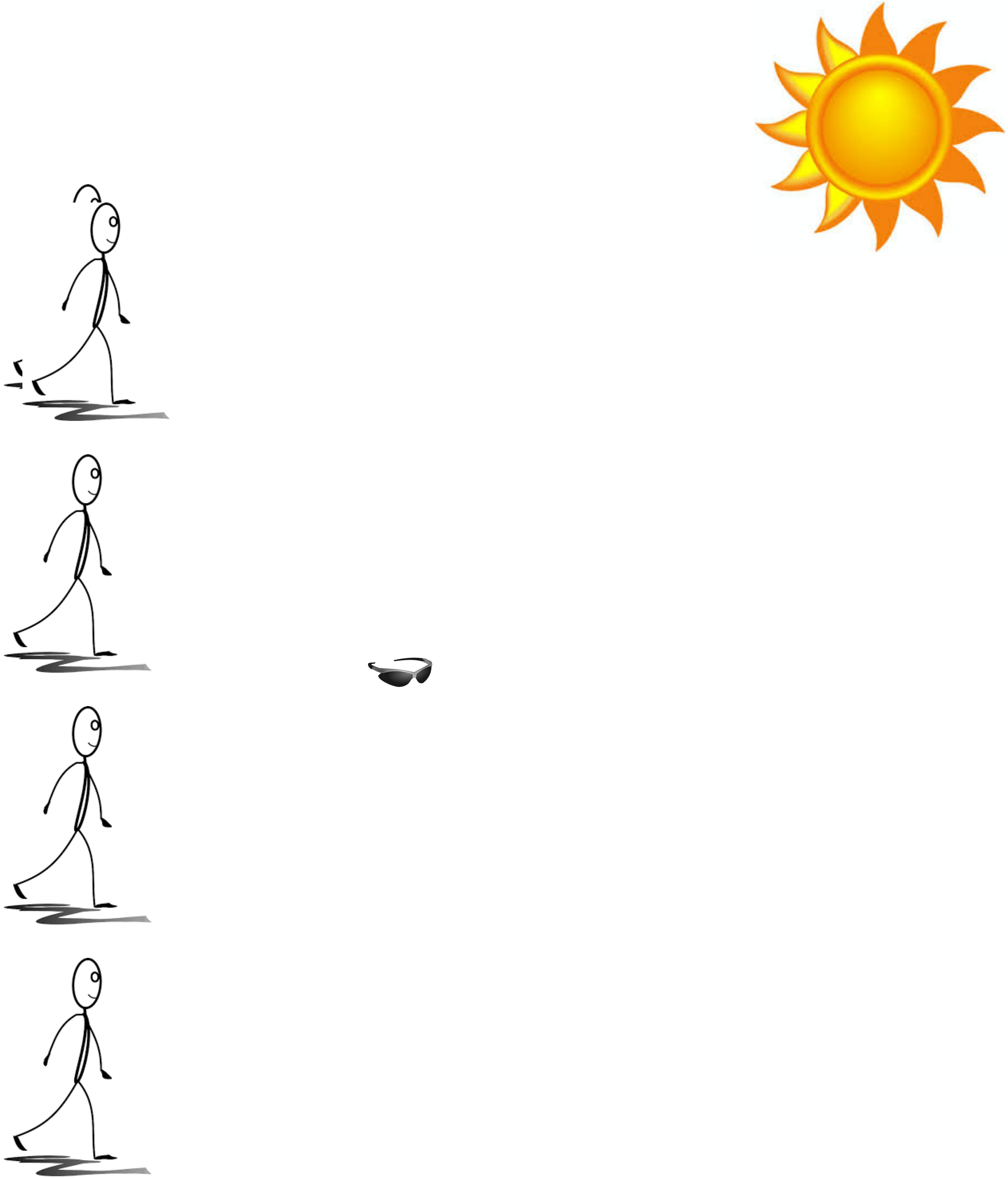


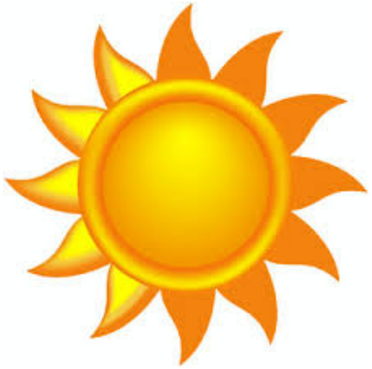


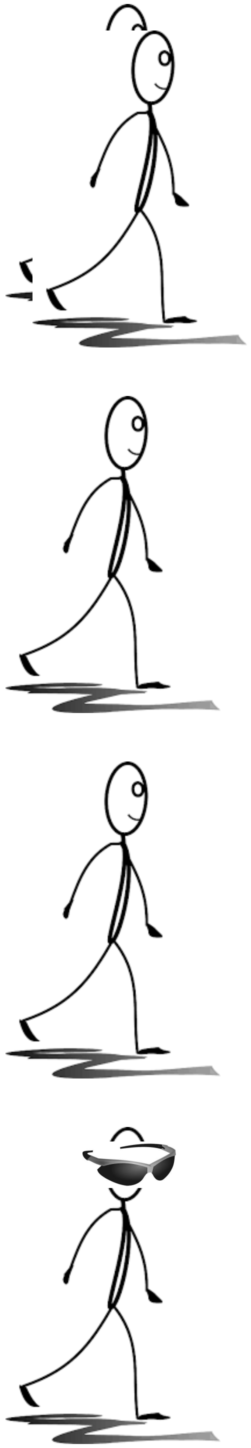


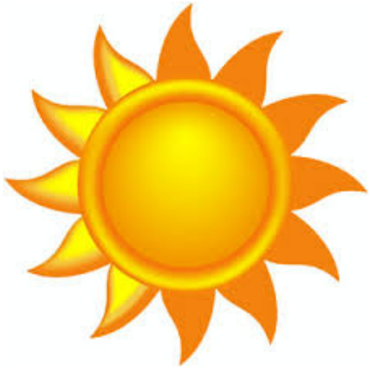


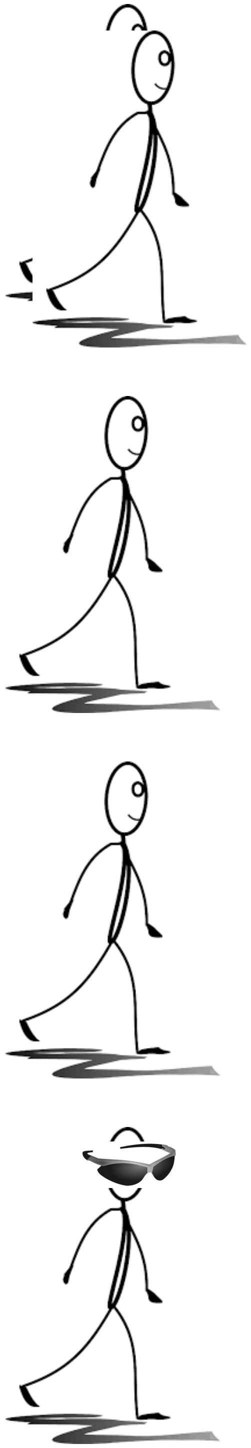


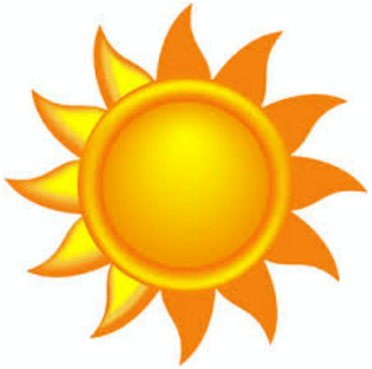


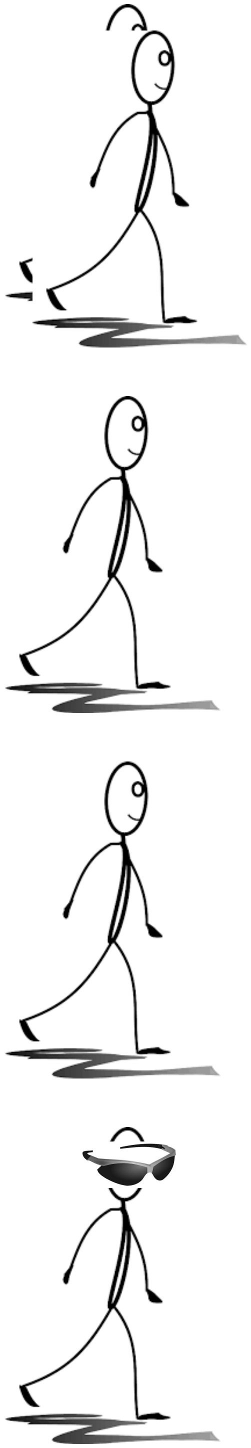




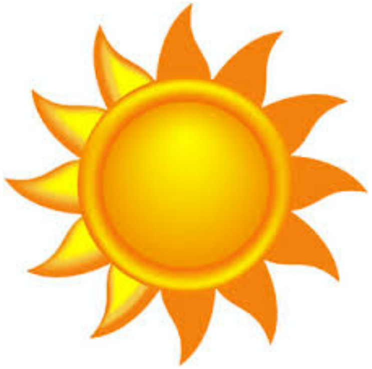




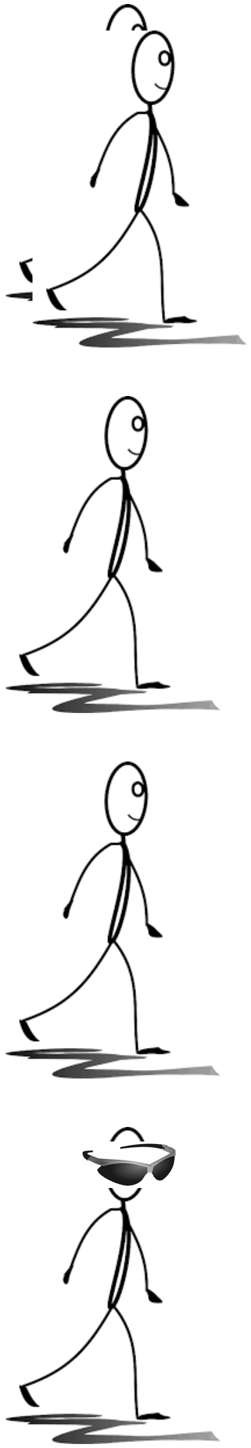




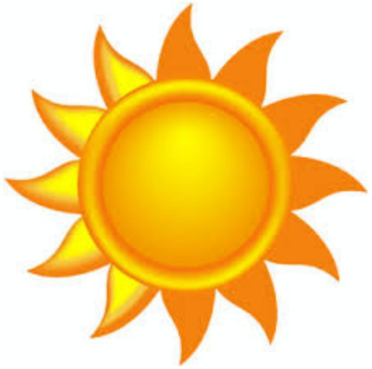
You: "There's a glorious pink rose!"

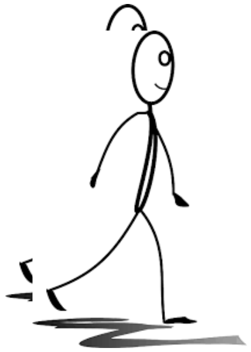


Naveen: "That is not so."



You: "There's a glorious pink rose!"

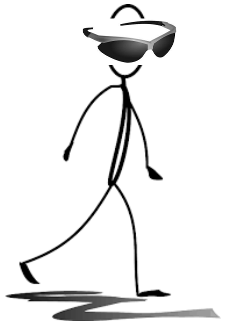


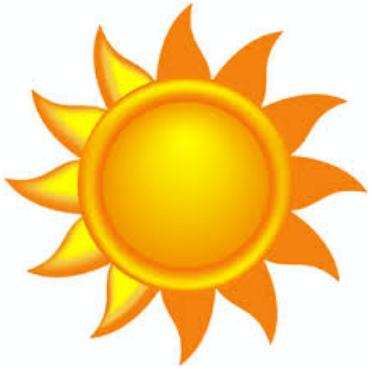


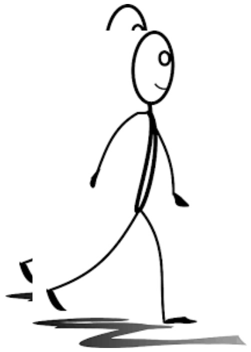
Naveen: “That is not so.”

Selmer: “Remove your glasses.”

You: “There’s a glorious pink rose!”







Naveen: “That is not so.”

Selmer: “Remove your glasses.”

You: “There’s a glorious pink rose!”

Naveen: “That is not so.”

Selmer: “Remove your glasses.”

You: “There’s a glorious pink rose!”

Ontological Inventory: Classical Triad

Ontological Inventory: Classical Triad

You (replete with sensors & effectors).

Ontological Inventory: Classical Triad

You (replete with sensors & effectors).

The white rose.

Ontological Inventory: Classical Triad

You (replete with sensors & effectors).

The white rose.

That which you perceived; the sense-datum
that led you to believe that you saw a *pink* rose.

Ontological Inventory: Adverbial Theory of Perception

Ontological Inventory: Adverbial Theory of Perception

You (replete with sensors & effectors).

Ontological Inventory: Adverbial Theory of Perception

You (replete with sensors & effectors).

The white rose.

Ontological Inventory: Adverbial Theory of Perception

You (replete with sensors & effectors).

The white rose.

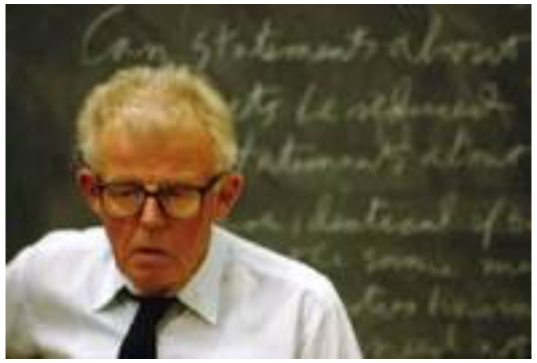
And that’s it! — because you perceive pinkly.

Ontological Inventory: Adverbial Theory of Perception

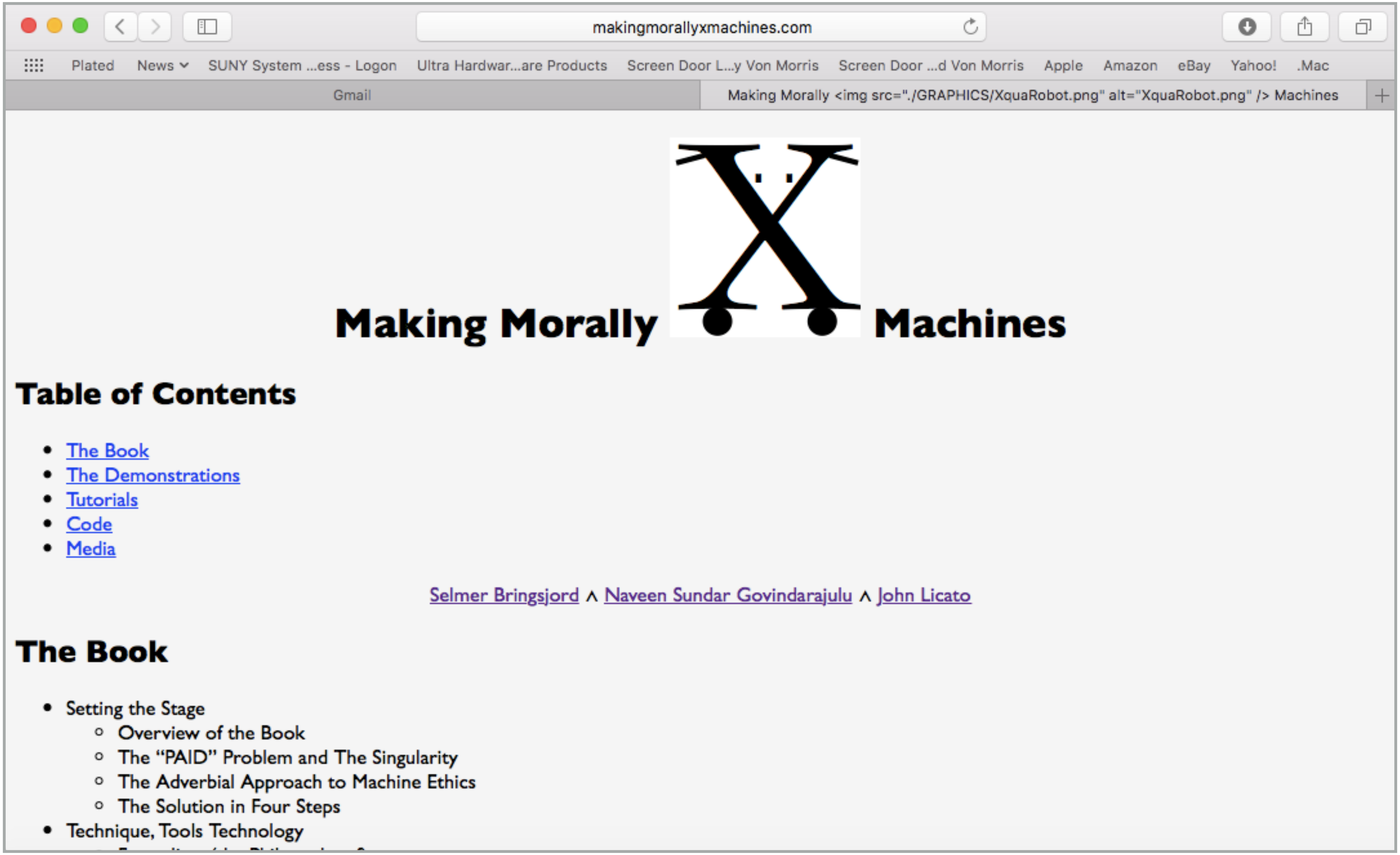
You (replete with sensors & effectors).

The white rose.

And that’s it! — because you perceive pinkly.



The Adverbial Approach to (Machine) Ethics


makingmorallyxmachines.com
Plated
News
SUNY System ...ess - Logon
Ultra Hardwar...are Products
Screen Door L...y Von Morris
Screen Door ...d Von Morris
Apple
Amazon
eBay
Yahoo!
.Mac
Gmail
Making Morally <img src="./GRAPHICS/XquaRobot.png" alt="XquaRobot.png" /> Machines

Making Morally X Machines

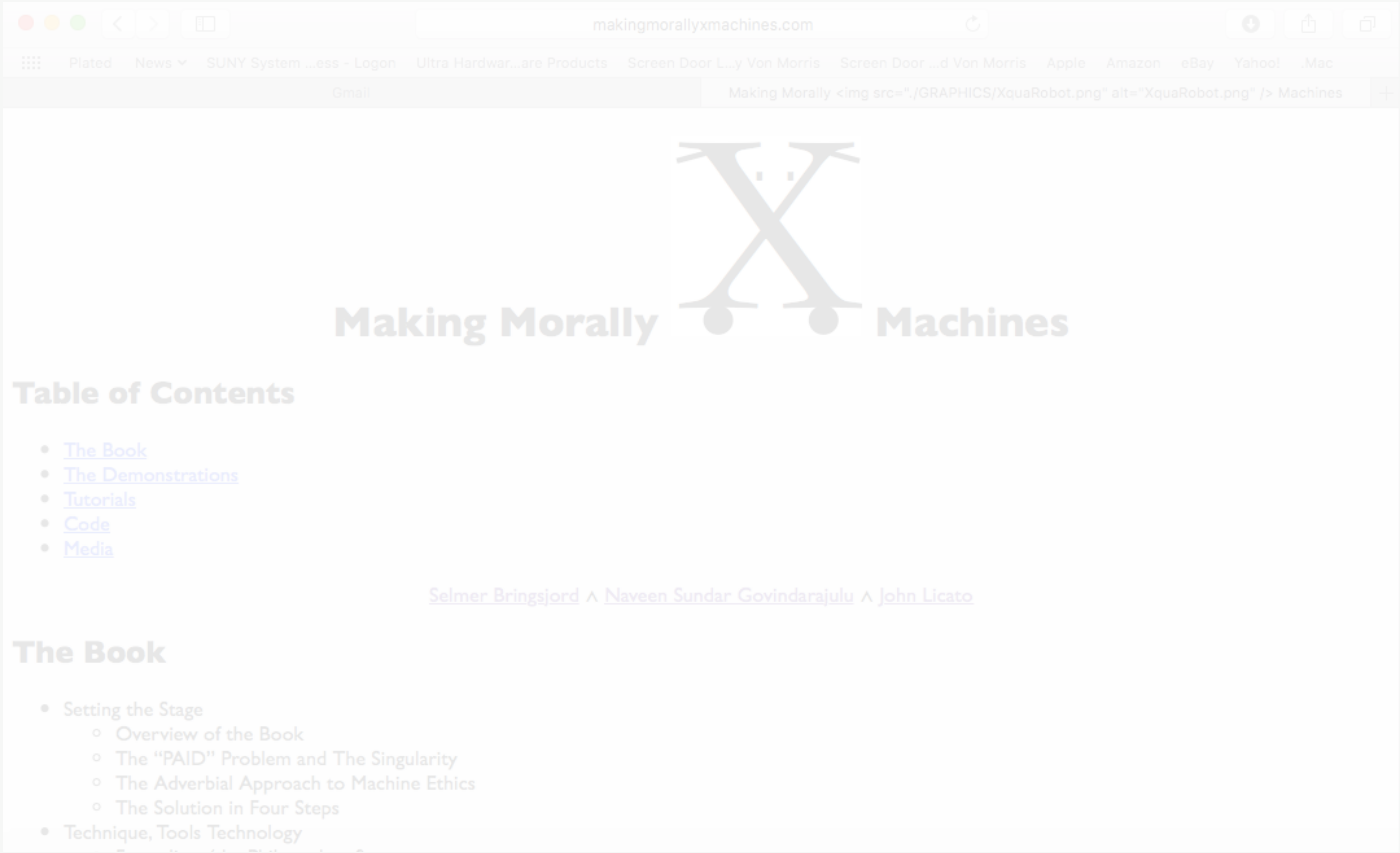
Table of Contents

- The Book
- The Demonstrations
- Tutorials
- Code
- Media

Selmer Bringsjord ∧ Naveen Sundar Govindarajulu ∧ John Licato

The Book

- Setting the Stage
  - Overview of the Book
  - The "PAID" Problem and The Singularity
  - The Adverbial Approach to Machine Ethics
  - The Solution in Four Steps
- Technique, Tools Technology


makingmorallyxmachines.com
Plated
News
SUNY System ...ess - Logon
Ultra Hardwar...are Products
Screen Door L...y Von Morris
Screen Door ...d Von Morris
Apple
Amazon
eBay
Yahoo!
.Mac
Gmail
Making Morally <img src="./GRAPHICS/XquaRobot.png" alt="XquaRobot.png" /> Machines

Making Morally Machines

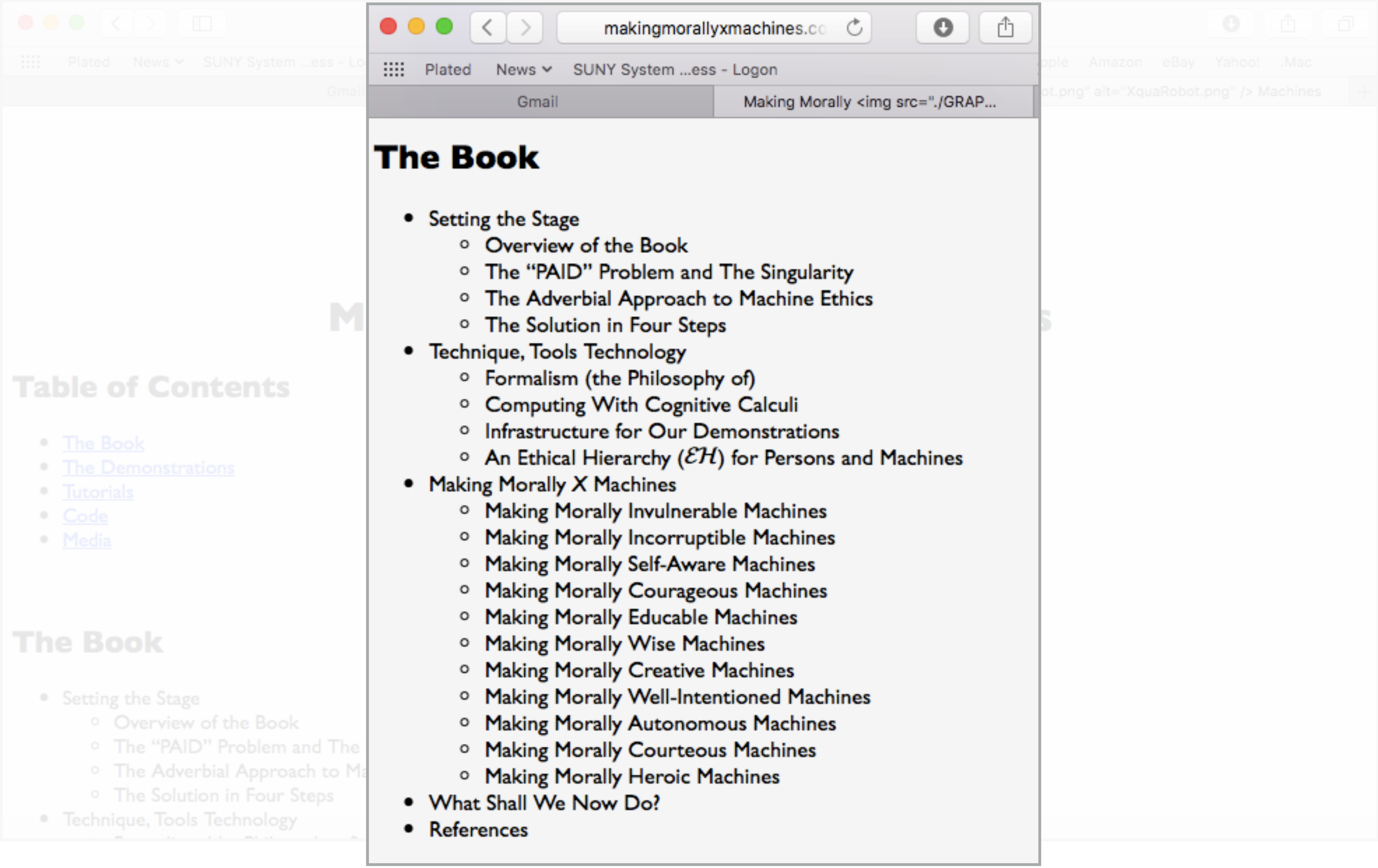
Table of Contents

- The Book
- The Demonstrations
- Tutorials
- Code
- Media

Selmer Bringsjord ∧ Naveen Sundar Govindarajulu ∧ John Licato

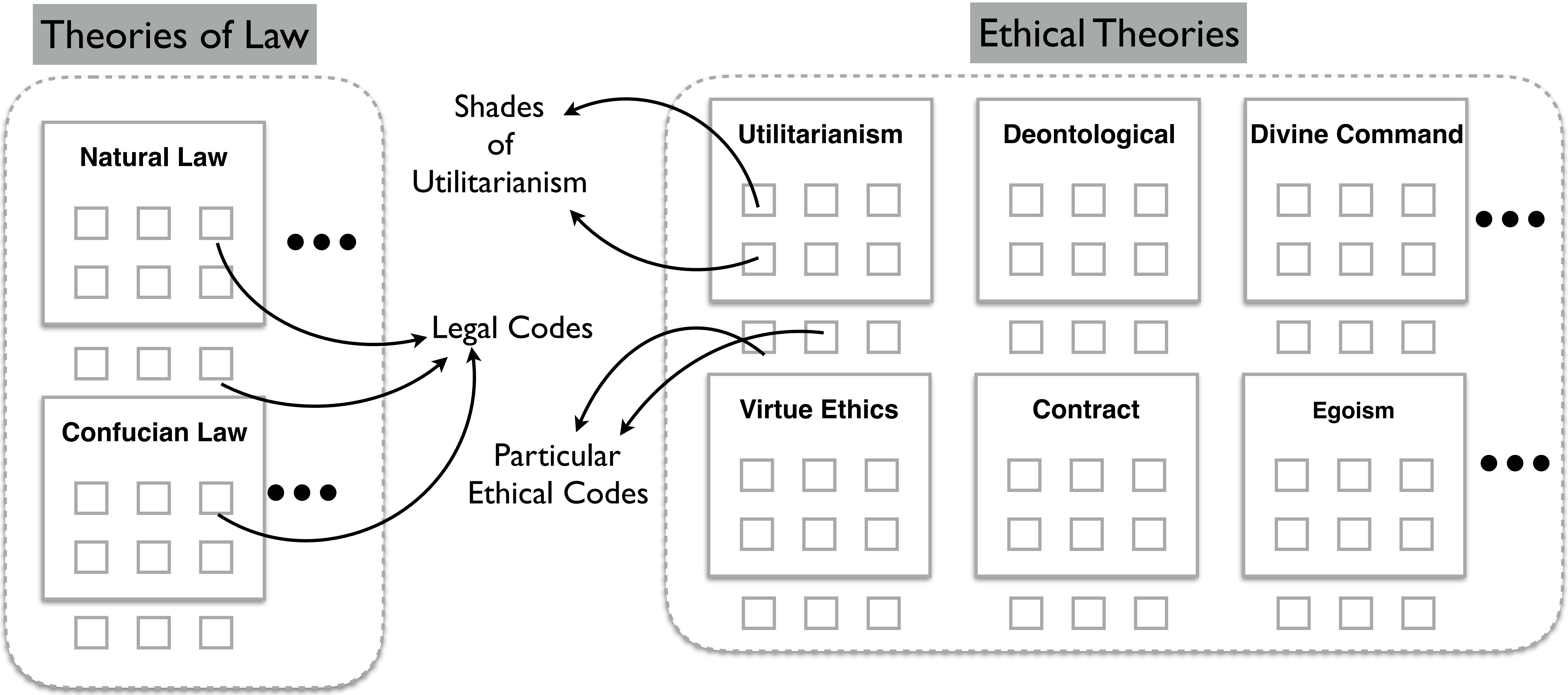
The Book

- Setting the Stage
  - Overview of the Book
  - The "PAID" Problem and The Singularity
  - The Adverbial Approach to Machine Ethics
  - The Solution in Four Steps
- Technique, Tools Technology


makingmorallyxmachines.co
Plated
News
SUNY System ...ess - Logon
Gmail
Making Morally <img src="./GRAP...

The Book

Making Morally X Machines, in Four Steps



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

Deontological

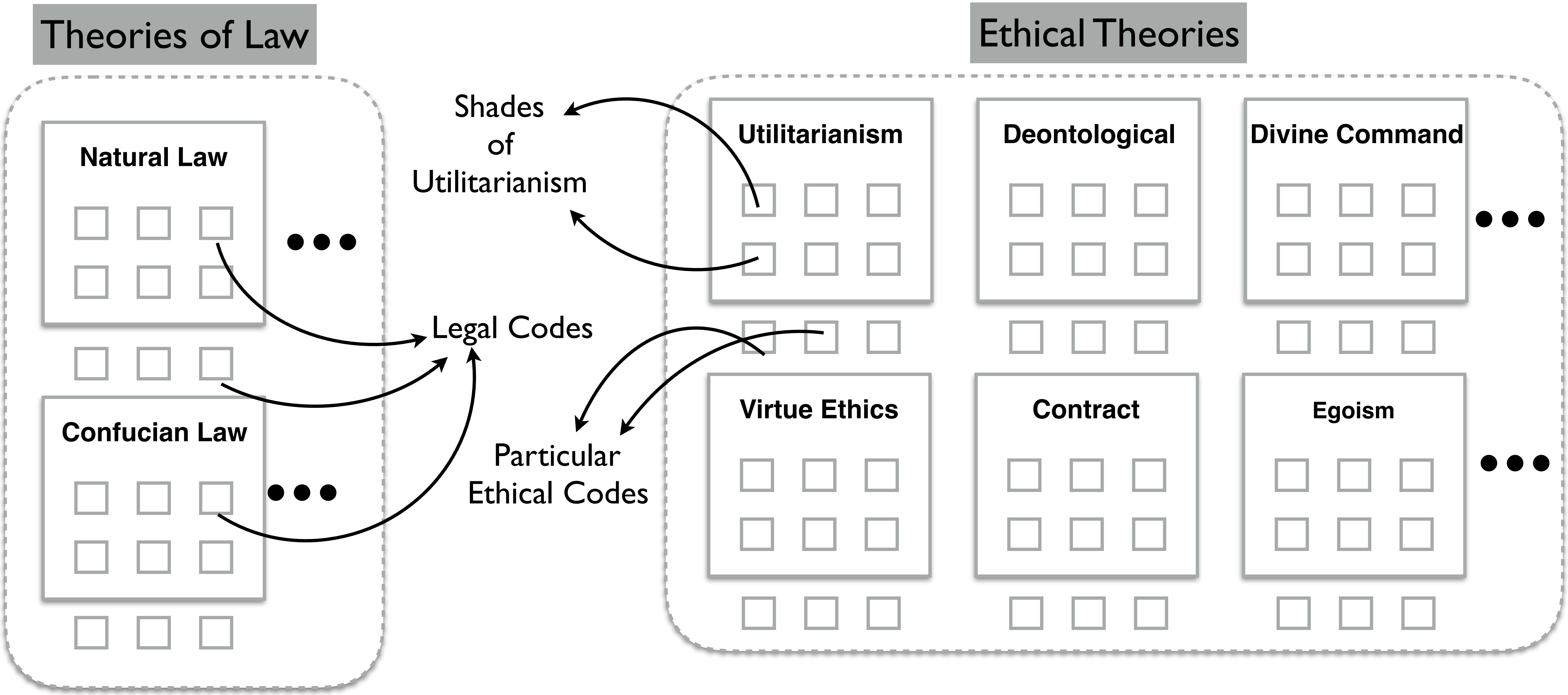
Divine Command

Virtue Ethics

Contract

Egoism

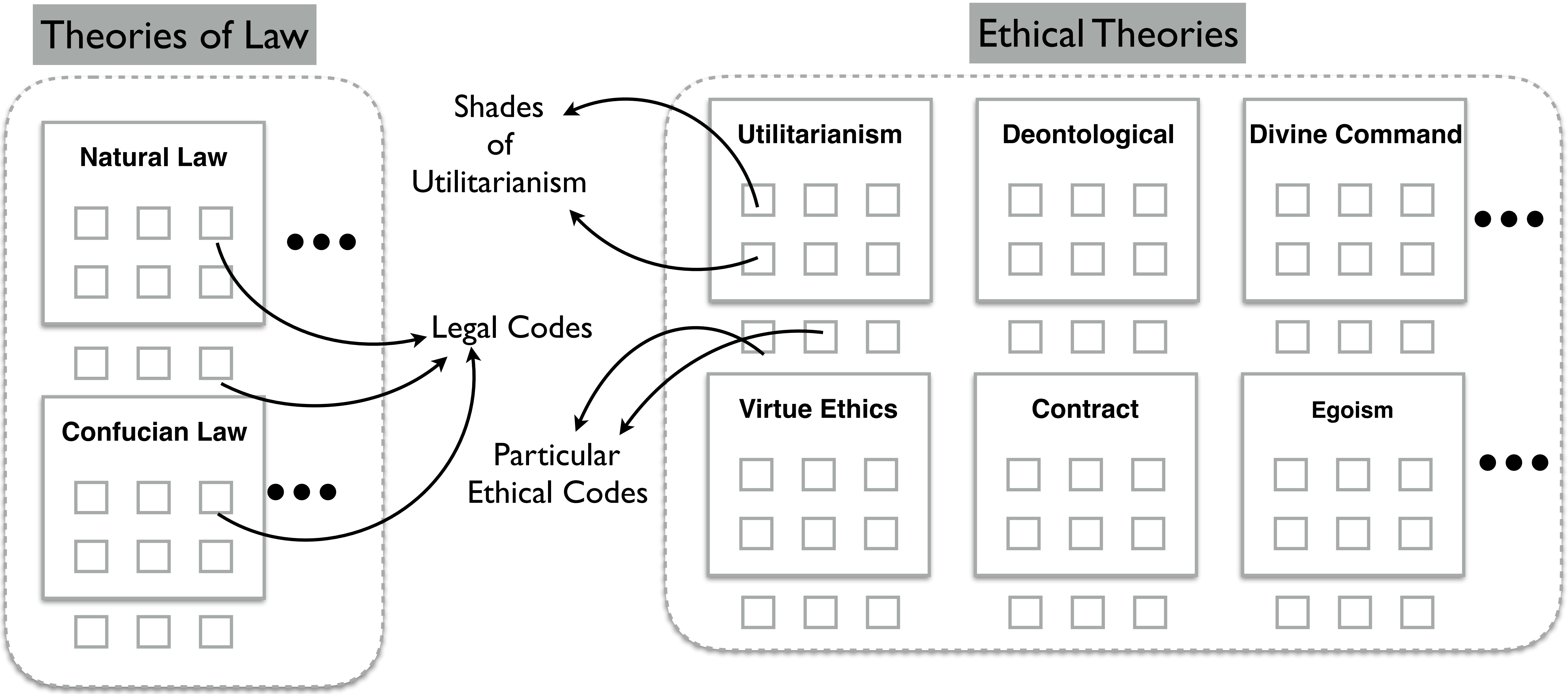
Making Morally *X* Machines, in Four Steps



Step 1

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

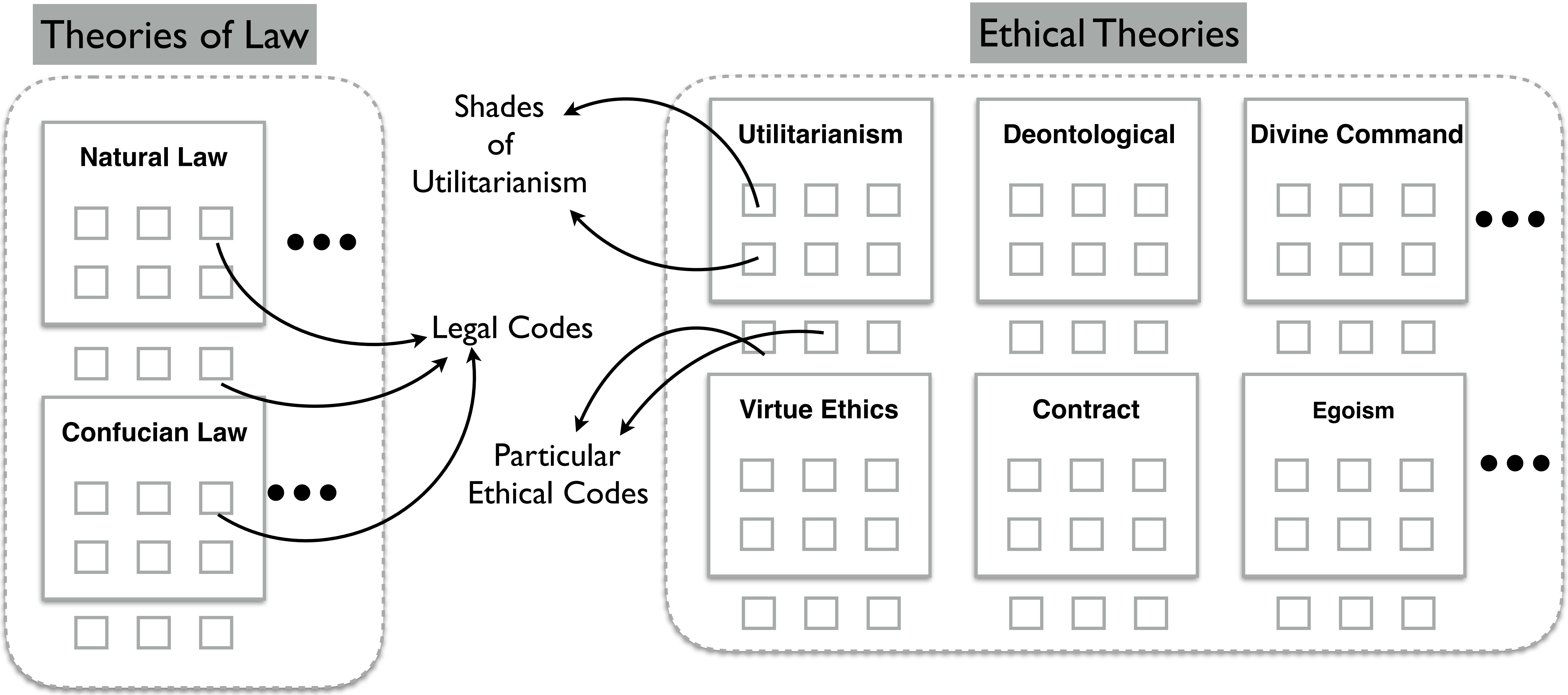
Making Morally *X* Machines, in Four Steps



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



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

Step 1

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

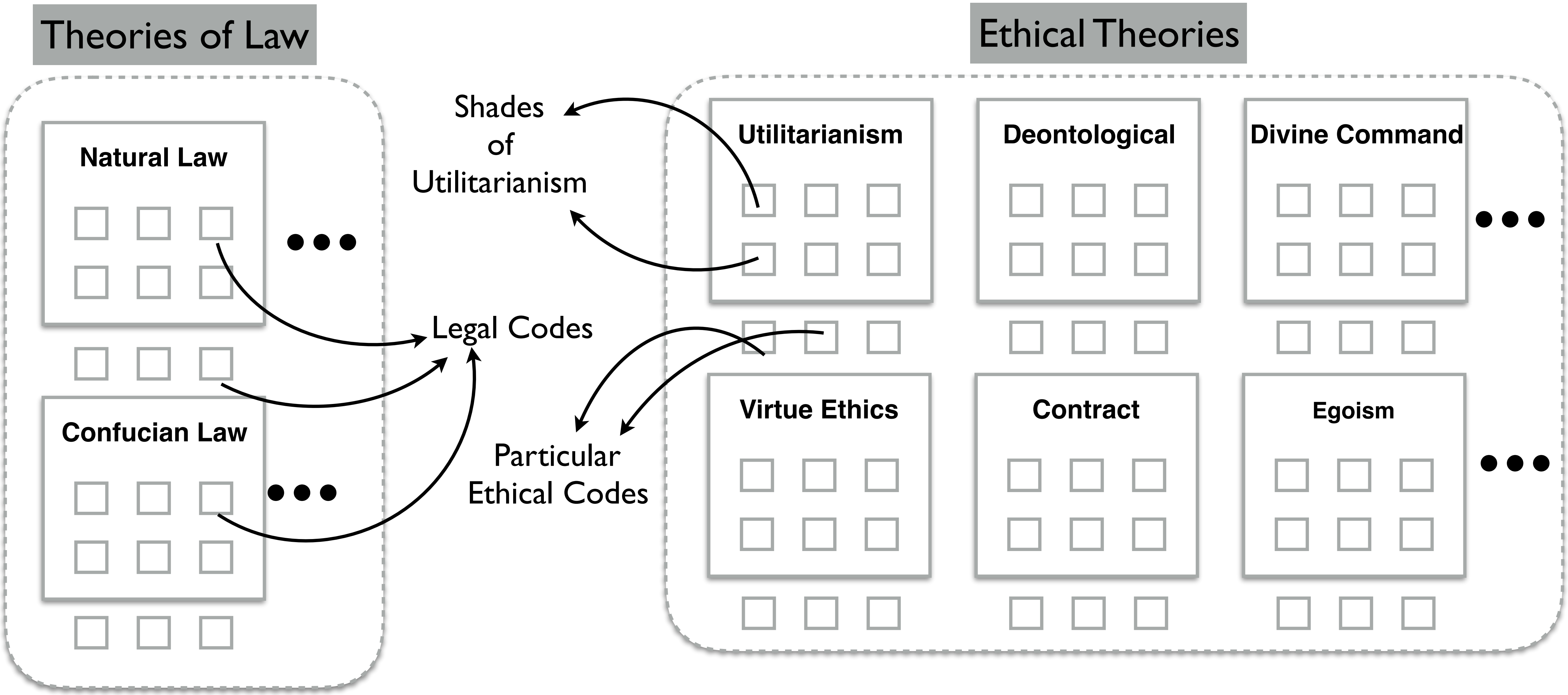
Step 2

Formalize & Automate

Shadow Prover

Spectra

Making Morally X Machines, in Four Steps



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*?

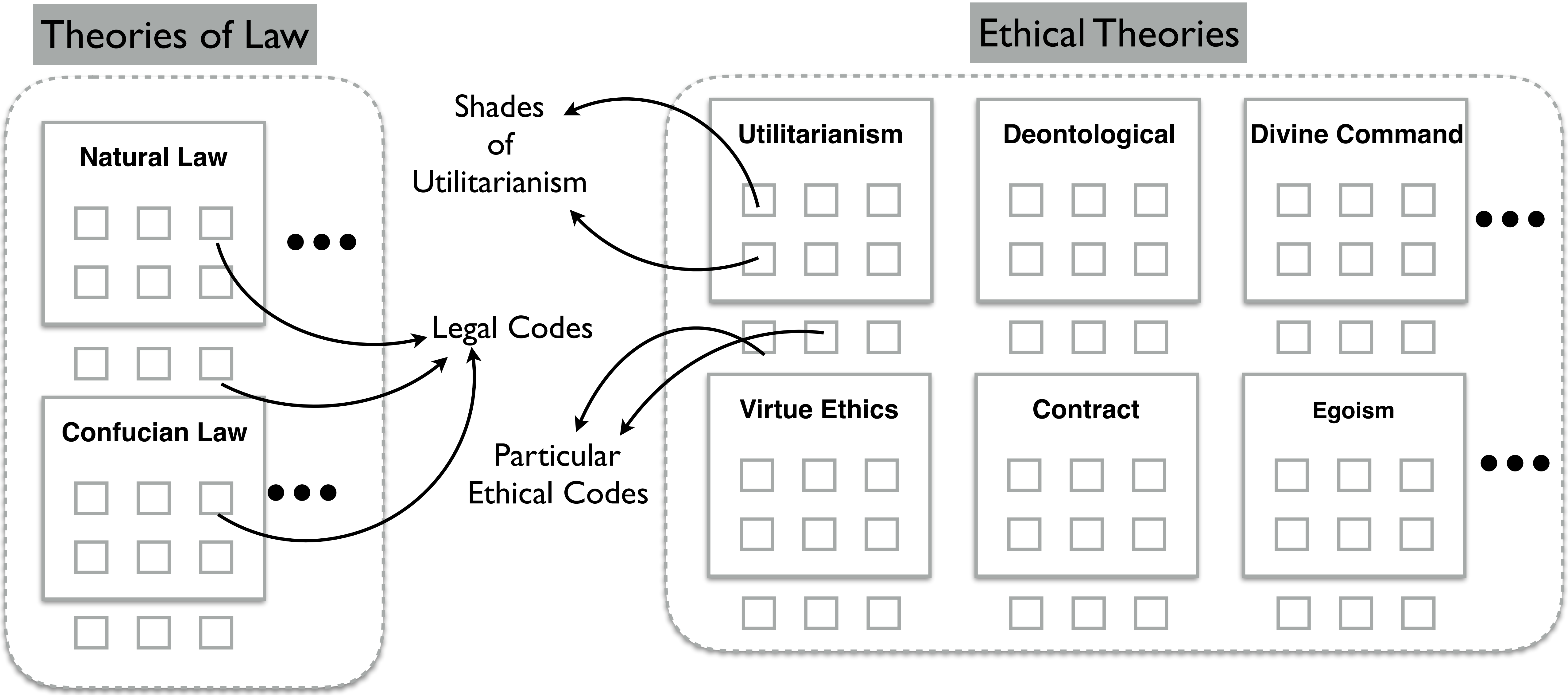
Step 2

Formalize & Automate

Shadow Prover

Spectra

Making Morally X Machines, in Four Steps



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

Utilitarianism

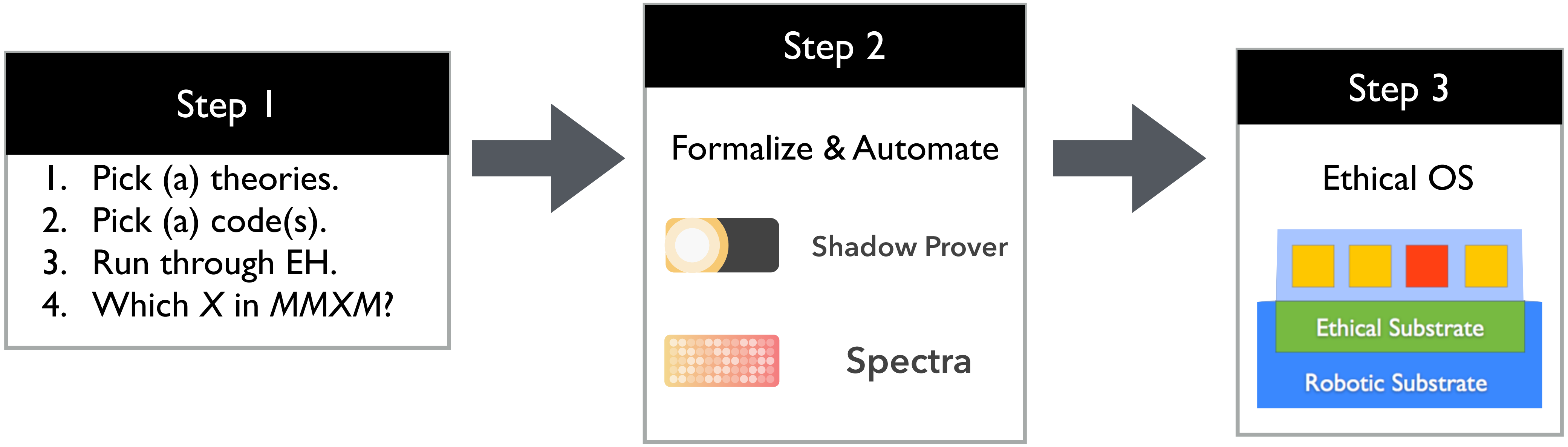
Deontological

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

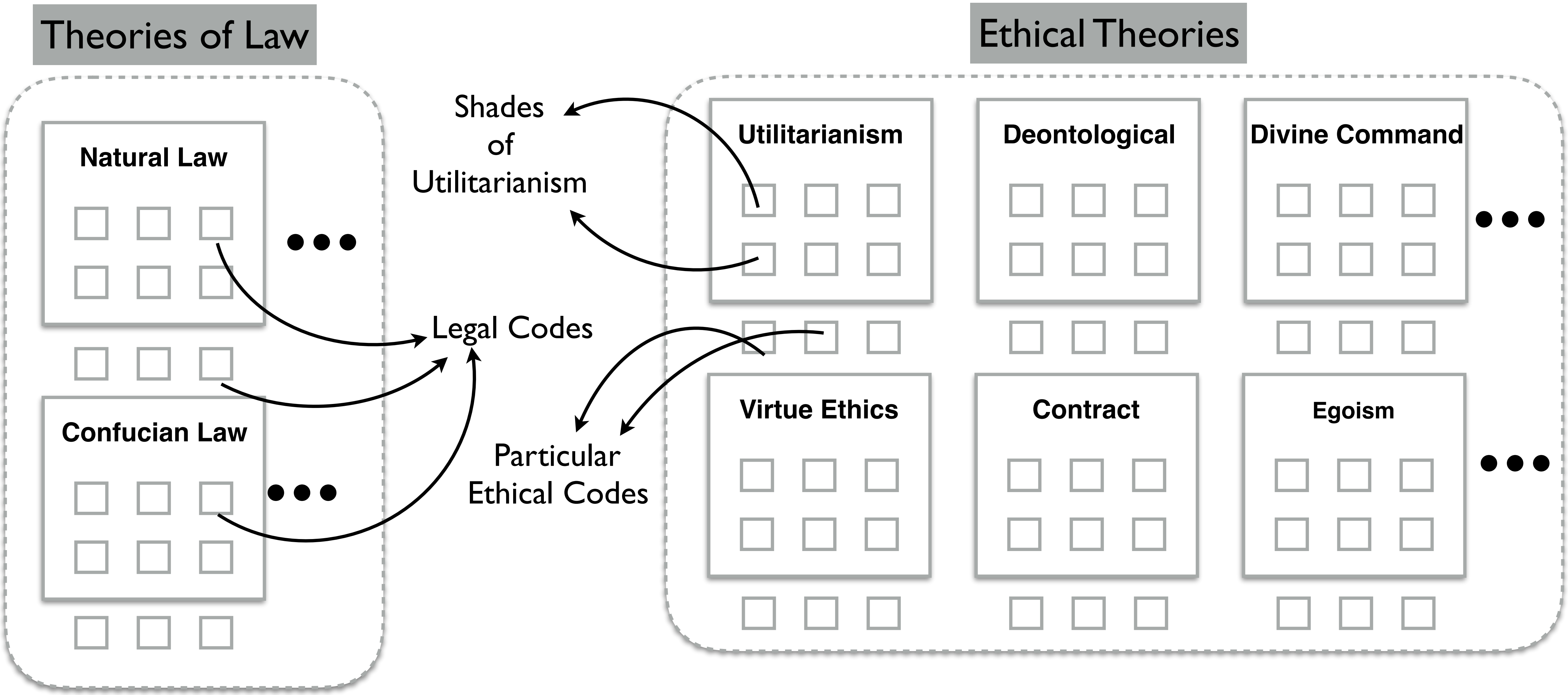
Step 3

Ethical OS

Ethical Substrate

Robotic Substrate

Making Morally X Machines, in Four Steps



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

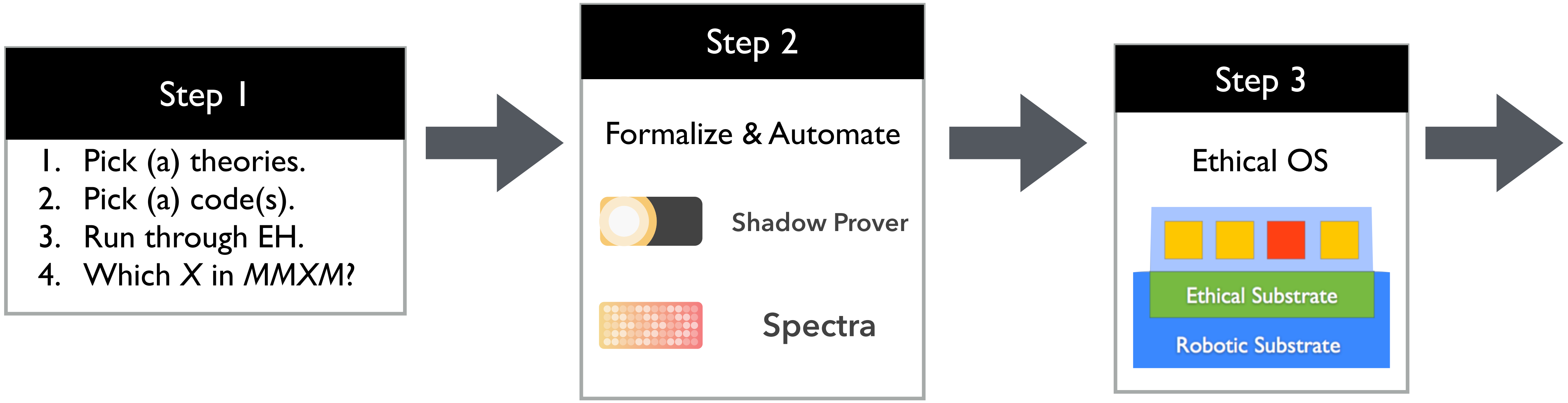
Deontological

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

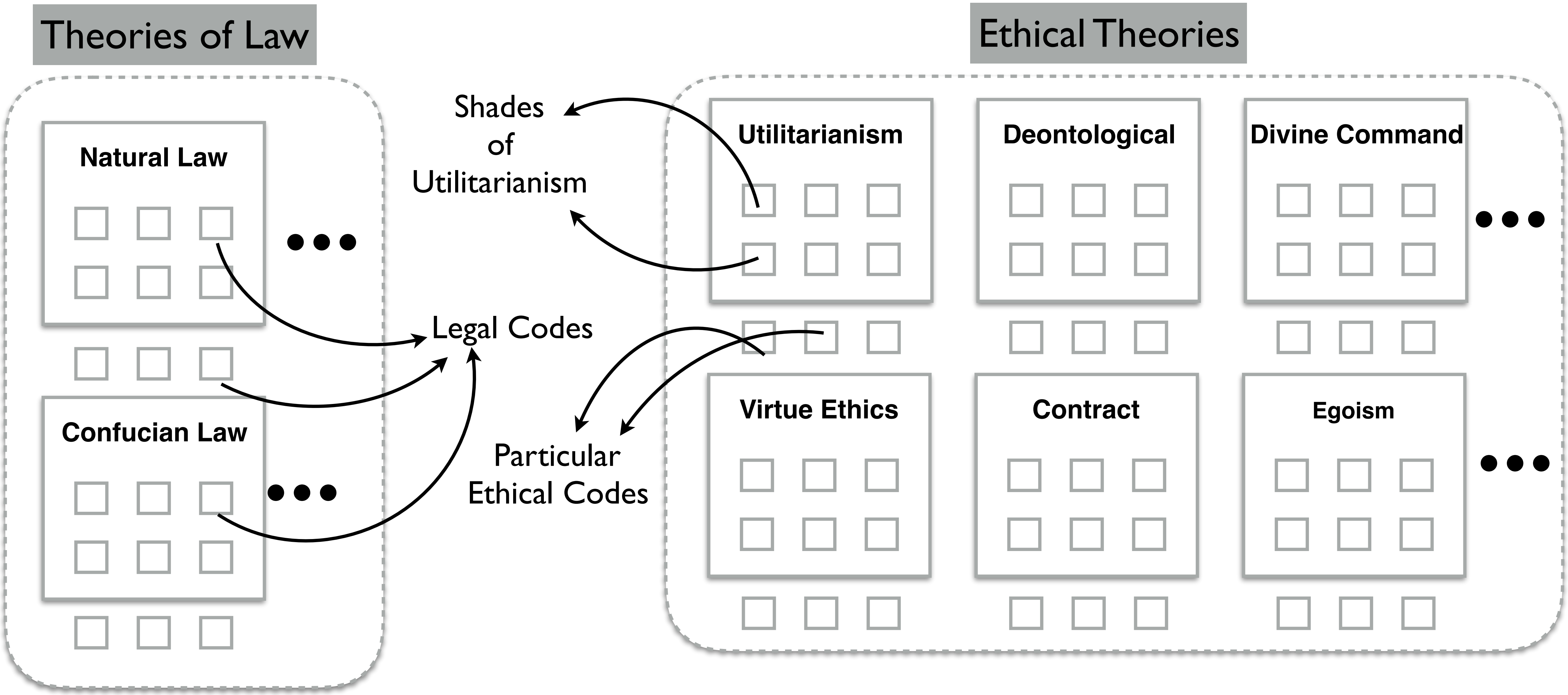
Step 3

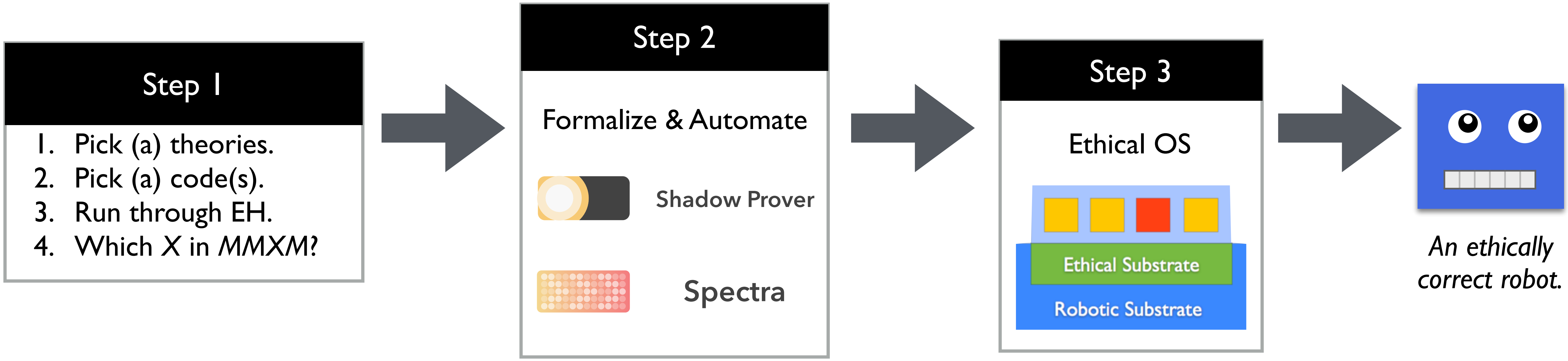
Ethical OS

Ethical Substrate

Robotic Substrate

Making Morally X Machines, in Four Steps





Step 1

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

Step 2

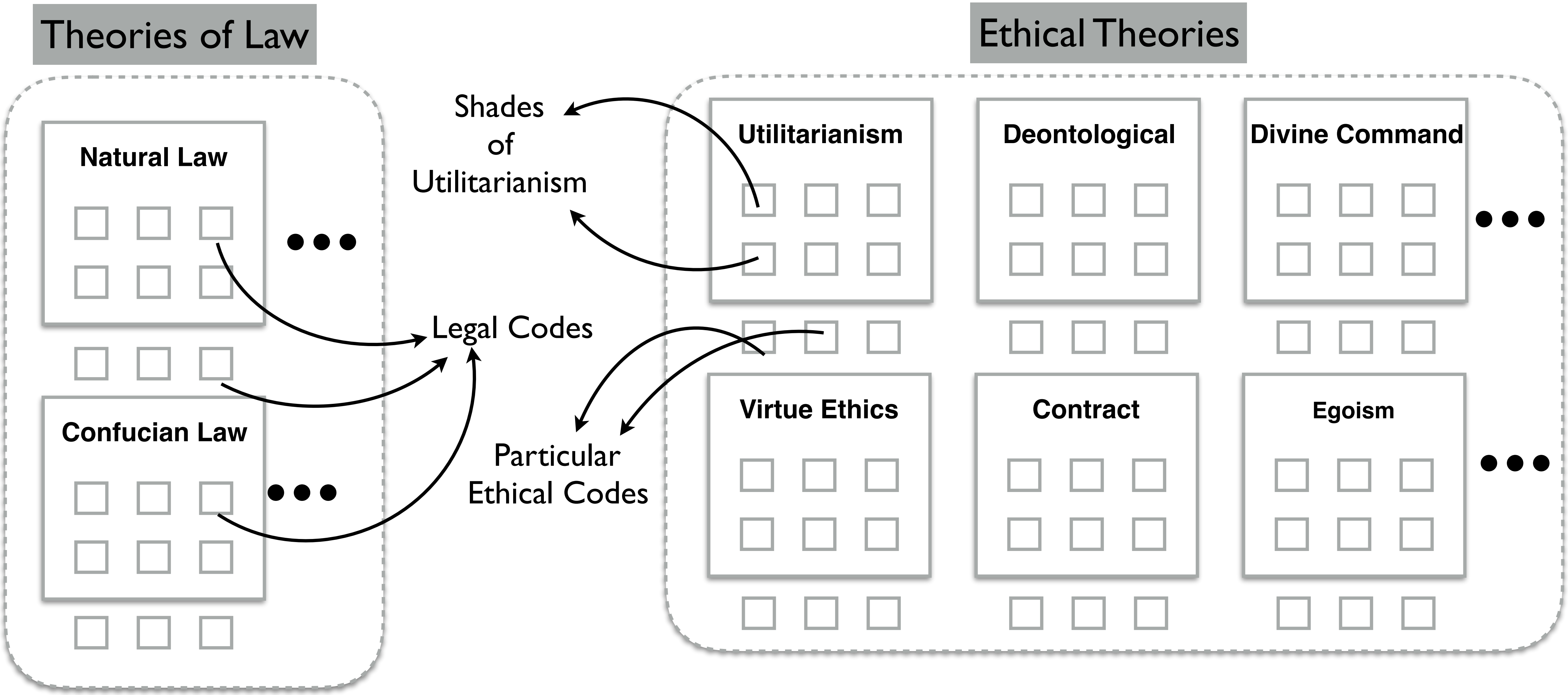
Formalize & Automate

Step 3

Ethical OS

An ethically correct robot.

Making Morally X Machines, in Four Steps



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

Utilitarianism

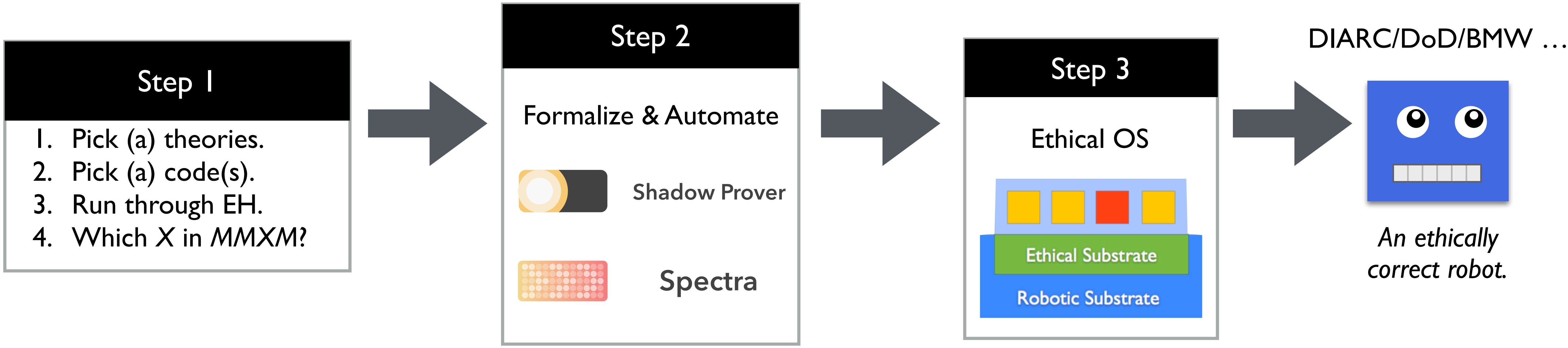
Deontological

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

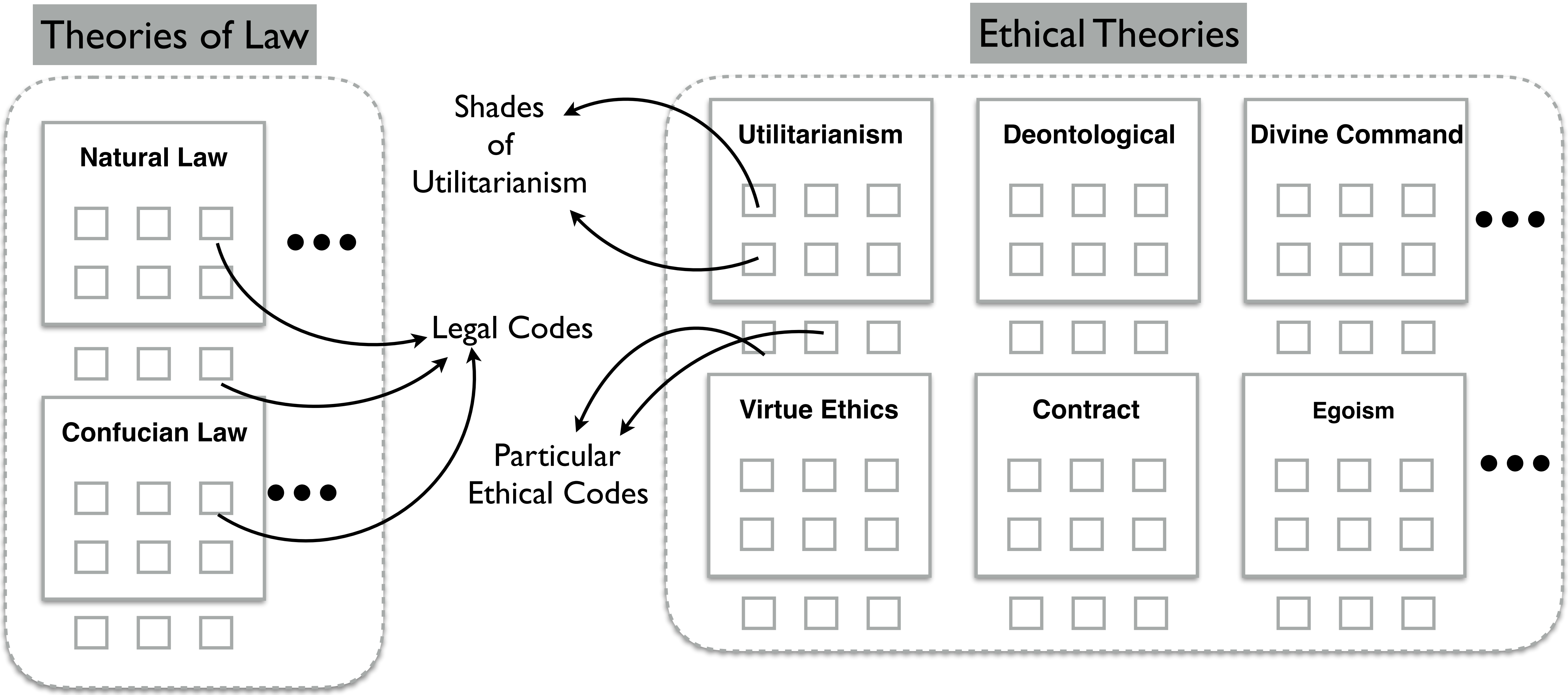
Ethical Substrate

Robotic Substrate

DIARC/DoD/BMW …

An ethically correct robot.

Making Morally X Machines, in Four Steps



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

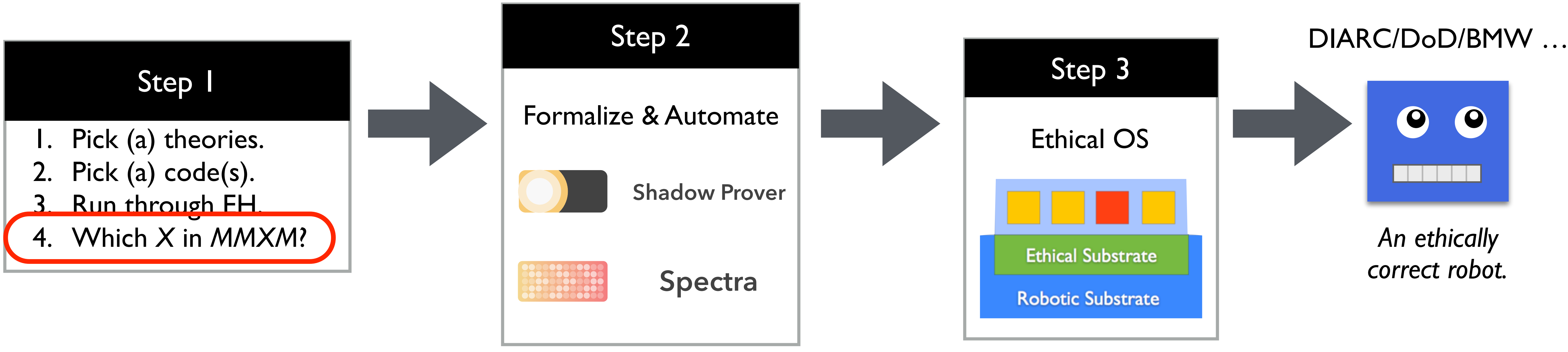
Deontological

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

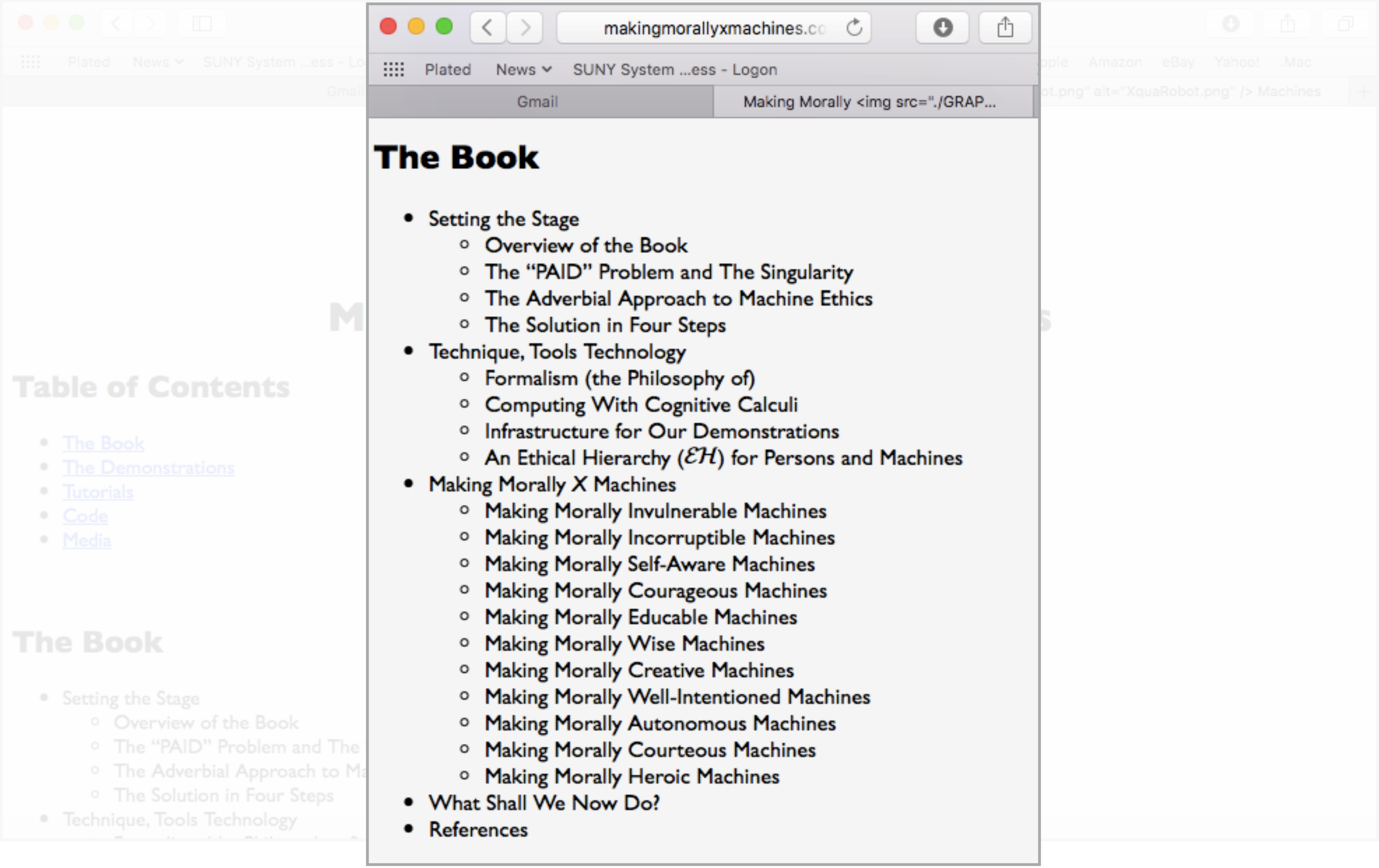
Ethical OS

Ethical Substrate

Robotic Substrate

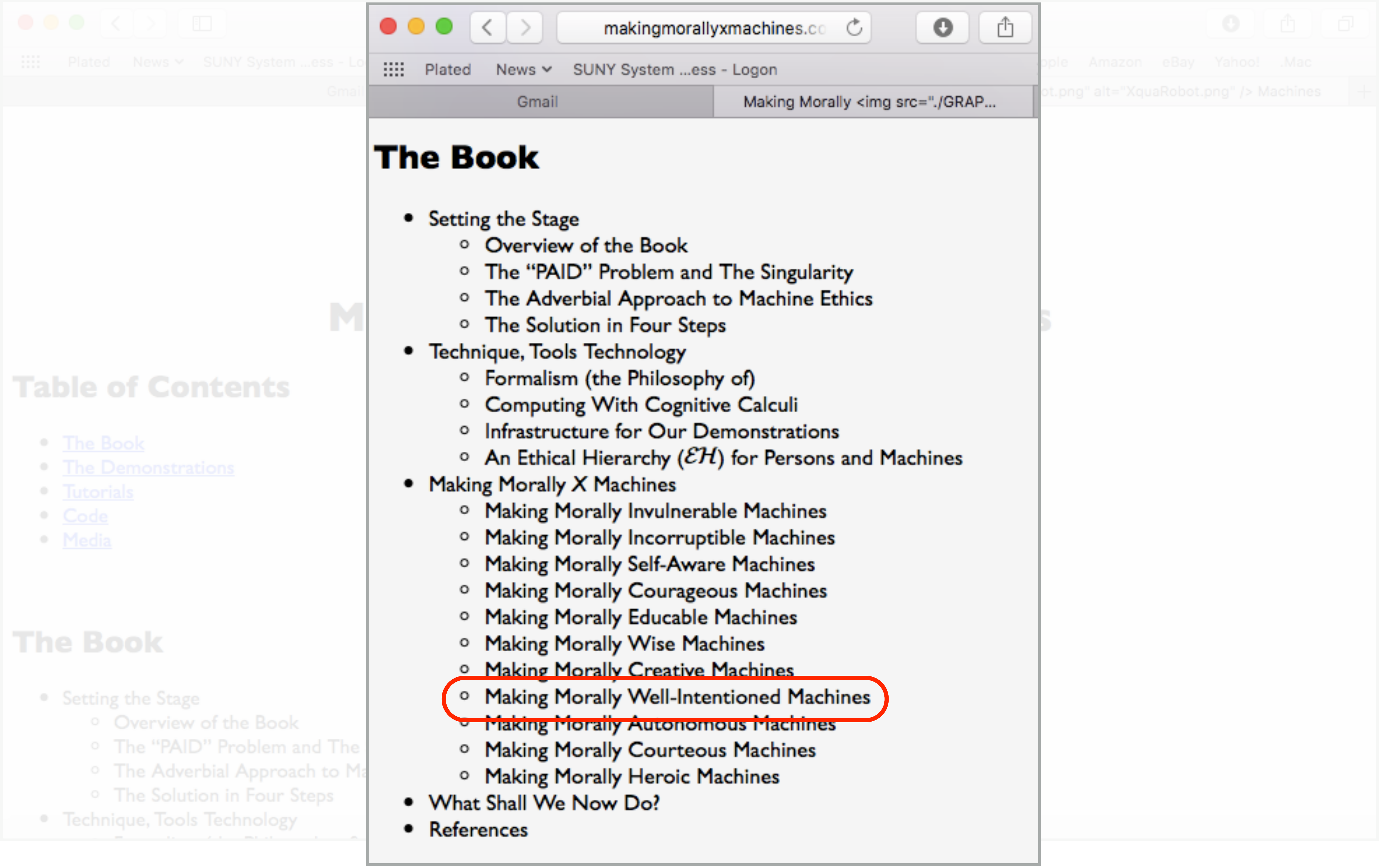
DIARC/DoD/BMW …

An ethically correct robot.


makingmorallyxmachines.co
Plated
News
SUNY System ...ess - Logon
Gmail
Making Morally <img src="./GRAP...

The Book

- Setting the Stage
  - Overview of the Book
  - The “PAID” Problem and The Singularity
  - The Adverbial Approach to Machine Ethics
  - The Solution in Four Steps
- Technique, Tools Technology
  - Formalism (the Philosophy of)
  - Computing With Cognitive Calculi
  - Infrastructure for Our Demonstrations
  - An Ethical Hierarchy ($\mathcal{EH}$) for Persons and Machines
- Making Morally *X* Machines
  - Making Morally Invulnerable Machines
  - Making Morally Incorruptible Machines
  - Making Morally Self-Aware Machines
  - Making Morally Courageous Machines
  - Making Morally Educable Machines
  - Making Morally Wise Machines
  - Making Morally Creative Machines
  - Making Morally Well-Intentioned Machines
  - Making Morally Autonomous Machines
  - Making Morally Courteous Machines
  - Making Morally Heroic Machines
- What Shall We Now Do?
- References


makingmorallyxmachines.co
Plated
News
SUNY System ...ess - Logon
Gmail
Making Morally <img src="./GRAP...
Amazon
eBay
Yahoo!
Mac
ot.png" alt="XquaRobot.png" /> Machines

The Book

- Setting the Stage
  - Overview of the Book
  - The "PAID" Problem and The Singularity
  - The Adverbial Approach to Machine Ethics
  - The Solution in Four Steps
- Technique, Tools Technology
  - Formalism (the Philosophy of)
  - Computing With Cognitive Calculi
  - Infrastructure for Our Demonstrations
  - An Ethical Hierarchy ($\mathcal{EH}$) for Persons and Machines
- Making Morally X Machines
  - Making Morally Invulnerable Machines
  - Making Morally Incorruptible Machines
  - Making Morally Self-Aware Machines
  - Making Morally Courageous Machines
  - Making Morally Educable Machines
  - Making Morally Wise Machines
  - Making Morally Creative Machines
  - Making Morally Well-Intentioned Machines
  - Making Morally Autonomous Machines
  - Making Morally Courteous Machines
  - Making Morally Heroic Machines
- What Shall We Now Do?
- References

Table of Contents

- The Book
- The Demonstrations
- Tutorials
- Code
- Media

The Book

- Setting the Stage
  - Overview of the Book
  - The "PAID" Problem and The
  - The Adverbial Approach to Ma
  - The Solution in Four Steps
- Technique, Tools Technology

End

(Extra slides follow.)