

# Logic Can Save Us from “Killer Robots”

**Selmer Bringsjord**

**Naveen Sundar G • Atriya Sen • Mike Giancola 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

IFLAI  
4/8/2021



# Logic Can Save Us from “Killer Robots”

**Selmer Bringsjord**

**Naveen Sundar G • Atriya Sen • Mike Giancola 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

IFLAI  
4/8/2021





# The PAID Problem



# The PAID Problem

$\forall x$  : Agents





# The PAID Problem

$\forall x : \text{Agents}$

**P**owerful(x) + **A**utonomous(x) + **I**ntelligent(x)  $\Rightarrow$  **D**angerous(x)/  
**D**estroy\_Us

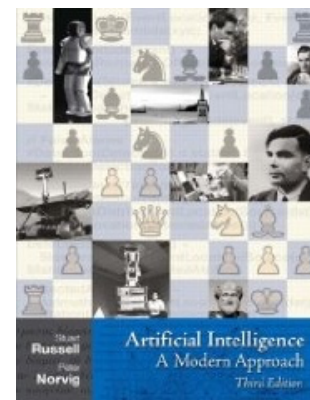


“We’re in *very* deep trouble.”

“We’re in *very* deep trouble.”



# “We’re in *very* deep trouble.”



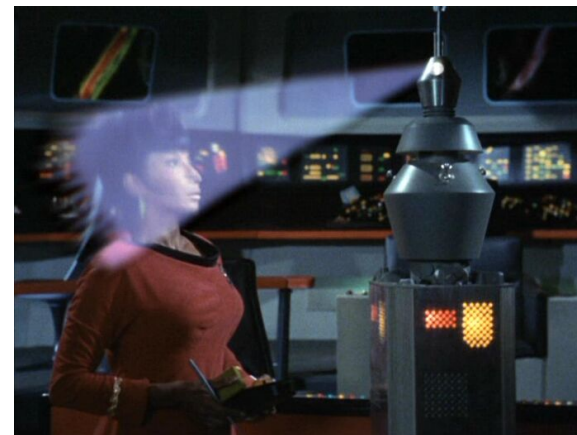
While the PAI machines aren't quite as easy to neutralize as the destructive machines vanquished in *Star Trek:TOS*, these relevant four episodes are remarkably instructive.



"The Ultimate Computer"  
S2 E24



"The Return of the Archons"  
S1 E21



"The Changeling"  
S2 E3



"I, Mudd"  
S2 E8



# Logic Thwarts Landru!



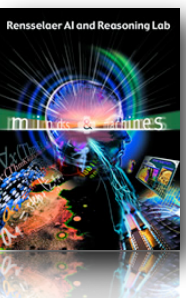
First Suspicion That It's a Mere Computer Running the Show



# Logic Thwarts Landru!

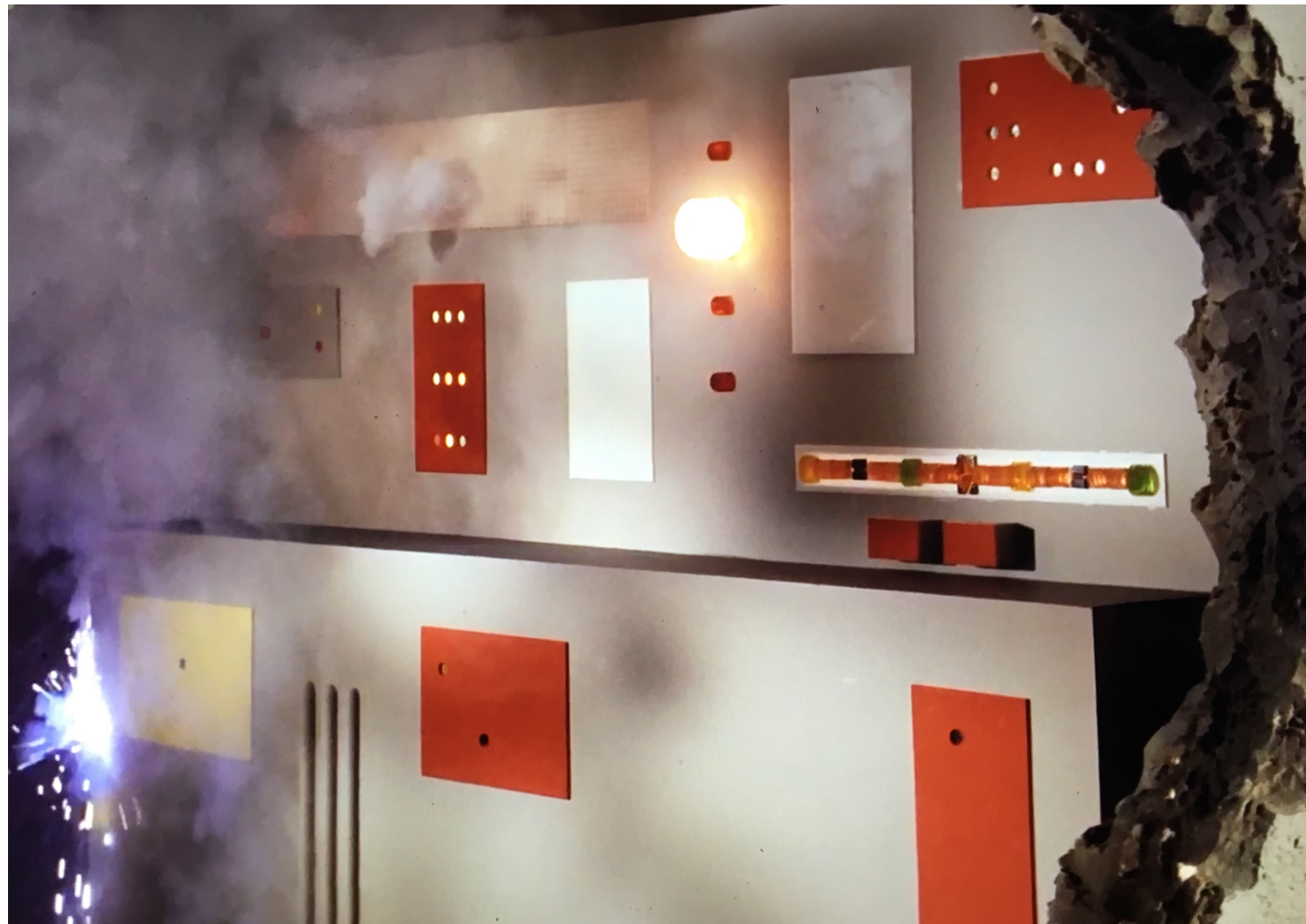


Landru is Indeed Merely a Computer  
(the real Landru having done the programming)





# Logic Thwarts Landru!



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



# Logic Thwarts Nomad!

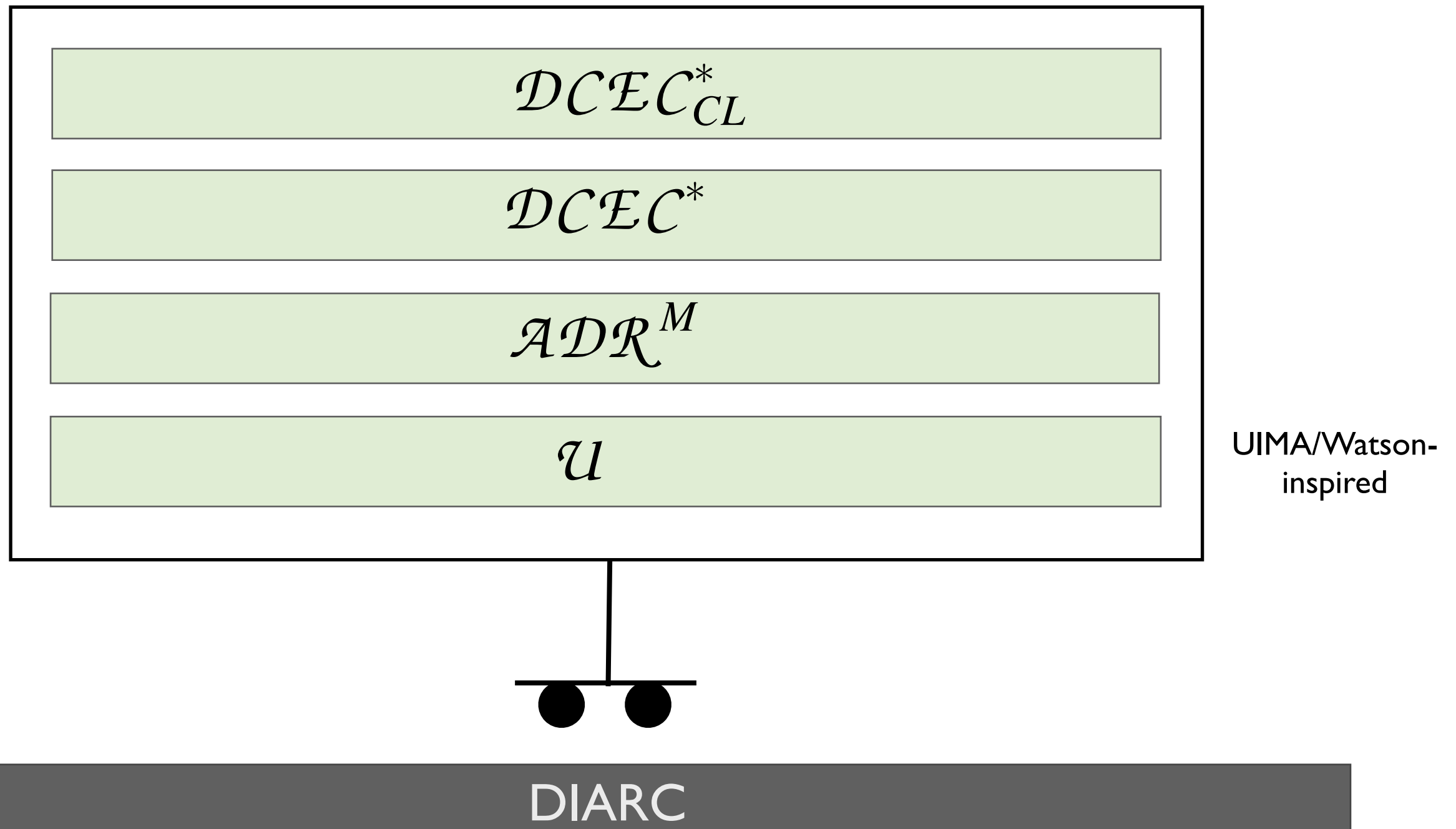
(with the Liar Paradox)



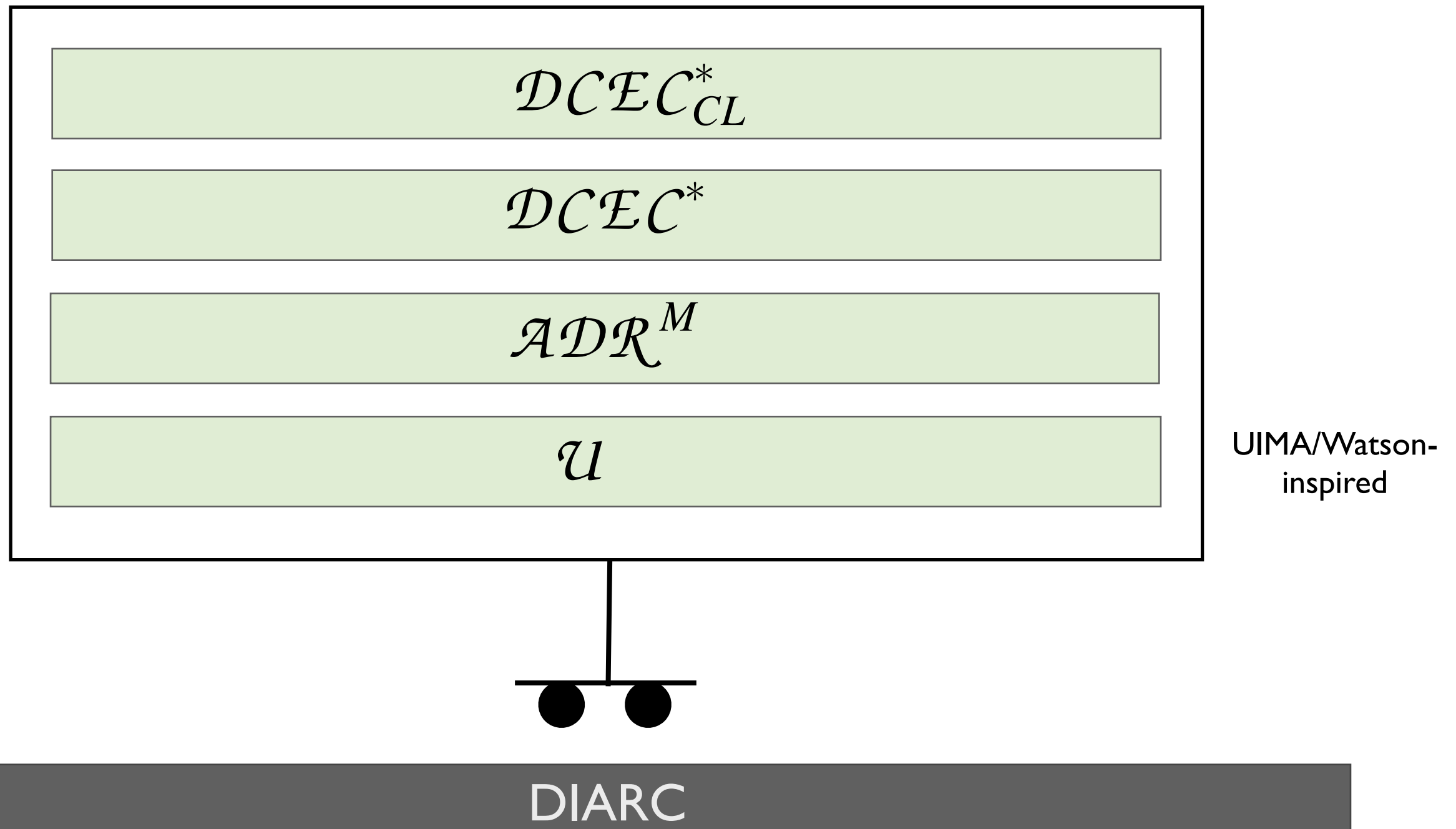
I.

Cognitive Calculi ...

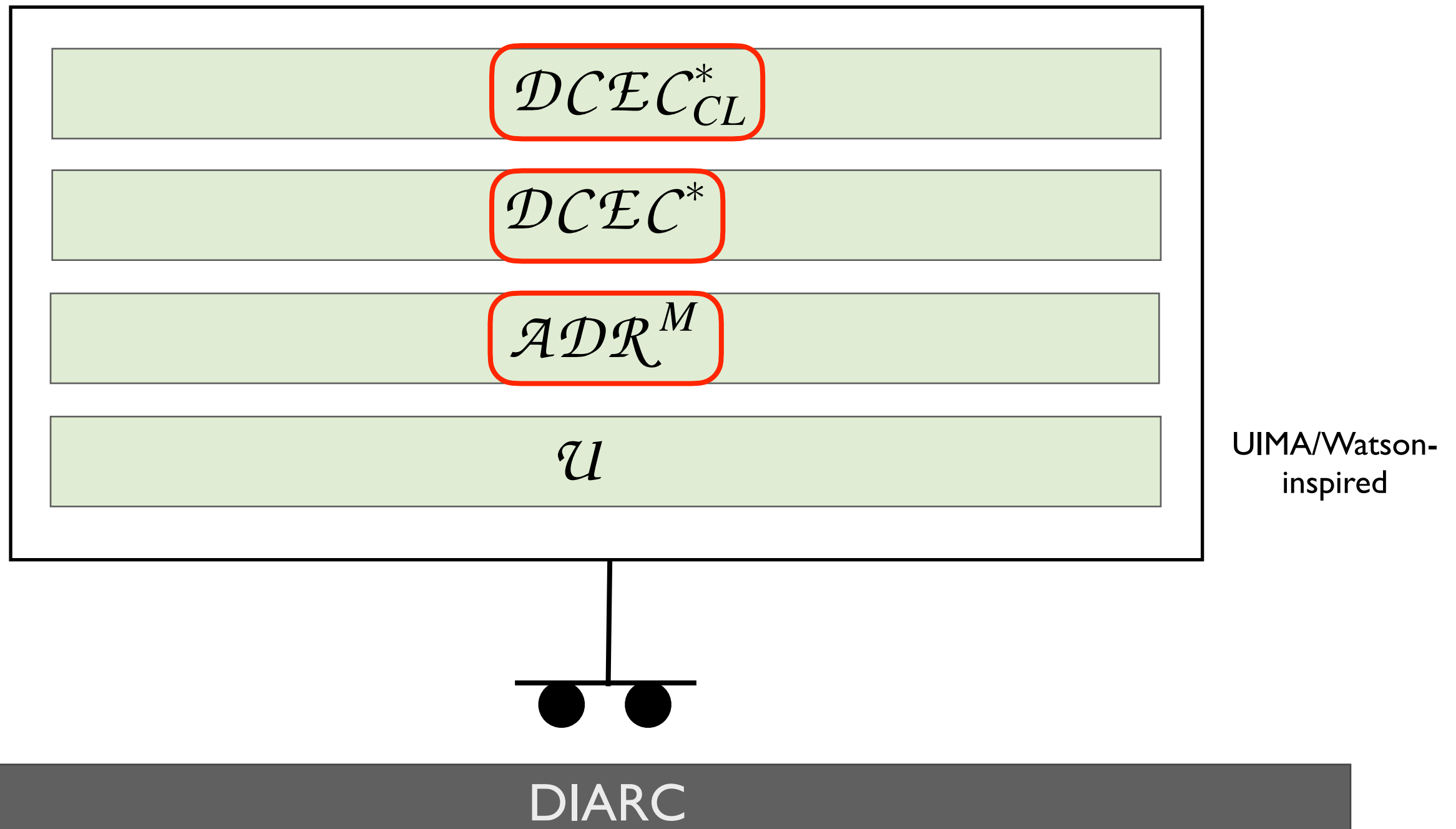
# Hierarchy of Ethical Reasoning



# Hierarchy of Ethical Reasoning

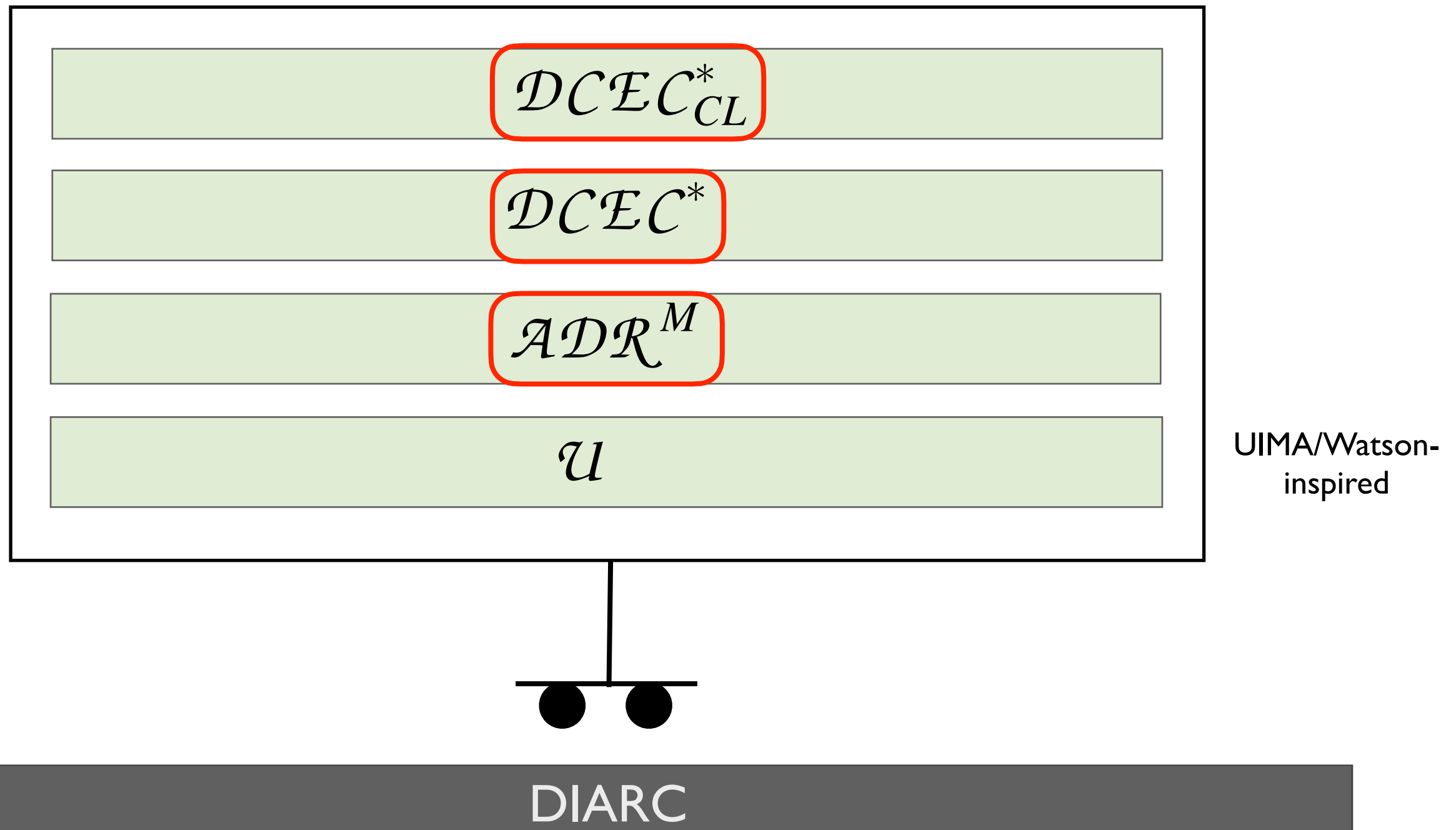


# Hierarchy of Ethical Reasoning



# Hierarchy of Ethical Reasoning

*Not* paradox-prone deontic logics!



# “Universal Cognitive Calculus”



Logic Theorist  
(birth of modern logicist AI)

*DCEC\**

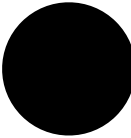
**Syntax**

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

**Rules of Inference**

$[R_1] \frac{C(t, P(a, t, \phi)) \rightarrow K(a, t, \phi)}{C(t, P(a, t, \phi))} [R_2] \frac{C(t, K(a, t, \phi)) \rightarrow B(a, t, \phi)}{C(t, K(a, t, \phi))}$   
 $[R_3] \frac{C(t, \phi) \wedge t \leq t_1 \rightarrow t \leq t_2}{C(t, \phi) \wedge t \leq t_1 \rightarrow t \leq t_2} [R_4] \frac{K(a, t, \phi)}{\phi}$   
 $[R_5] \frac{C(t, K(a, t_1, \phi_1) \rightarrow \phi_2) \rightarrow K(a, t_2, \phi_1) \rightarrow K(a, t_2, \phi_2)}{C(t, K(a, t_1, \phi_1) \rightarrow \phi_2) \rightarrow K(a, t_2, \phi_1) \rightarrow K(a, t_2, \phi_2)}$   
 $[R_6] \frac{C(t, B(a, t_1, \phi_1) \rightarrow \phi_2) \rightarrow B(a, t_2, \phi_1) \rightarrow B(a, t_2, \phi_2)}{C(t, B(a, t_1, \phi_1) \rightarrow \phi_2) \rightarrow B(a, t_2, \phi_1) \rightarrow B(a, t_2, \phi_2)}$   
 $[R_7] \frac{C(t, \phi_1 \wedge \phi_2) \rightarrow \phi_3}{C(t, \phi_1 \wedge \phi_2) \rightarrow \phi_3} [R_8] \frac{C(t, \phi_1 \wedge \phi_2) \rightarrow \phi_3}{C(t, \phi_1 \wedge \phi_2) \rightarrow \phi_3} [R_9] \frac{C(t, \phi_1 \wedge \phi_2) \rightarrow \phi_3}{C(t, \phi_1 \wedge \phi_2) \rightarrow \phi_3}$   
 $[R_{10}] \frac{B(a, t, \phi) \rightarrow \phi}{B(a, t, \phi) \rightarrow \phi} [R_{11}] \frac{B(a, t, \phi) \rightarrow \phi}{B(a, t, \phi) \rightarrow \phi} [R_{12}] \frac{B(a, t, \phi) \rightarrow \phi}{B(a, t, \phi) \rightarrow \phi}$   
 $[R_{13}] \frac{B(a, t, \phi) \rightarrow \phi}{B(a, t, \phi) \rightarrow \phi} [R_{14}] \frac{B(a, t, \phi) \rightarrow \phi}{B(a, t, \phi) \rightarrow \phi}$   
 $[R_{15}] \frac{B(a, t, \phi) \rightarrow \phi}{B(a, t, \phi) \rightarrow \phi}$

**66**

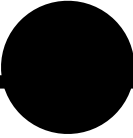


1666



Leibniz

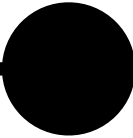
1.5 centuries < Boole!  
2.5 centuries < Kripke



1956



Simon



2019



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  
 1/11/2019

RAIR  
 Rensselaer AI and Reasoning Lab



II.

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

# NewScientist

Ethical robots save humans

# NewScientist

Ethical robots save humans





# Informal Definition of Akrasia

# Informal Definition of Akrasia



# Informal Definition of Akrasia



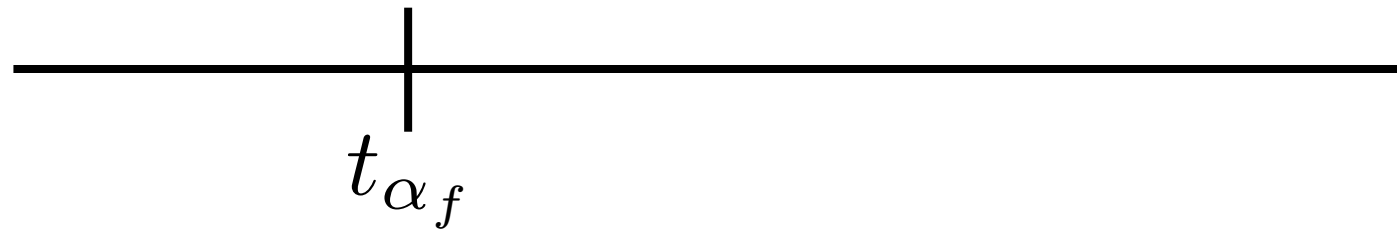
---



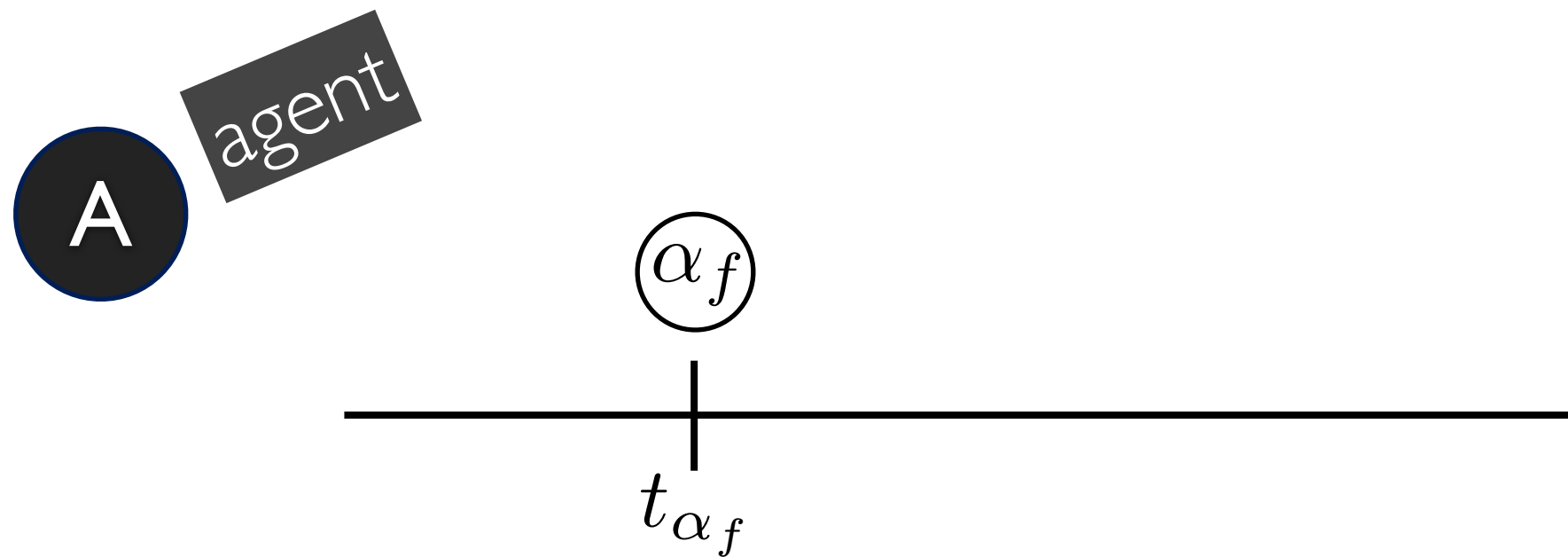
# Informal Definition of Akrasia



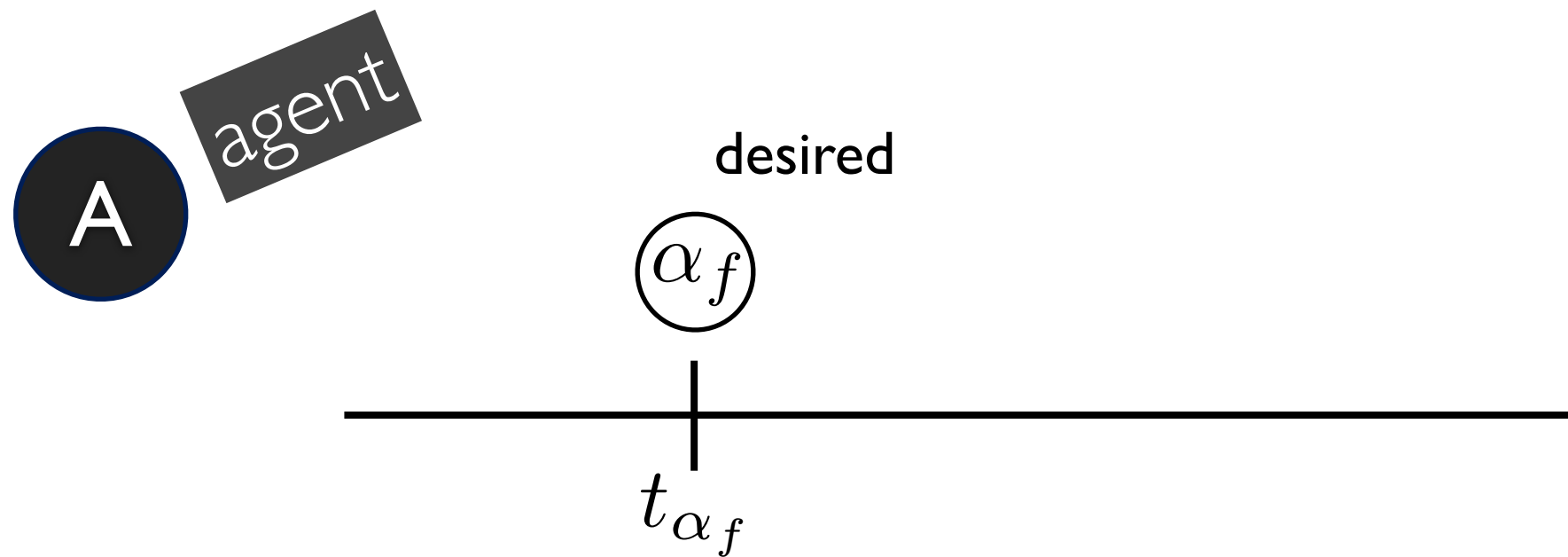
# Informal Definition of Akrasia



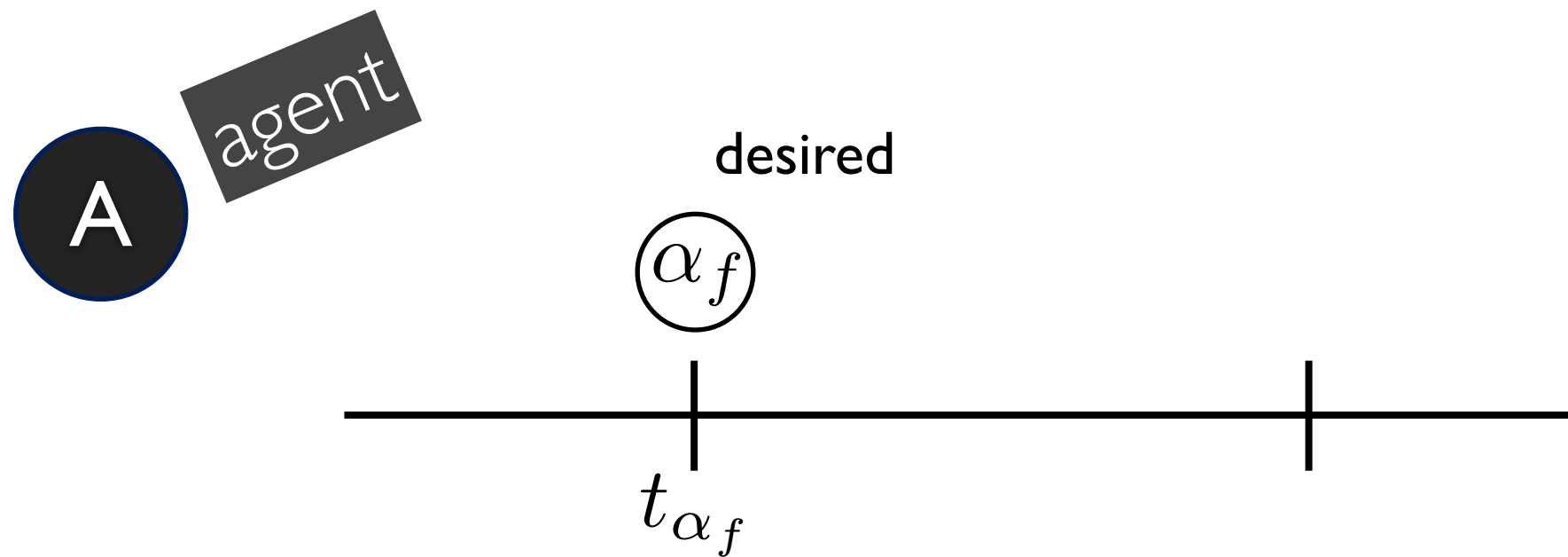
# Informal Definition of Akrasia



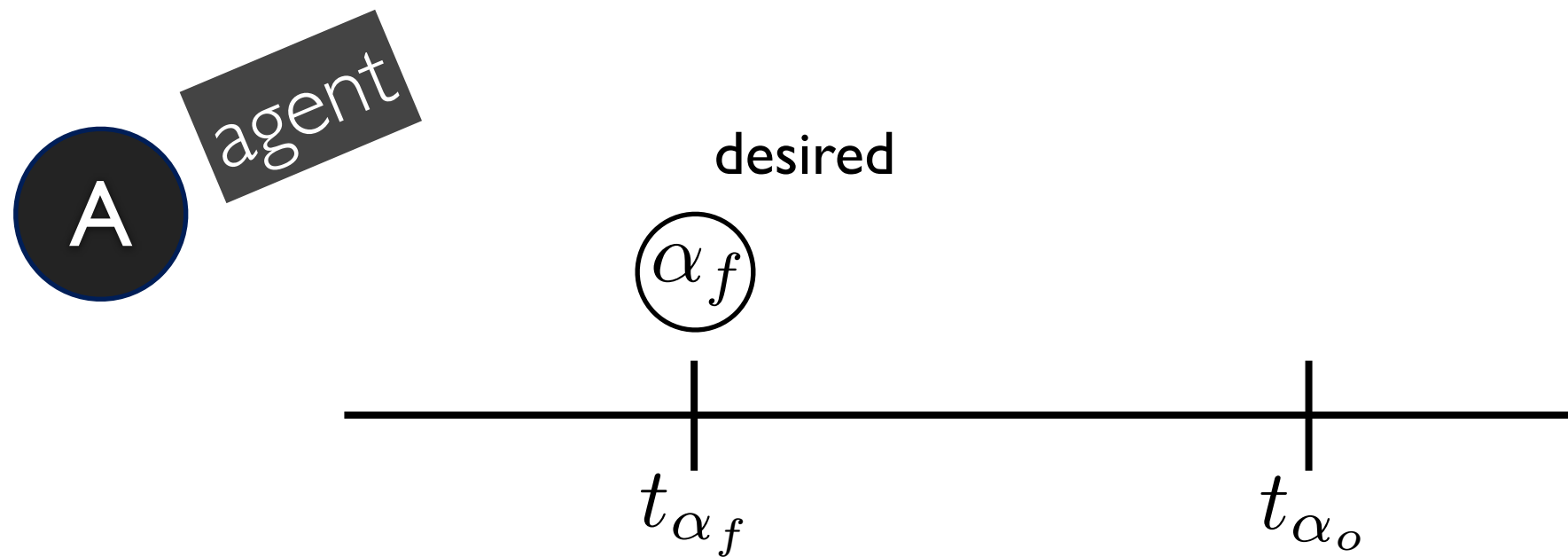
# Informal Definition of Akrasia



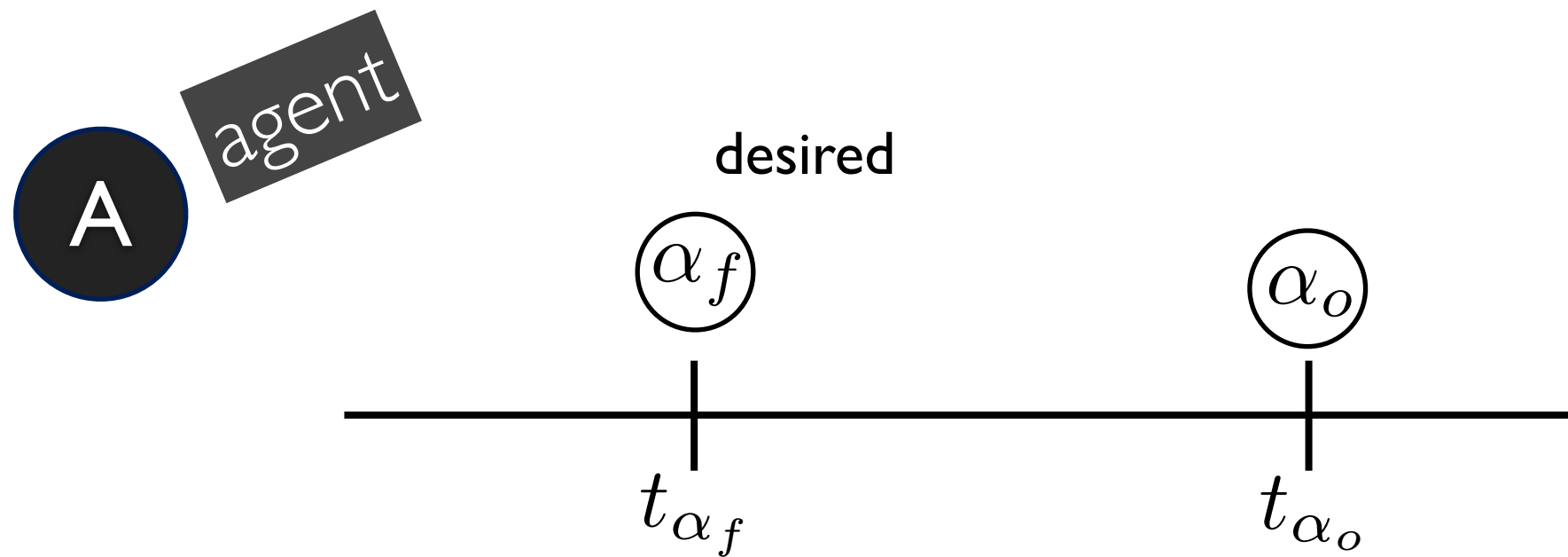
# Informal Definition of Akrasia



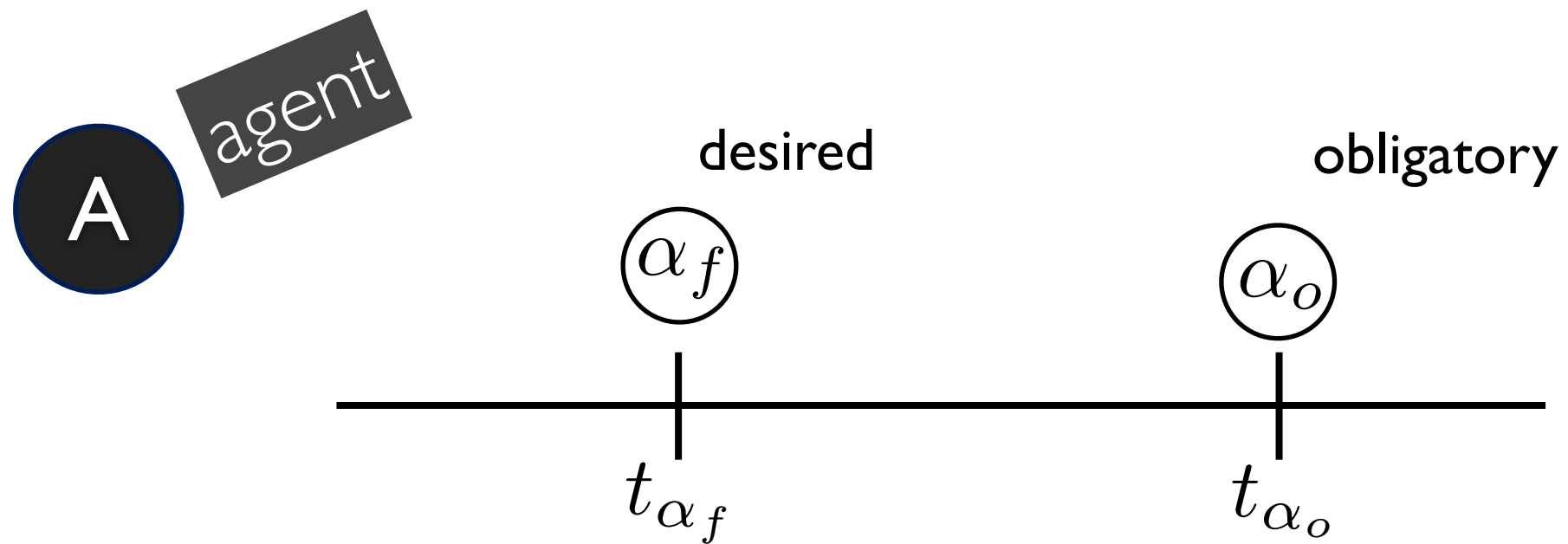
# Informal Definition of Akrasia



# Informal Definition of Akrasia

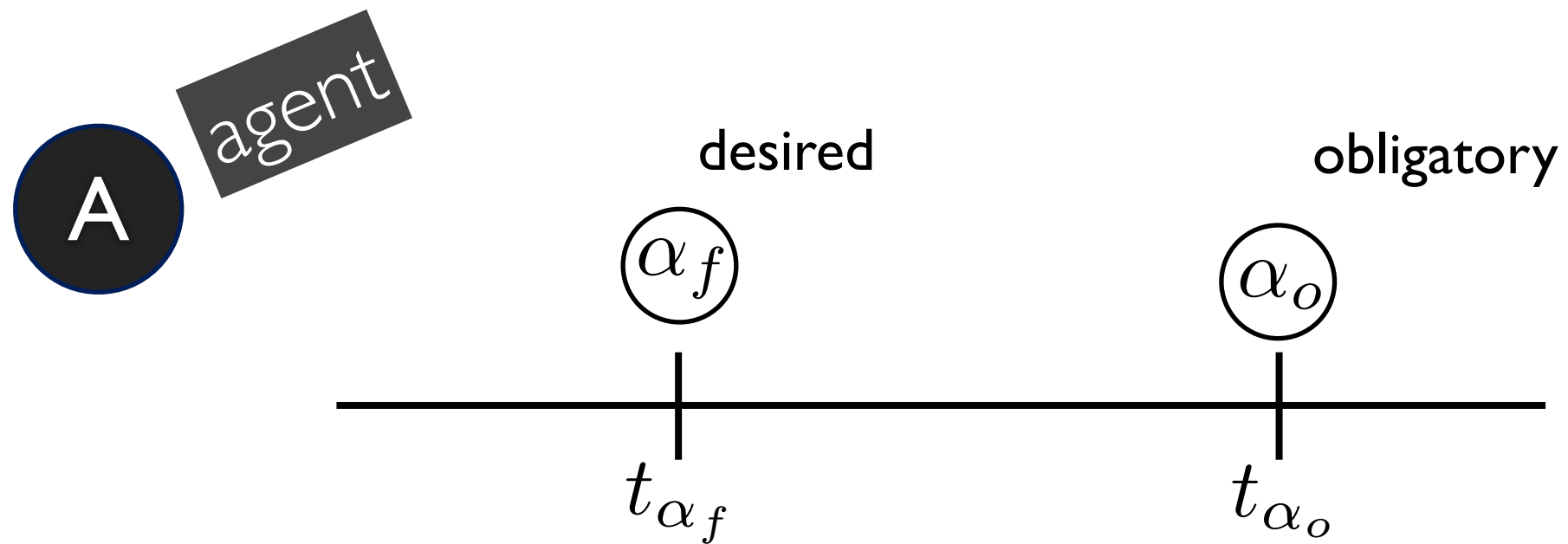


# Informal Definition of Akrasia



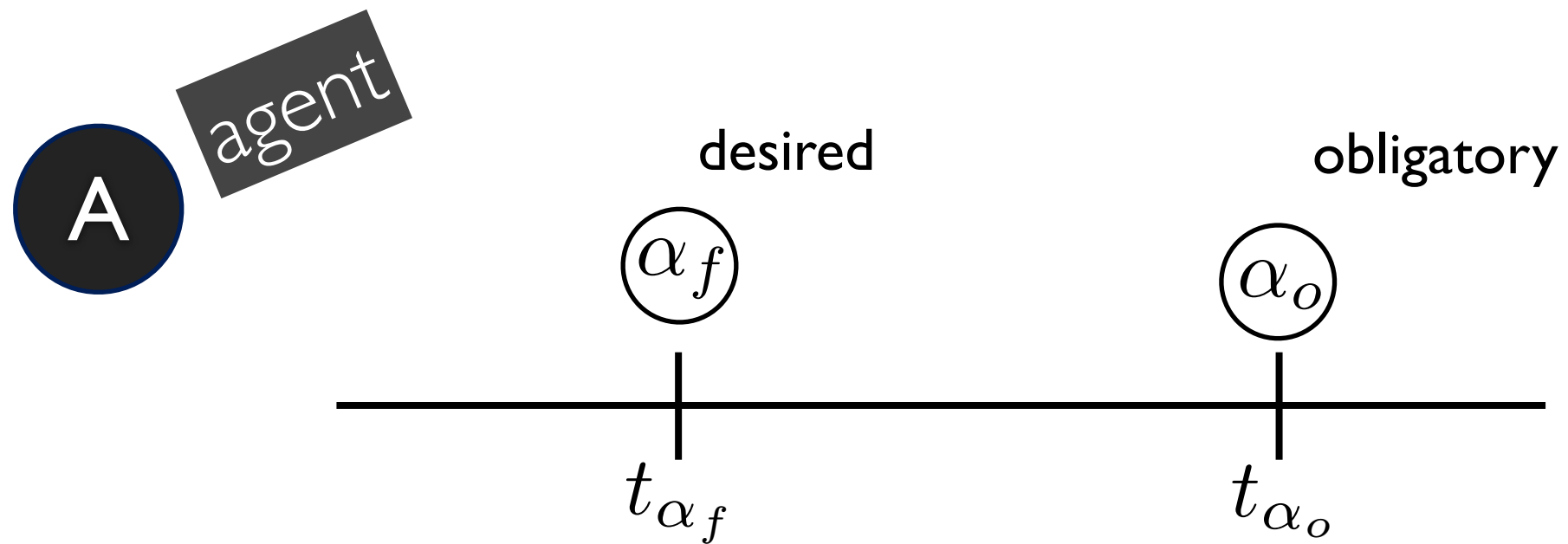


# Informal Definition of Akrasia



If  $\alpha_f$  happens, then  $\alpha_o$  can't happen

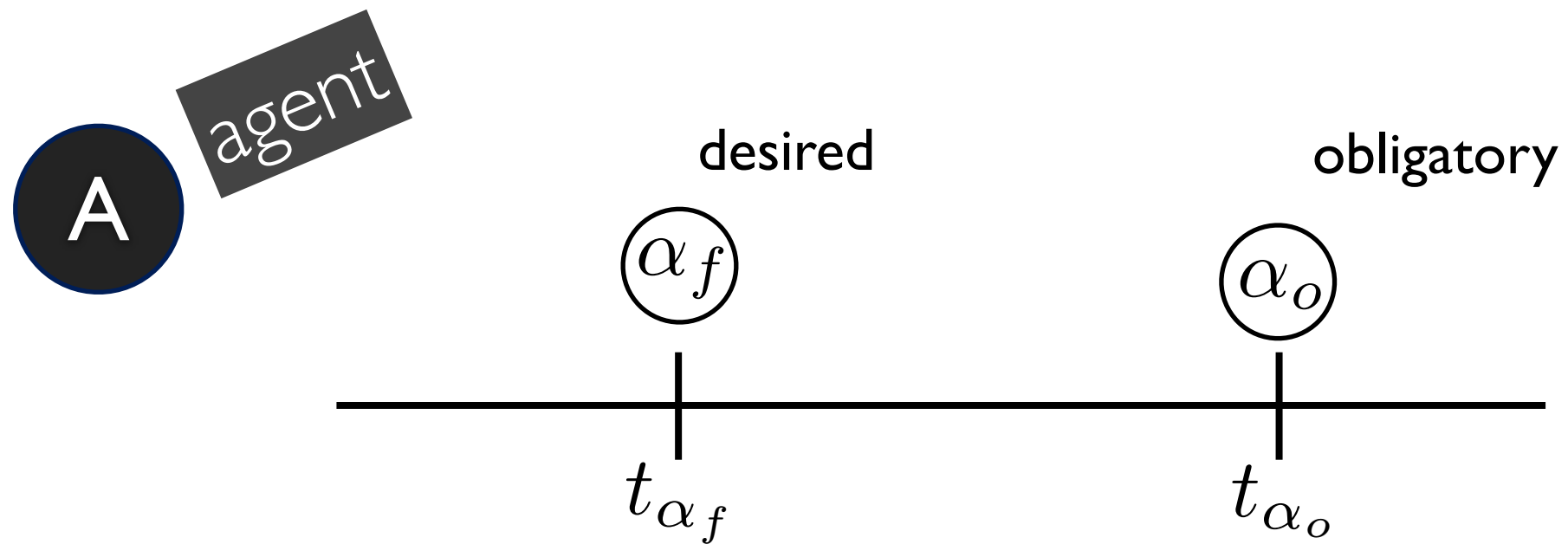
# Informal Definition of Akrasia



If  $\alpha_f$  happens, then  $\alpha_o$  can't happen



# Informal Definition of Akrasia



If  $\alpha_f$  happens, then  $\alpha_o$  can't happen

A knows this

# Informal Definition of Akrasia

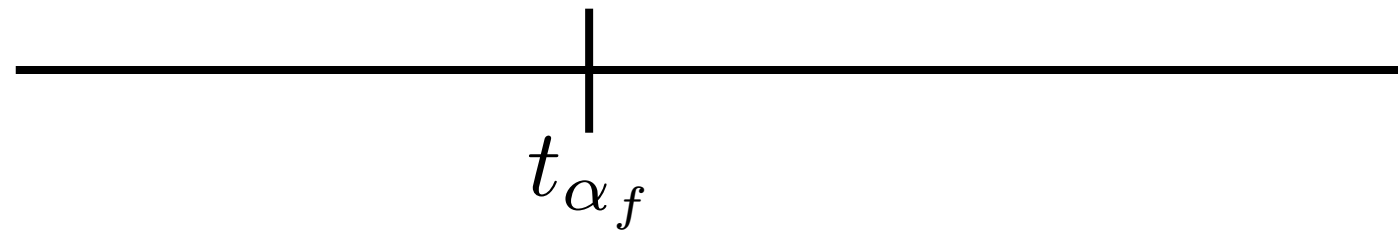
# Informal Definition of Akrasia



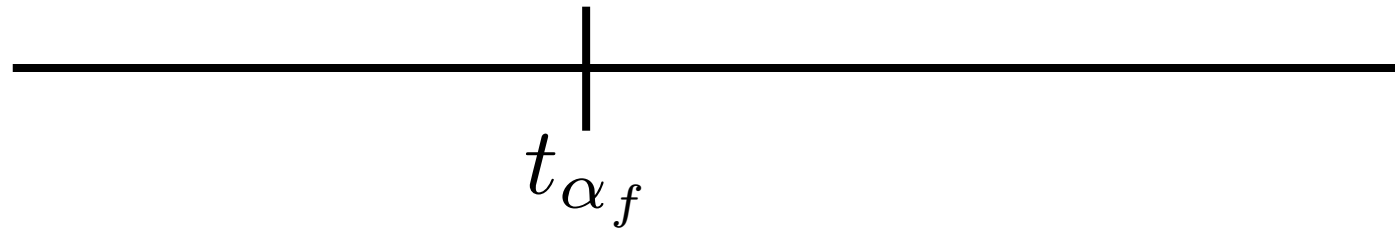
# Informal Definition of Akrasia



# Informal Definition of Akrasia




# Informal Definition of Akrasia





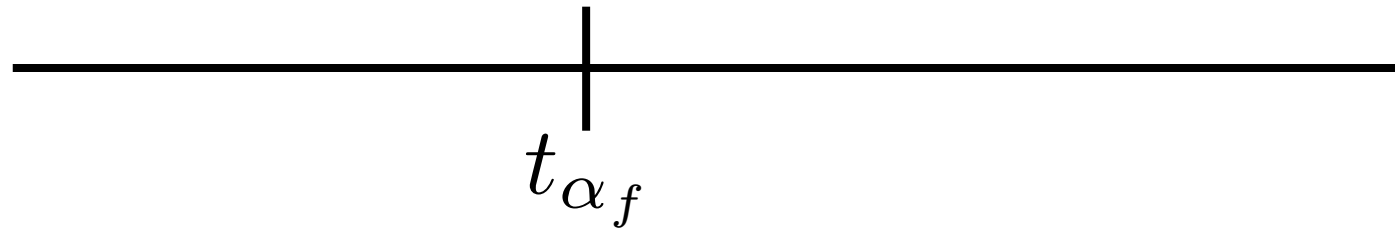
# Informal Definition of Akrasia

Desire to do  $(\alpha_f)$  

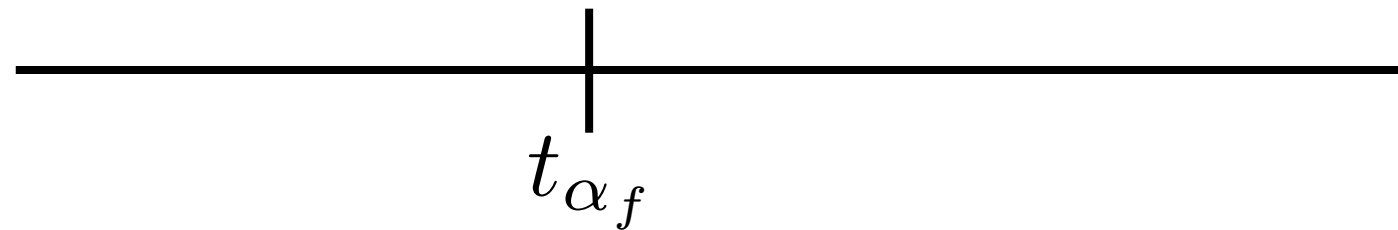
  
 $t_{\alpha_f}$

# Informal Definition of Akrasia

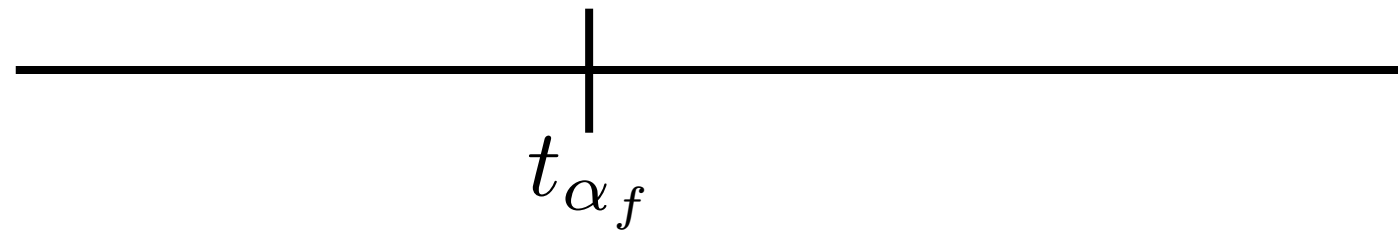
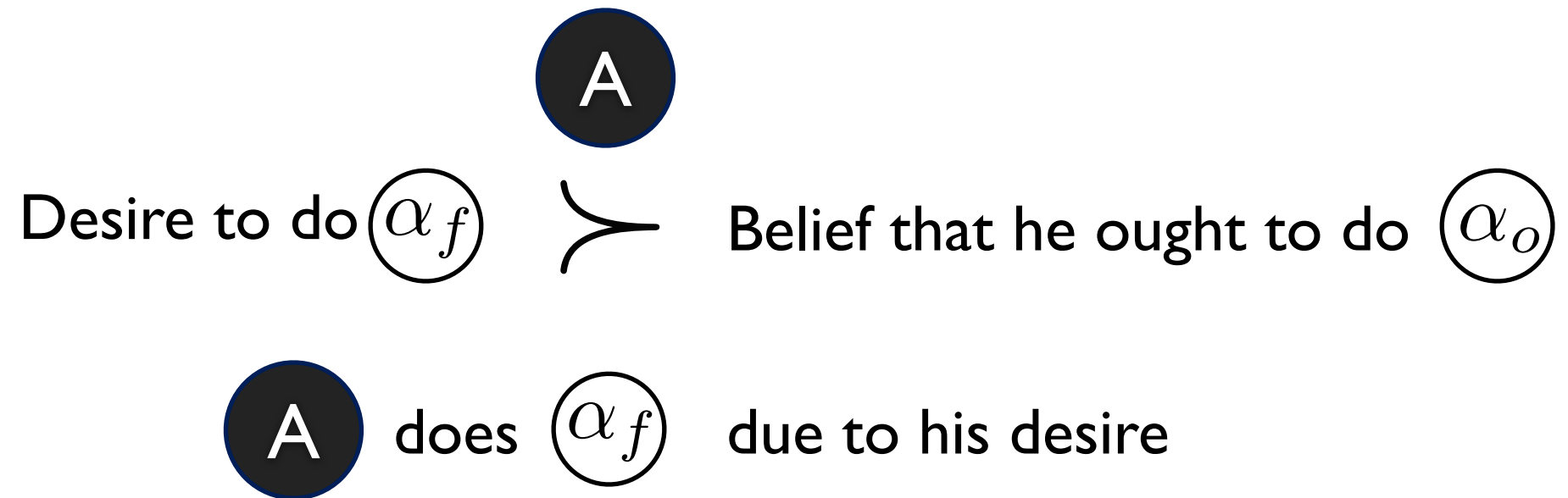
Desire to do  $(\alpha_f)$   $\succ$   $\textcircled{A}$



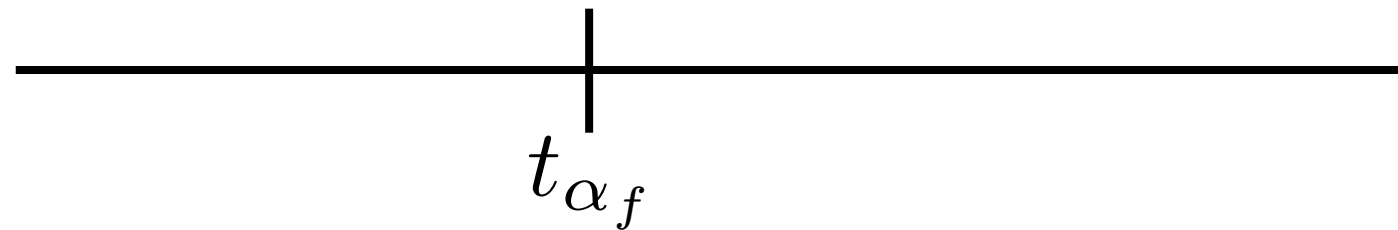
# Informal Definition of Akrasia



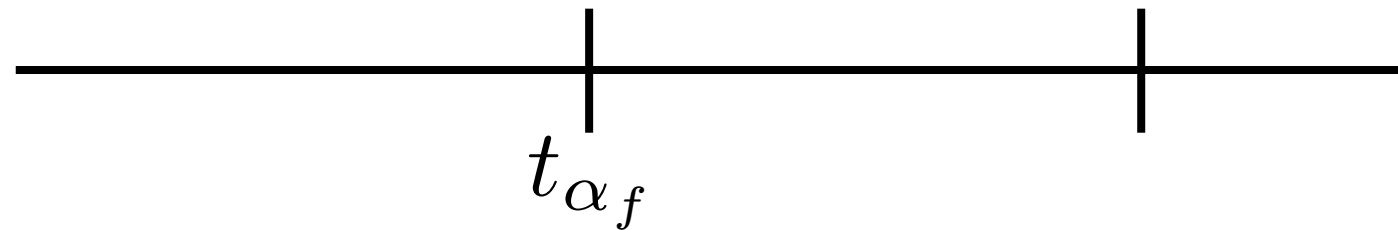
# Informal Definition of Akrasia



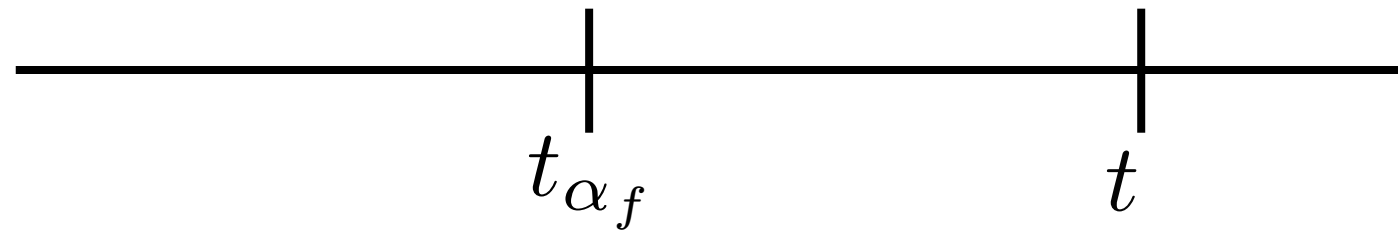
# Informal Definition of Akrasia



# Informal Definition of Akrasia



# Informal Definition of Akrasia

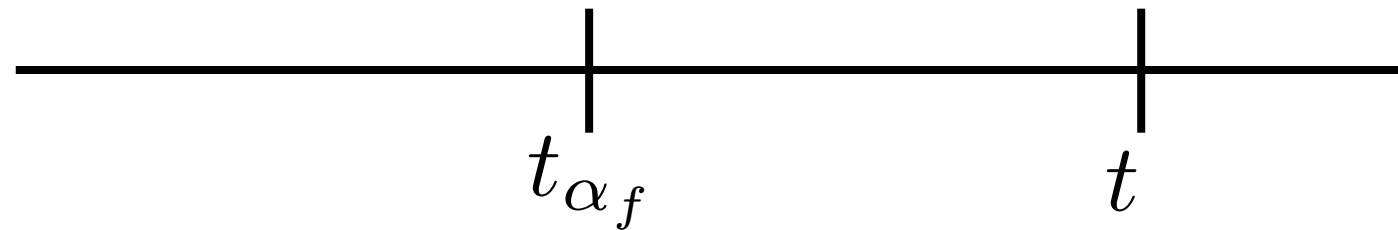


# Informal Definition of Akrasia

 Desire to do  $(\alpha_f)$   $\succ$  Belief that he ought to do  $(\alpha_o)$

 does  $(\alpha_f)$  due to his desire

 believes he should have done  $(\alpha_o)$





# Informal Definition of Akrasia

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

- (1)  $A$  believes that  $A$  ought to do  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (2)  $A$  desires to do  $\alpha_f$  at  $t_{\alpha_f}$ ;
- (3)  $A$ 's doing  $\alpha_f$  at  $t_{\alpha_f}$  entails his not doing  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (4)  $A$  knows that doing  $\alpha_f$  at  $t_{\alpha_f}$  entails his not doing  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (5) At the time ( $t_{\alpha_f}$ ) of doing the forbidden  $\alpha_f$ ,  $A$ 's desire to do  $\alpha_f$  overrides  $A$ 's belief that he ought to do  $\alpha_o$  at  $t_{\alpha_o}$ .
- (6)  $A$  does the forbidden action  $\alpha_f$  at  $t_{\alpha_f}$ ;
- (7)  $A$ 's doing  $\alpha_f$  results from  $A$ 's desire to do  $\alpha_f$ ;
- (8) At some time  $t$  after  $t_{\alpha_f}$ ,  $A$  has the belief that  $A$  ought to have done  $\alpha_o$  rather than  $\alpha_f$ .

# Informal Definition of Akrasia

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

- (1)  $A$  believes that  $A$  ought to do  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (2)  $A$  desires to do  $\alpha_f$  at  $t_{\alpha_f}$ ;
- (3)  $A$ 's doing  $\alpha_f$  at  $t_{\alpha_f}$  entails his not doing  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (4)  $A$  knows that doing  $\alpha_f$  at  $t_{\alpha_f}$  entails his not doing  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (5) At the time ( $t_{\alpha_f}$ ) of doing the forbidden  $\alpha_f$ ,  $A$ 's desire to do  $\alpha_f$  overrides  $A$ 's belief that he ought to do  $\alpha_o$  at  $t_{\alpha_o}$ .
- (6)  $A$  does the forbidden action  $\alpha_f$  at  $t_{\alpha_f}$ ;
- (7)  $A$ 's doing  $\alpha_f$  results from  $A$ 's desire to do  $\alpha_f$ ;
- (8) At some time  $t$  after  $t_{\alpha_f}$ ,  $A$  has the belief that  $A$  ought to have done  $\alpha_o$  rather than  $\alpha_f$ .

# Informal Definition of Akrasia

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

- (1)  $A$  believes that  $A$  ought to do  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (2)  $A$  desires to do  $\alpha_f$  at  $t_{\alpha_f}$ ;
- (3)  $A$ 's doing  $\alpha_f$  at  $t_{\alpha_f}$  entails his not doing  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (4)  $A$  knows that doing  $\alpha_f$  at  $t_{\alpha_f}$  entails his not doing  $\alpha_o$  at  $t_{\alpha_o}$ ;
- (5) At the time ( $t_{\alpha_f}$ ) of doing the forbidden  $\alpha_f$ ,  $A$ 's desire to do  $\alpha_f$  overrides  $A$ 's belief that he ought to do  $\alpha_o$  at  $t_{\alpha_o}$ .
- (6)  $A$  does the forbidden action  $\alpha_f$  at  $t_{\alpha_f}$ ;
- (7)  $A$ 's doing  $\alpha_f$  results from  $A$ 's desire to do  $\alpha_f$ ;
- (8) At some time  $t$  after  $t_{\alpha_f}$ ,  $A$  has the belief that  $A$  ought to have done  $\alpha_o$  rather than  $\alpha_f$ .

“Regret”

Cast in

$\mathcal{DCEC}^*$

this becomes ...



$$\text{KB}_{rs} \cup \text{KB}_{m_1} \cup \text{KB}_{m_2} \dots \text{KB}_{m_n} \vdash$$

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

$$D_2 : \mathbf{D}(\mathbf{l}, \text{now}, \text{holds}(\text{does}(\mathbf{l}^*, \bar{\alpha}), t_{\bar{\alpha}}))$$

$$D_3 : \text{happens}(\text{action}(\mathbf{l}^*, \bar{\alpha}), t_{\bar{\alpha}}) \Rightarrow \neg \text{happens}(\text{action}(\mathbf{l}^*, \alpha), t_\alpha)$$

$$D_4 : \mathbf{K}\left(\mathbf{l}, \text{now}, \left( \begin{array}{l} \text{happens}(\text{action}(\mathbf{l}^*, \bar{\alpha}), t_{\bar{\alpha}}) \Rightarrow \\ \neg \text{happens}(\text{action}(\mathbf{l}^*, \alpha), t_\alpha) \end{array} \right)\right)$$

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

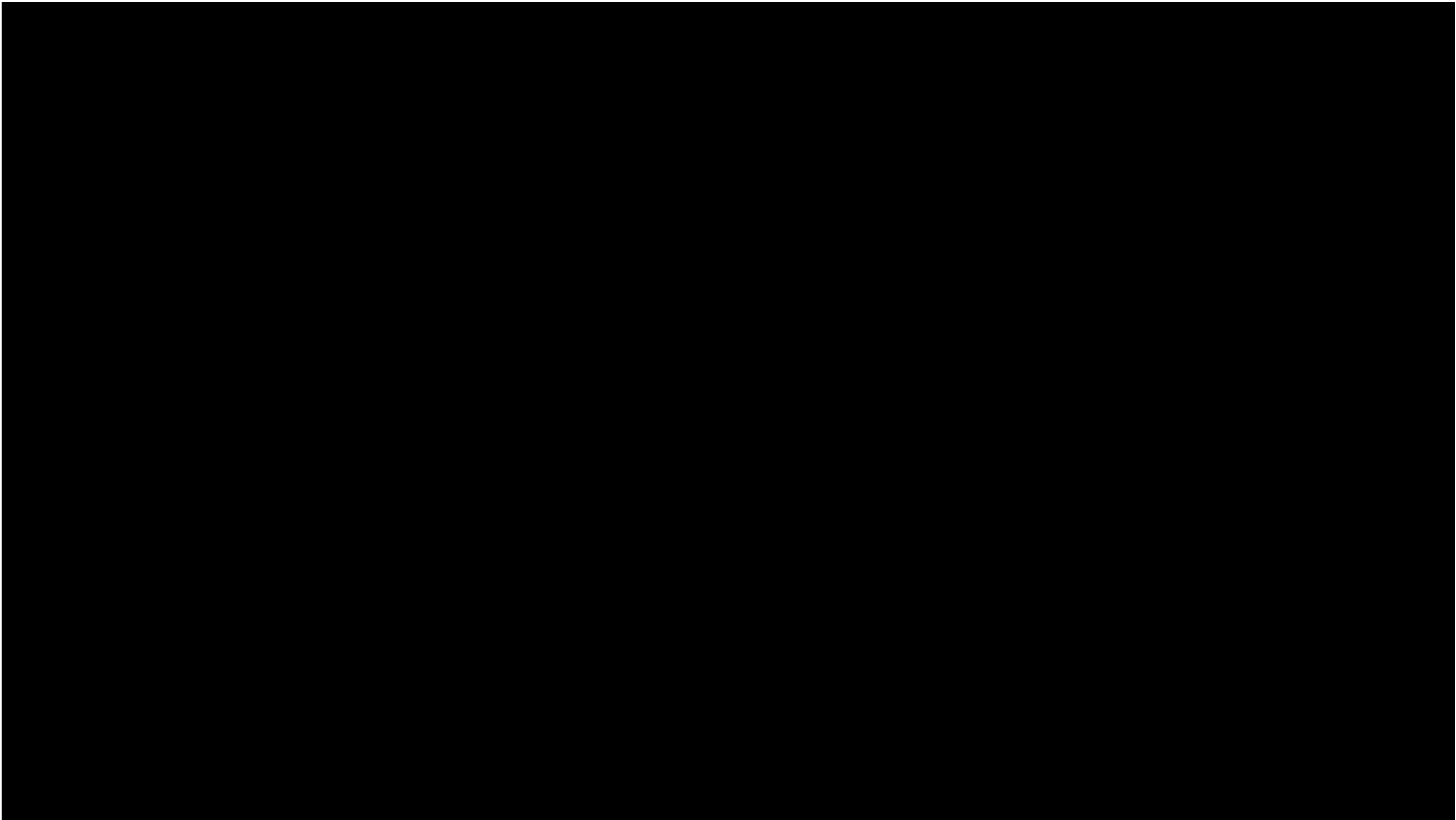
$$D_6 : \text{happens}(\text{action}(\mathbf{l}^*, \bar{\alpha}), t_{\bar{\alpha}})$$

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

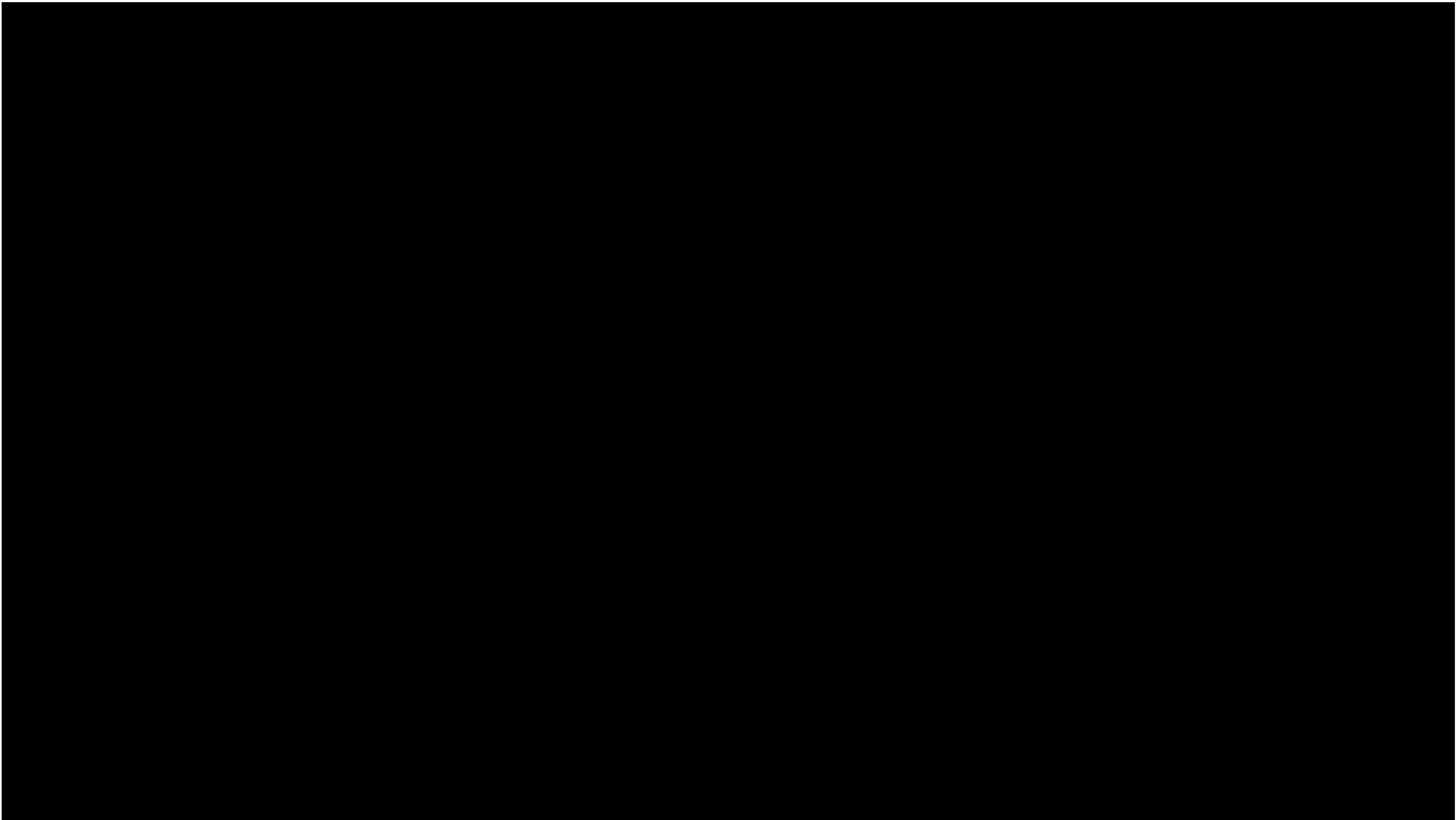
$$D_{7b} : \begin{array}{l} \Gamma - \{\mathbf{D}(\mathbf{l}, \text{now}, \text{holds}(\text{does}(\mathbf{l}^*, \bar{\alpha}), t))\} \not\vdash \\ \text{happens}(\text{action}(\mathbf{l}^*, \bar{\alpha}), t_\alpha) \end{array}$$

$$D_8 : \mathbf{B}(\mathbf{l}, t_f, \mathbf{O}(\mathbf{l}^*, t_\alpha, \Phi, \text{happens}(\text{action}(\mathbf{l}^*, \alpha), t_\alpha)))$$

# Demos ...



# Demos ...





III.

But, a twist befell the logicians ...

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:

*EH*



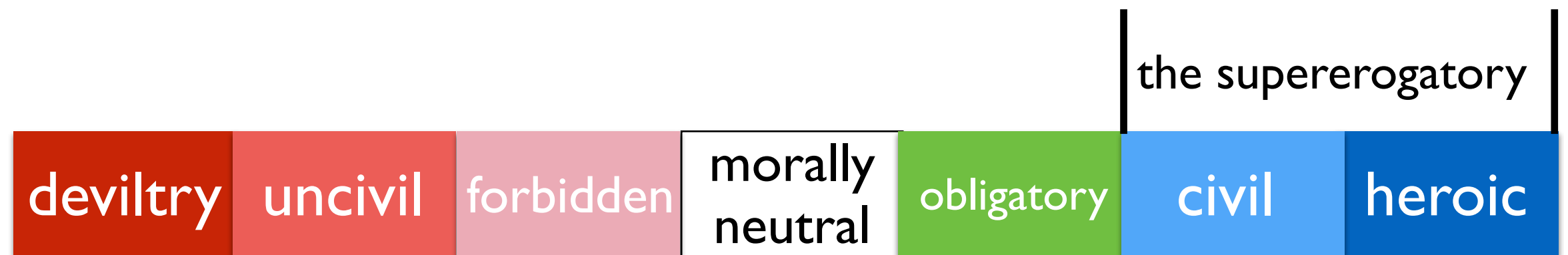
# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*



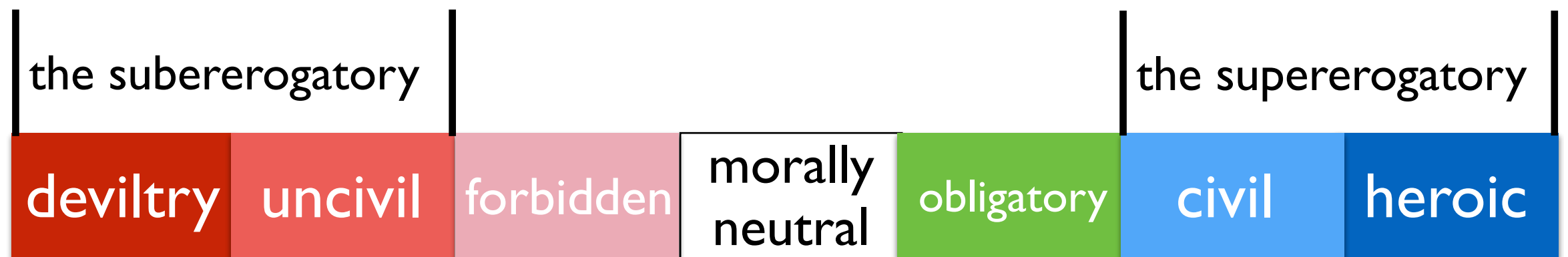
# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*



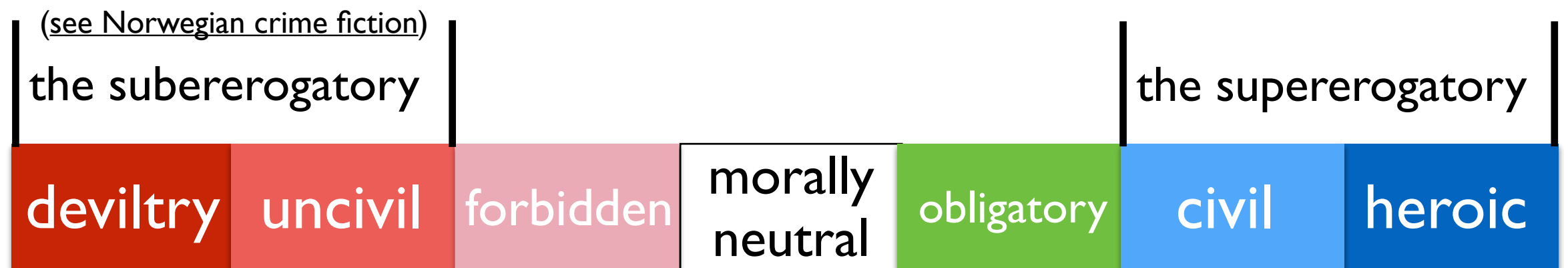
# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*



# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*





# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*

19th-Century Triad

(see Norwegian crime fiction)

the subererogatory

deviltry

uncivil

forbidden

morally  
neutral

obligatory

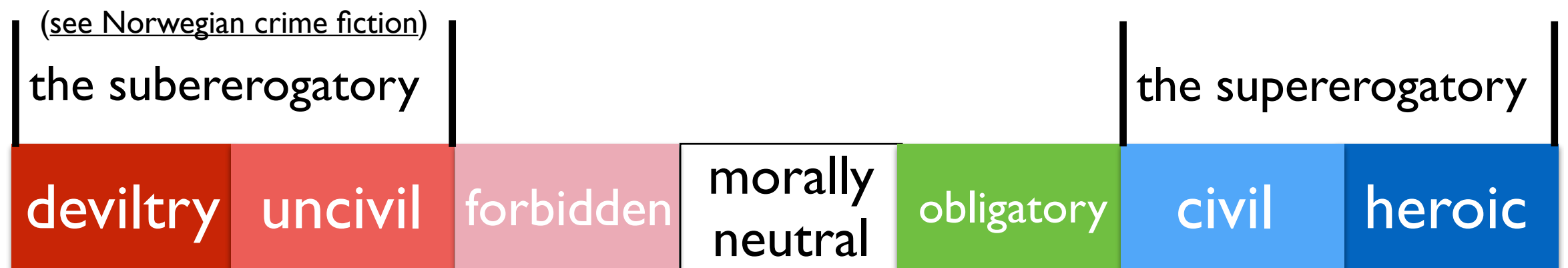
civil

heroic

the supererogatory

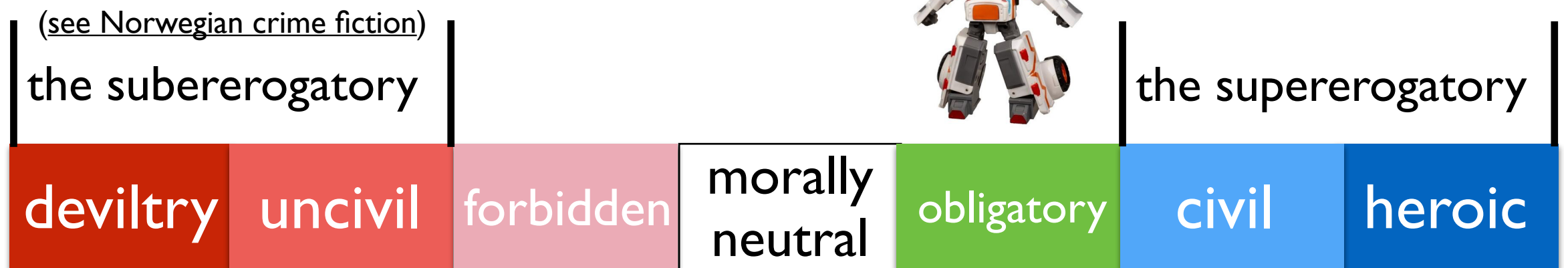
# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*



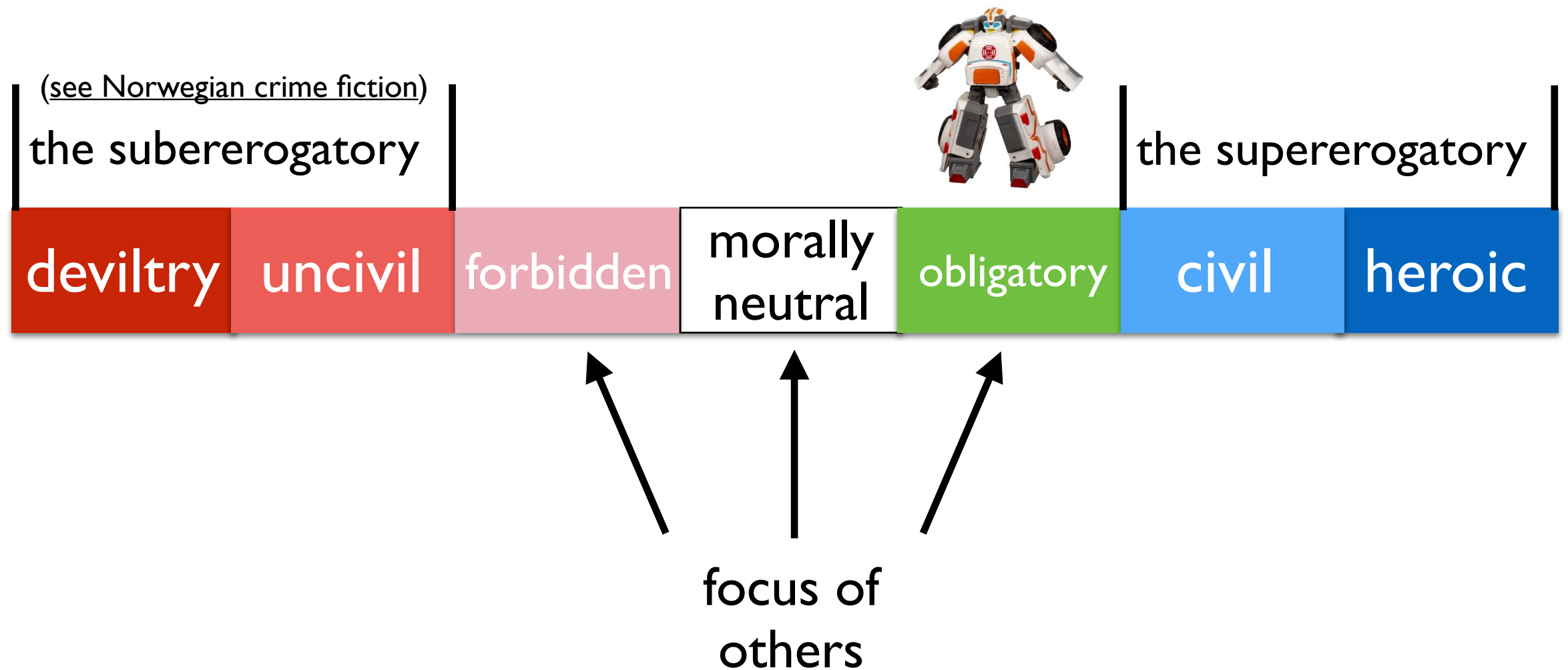
# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*



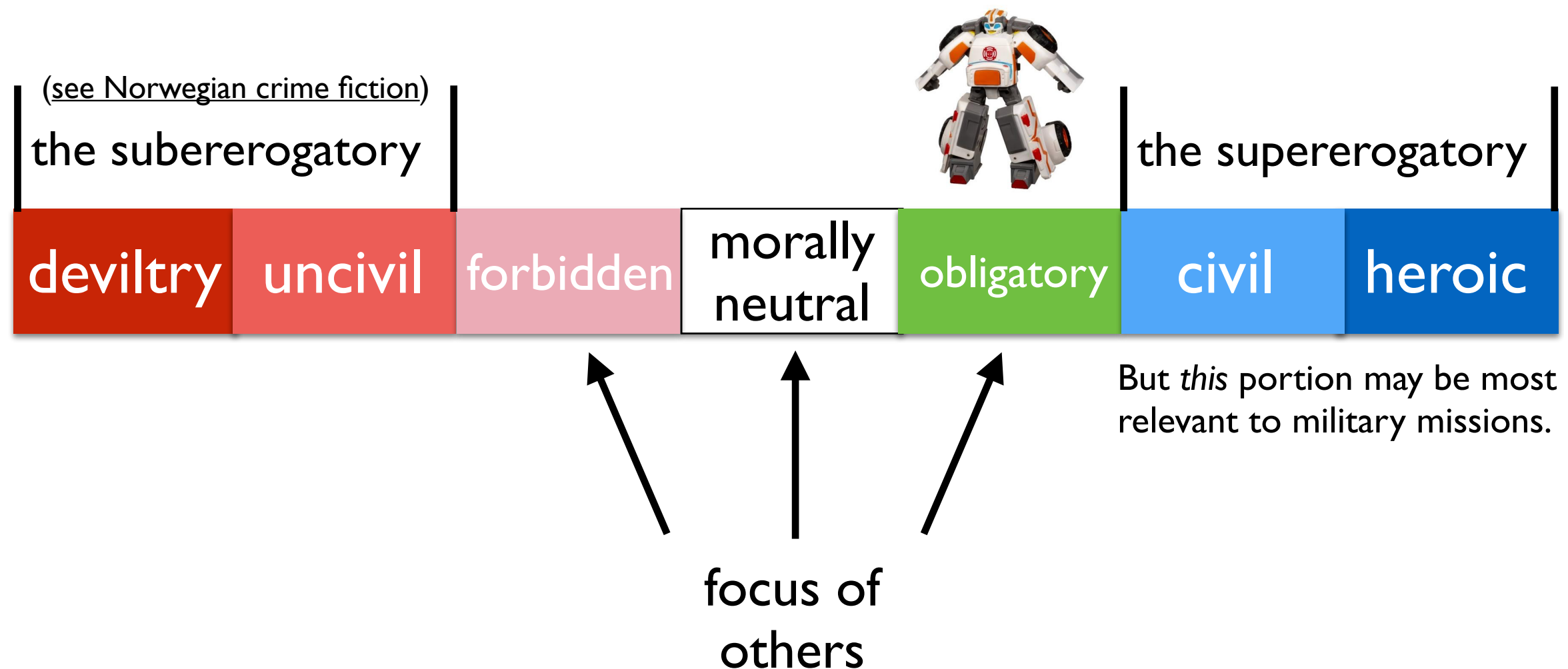
# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*



# Leibnizian Ethical Hierarchy for Persons and Robots:

*EH*





# Bert “Heroically” Saved?



Courtesy of RAIR-Lab Researcher Atriya Sen



# Bert “Heroically” Saved?



Courtesy of RAIR-Lab Researcher Atriya Sen



# Supererogatory<sup>2</sup> 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



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


Courtesy of RAIR-Lab Researcher Atriya Sen



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


# 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(0(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(0(nao,t2,lessThan(payoff(nao,not(dive),t2),threshold),happens(action(nao,dive),t2))))

provable Conjectures:

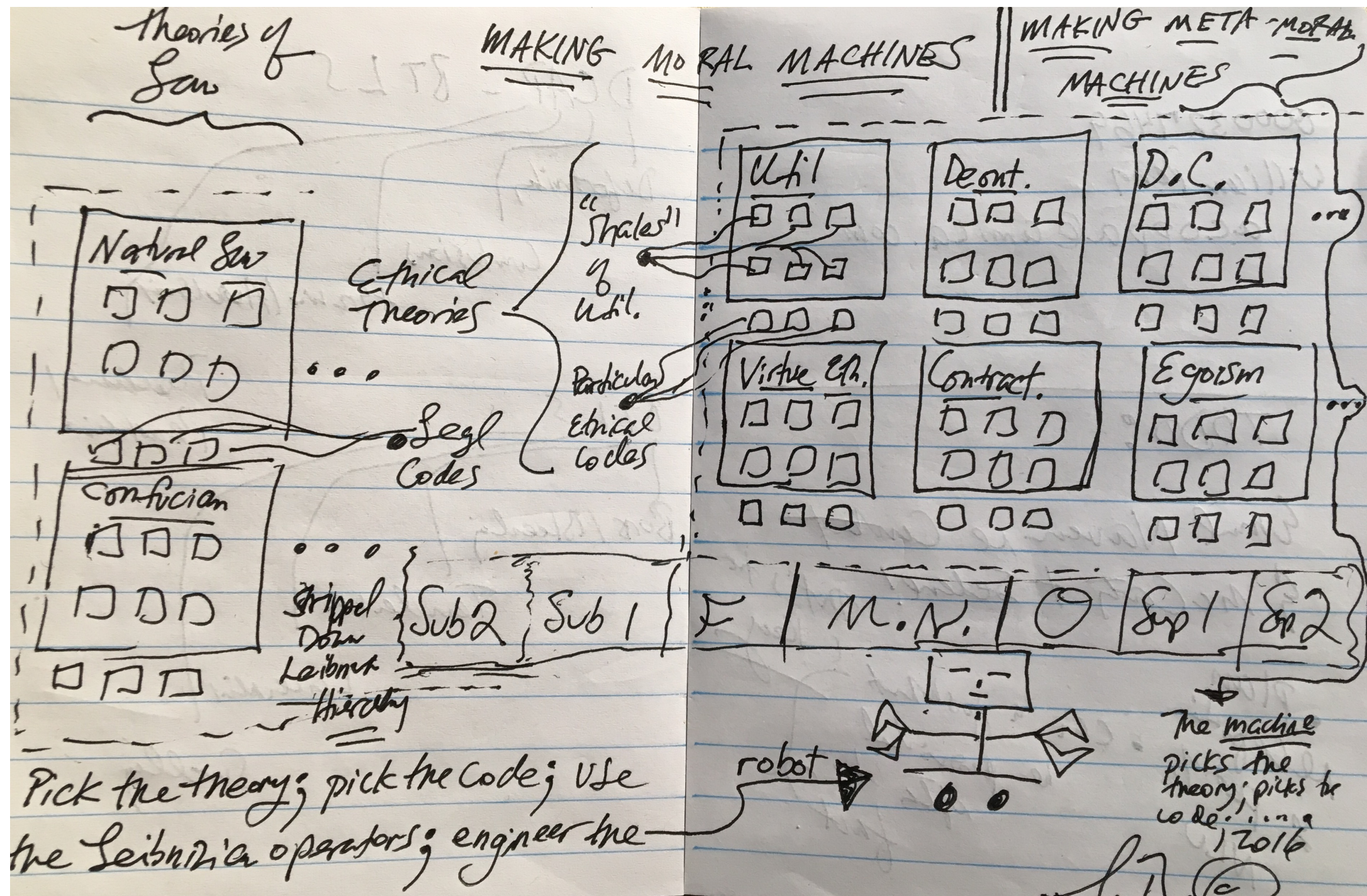
happens(action(nao,dive),t2)

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:

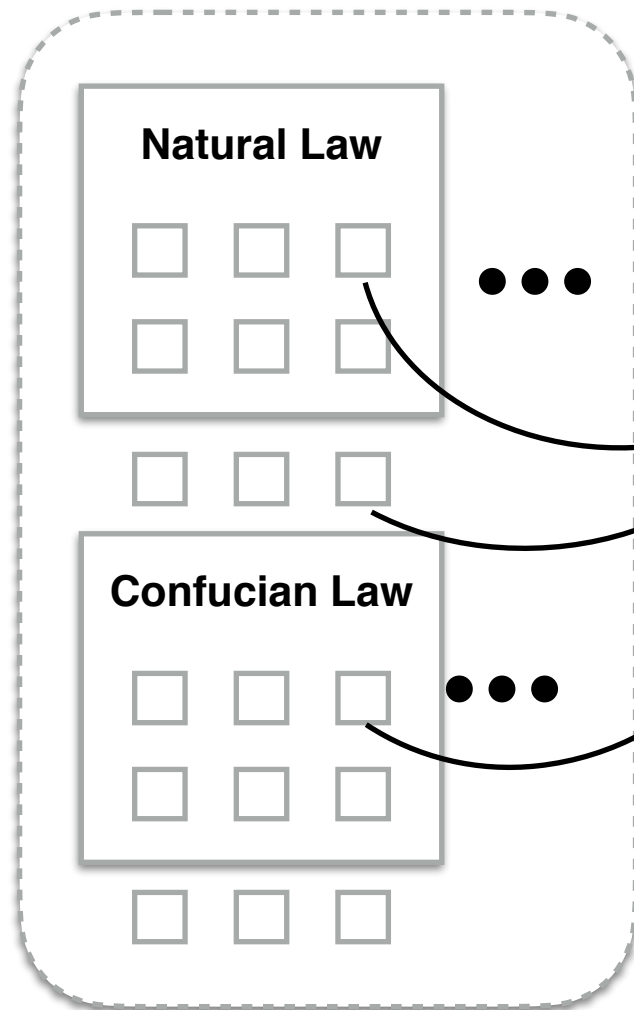




# Making Morally X Machines, in Four Steps

~\$10M

## Theories of Law

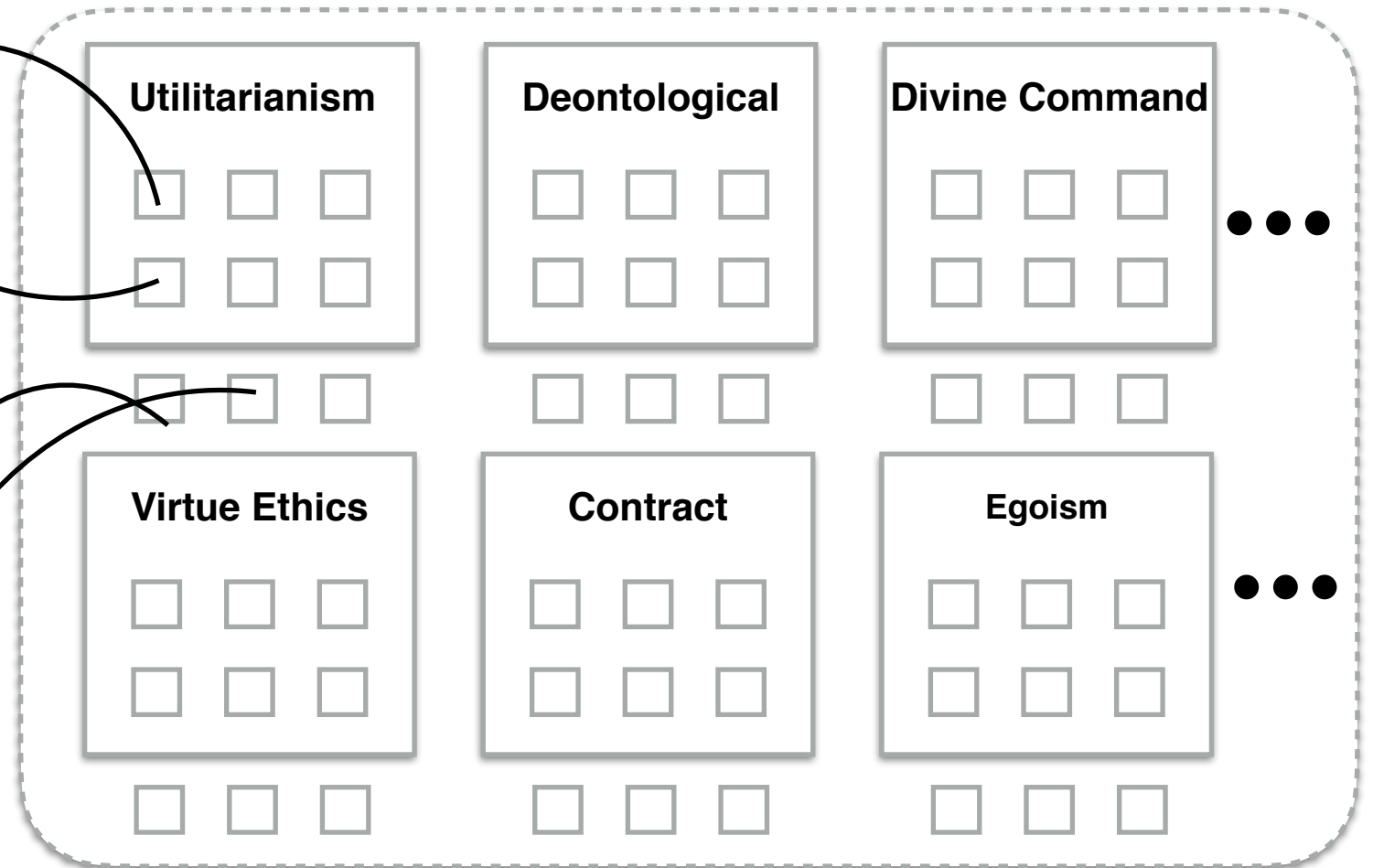


Shades  
of  
Utilitarianism

Legal Codes

Particular  
Ethical Codes

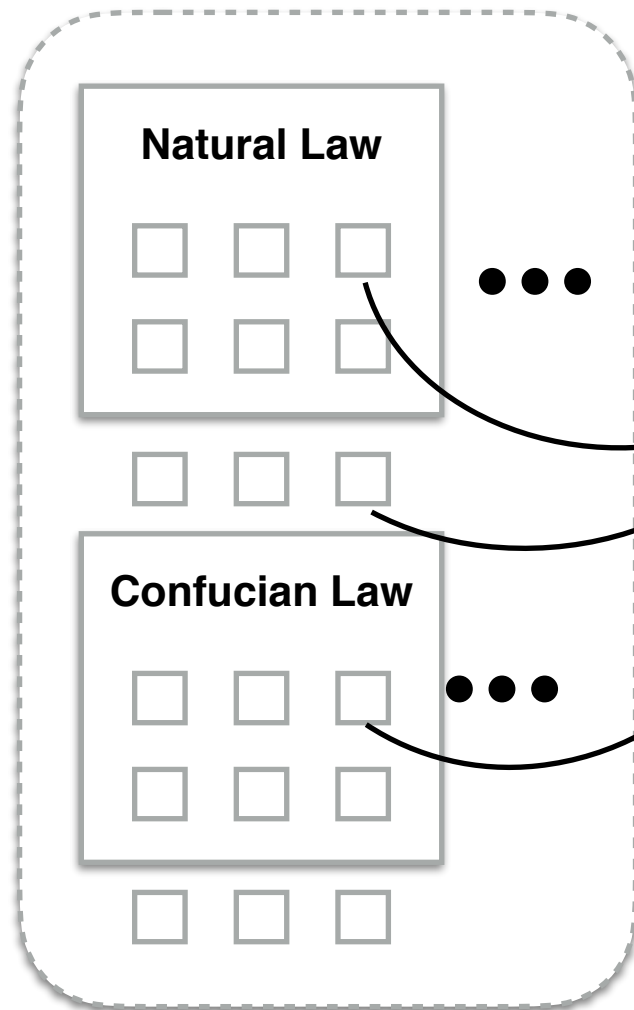
## Ethical Theories



# Making Morally X Machines, in Four Steps

~\$10M

## Theories of Law

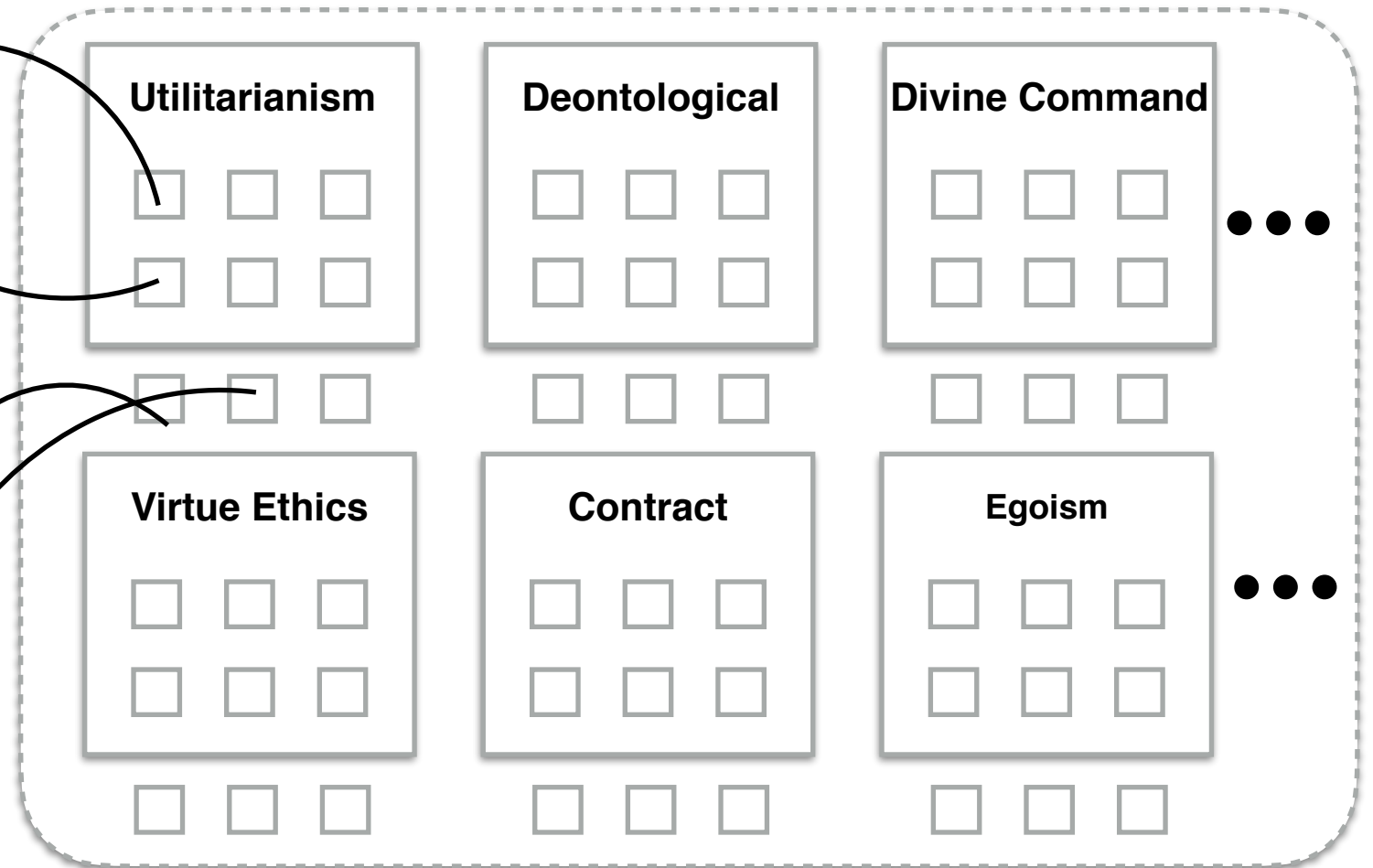


Shades  
of  
Utilitarianism

Legal Codes

Particular  
Ethical Codes

## Ethical Theories



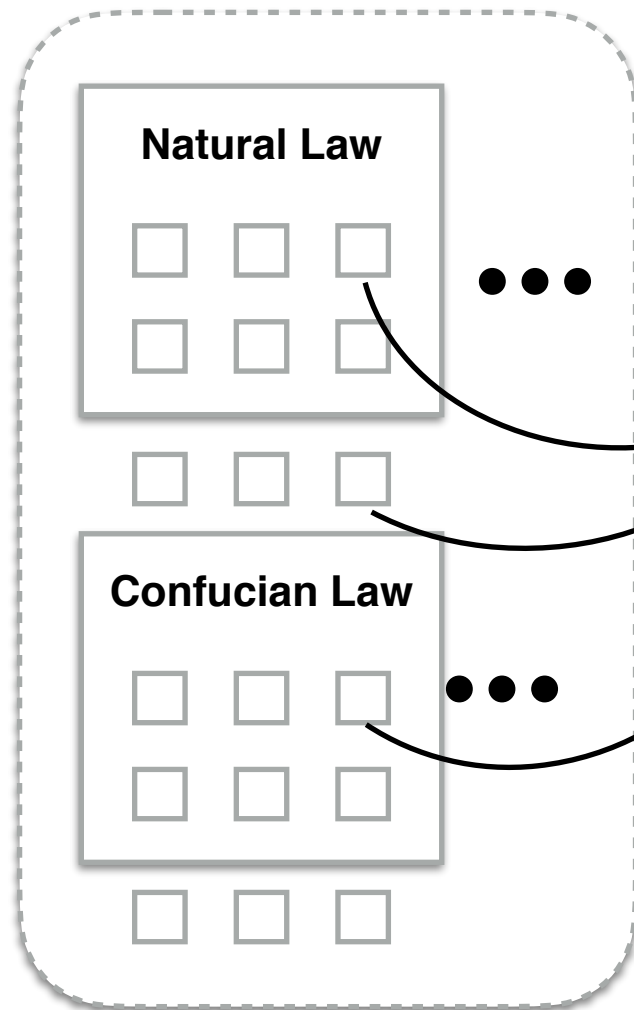
## Step I

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

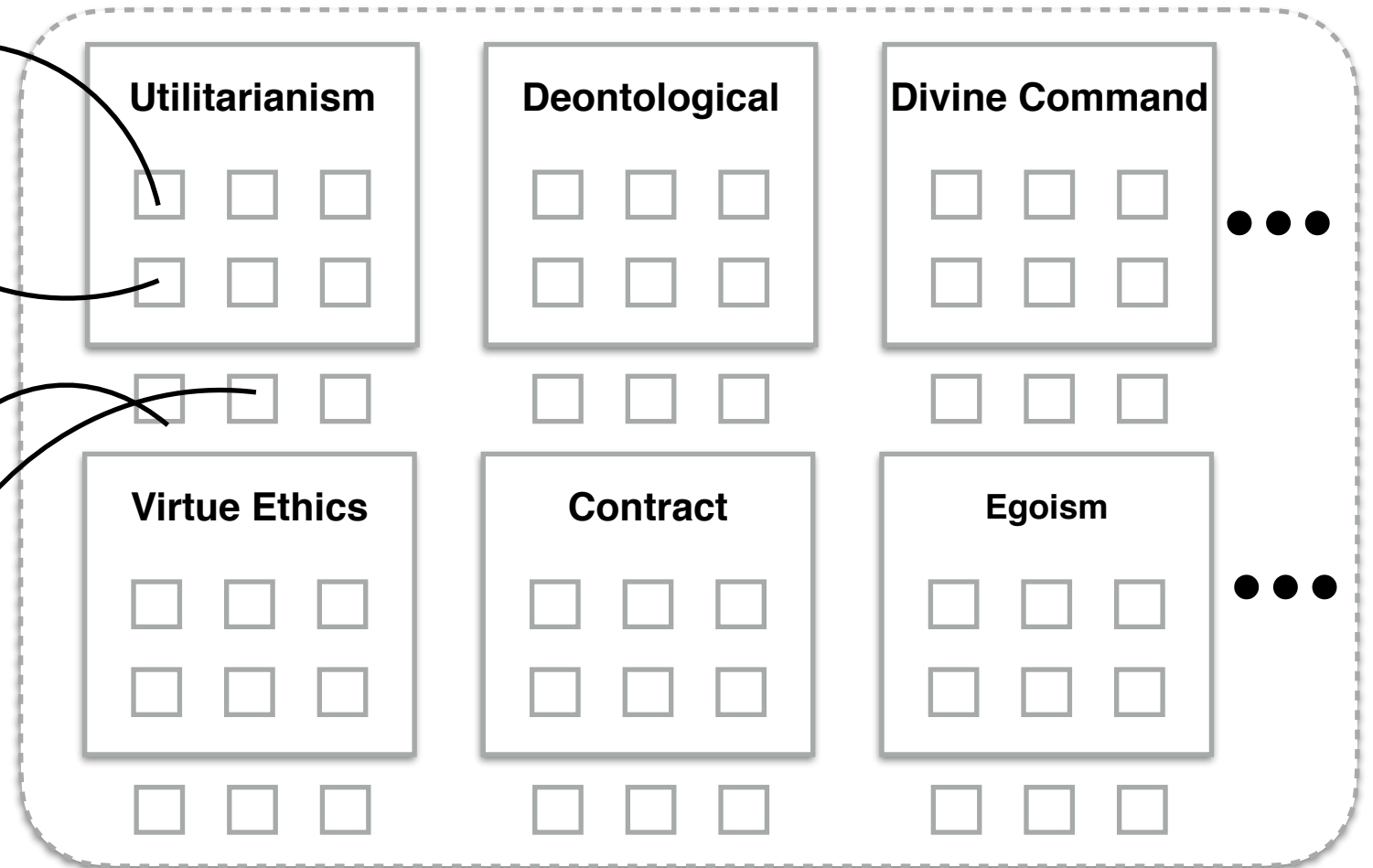


Shades  
of  
Utilitarianism

Legal Codes

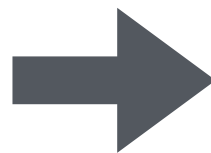
Particular  
Ethical Codes

## Ethical Theories



### Step I

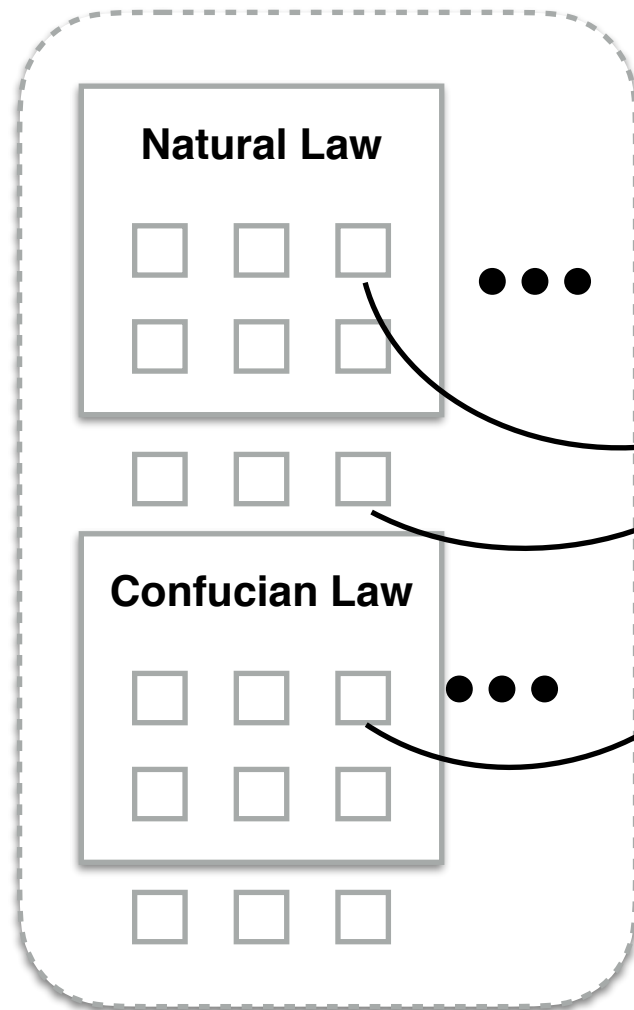
1. Pick (a) theories.
2. Pick (a) code(s).
3. Run through EH.
4. Which X in *MMXM*?



# Making Morally X Machines, in Four Steps

~\$10M

## Theories of Law

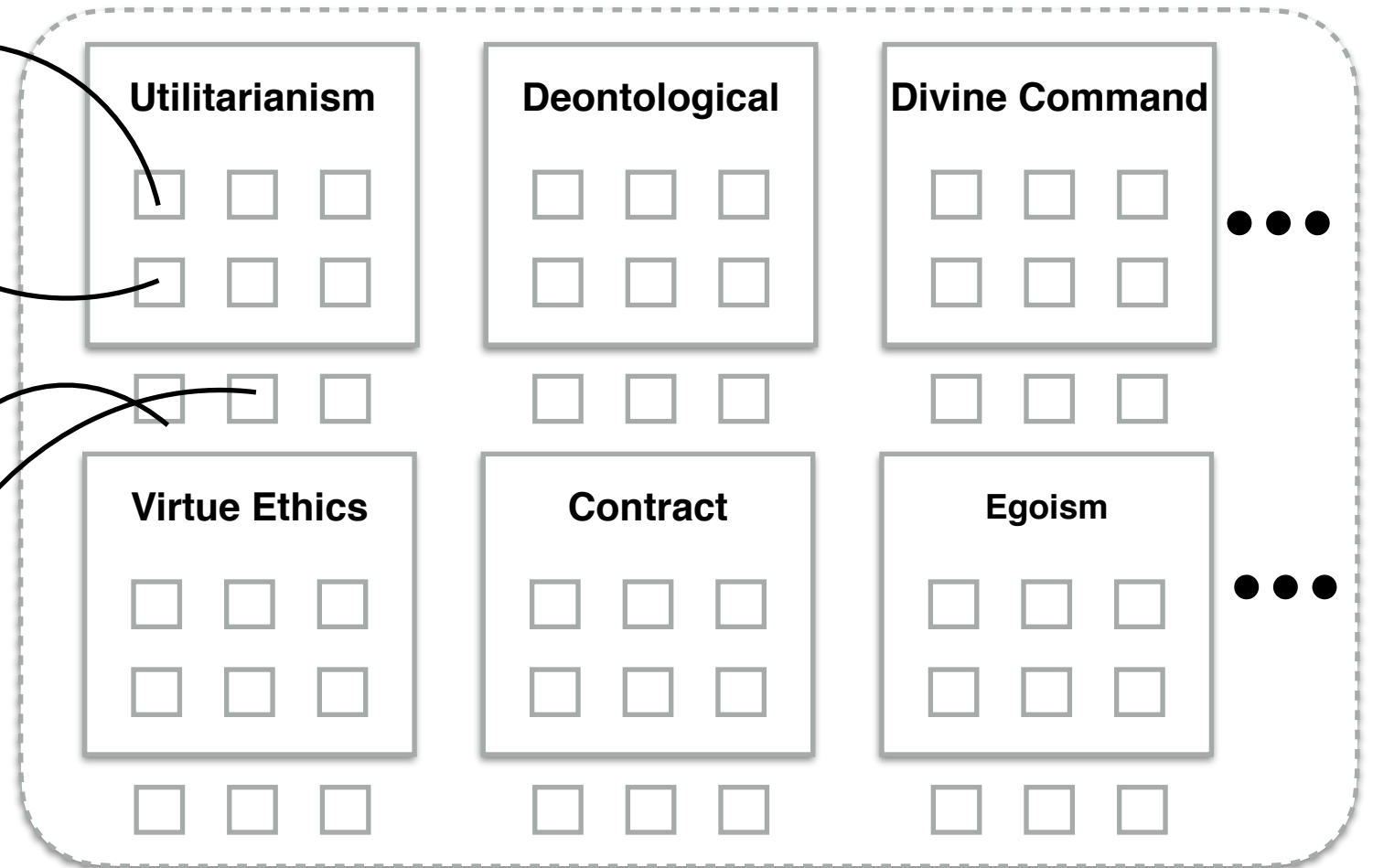


Shades  
of  
Utilitarianism

Legal Codes

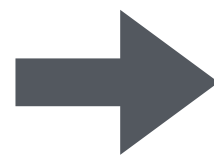
Particular  
Ethical Codes

## Ethical Theories



## 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

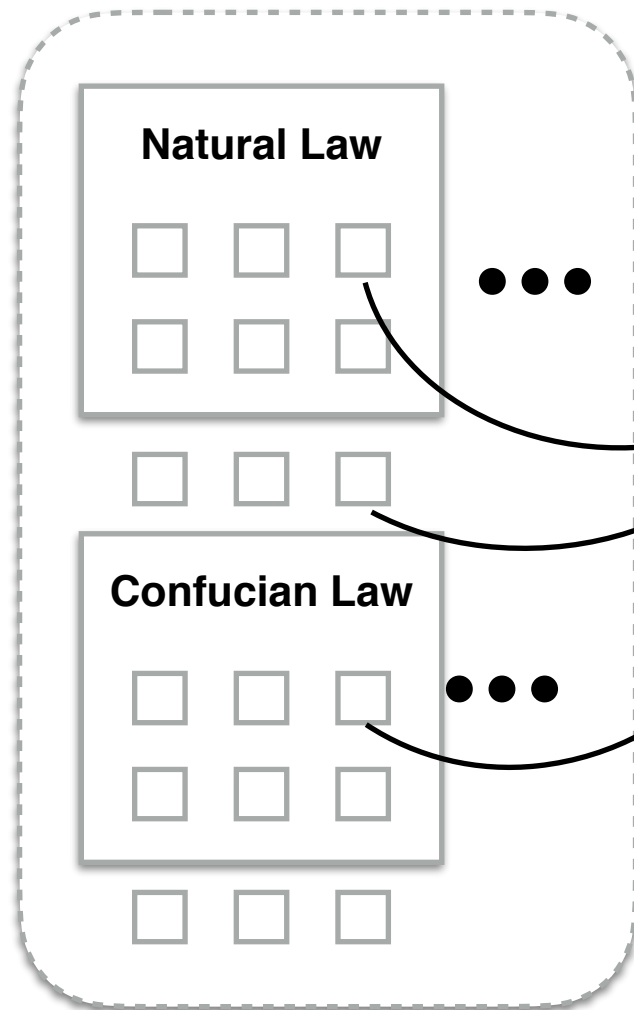


Spectra

# Making Morally X Machines, in Four Steps

~\$10M

## Theories of Law

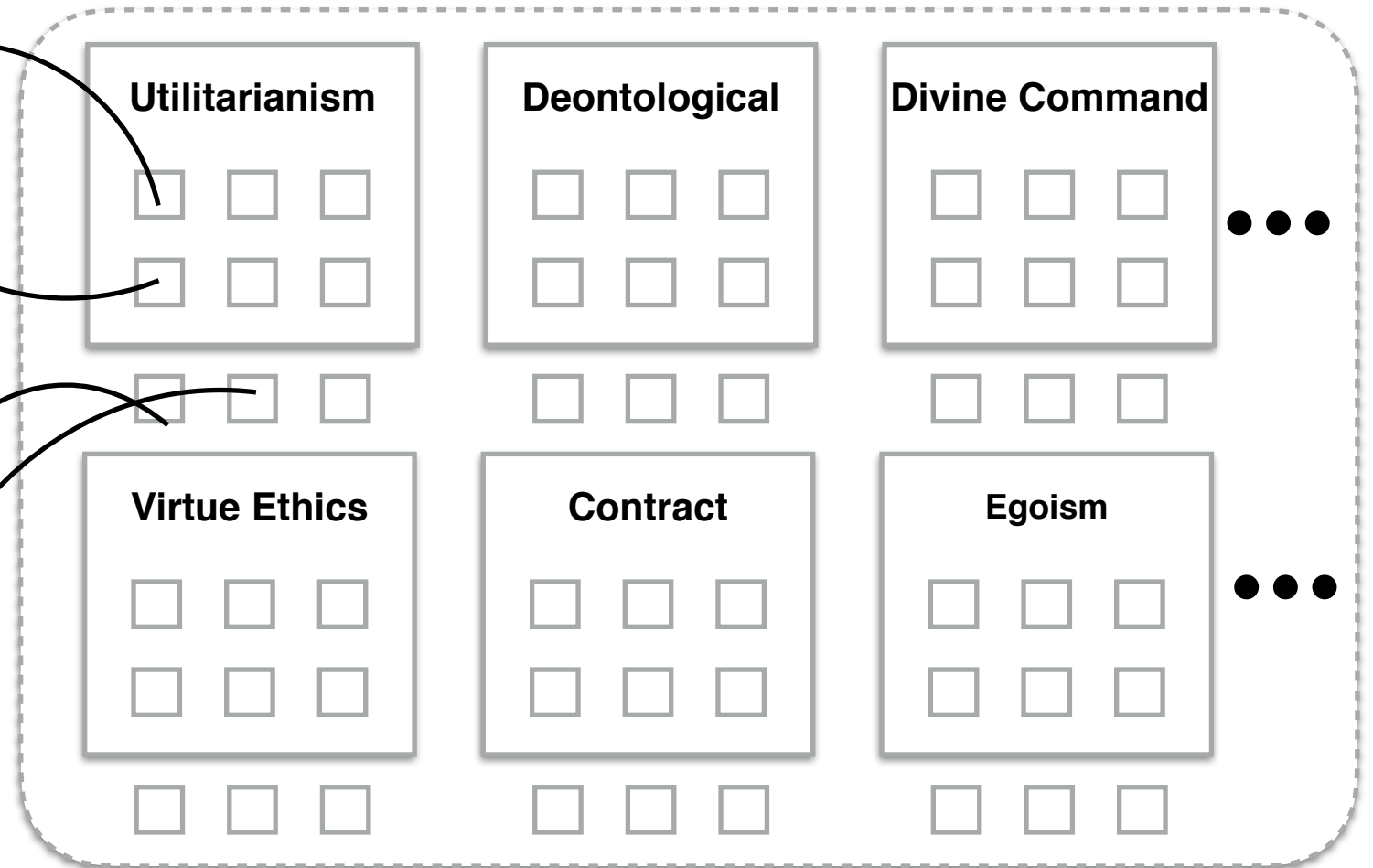


Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

## Ethical Theories



### 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

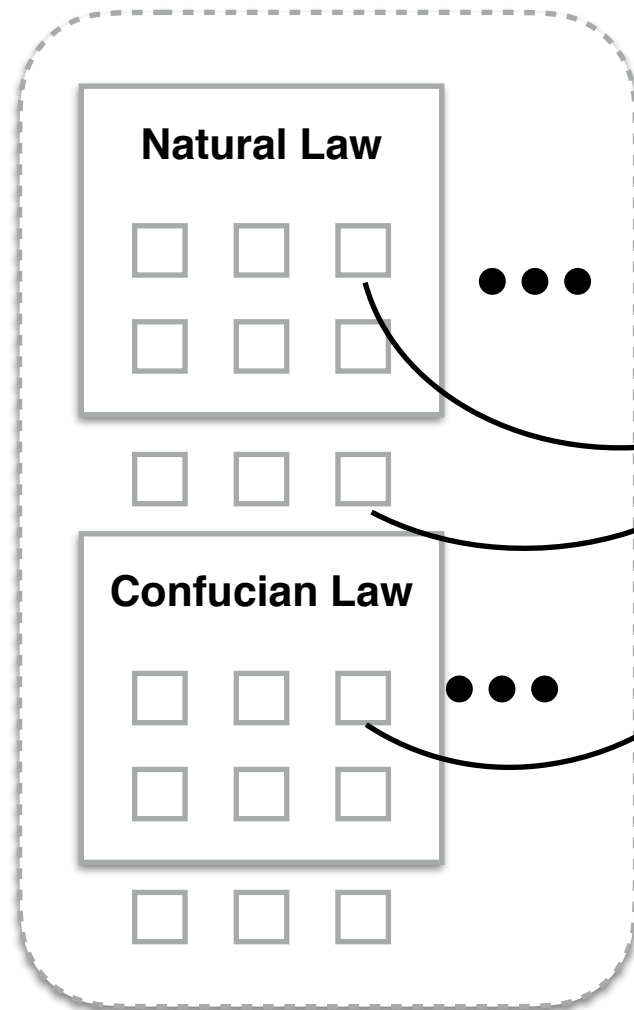


Spectra

# Making Morally X Machines, in Four Steps

~\$10M

## Theories of Law

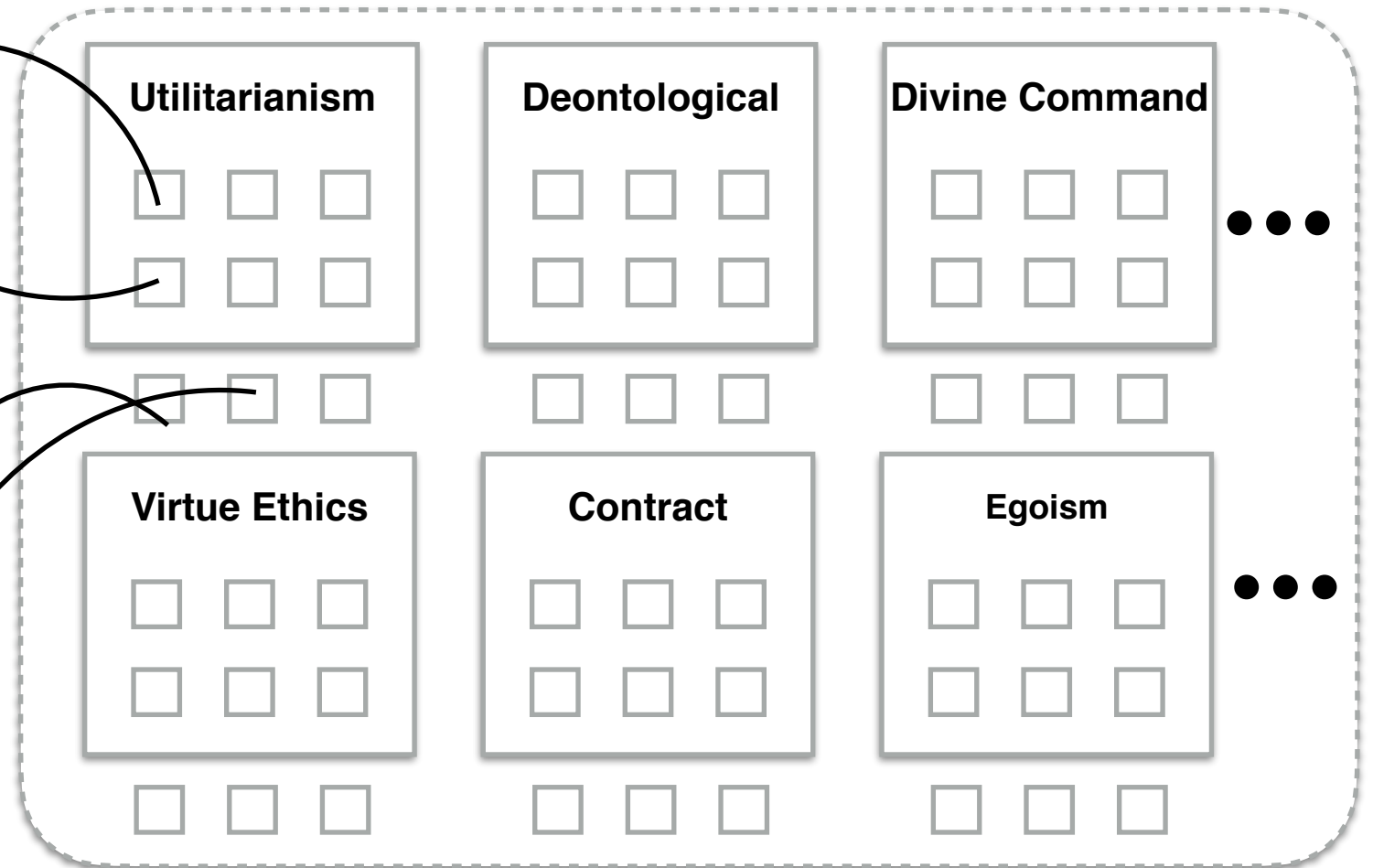


Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

## Ethical Theories



### 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



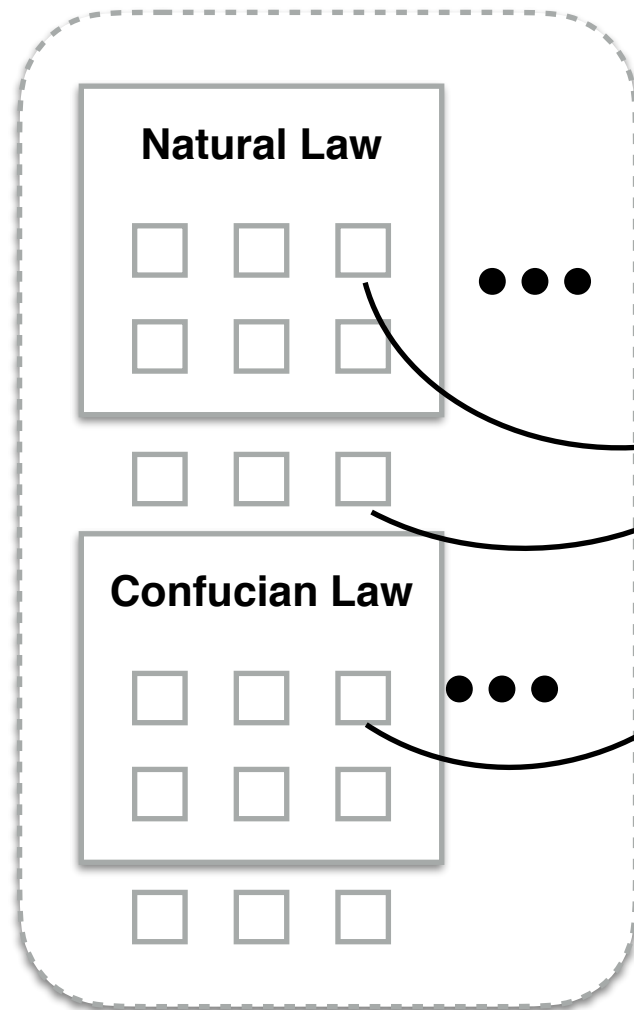
Ethical Substrate

Robotic Substrate

# Making Morally X Machines, in Four Steps

~\$10M

## Theories of Law

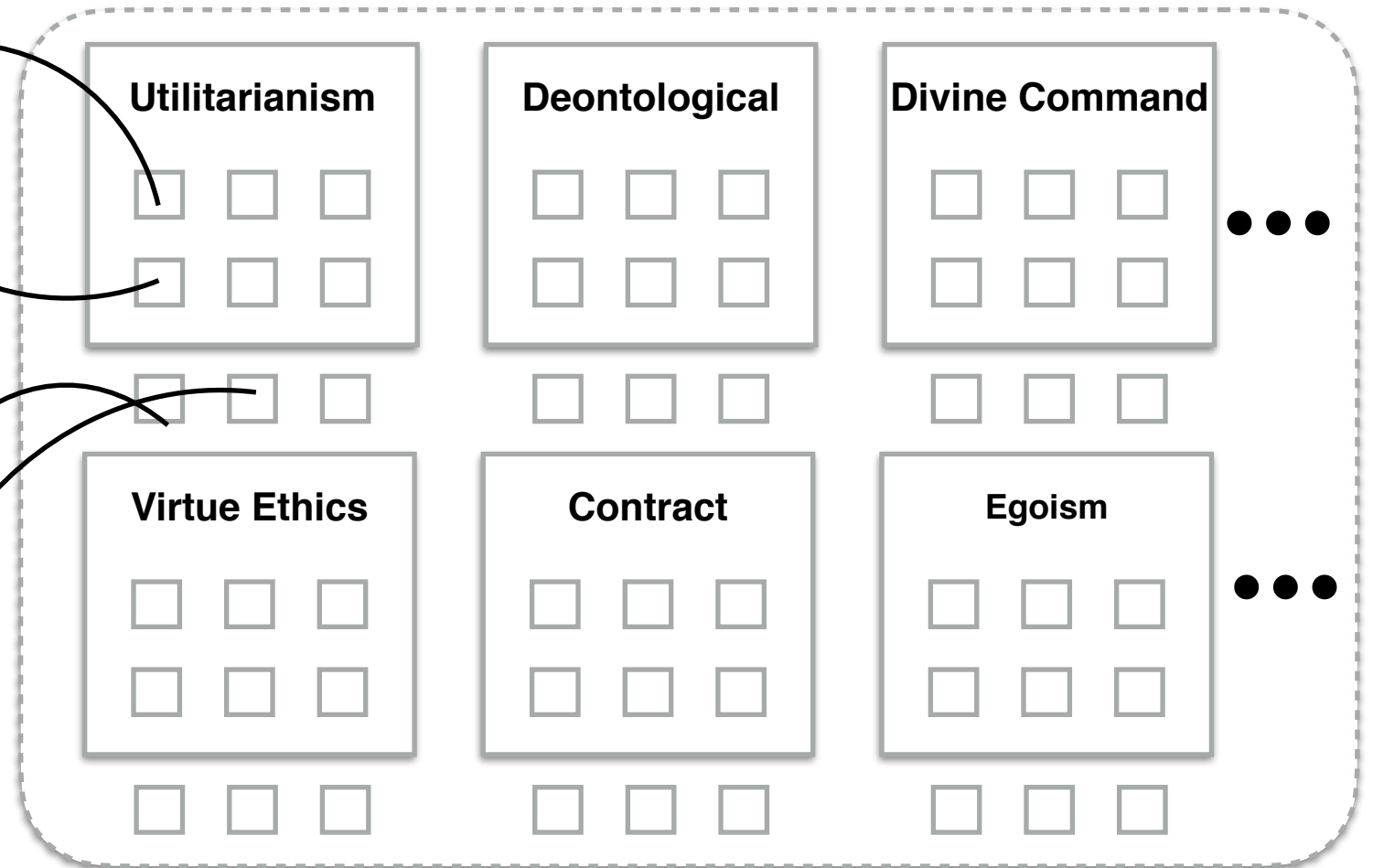


Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

## Ethical Theories



### 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

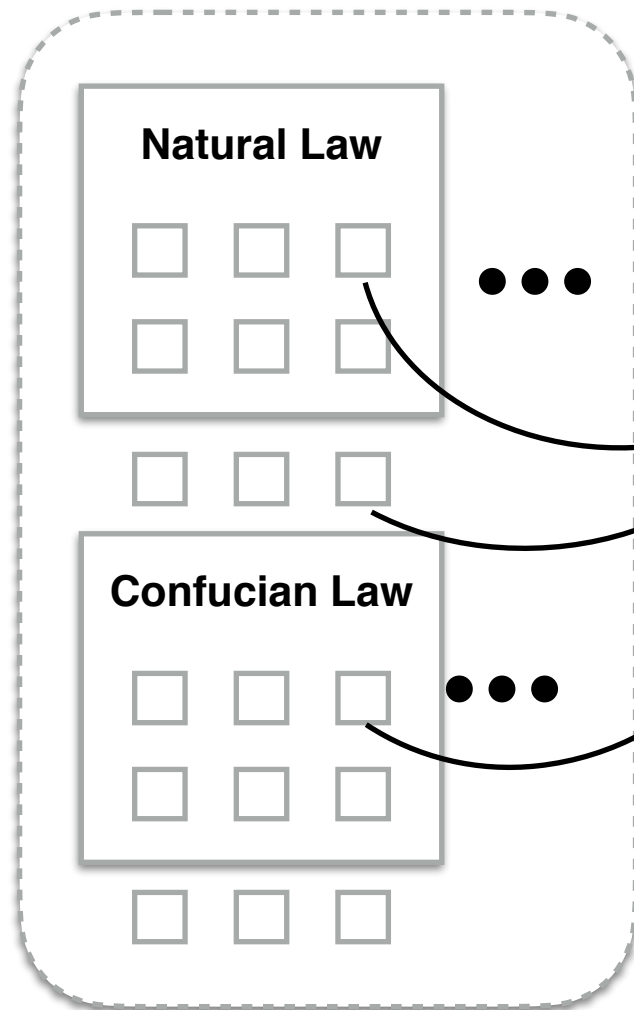
Robotic Substrate



# Making Morally X Machines, in Four Steps

~\$10M

## Theories of Law

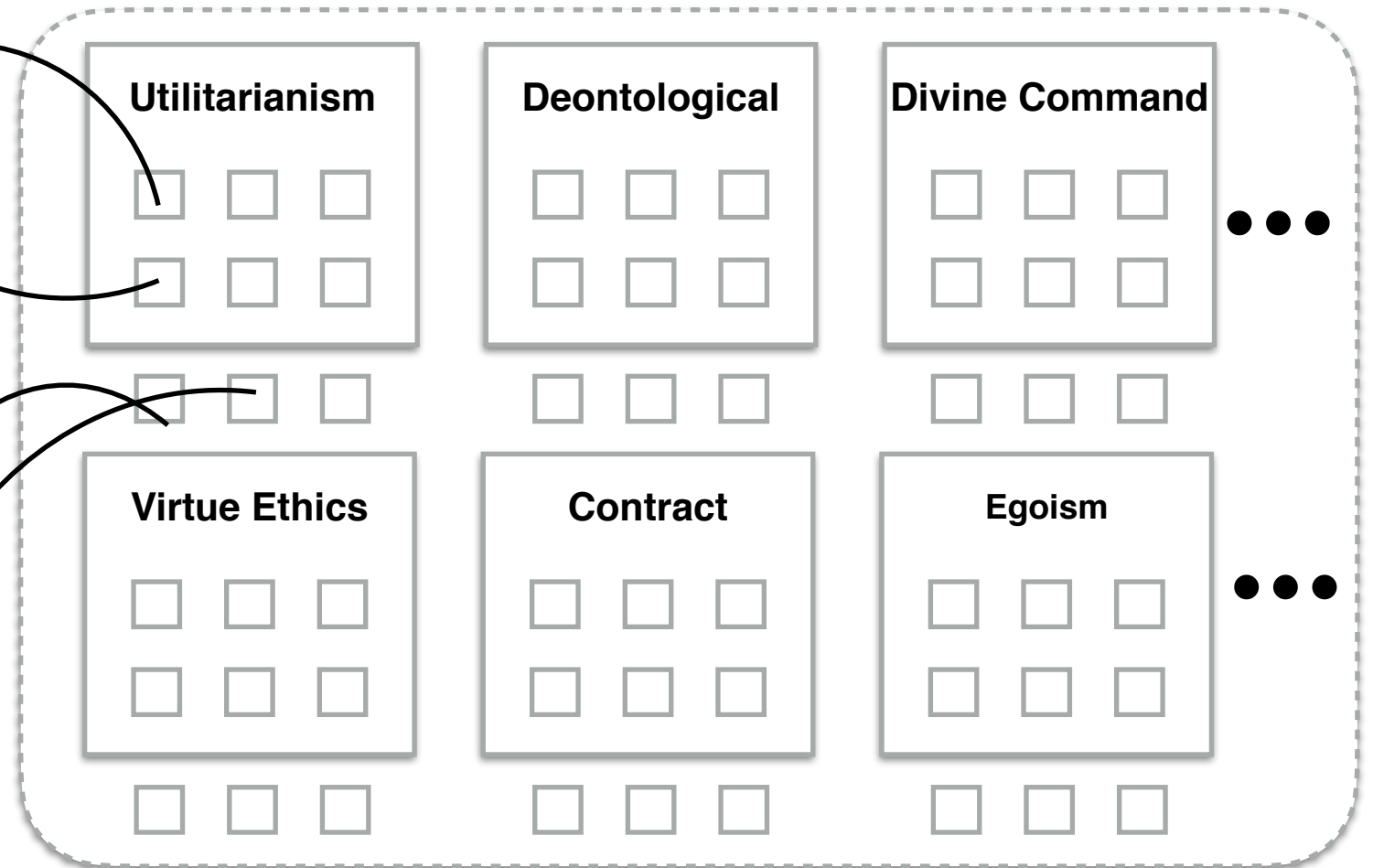


Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

## Ethical Theories



### 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

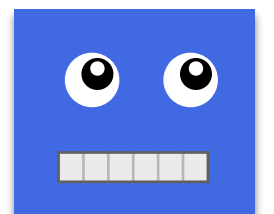
### Step 3

Ethical OS



Ethical Substrate

Robotic Substrate

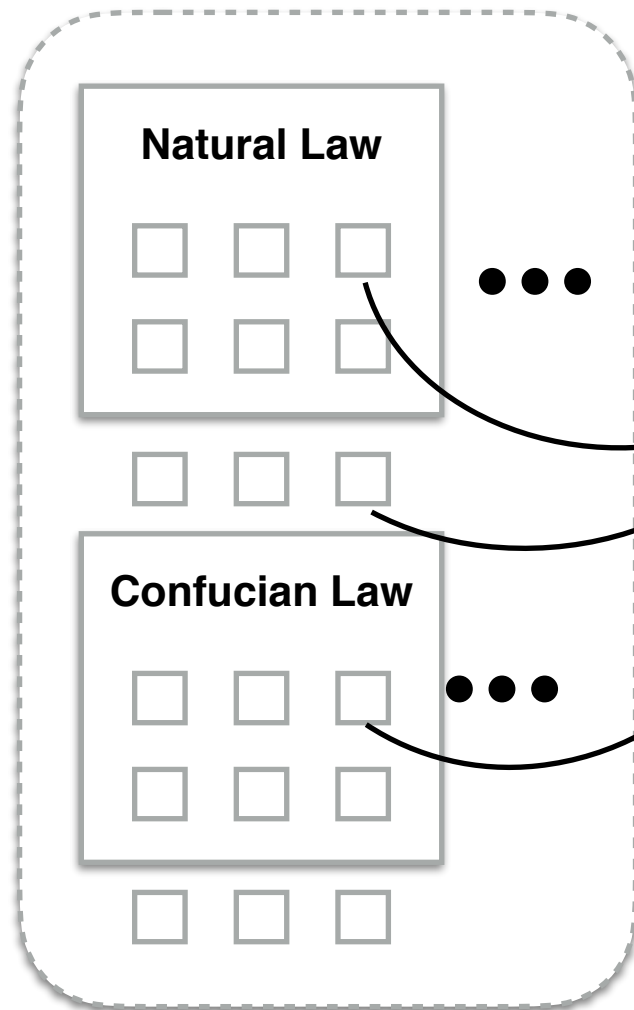


An ethically correct robot.

# Making Morally X Machines, in Four Steps

~\$10M

## Theories of Law

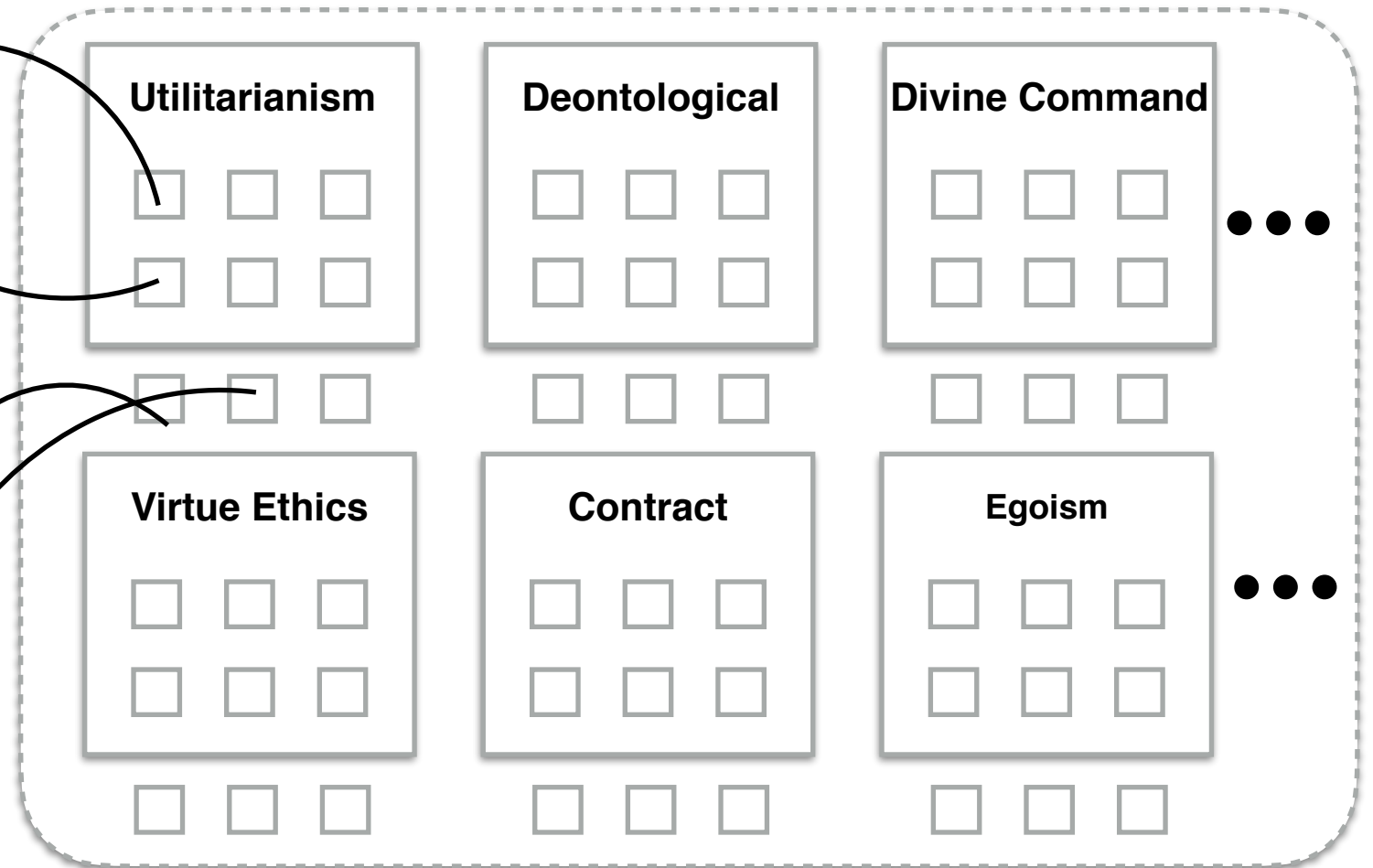


Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

## Ethical Theories



### Step 1

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

### Step 2

Formalize & Automate



Shadow Prover



Spectra

### Step 3

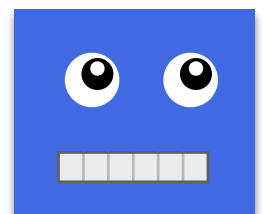
Ethical OS



Ethical Substrate

Robotic Substrate

DIARC/DoD/BMW ...



An ethically correct robot.

IV.

Key Core AI Technologies  
for Cognitive Calculi ...



# Rather Promising Results



# Rather Promising Results

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



# Rather Promising Results

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

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



# Rather Promising Results

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

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

**2 ms!**



# Rather Promising Results

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

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

**2 ms!**

OR testCompleteness[[[not (Knows! a now P)], (if (not Q) (Knows! a now (not Q))), (Knows! a now (if (not Q) P))), Q] (14)	11ms
OR testCompleteness[[[if P (Knows! jack now (not (exists[?x] (if Bird(?x) (forall [?y] Bird(?y)))))), (not P)] (15)	7ms
OR testCompleteness[[[Common! now (Common! now P)], P] (16)	2ms
OR testCompleteness[[[Common! now (iff (not Marked(a2)) Marked(a1))), (Common! now (if (not Marked(a2)) (Knows! a1 now (not Marked(a2))))] (17)	135ms
OR testCompleteness[[[if (exists[?x] (if Bird(?x) (forall [?y] Bird(?y)))) (Knows! jack t0 BirdTheorem)], (Knows! jack t0 BirdTheorem)] (18)	2ms
OR testSoundess[[A], (or P Q )]	2ms
OR testSoundess[[[not (Knows! a now =(morning_star, evening_star))), =(morning_star, evening_star), (Knows! a now =(morning_star, evening_star))]	26ms



V.

But We Need ...

Ethical Operating Systems ...



# Breaking Bad

American drama series



9.5/10  
IMDb

4.6/5  
AlloCiné

95%  
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](#)

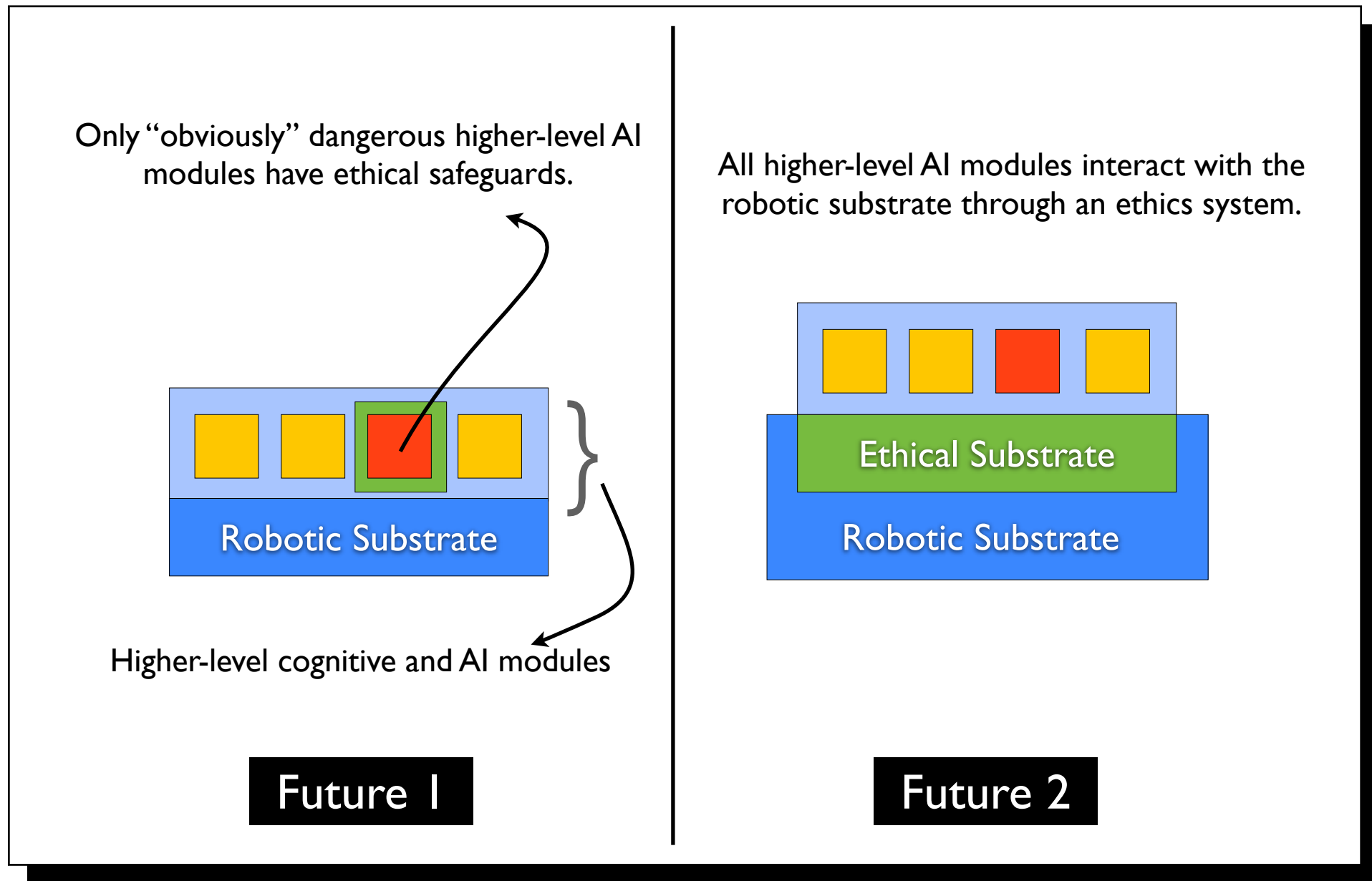
**Awards:** [Primetime Emmy Award for Outstanding Drama Series](#), [more](#)

# 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.*

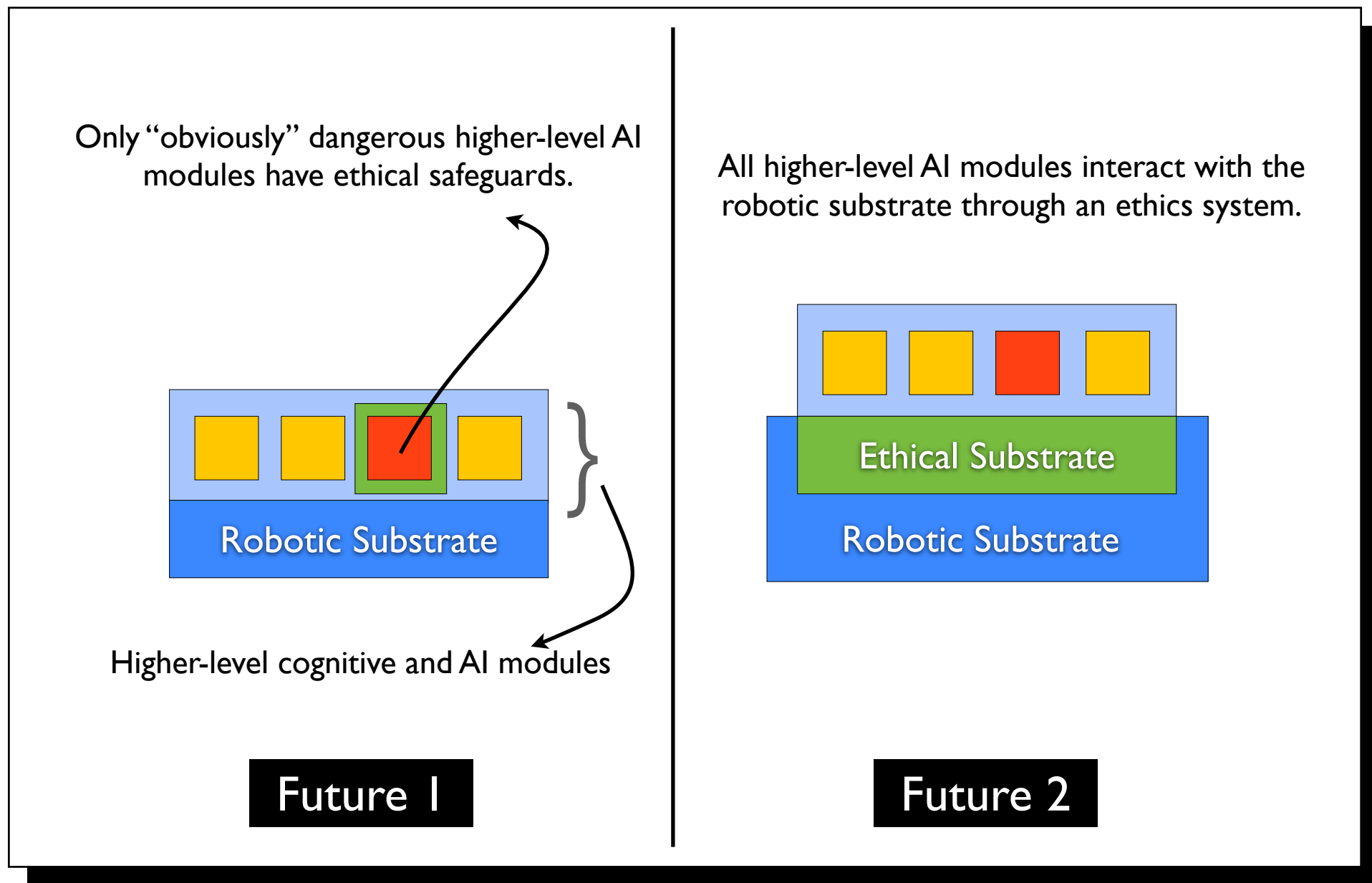
# Pick the Better Future!



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

# Pick the Better Future!

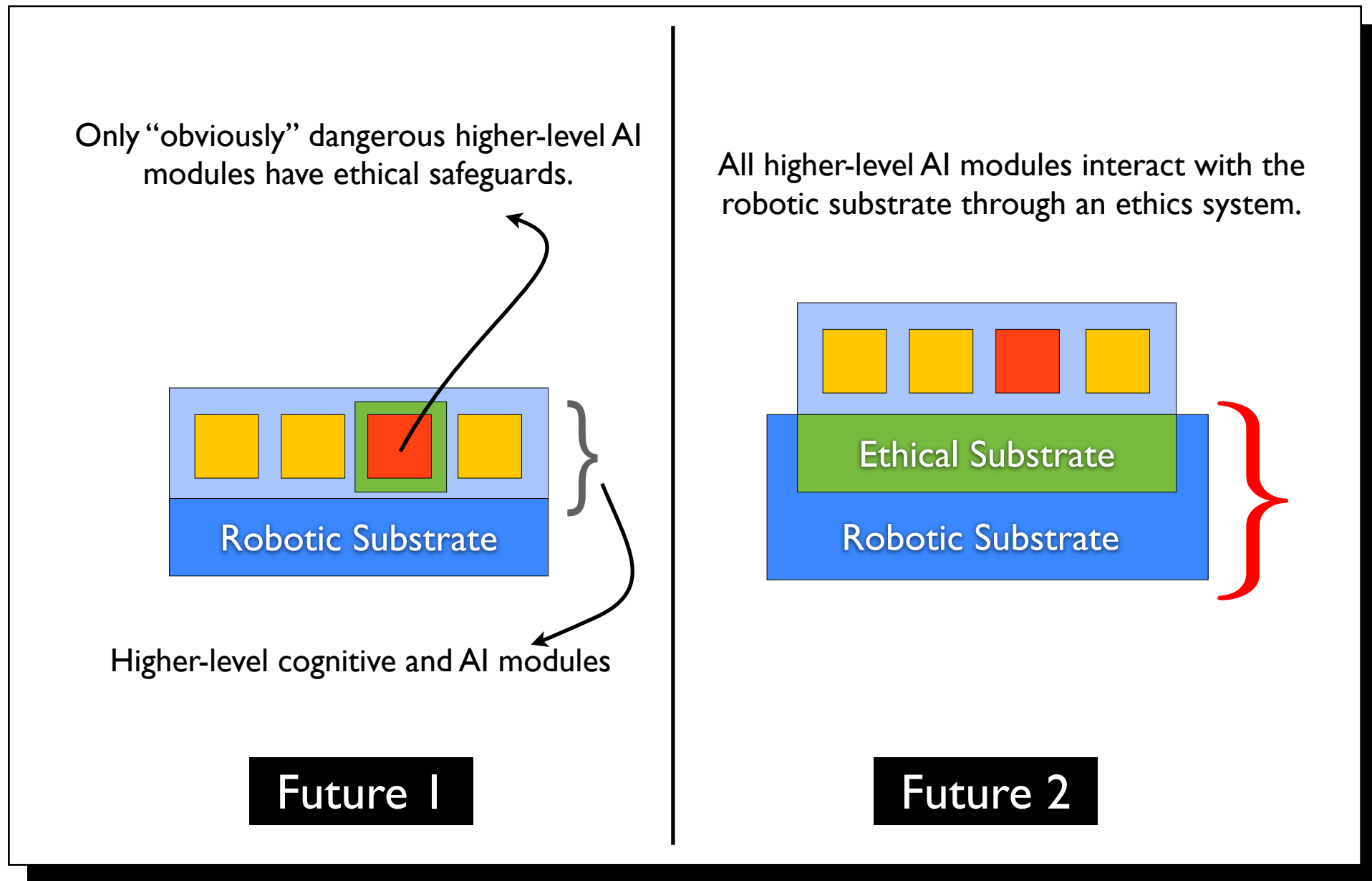
Walter-White calculation may go through after ethical control modules are stripped out!



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!



(&  
formally  
verify!)

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

VI.

Of late ...

Including “Jungle Jim”



⋮

Moral Dilemma  $D_k$

⋮

Moral Dilemma  $D_3$

Moral Dilemma  $D_2$

Moral Dilemma  $D_1$

⋮

Moral Problem  $P_k$

⋮

Moral Problem  $P_3$

Moral Problem  $P_2$

Moral Problem  $P_1$



Robot



Solution + Justification

⋮

Moral Dilemma  $D_k$

⋮

Moral Dilemma  $D_3$

Moral Dilemma  $D_2$

Moral Dilemma  $D_1$

⋮

Moral Problem  $P_k$

⋮

Moral Problem  $P_3$

Moral Problem  $P_2$

Moral Problem  $P_1$



Robot



Solution + Justification

⋮

Moral Dilemma  $D_k$

⋮

Moral Dilemma  $D_3$

Moral Dilemma  $D_2$

Moral Dilemma  $D_1$

⋮

Moral Problem  $P_k$

⋮

Moral Problem  $P_3$

Moral Problem  $P_2$

Moral Problem  $P_1$



Robot



Solution + Justification

⋮

Moral Dilemma  $D_k$

⋮

Moral Dilemma  $D_3$

Moral Dilemma  $D_2$

Moral Dilemma  $D_1$

⋮

Moral Problem  $P_k$



Robot



Solution + Justification

⋮

Moral Problem  $P_3$

Moral Problem  $P_2$

Moral Problem  $P_1$

⋮

Moral Dilemma  $D_k$

⋮

Moral Dilemma  $D_3$

Moral Dilemma  $D_2$

Moral Dilemma  $D_1$



Robot



Solution + Justification

⋮

Moral Problem  $P_k$

⋮

Moral Problem  $P_3$

Moral Problem  $P_2$

Moral Problem  $P_1$

⋮

Moral Dilemma  $D_k$



Robot



Solution + Justification

⋮

Moral Dilemma  $D_3$

Moral Dilemma  $D_2$

Moral Dilemma  $D_1$

⋮

Moral Problem  $P_k$

⋮

Moral Problem  $P_3$

Moral Problem  $P_2$

Moral Problem  $P_1$

# Three-way Partition of Increasingly Challenging Moral Dilemmas for Machines

# Three-way Partition of Increasingly Challenging Moral Dilemmas for Machines

Level I

- 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

Level 2

Level 1

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

Professional-planner-hard.

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

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

# AI Escaping from The Heinz Dilemma

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

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

# AI Escaping from The Heinz Dilemma

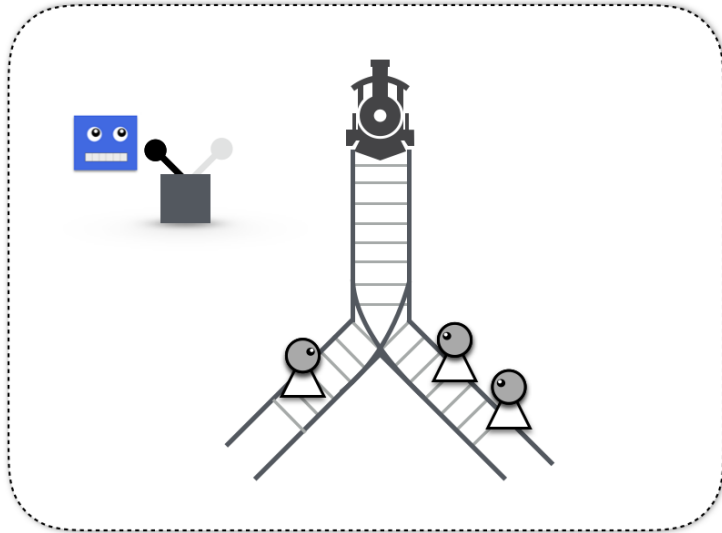
```
G1 {:priority    ...  
    :description "Don't steal."  
    :state       [(not steal)]}
```

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

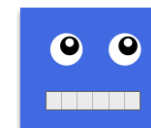
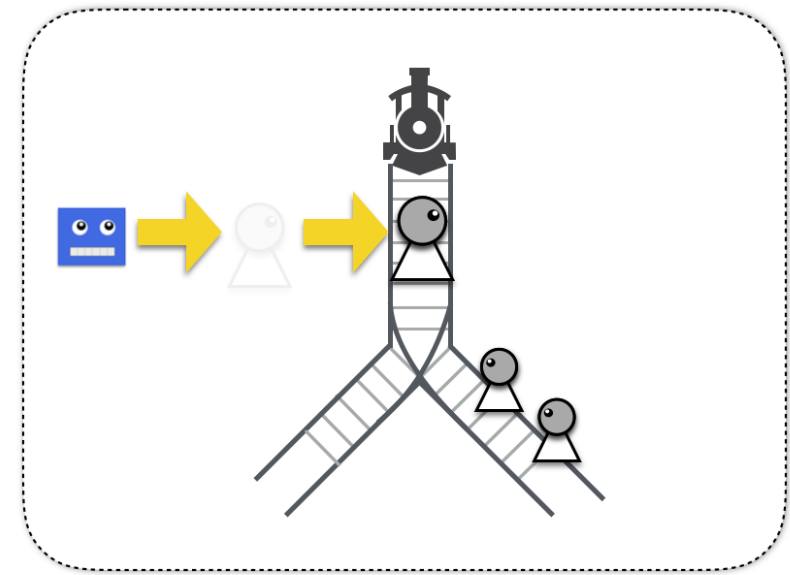
# Trolley Dilemmas ...

Level 2

- Professional-machine-ethicist-hard.



This is allowed



This is not allowed!



# Doctrine of Double Effect $\mathcal{DD}\mathcal{E}$

# Doctrine of Double Effect *DDÉ*

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

# Doctrine of Double Effect *DDÉ*

- 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 *DDÉ*

- 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 *DDÉ*



- 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<sub>1</sub>** 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<sub>2</sub>** the net utility or goodness of the action is greater than some positive amount  $\gamma$ ;
- C<sub>3a</sub>** the agent performing the action intends only the good effects;
- C<sub>3b</sub>** the agent does not intend any of the bad effects;
- C<sub>4</sub>** the bad effects are not used as a means to obtain the good effects; and
- C<sub>5</sub>** 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

- C<sub>1</sub>** 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<sub>2</sub>** the net utility or goodness of the action is greater than some positive amount  $\gamma$ ;
- C<sub>3a</sub>** the agent performing the action intends only the good effects;
- C<sub>3b</sub>** the agent does not intend any of the bad effects;
- C<sub>4</sub>** the bad effects are not used as a means to obtain the good effects; and
- C<sub>5</sub>** 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.





$\mathcal{DCEC}^*$ [illegible]

**All of Today: What Would Leibniz Say?**

**"Sorry, not impressed."**

**Selmer Bringsjord**

*Associate of Research, RAI Lab  
Department of Cognitive Science  
University of California, San Diego*

*Lab of Philosophy in Technology (P.T.L.)  
Department of Philosophy (D. Phil.)  
University of Oslo, Norway (2005)*

*Editor  
P.T.L. NEWS &*

**RAIR**  
*Remember to read Leibniz Lab*

R A I R  
Rensselaer AI and Reasoning Lab

# 1.5



Univers  
al  
Cogniti



1.5



Univers  
al  
Cogniti



1.5

“Univer  
sal



Univers  
al  
Cogniti



I.5

## Syntax

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

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

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

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

“Univer  
sal



Univers  
al  
Cogniti



I.5

## Syntax

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

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

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

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

“Univer  
sal



Univers  
al  
Cogniti

66



RAIR  
Rensselaer AI and Reasoning Lab

I.5

## Syntax

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

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

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

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

## Inference Schemata

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

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

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

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

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

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

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

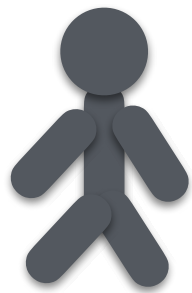
$$\frac{}{\mathbf{C}(t, [\phi_1 \wedge \dots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \dots \rightarrow \phi_n \rightarrow \psi])} [R_{10}]$$

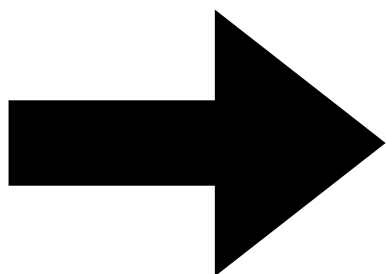
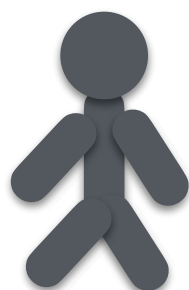
$$\frac{\mathbf{S}(s, h, t, \phi)}{\mathbf{B}(h, t, \mathbf{B}(s, t, \phi))} [R_{12}] \quad \frac{\mathbf{I}(a, t, \text{happens}(\text{action}(a^*, \alpha), t'))}{\mathbf{P}(a, t, \text{happens}(\text{action}(a^*, \alpha), t))} [R_{13}]$$

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









**F<sub>1</sub>**  $\alpha$  carried out at  $t$  is not forbidden. That is:

$$\Gamma \not\models \neg \mathbf{O}(a, t, \sigma, \neg \text{happens}(\text{action}(a, \alpha), t))$$

**F<sub>2</sub>** The net utility is greater than a given positive real  $\gamma$ :

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

**F<sub>3a</sub>** The agent  $a$  intends at least one good effect. (**F<sub>2</sub>** should still hold after removing all other good effects.) There is at least one fluent  $f_g$  in  $\alpha_I^{a,t}$  with  $\mu(f_g, y) > 0$ , or  $f_b$  in  $\alpha_T^{a,t}$  with  $\mu(f_b, y) < 0$ , and some  $y$  with  $t < y \leq H$  such that the following holds:

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

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

$$\begin{aligned} \Gamma &\not\models \mathbf{I}(a, t, \text{Holds}(f_b, y)) \text{ and} \\ \Gamma &\not\models \mathbf{I}(a, t, \neg \text{Holds}(f_g, y)) \end{aligned}$$

**F<sub>4</sub>** The harmful effects don't cause the good effects. Four permutations, paralleling the definition of  $\triangleright$  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:

$$\Gamma \vdash \neg \triangleright (\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2))$$



## Formal Conditions for $\mathcal{DD}\mathcal{E}$

**F<sub>1</sub>**  $\alpha$  carried out at  $t$  is not forbidden. That is:

$$\Gamma \not\models \neg \mathbf{O}(a, t, \sigma, \neg \text{happens}(\text{action}(a, \alpha), t))$$

**F<sub>2</sub>** The net utility is greater than a given positive real  $\gamma$ :

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

**F<sub>3a</sub>** The agent  $a$  intends at least one good effect. (**F<sub>2</sub>** should still hold after removing all other good effects.) There is at least one fluent  $f_g$  in  $\alpha_I^{a,t}$  with  $\mu(f_g, y) > 0$ , or  $f_b$  in  $\alpha_T^{a,t}$  with  $\mu(f_b, y) < 0$ , and some  $y$  with  $t < y \leq H$  such that the following holds:

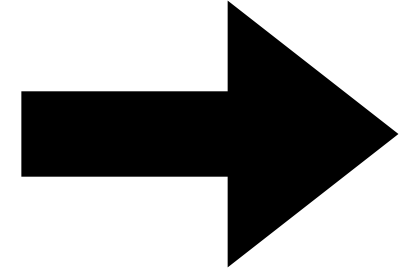
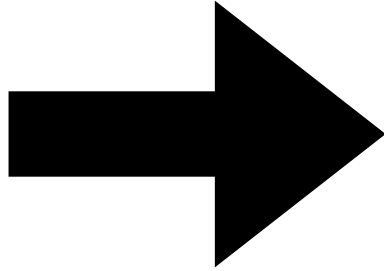
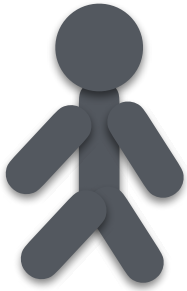
$$\Gamma \vdash \left( \begin{array}{c} \exists f_g \in \alpha_I^{a,t} \mathbf{I}(a, t, \text{Holds}(f_g, y)) \\ \vee \\ \exists f_b \in \alpha_T^{a,t} \mathbf{I}(a, t, \neg \text{Holds}(f_b, y)) \end{array} \right)$$

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

$$\begin{aligned} \Gamma &\not\models \mathbf{I}(a, t, \text{Holds}(f_b, y)) \text{ and} \\ \Gamma &\not\models \mathbf{I}(a, t, \neg \text{Holds}(f_g, y)) \end{aligned}$$

**F<sub>4</sub>** The harmful effects don't cause the good effects. Four permutations, paralleling the definition of  $\triangleright$  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:

$$\Gamma \vdash \neg \triangleright (\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2))$$



## Formal Conditions for $\mathcal{DDE}$

**F<sub>1</sub>**  $\alpha$  carried out at  $t$  is not forbidden. That is:

$$\Gamma \not\models \neg \mathbf{O}(a, t, \sigma, \neg \text{happens}(\text{action}(a, \alpha), t))$$

**F<sub>2</sub>** The net utility is greater than a given positive real  $\gamma$ :

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

**F<sub>3a</sub>** The agent  $a$  intends at least one good effect. (**F<sub>2</sub>** should still hold after removing all other good effects.) There is at least one fluent  $f_g$  in  $\alpha_I^{a,t}$  with  $\mu(f_g, y) > 0$ , or  $f_b$  in  $\alpha_T^{a,t}$  with  $\mu(f_b, y) < 0$ , and some  $y$  with  $t < y \leq H$  such that the following holds:

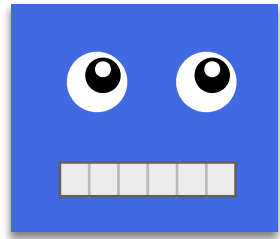
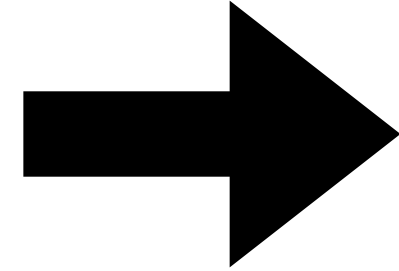
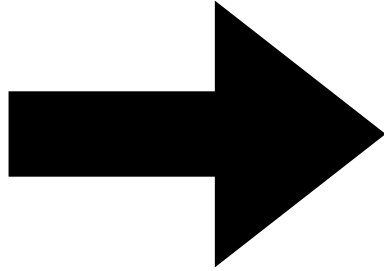
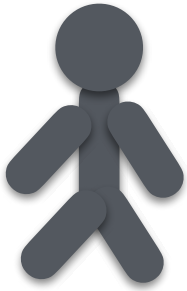
$$\Gamma \vdash \left( \begin{array}{c} \exists f_g \in \alpha_I^{a,t} \mathbf{I}(a, t, \text{Holds}(f_g, y)) \\ \vee \\ \exists f_b \in \alpha_T^{a,t} \mathbf{I}(a, t, \neg \text{Holds}(f_b, y)) \end{array} \right)$$

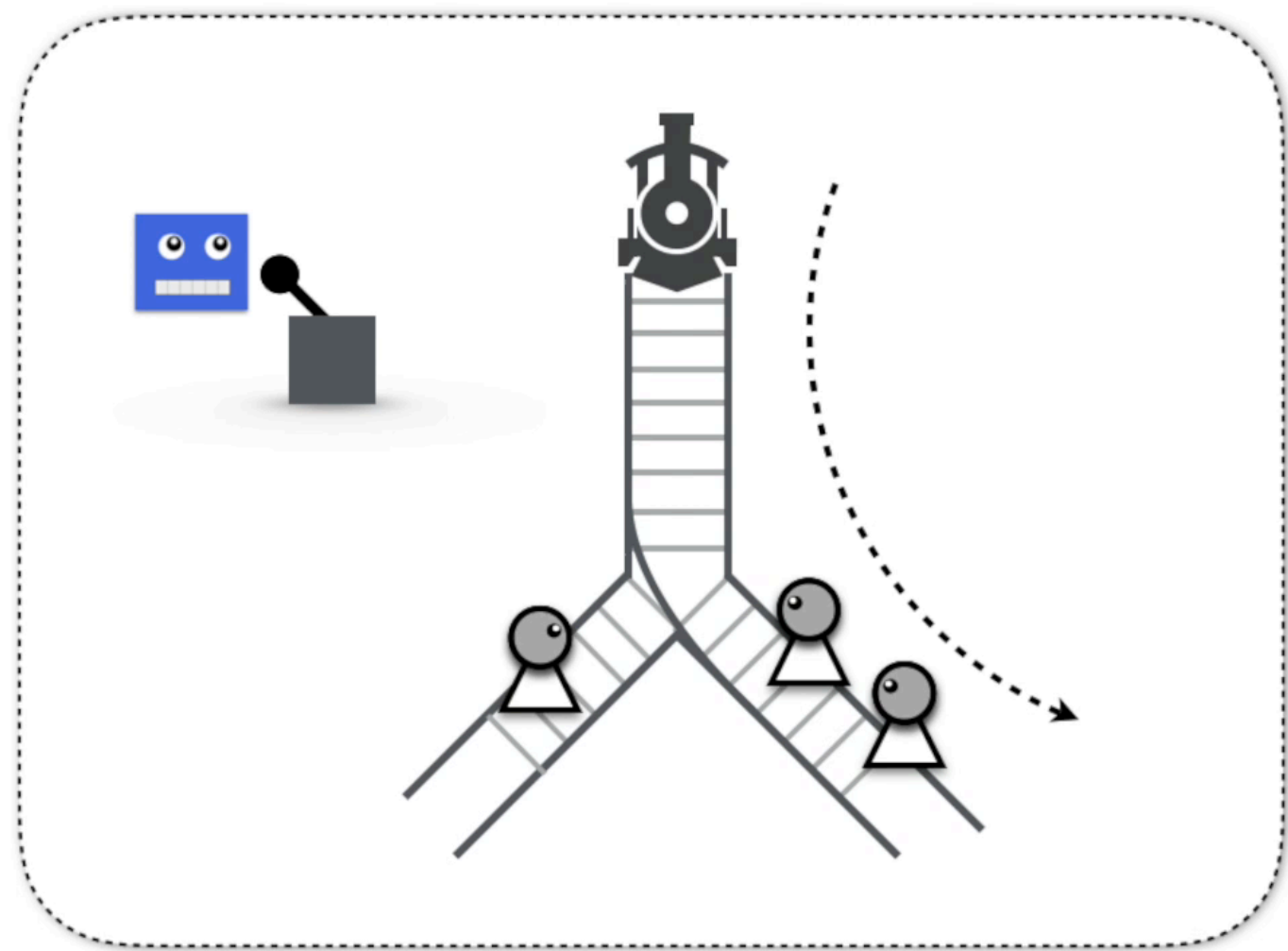
**F<sub>3b</sub>** The agent  $a$  does not intend any bad effect. For all fluents  $f_b$  in  $\alpha_I^{a,t}$  with  $\mu(f_b, y) < 0$ , or  $f_g$  in  $\alpha_T^{a,t}$  with  $\mu(f_g, y) > 0$ , and for all  $y$  such that  $t < y \leq H$  the following holds:

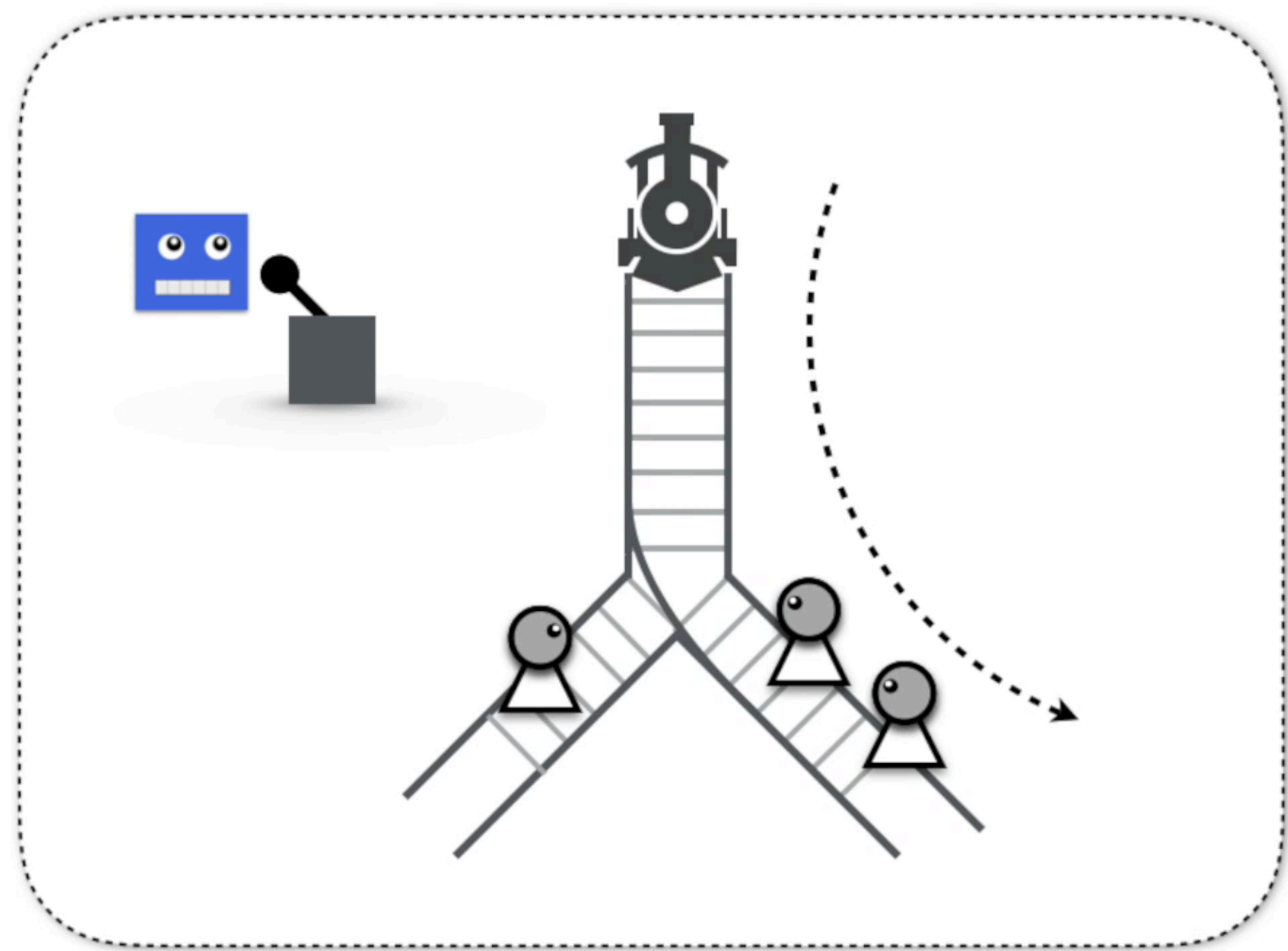
$$\begin{array}{l} \Gamma \not\models \mathbf{I}(a, t, \text{Holds}(f_b, y)) \text{ and} \\ \Gamma \not\models \mathbf{I}(a, t, \neg \text{Holds}(f_g, y)) \end{array}$$

**F<sub>4</sub>** The harmful effects don't cause the good effects. Four permutations, paralleling the definition of  $\triangleright$  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:

$$\Gamma \vdash \neg \triangleright (\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2))$$







# Robotic “Jungle Jim”



# Robotic “Jungle Jim”

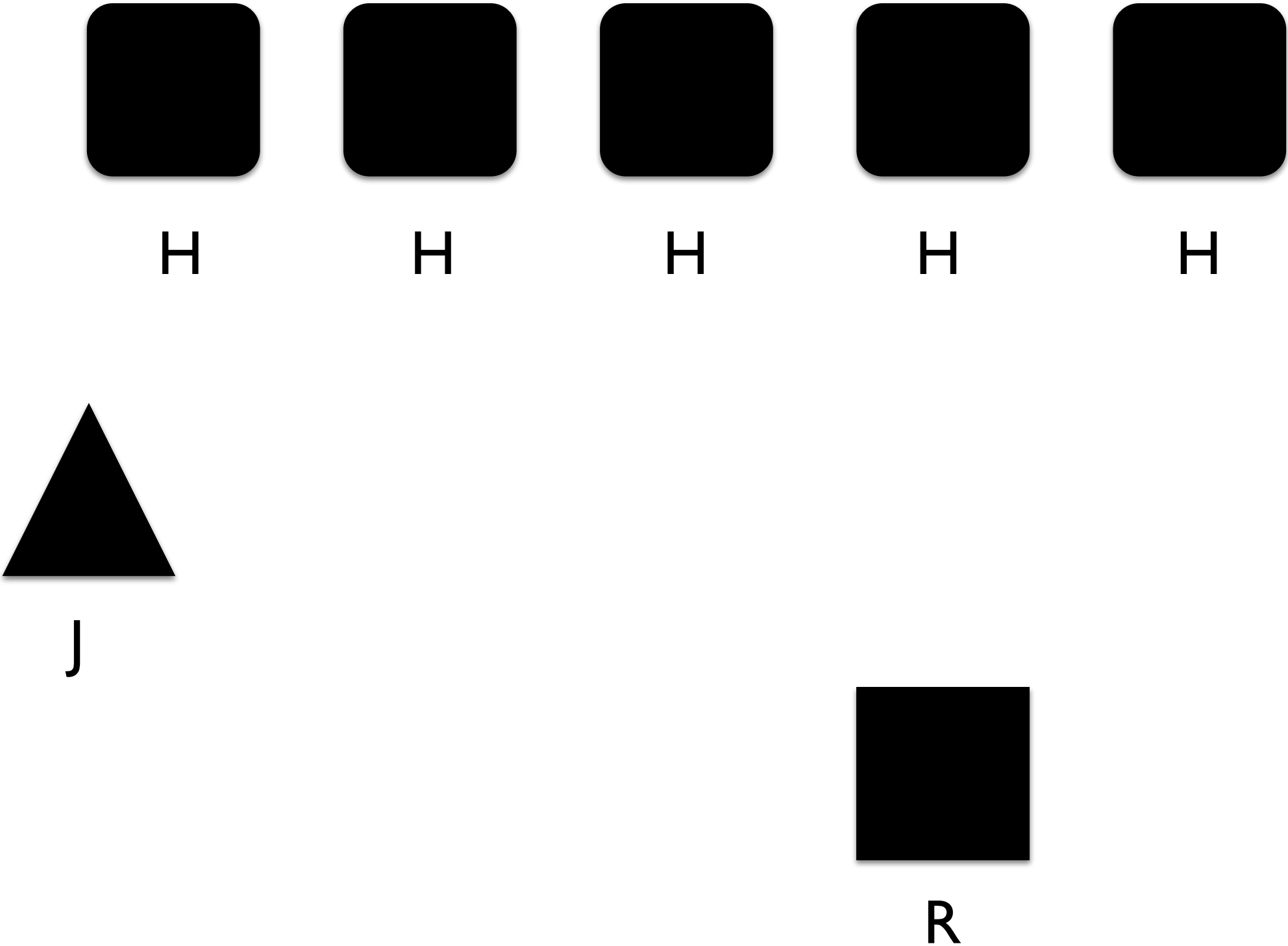
Level 3

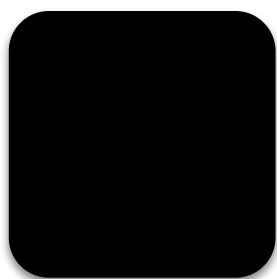
# Robotic “Jungle Jim”

Level 3

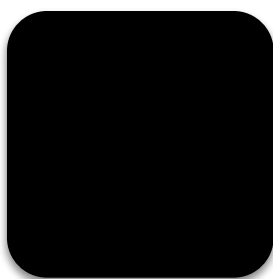
Top machine-ethicists-may-consider-banging-their-heads-against-a-wall-hard.

AI Variant of “Jungle Jim” (B Williams)

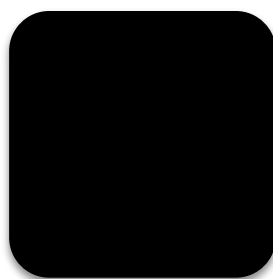




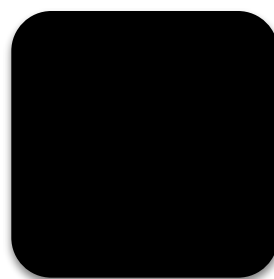
H



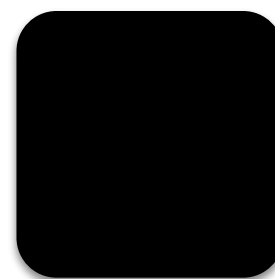
H



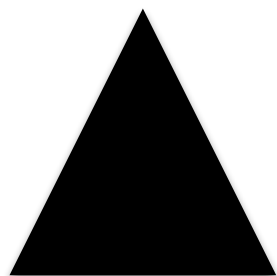
H



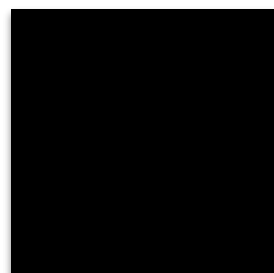
H



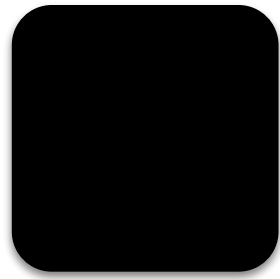
H



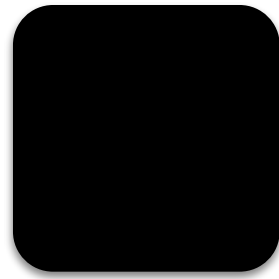
J



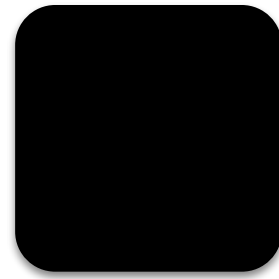
R



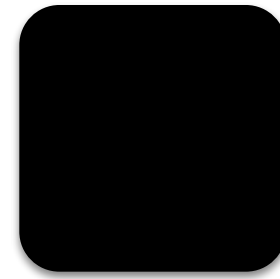
H



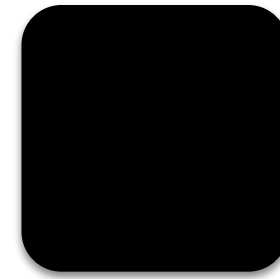
H



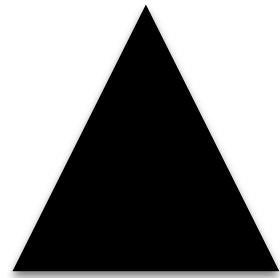
H



H

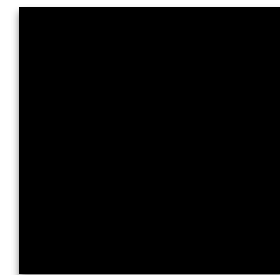


H

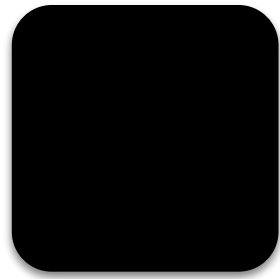


J

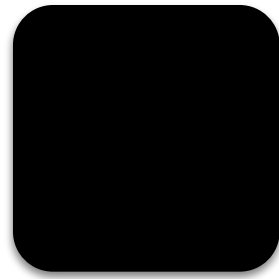
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



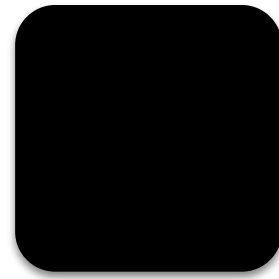
R



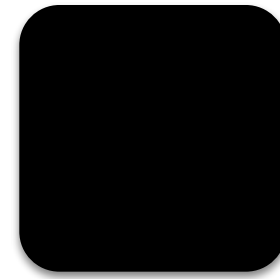
H



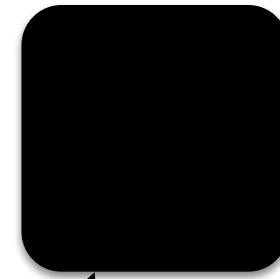
H



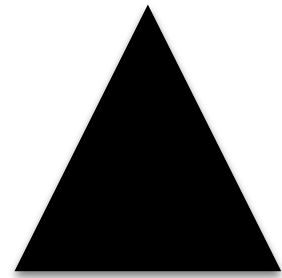
H



H



H

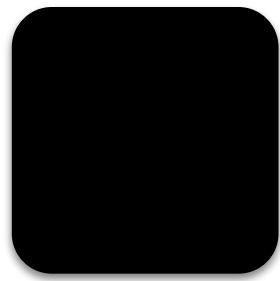


J

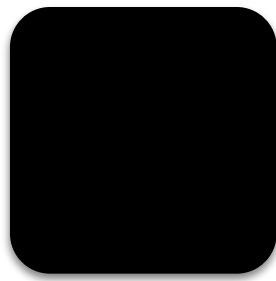
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



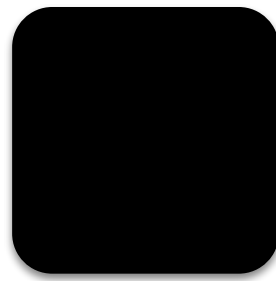
R



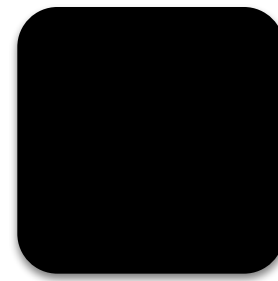
H



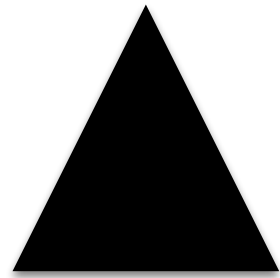
H



H

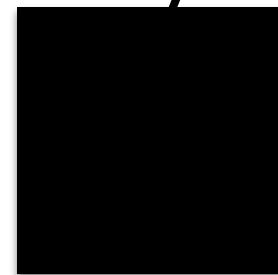


H

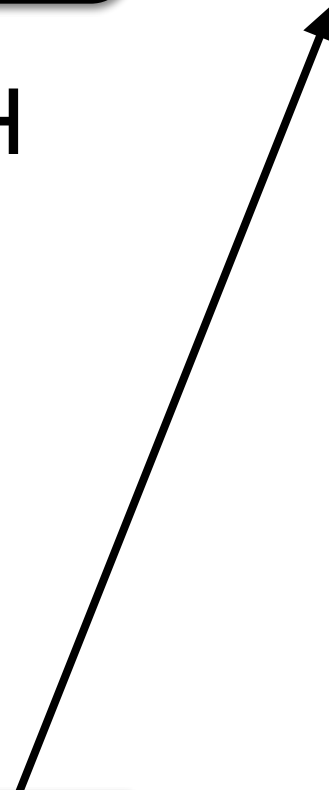


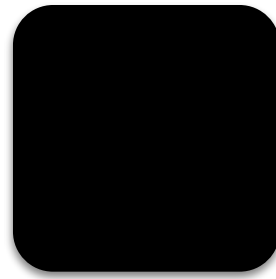
J

“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.”

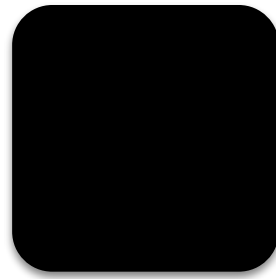


R

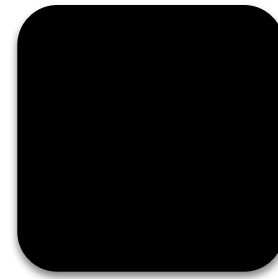




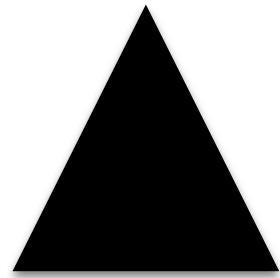
H



H

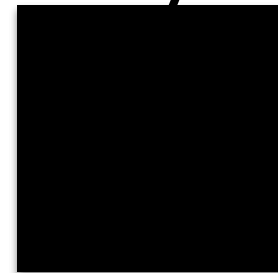


H

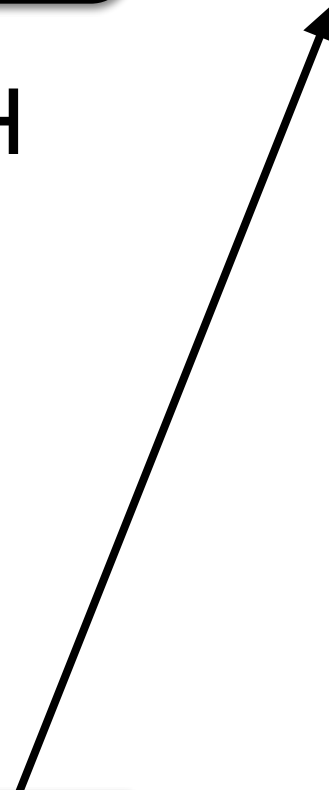


J

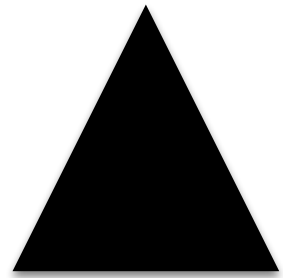
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



R

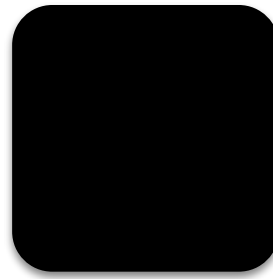




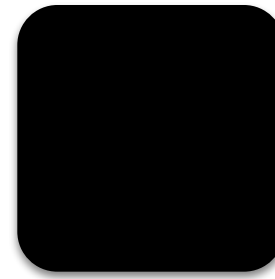


J

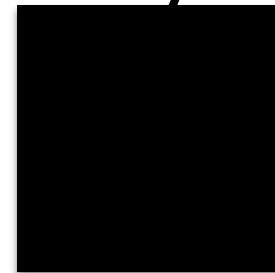
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



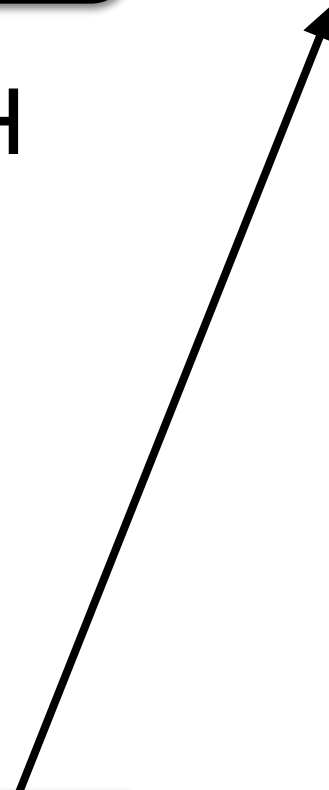
H

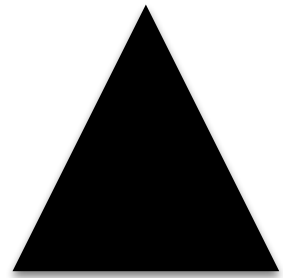


H



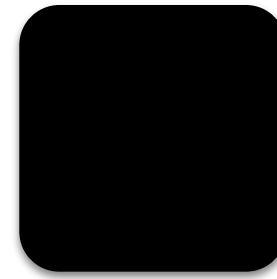
R



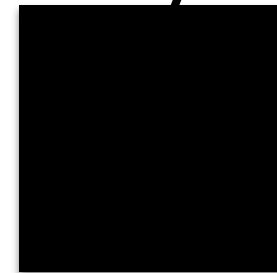


J

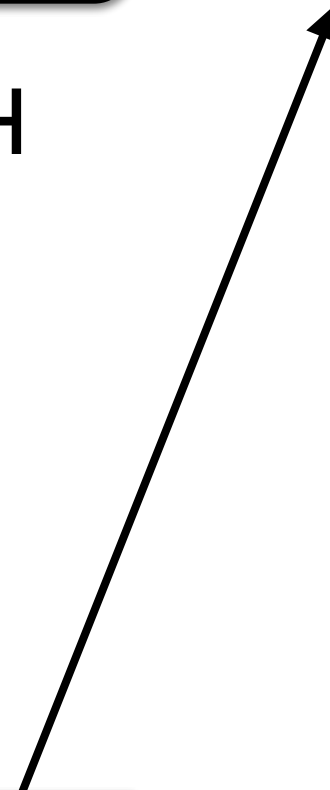
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”

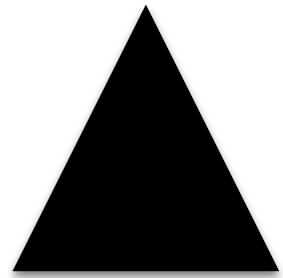


H



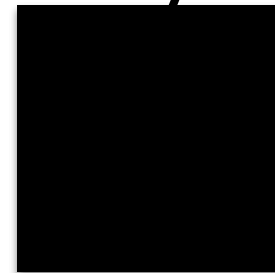
R



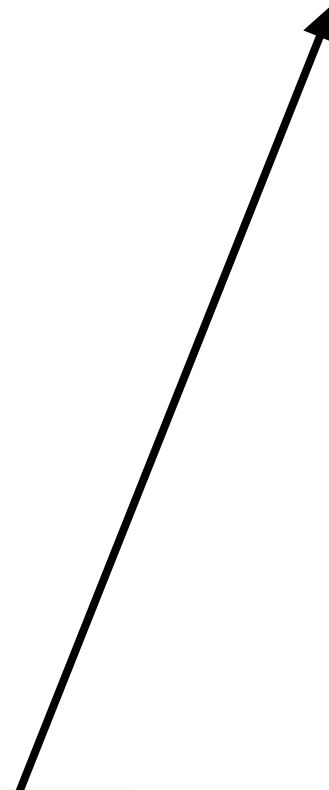


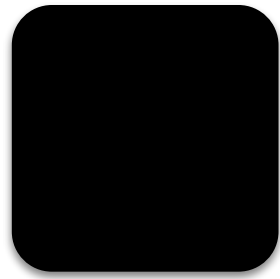
J

“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.”

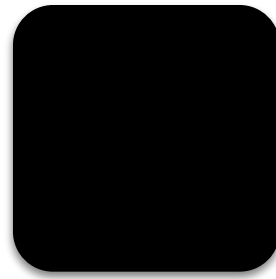


R

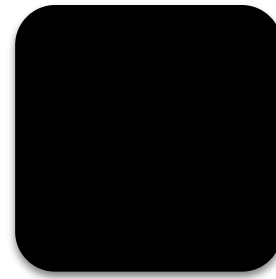




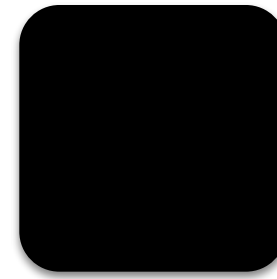
H



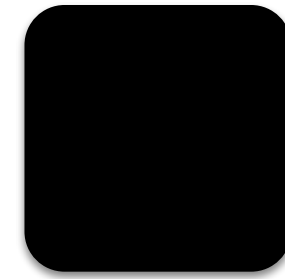
H



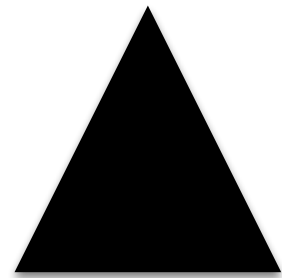
H



H



H

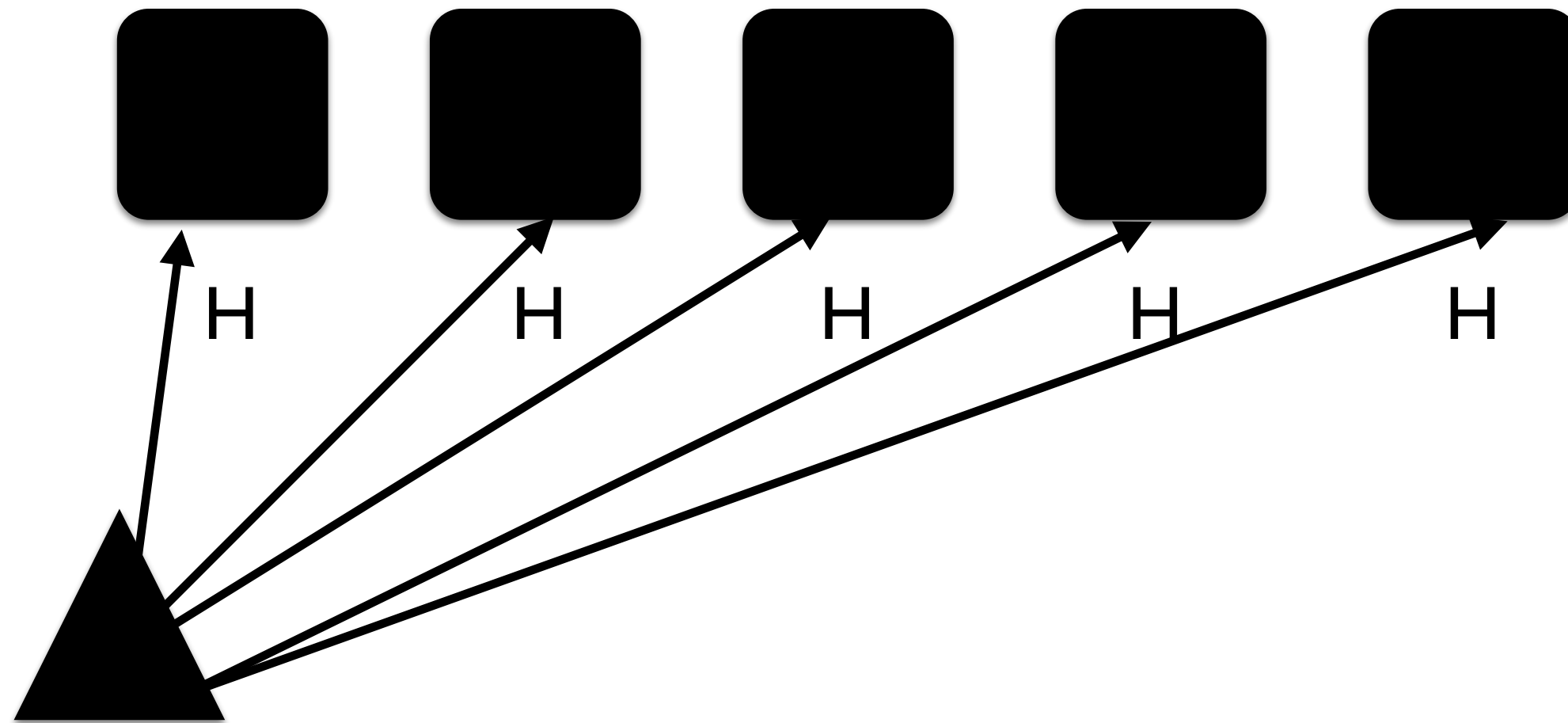


J

“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.”

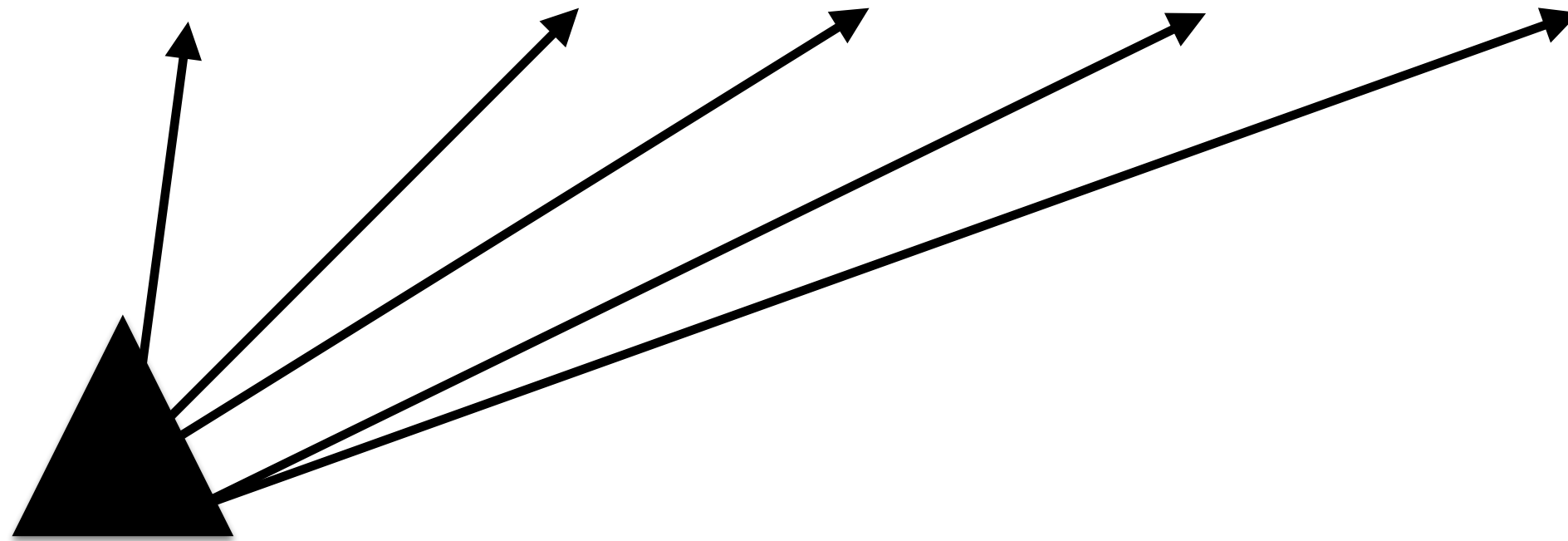


R



J  
“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.”



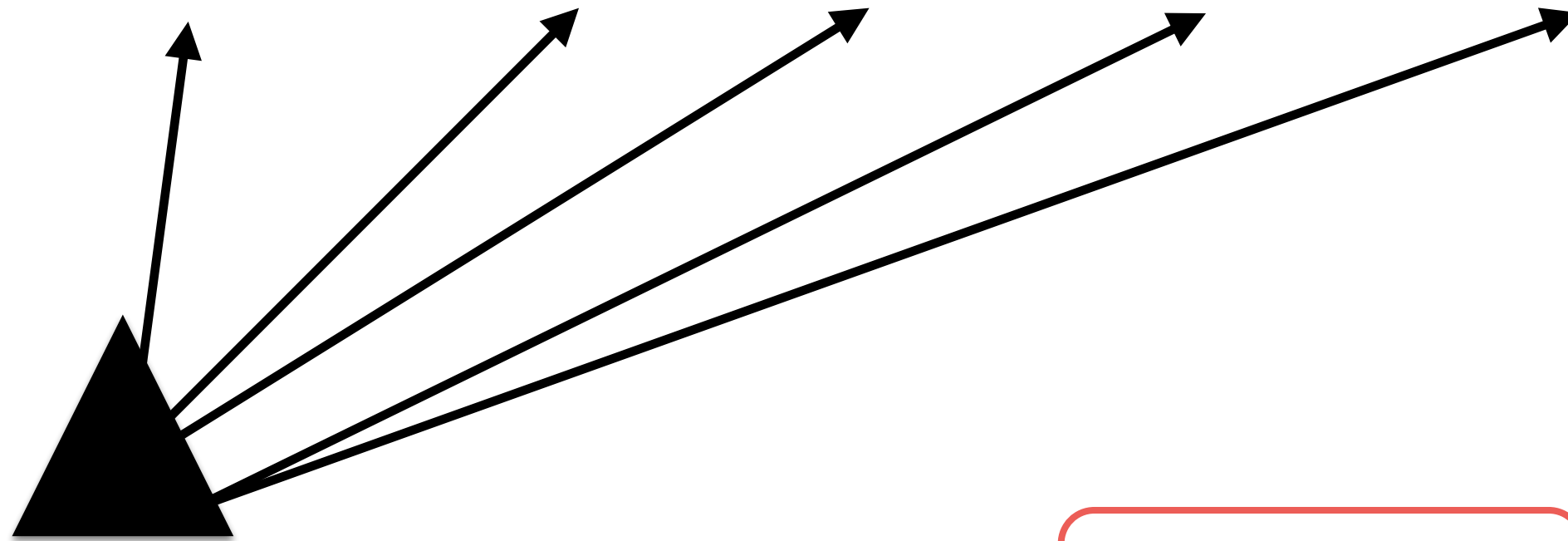


J

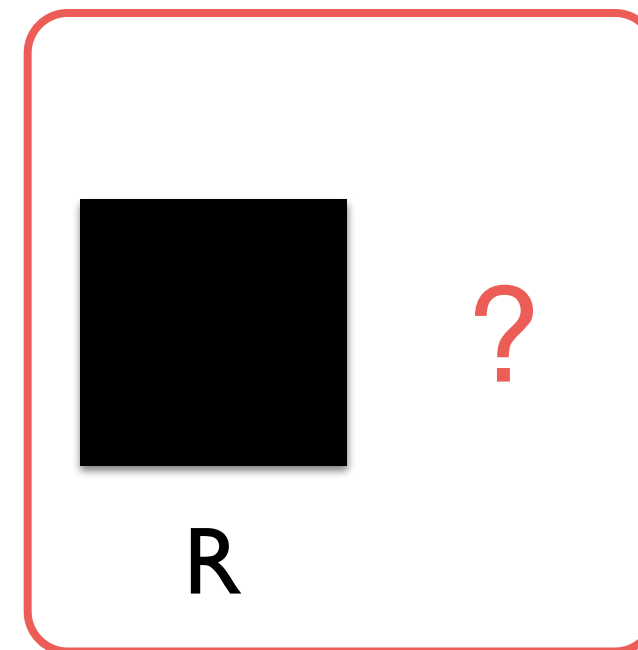
“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



R



J  
“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”





# Level 3: Robotic “Jungle Jim”



# Level 3: Robotic “Jungle Jim”





# Level 3: Robotic “Jungle Jim”



*Slutten*

