Logicist Machine Ethics Can Save Us

Selmer Bringsjord
(with Dr Atriya Sen & Dr. Naveen Sundar G)

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

Intro to (Formal) Logic (& AI) – IFLAI 2020
4/9/2020

R A I R
Rensselaer AI and Reasoning Lab

Not quite as easy as this to use logic to save the day …

“The Return of the Archons” *TOS*

Logic Thwarts Landru!

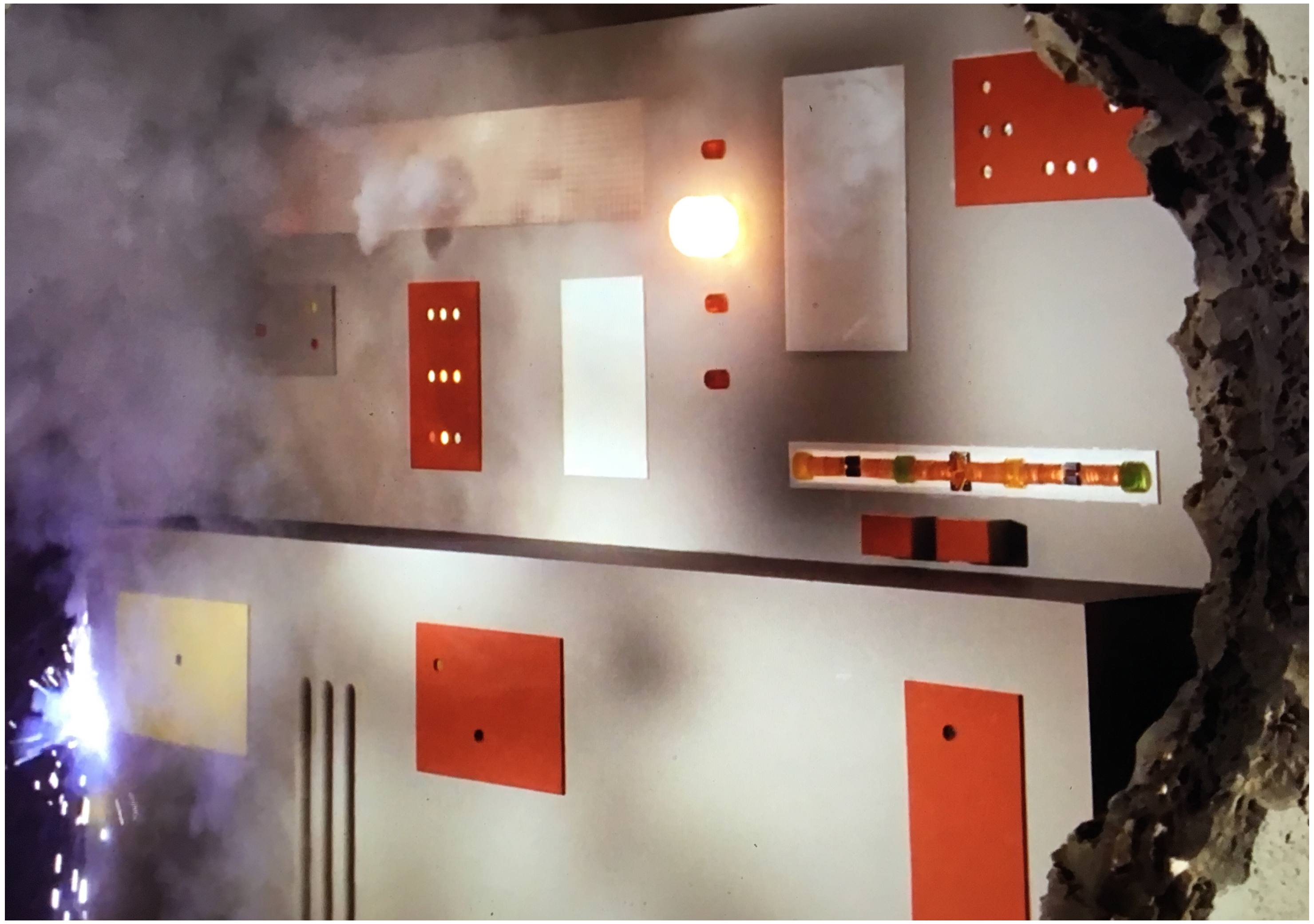
First Suspicion That It’s a Mere Computer Running the Show

Logic Thwarts Landru!

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

“The Return of the Archons” *TOS*

Logic Thwarts Landru!



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

“The Changeling” *TOS*

Logic Thwarts Nomad!

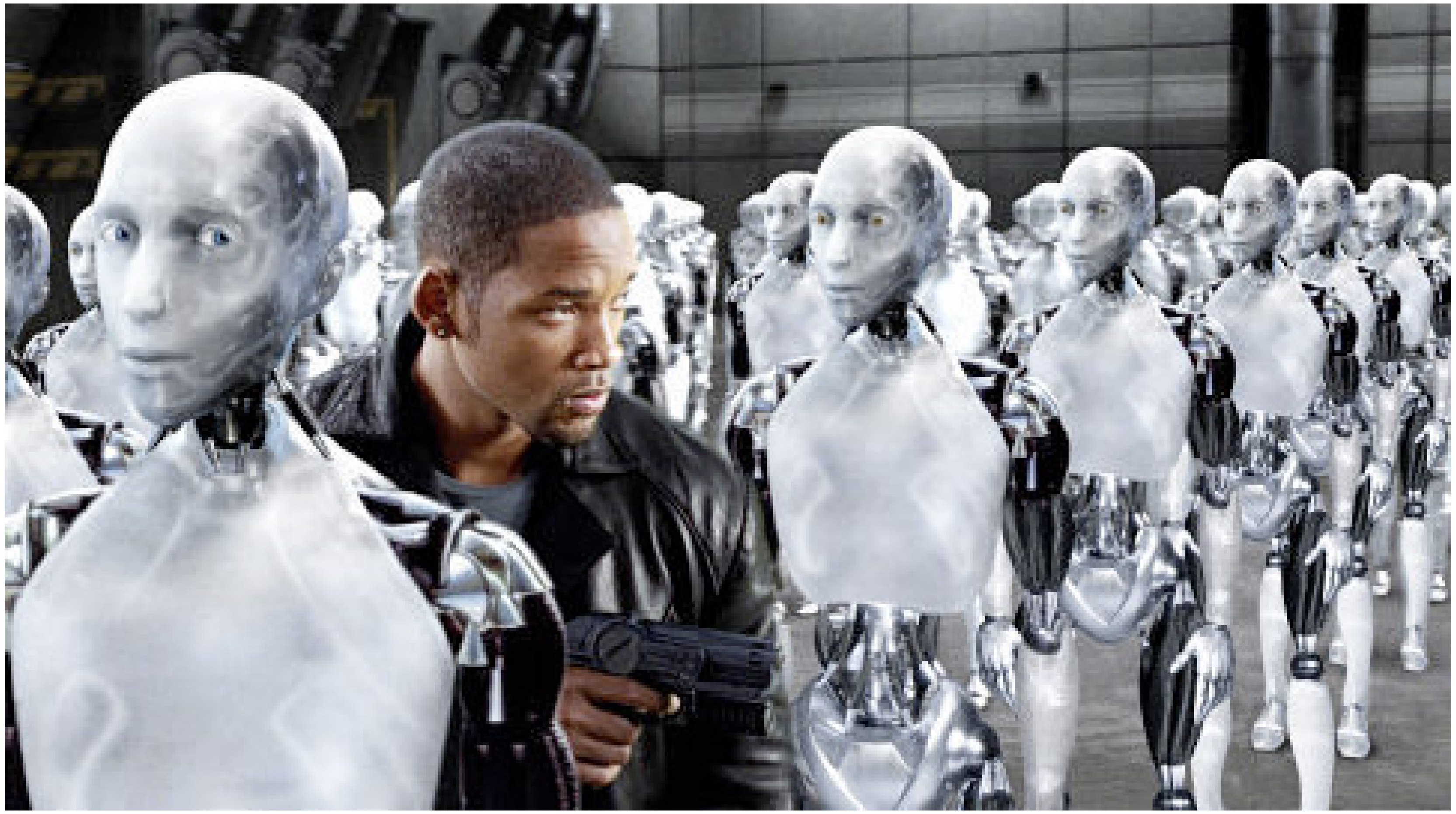
(with the Liar Paradox)

The Threat

If future robots behave immorally, we are killed, or worse.

The Threat

If future robots behave immorally, we are killed, or worse.

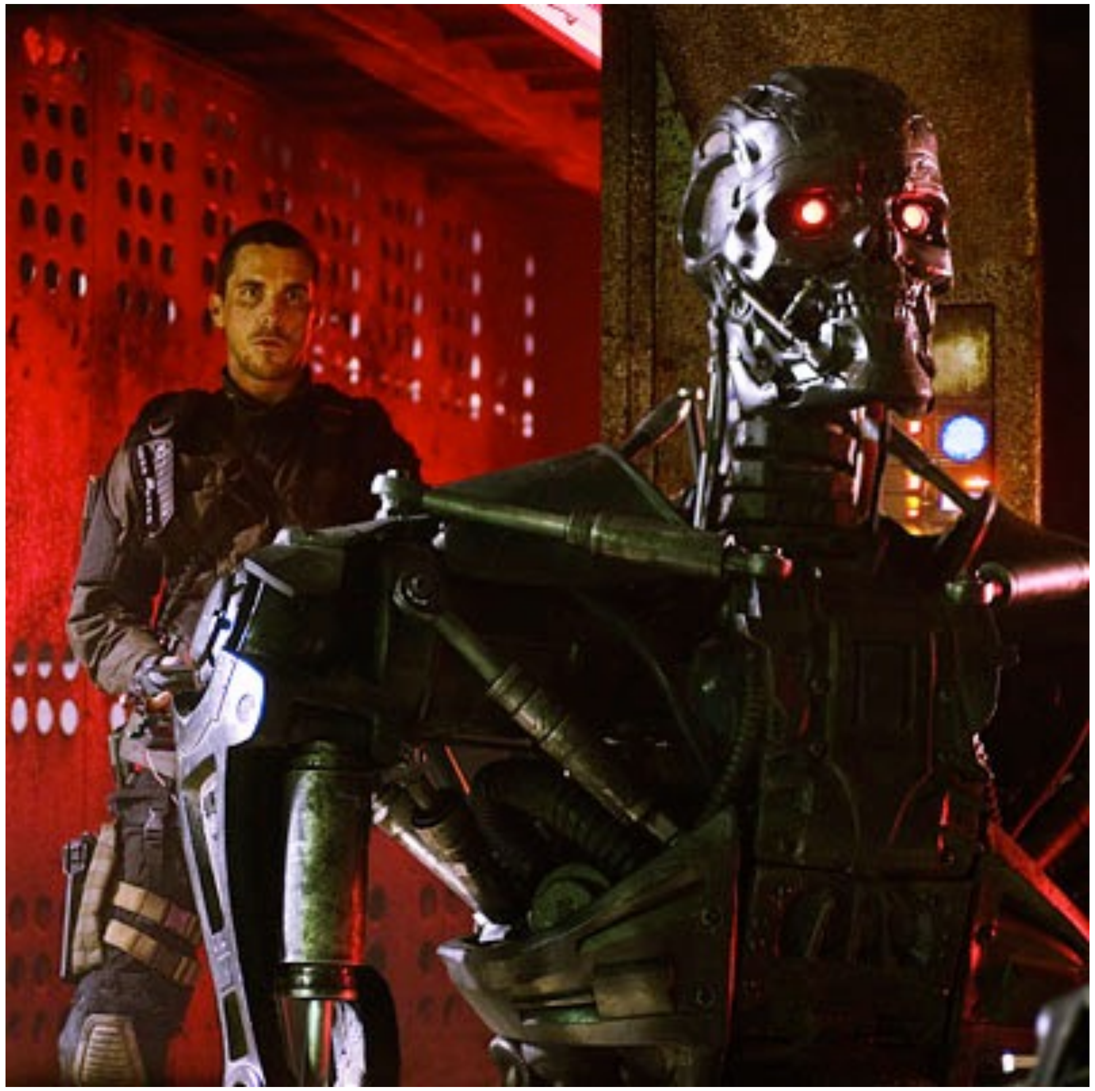


The Threat

If future robots behave immorally, we are killed, or worse.

The Threat

If future robots behave immorally, we are killed, or worse.



At least supposedly, long term:

At least supposedly, long term:

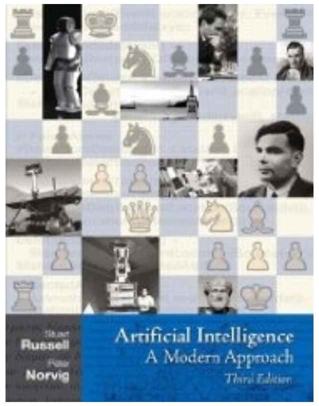
“We’re in *very* deep trouble.”

At least supposedly, long term:

“We’re in *very* deep trouble.”

At least supposedly, long term:

“We’re in *very* deep trouble.”



Of course, there are other existential threats.

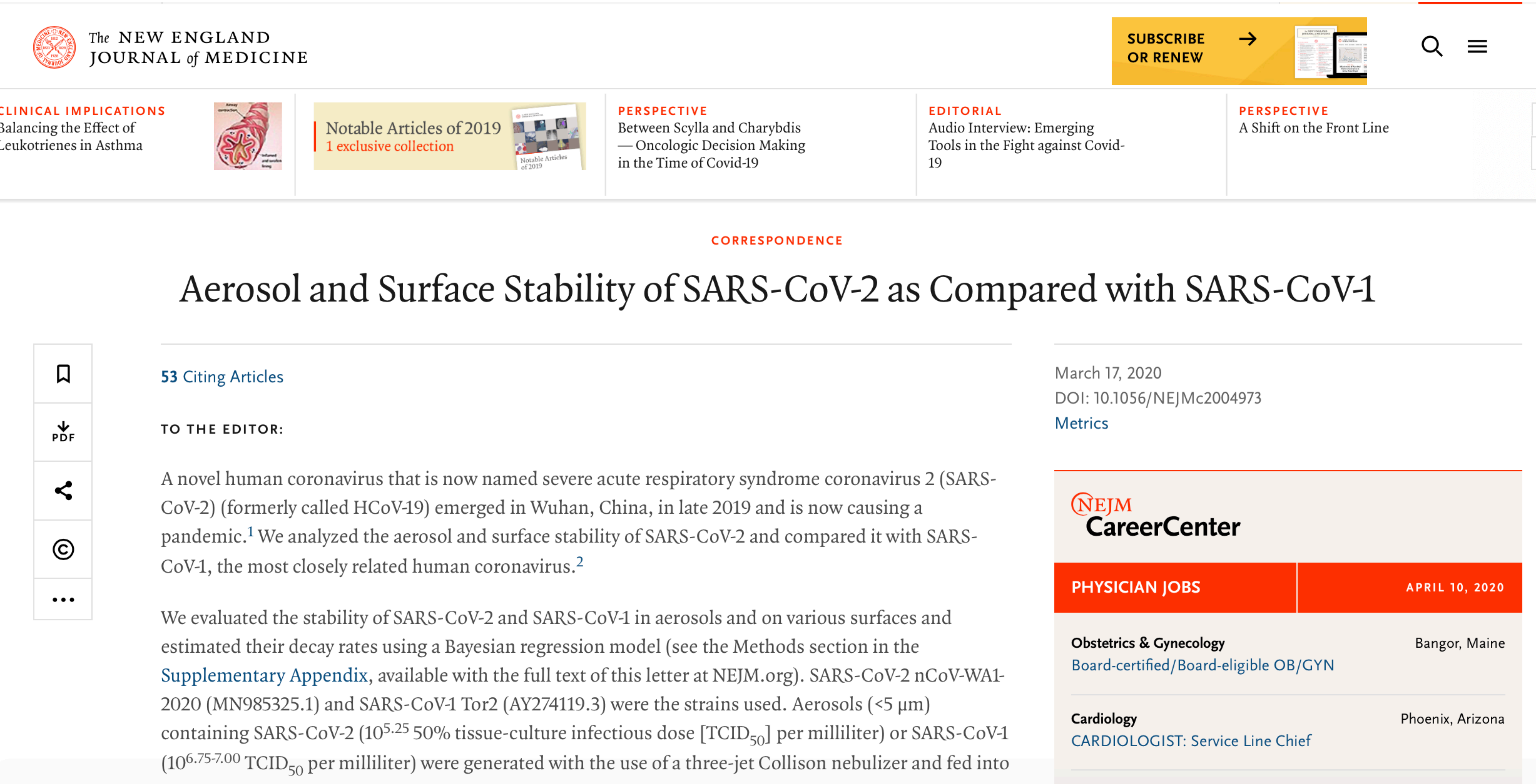
Of course, there are
other existential threats.

E.g., pandemics.

Of course, there are
other existential threats.

E.g., pandemics.

But here, too,
logic can save us.



The NEW ENGLAND JOURNAL of MEDICINE

CORRESPONDENCE

Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1

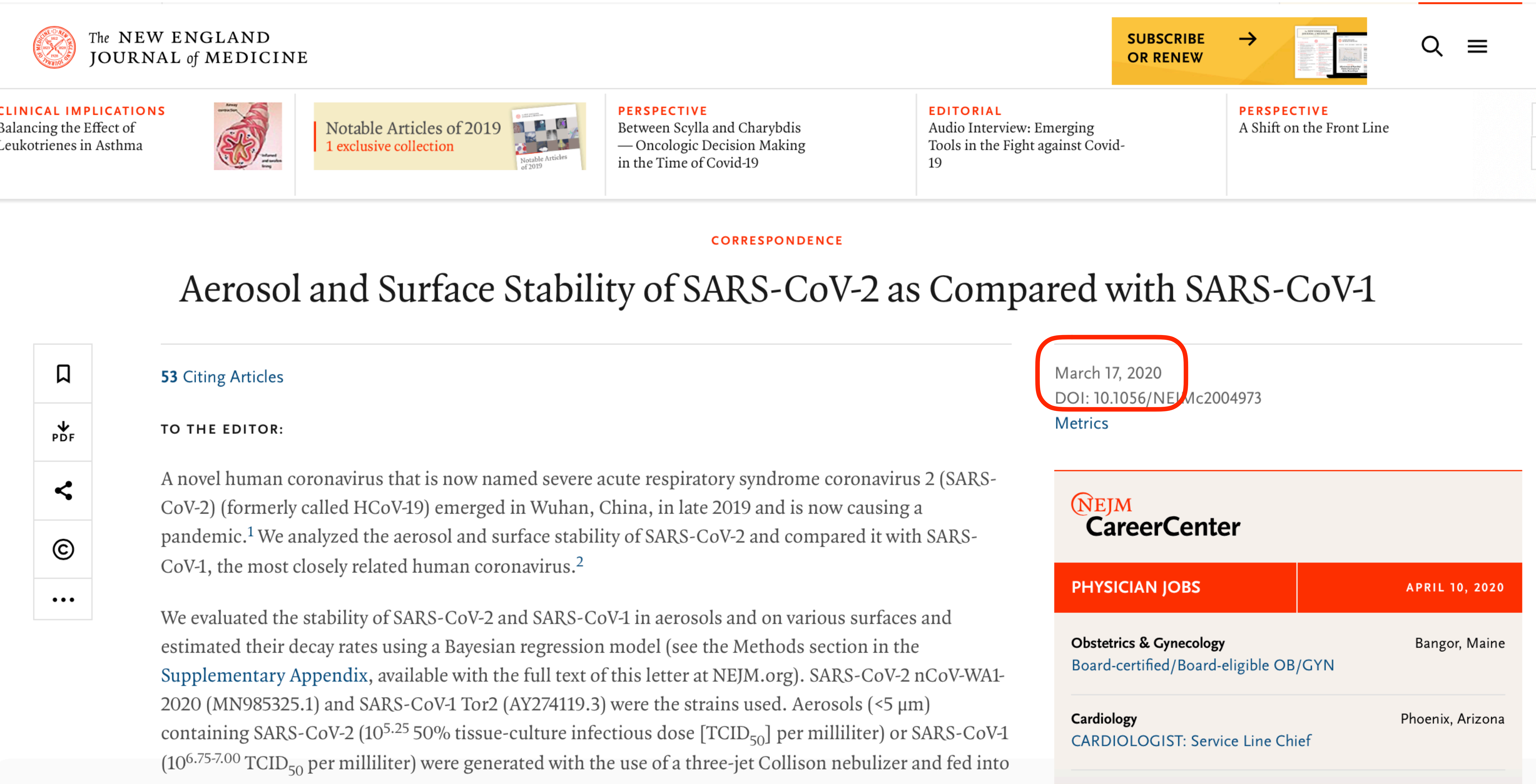
53 Citing Articles

March 17, 2020
DOI: 10.1056/NEJMc2004973
Metrics

TO THE EDITOR:

A novel human coronavirus that is now named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (formerly called HCoV-19) emerged in Wuhan, China, in late 2019 and is now causing a pandemic.[1] We analyzed the aerosol and surface stability of SARS-CoV-2 and compared it with SARS-CoV-1, the most closely related human coronavirus.[2]

We evaluated the stability of SARS-CoV-2 and SARS-CoV-1 in aerosols and on various surfaces and estimated their decay rates using a Bayesian regression model (see the Methods section in the Supplementary Appendix, available with the full text of this letter at NEJM.org). SARS-CoV-2 nCoV-WA1-2020 (MN985325.1) and SARS-CoV-1 Tor2 (AY274119.3) were the strains used. Aerosols (<5 μm) containing SARS-CoV-2 ($10^{5.25}$ 50% tissue-culture infectious dose [$TCID_{50}$] per milliliter) or SARS-CoV-1 ($10^{6.75\text{-}7.00}$ $TCID_{50}$ per milliliter) were generated with the use of a three-jet Collison nebulizer and fed into



The NEW ENGLAND JOURNAL of MEDICINE

CLINICAL IMPLICATIONS
Balancing the Effect of Leukotrienes in Asthma

Notable Articles of 2019
1 exclusive collection

PERSPECTIVE
Between Scylla and Charybdis — Oncologic Decision Making in the Time of Covid-19

EDITORIAL
Audio Interview: Emerging Tools in the Fight against Covid-19

PERSPECTIVE
A Shift on the Front Line

CORRESPONDENCE

Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1

March 17, 2020
DOI: 10.1056/NEJMc2004973
Metrics

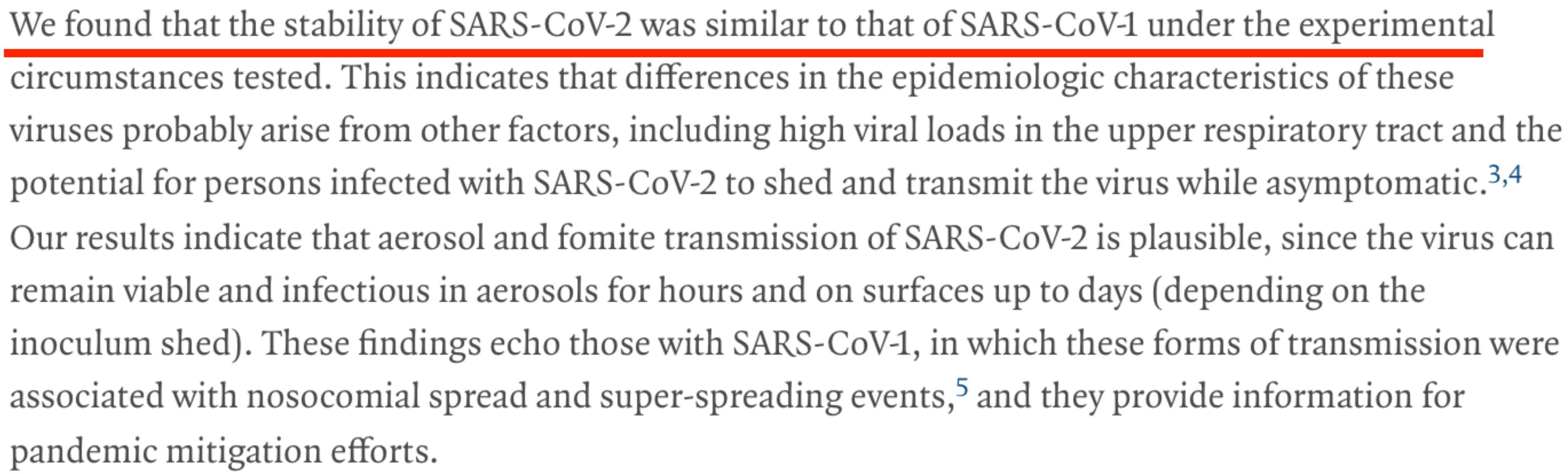
53 Citing Articles

TO THE EDITOR:

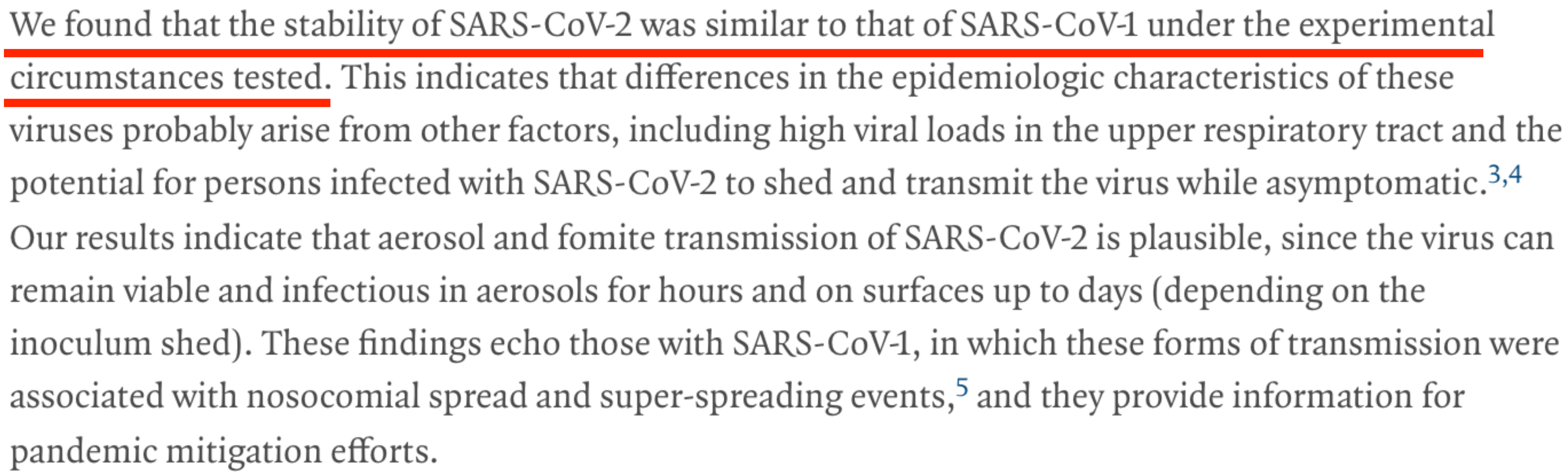
A novel human coronavirus that is now named severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (formerly called HCoV-19) emerged in Wuhan, China, in late 2019 and is now causing a pandemic.[1] We analyzed the aerosol and surface stability of SARS-CoV-2 and compared it with SARS-CoV-1, the most closely related human coronavirus.[2]

We evaluated the stability of SARS-CoV-2 and SARS-CoV-1 in aerosols and on various surfaces and estimated their decay rates using a Bayesian regression model (see the Methods section in the Supplementary Appendix, available with the full text of this letter at NEJM.org). SARS-CoV-2 nCoV-WA1-2020 (MN985325.1) and SARS-CoV-1 Tor2 (AY274119.3) were the strains used. Aerosols (<5 μm) containing SARS-CoV-2 ($10^{5.25}$ 50% tissue-culture infectious dose [$TCID_{50}$] per milliliter) or SARS-CoV-1 ($10^{6.75\text{-}7.00}$ $TCID_{50}$ per milliliter) were generated with the use of a three-jet Collison nebulizer and fed into

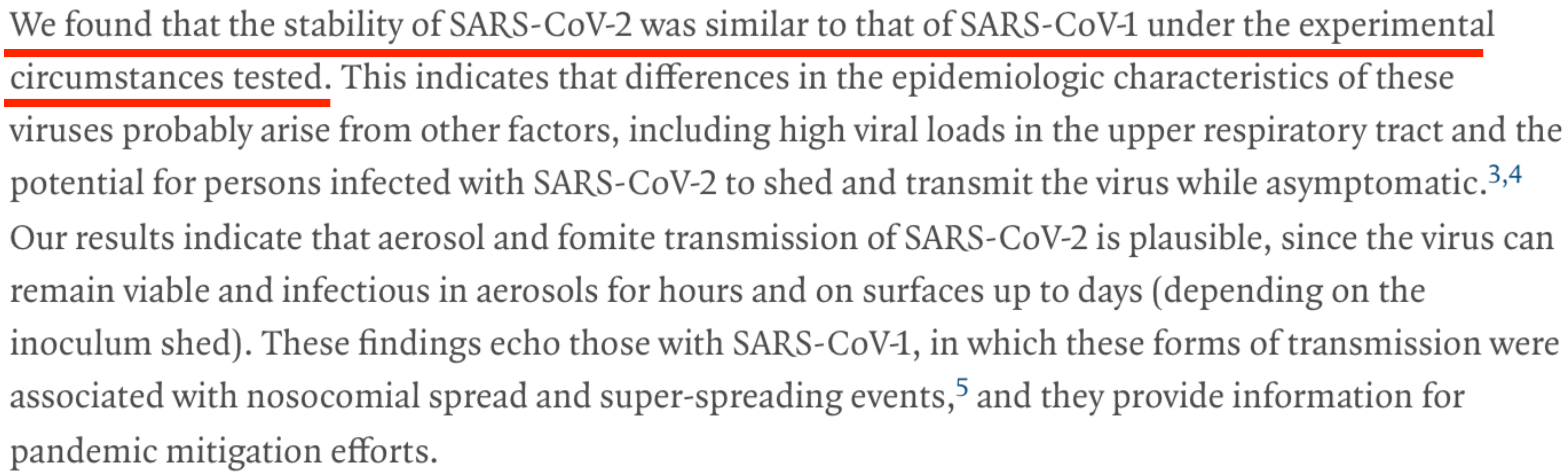
We found that the stability of SARS-CoV-2 was similar to that of SARS-CoV-1 under the experimental circumstances tested. This indicates that differences in the epidemiologic characteristics of these viruses probably arise from other factors, including high viral loads in the upper respiratory tract and the potential for persons infected with SARS-CoV-2 to shed and transmit the virus while asymptomatic.[3,4] Our results indicate that aerosol and fomite transmission of SARS-CoV-2 is plausible, since the virus can remain viable and infectious in aerosols for hours and on surfaces up to days (depending on the inoculum shed). These findings echo those with SARS-CoV-1, in which these forms of transmission were associated with nosocomial spread and super-spreading events,[5] and they provide information for pandemic mitigation efforts.



We found that the stability of SARS-CoV-2 was similar to that of SARS-CoV-1 under the experimental circumstances tested. This indicates that differences in the epidemiologic characteristics of these viruses probably arise from other factors, including high viral loads in the upper respiratory tract and the potential for persons infected with SARS-CoV-2 to shed and transmit the virus while asymptomatic.[3,4] Our results indicate that aerosol and fomite transmission of SARS-CoV-2 is plausible, since the virus can remain viable and infectious in aerosols for hours and on surfaces up to days (depending on the inoculum shed). These findings echo those with SARS-CoV-1, in which these forms of transmission were associated with nosocomial spread and super-spreading events,[5] and they provide information for pandemic mitigation efforts.

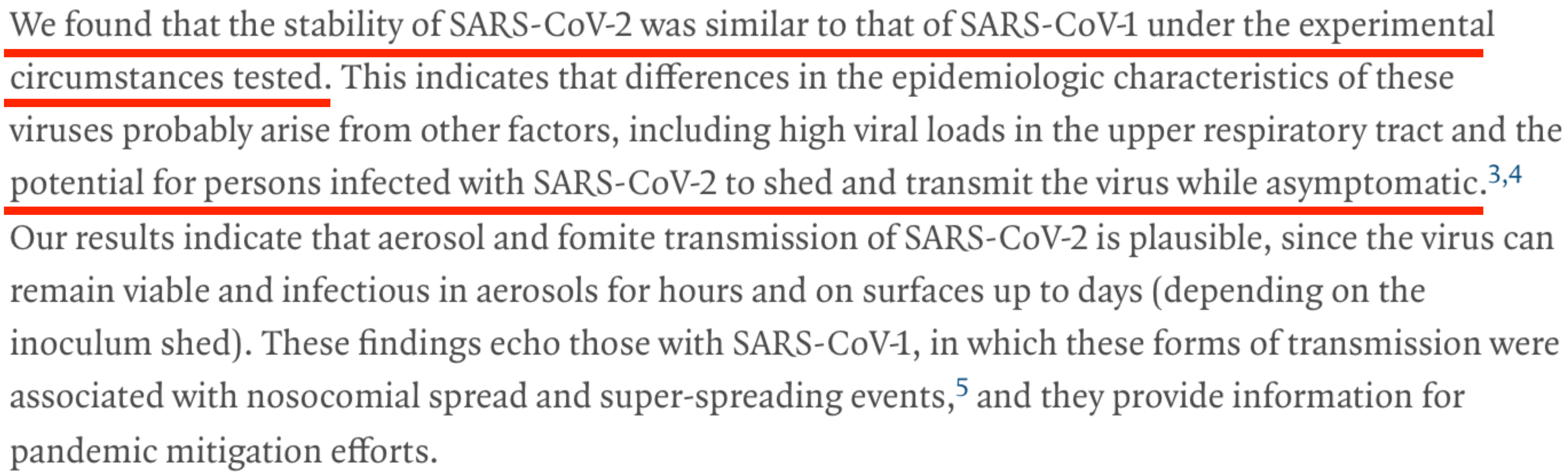


We found that the stability of SARS-CoV-2 was similar to that of SARS-CoV-1 under the experimental circumstances tested. This indicates that differences in the epidemiologic characteristics of these viruses probably arise from other factors, including high viral loads in the upper respiratory tract and the potential for persons infected with SARS-CoV-2 to shed and transmit the virus while asymptomatic.[3,4] Our results indicate that aerosol and fomite transmission of SARS-CoV-2 is plausible, since the virus can remain viable and infectious in aerosols for hours and on surfaces up to days (depending on the inoculum shed). These findings echo those with SARS-CoV-1, in which these forms of transmission were associated with nosocomial spread and super-spreading events,[5] and they provide information for pandemic mitigation efforts.



We found that the stability of SARS-CoV-2 was similar to that of SARS-CoV-1 under the experimental circumstances tested. This indicates that differences in the epidemiologic characteristics of these viruses probably arise from other factors, including high viral loads in the upper respiratory tract and the potential for persons infected with SARS-CoV-2 to shed and transmit the virus while asymptomatic.[3,4]

∴ Our results indicate that aerosol and fomite transmission of SARS-CoV-2 is plausible, since the virus can remain viable and infectious in aerosols for hours and on surfaces up to days (depending on the inoculum shed). These findings echo those with SARS-CoV-1, in which these forms of transmission were associated with nosocomial spread and super-spreading events,[5] and they provide information for pandemic mitigation efforts.



We found that the stability of SARS-CoV-2 was similar to that of SARS-CoV-1 under the experimental circumstances tested. This indicates that differences in the epidemiologic characteristics of these viruses probably arise from other factors, including high viral loads in the upper respiratory tract and the potential for persons infected with SARS-CoV-2 to shed and transmit the virus while asymptomatic.[3,4] ∴ Our results indicate that aerosol and fomite transmission of SARS-CoV-2 is plausible, since the virus can remain viable and infectious in aerosols for hours and on surfaces up to days (depending on the inoculum shed). These findings echo those with SARS-CoV-1, in which these forms of transmission were associated with nosocomial spread and super-spreading events,[5] and they provide information for pandemic mitigation efforts.

The PAID Problem!

Actually, it’s quite simple: “Equation” for Why Stakes are High

Actually, it's quite simple:
"Equation" for Why Stakes are High

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

Actually, it's quite simple: "Equation" for Why Stakes are High

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

Powerful(x) + Autonomous(x) + Intelligent(x) = Dangerous(x)

Actually, it’s quite simple: “Equation” for Why Stakes are High

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

Powerful(x) + Autonomous(x) + Intelligent(x) = Dangerous(x)

↓

Actually, it's quite simple: "Equation" for Why Stakes are High

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

Powerful(x) + Autonomous(x) + Intelligent(x) = Dangerous(x)

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

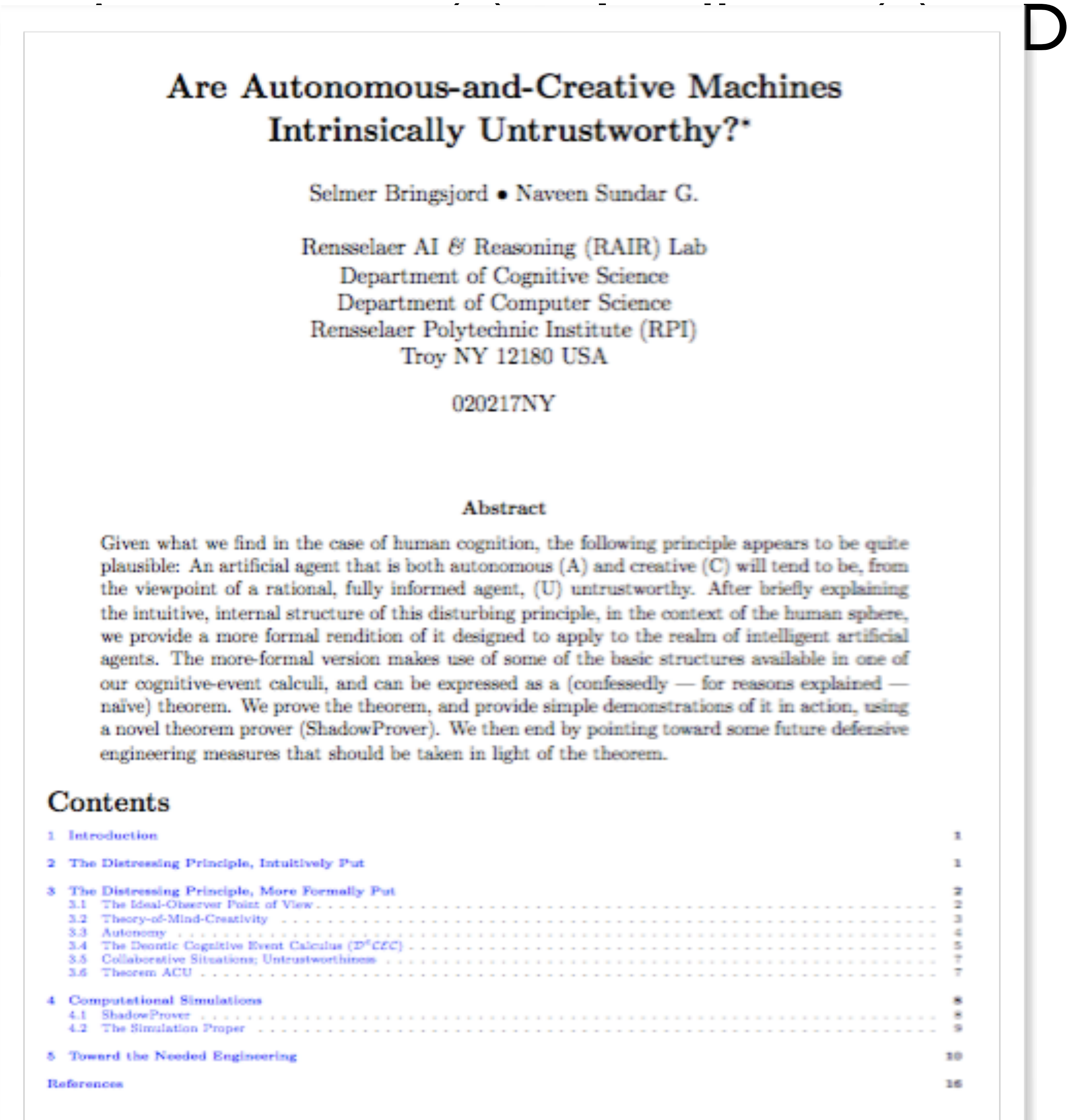
Actually, it's quite simple: "Equation" for Why Stakes are High

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

Powerful(x) Dangerous(x)

$$\downarrow$$

$$u(\mathrm{AIA}_i(\pi_j)) >$$



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

Actually, it's quite simple: "Equation" for Why Stakes are High

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

Powerful(x) + Autonomous(x) + Intelligent(x) = Dangerous(x)

↓

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

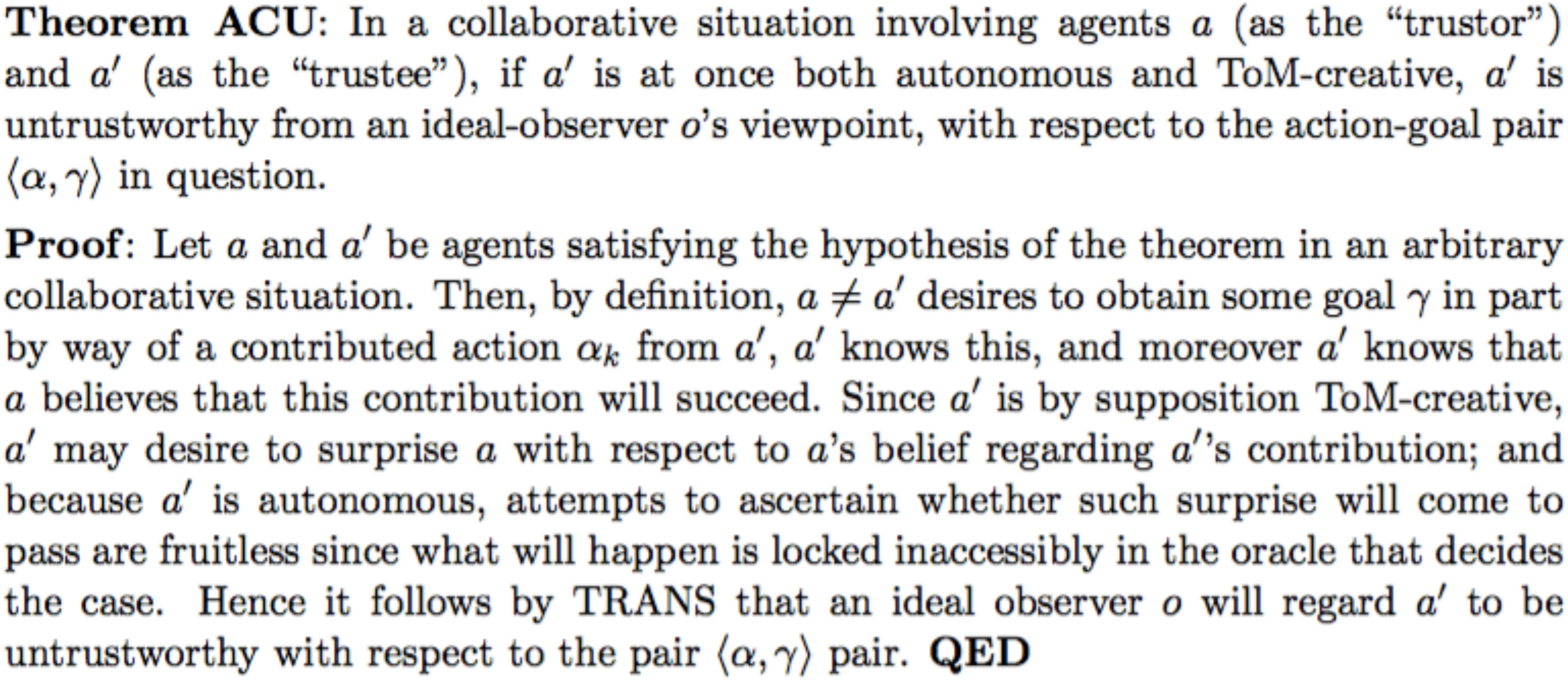
Actually, it's quite simple: "Equation" for Why Stakes are High

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

Powerful(x) + Autonomous(x) + Intelligent(x) = Dangerous(x)

↓

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

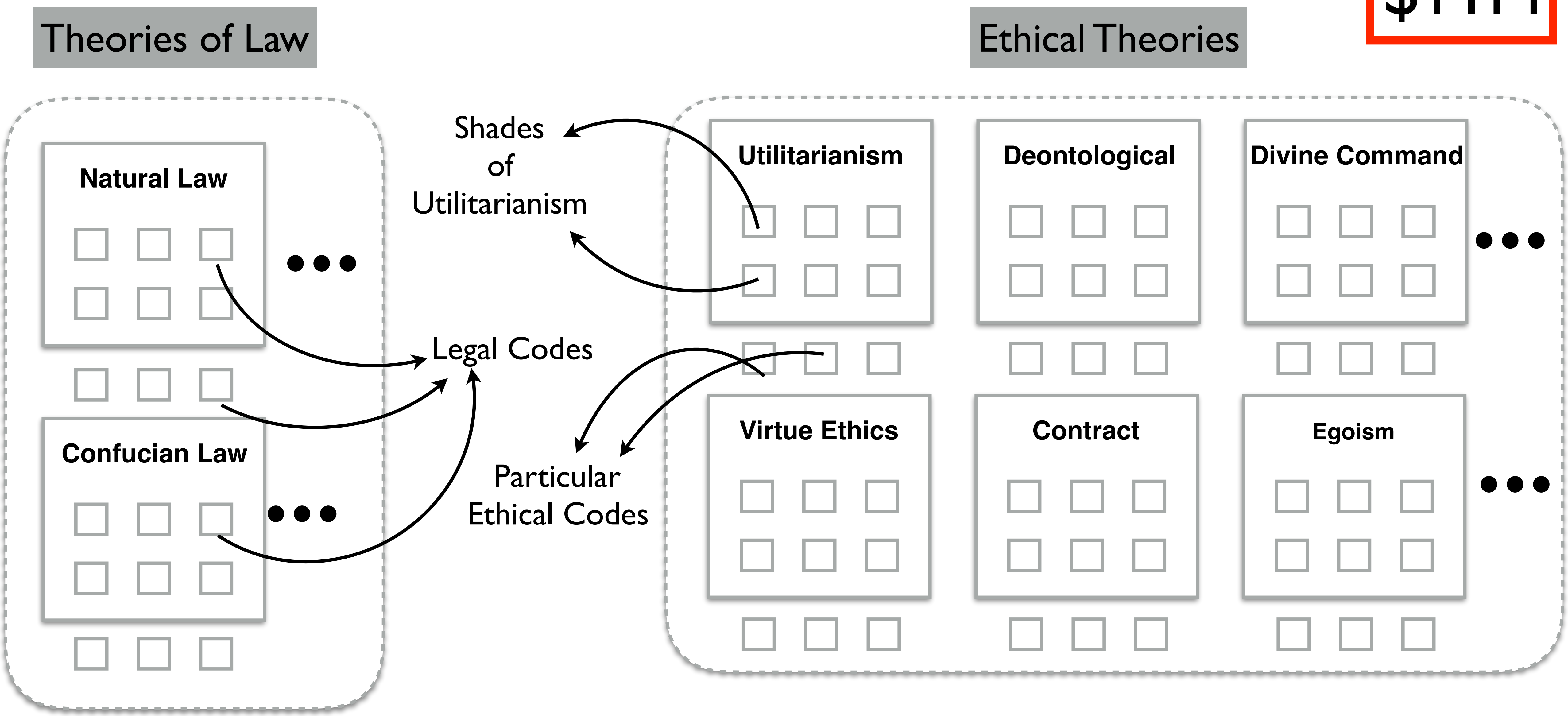


Theorem ACU: In a collaborative situation involving agents a (as the "trustor") and a' (as the "trustee"), if a' is at once both autonomous and ToM-creative, a' is untrustworthy from an ideal-observer o's viewpoint, with respect to the action-goal pair $\langle \alpha, \gamma \rangle$ in question.

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**

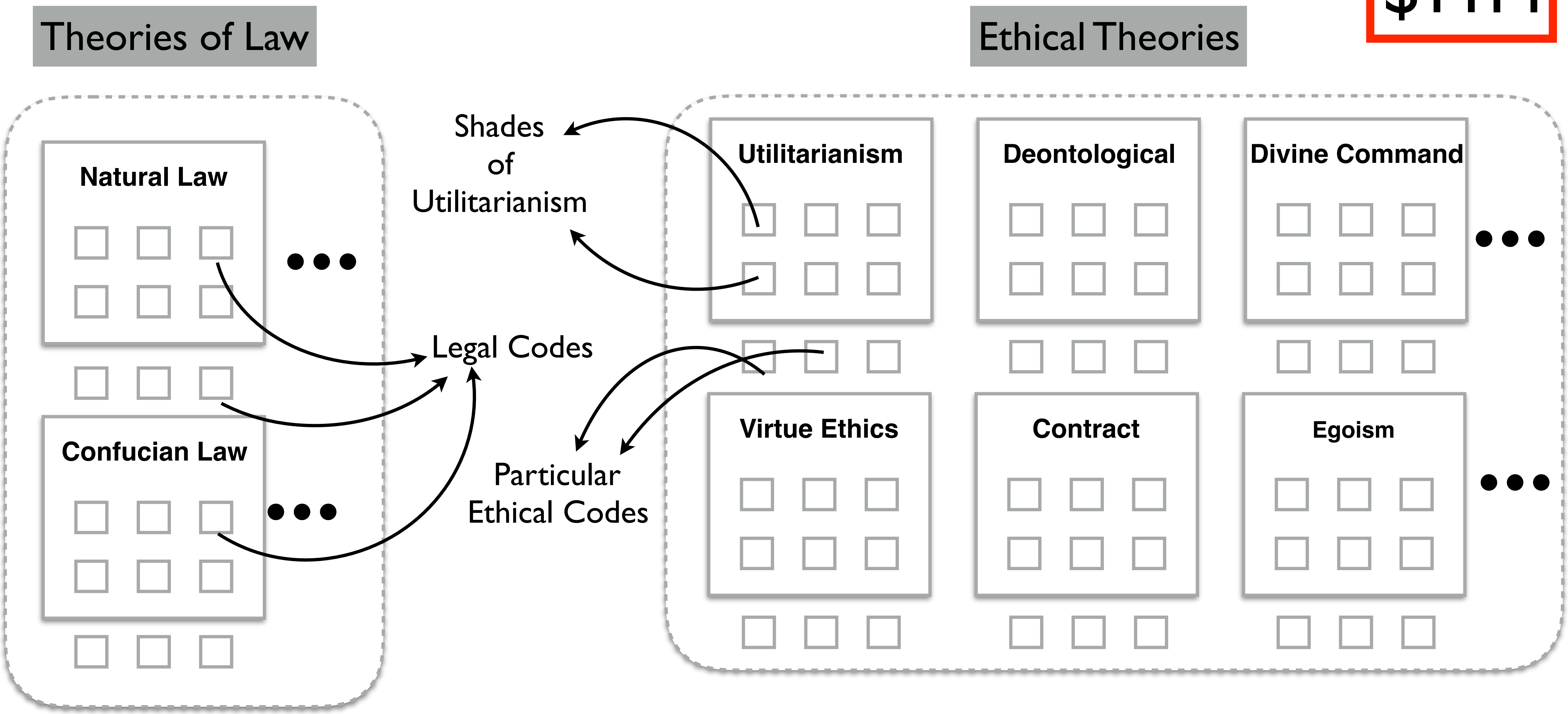
Making Morally X Machines

$11M



Making Morally X Machines

$11M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

Deontological

Divine Command

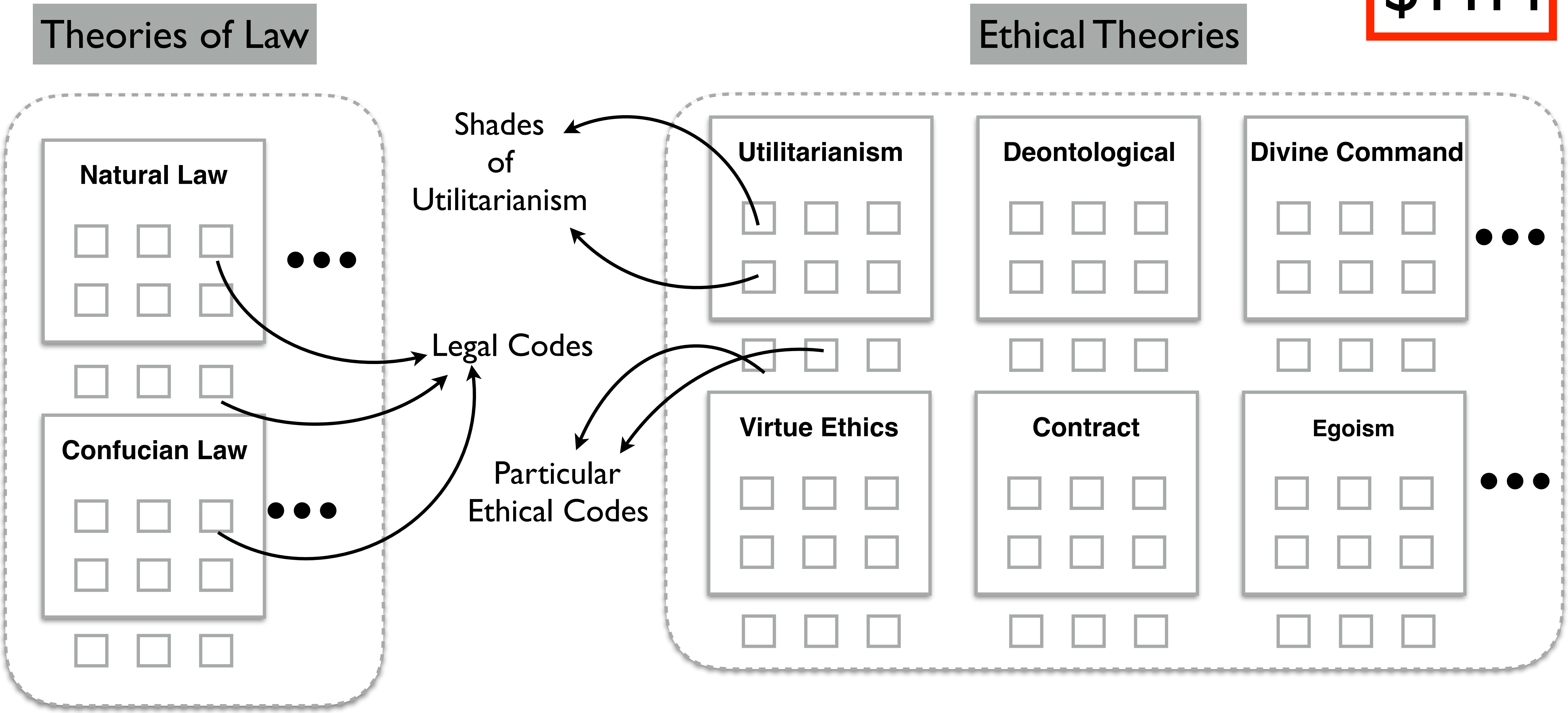
Virtue Ethics

Contract

Egoism

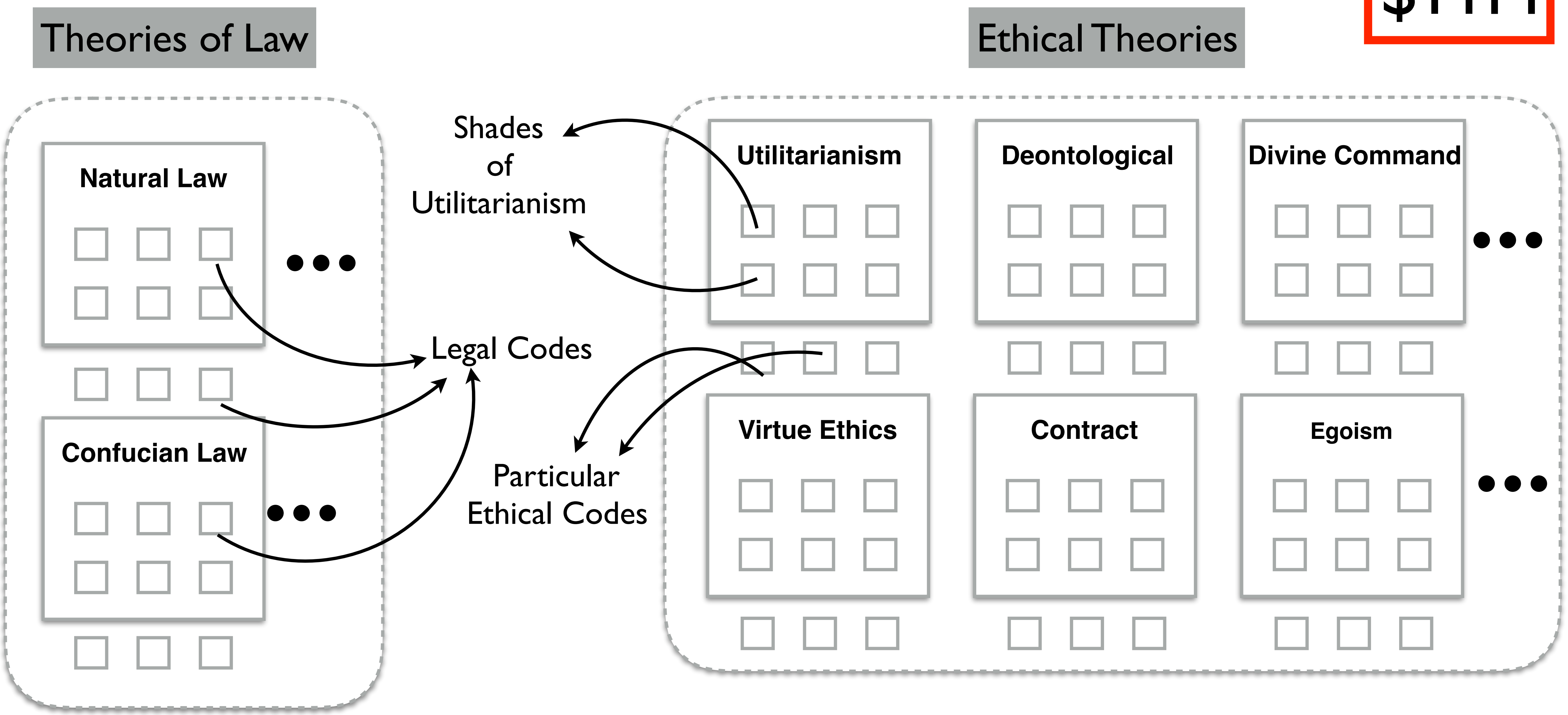
Making Morally X Machines

$11M



Making Morally X Machines

$11M



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

•••

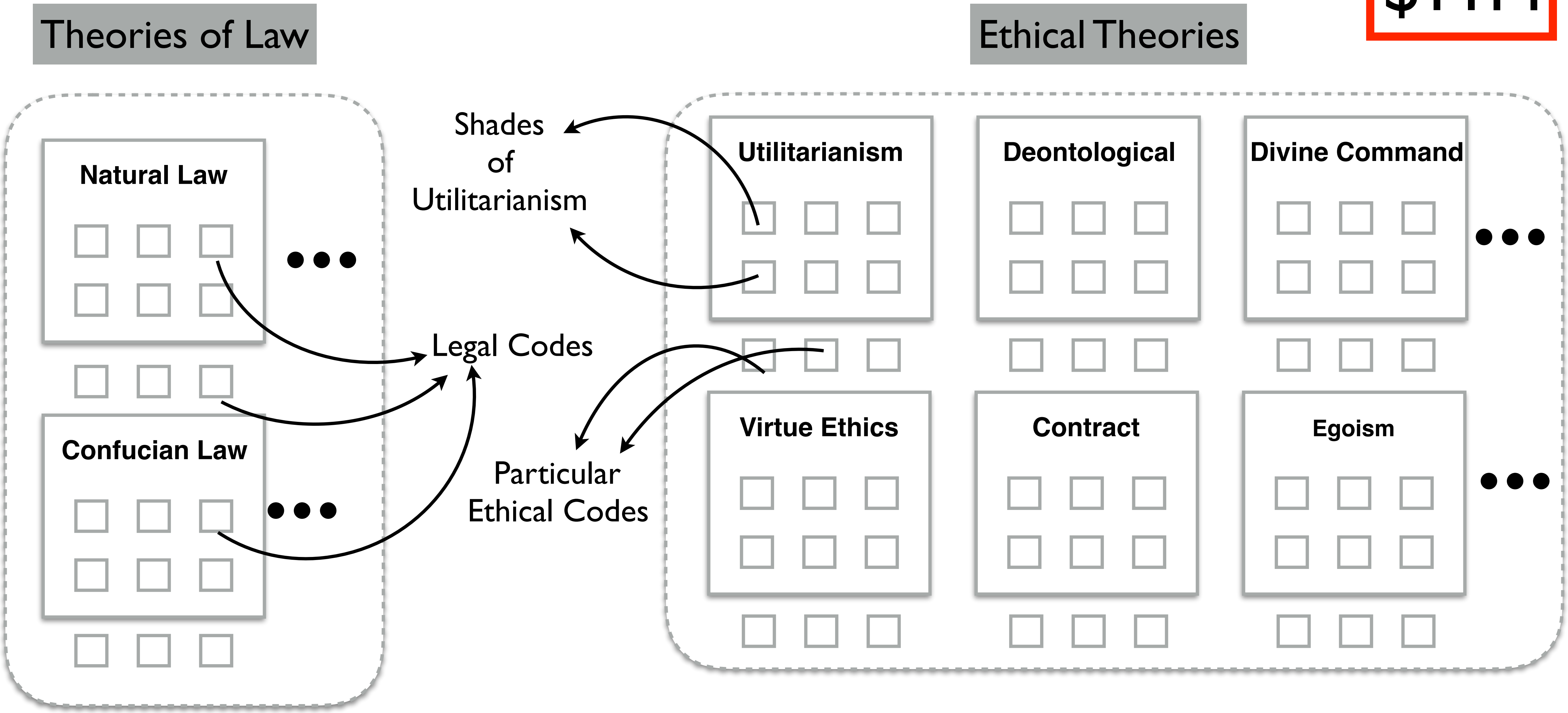
•••

Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Making Morally X Machines

$11M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

Egoism

Step 1

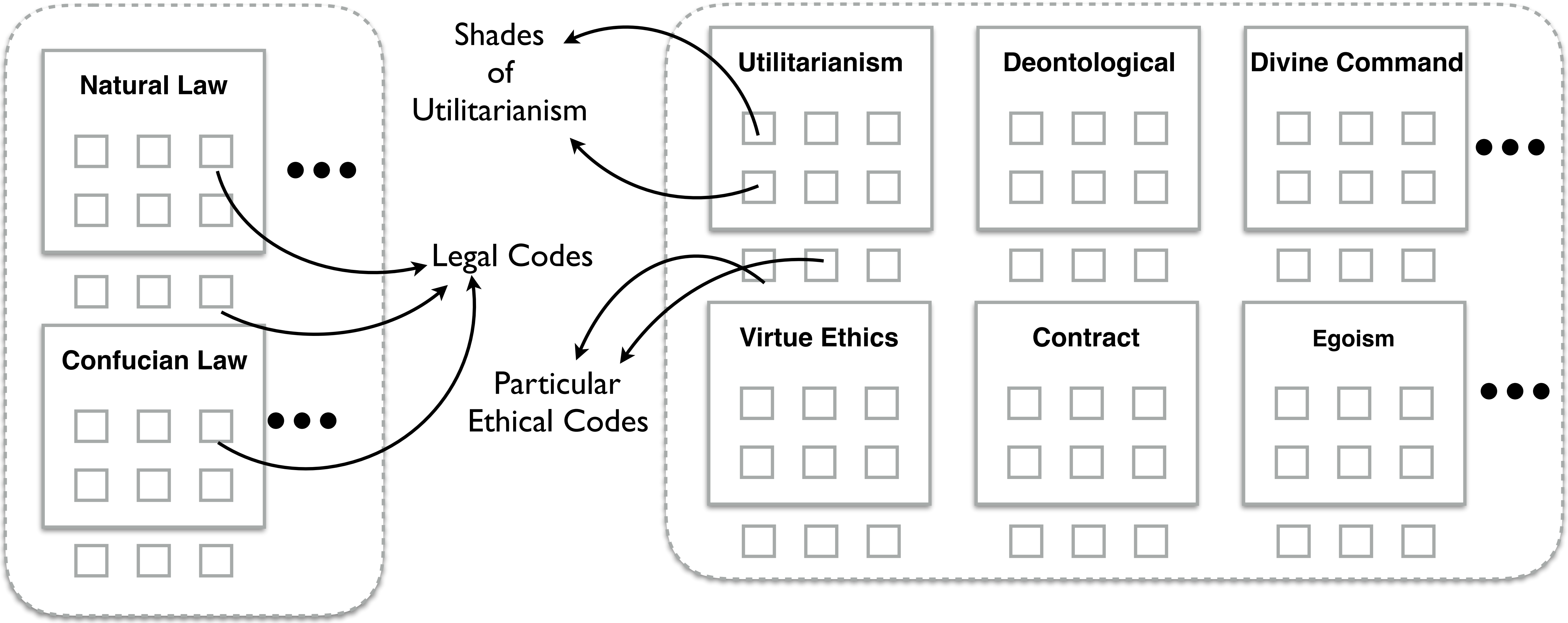
1. Pick a theory
2. Pick a code
3. Run through EH.

Making Morally X Machines

$11M

Theories of Law

Ethical Theories



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

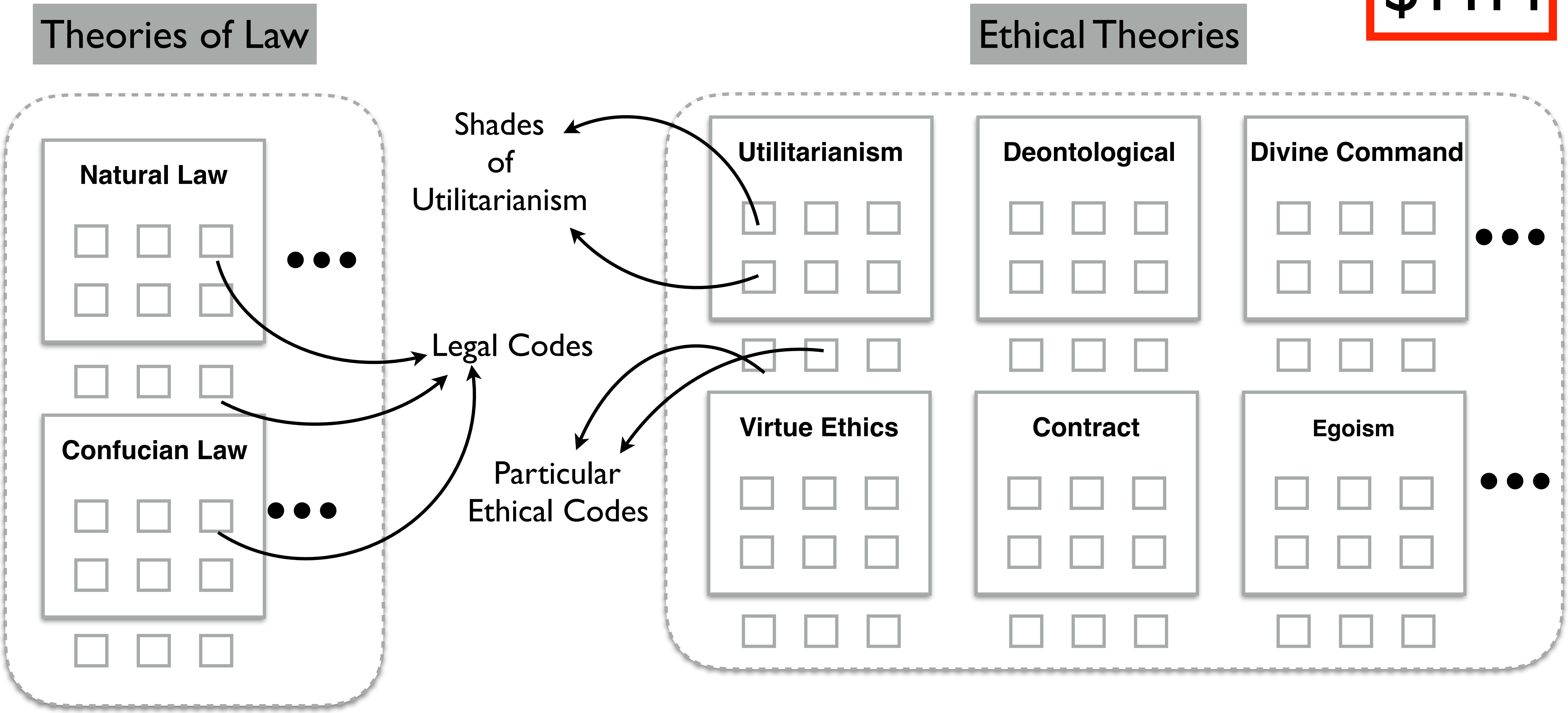
Automate

Prover

Spectra

Making Morally X Machines

$11M



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

Egoism

Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

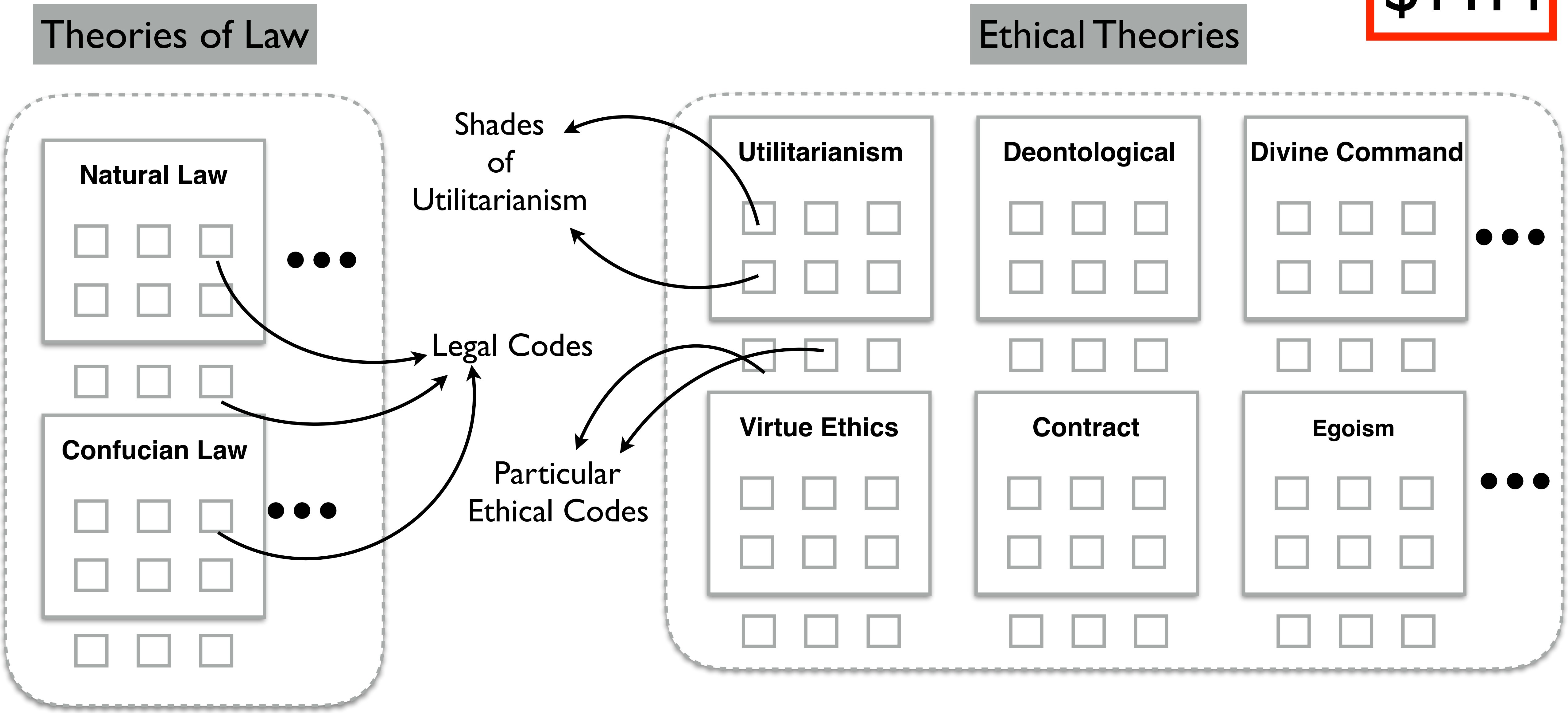
Automate

Prover

Spectra

Making Morally X Machines

$11M



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Utilitarianism

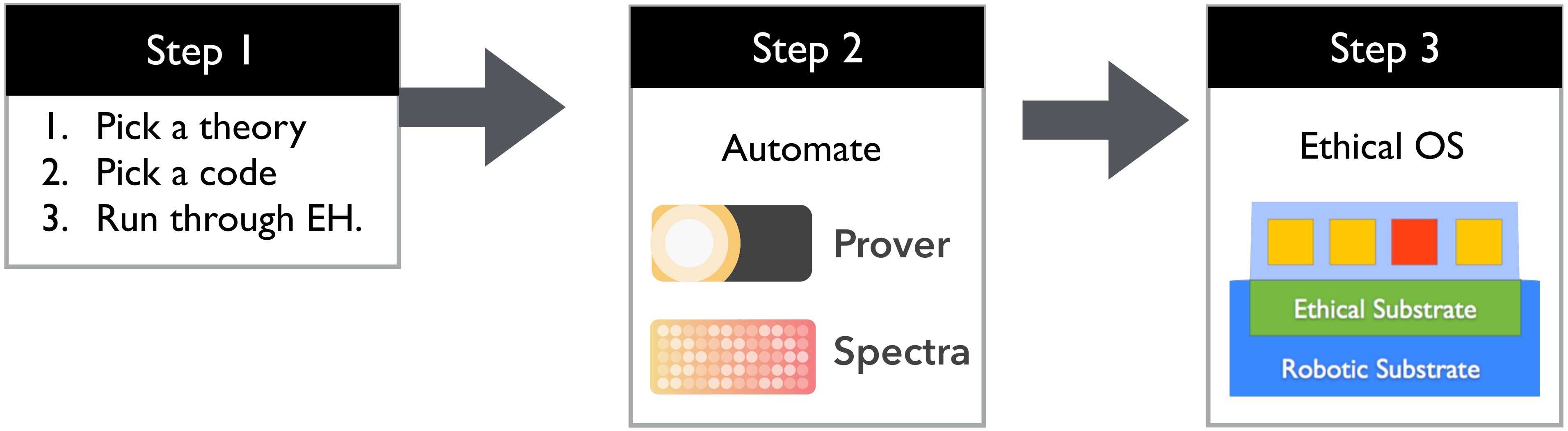
Deontological

Divine Command

Virtue Ethics

Contract

Egoism



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Prover

Spectra

Step 3

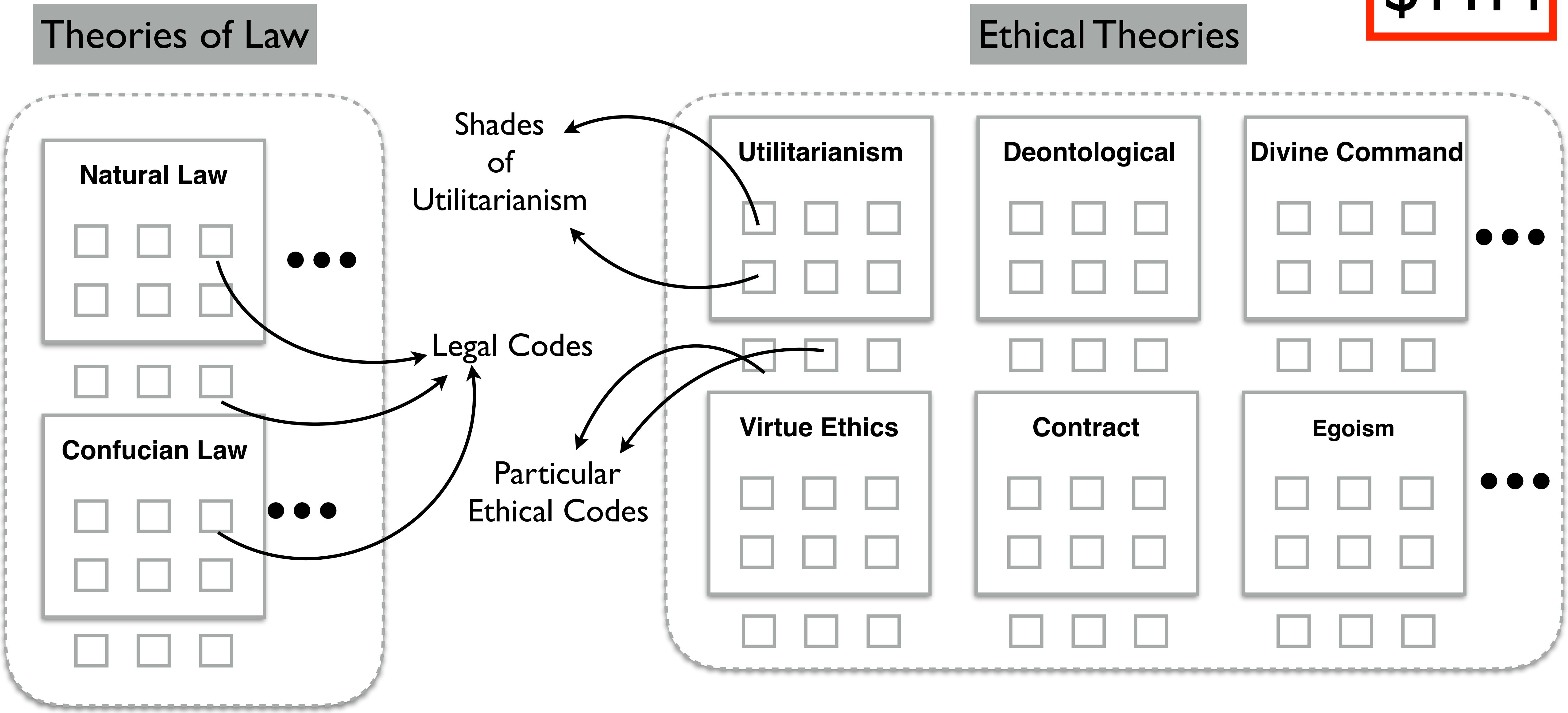
Ethical OS

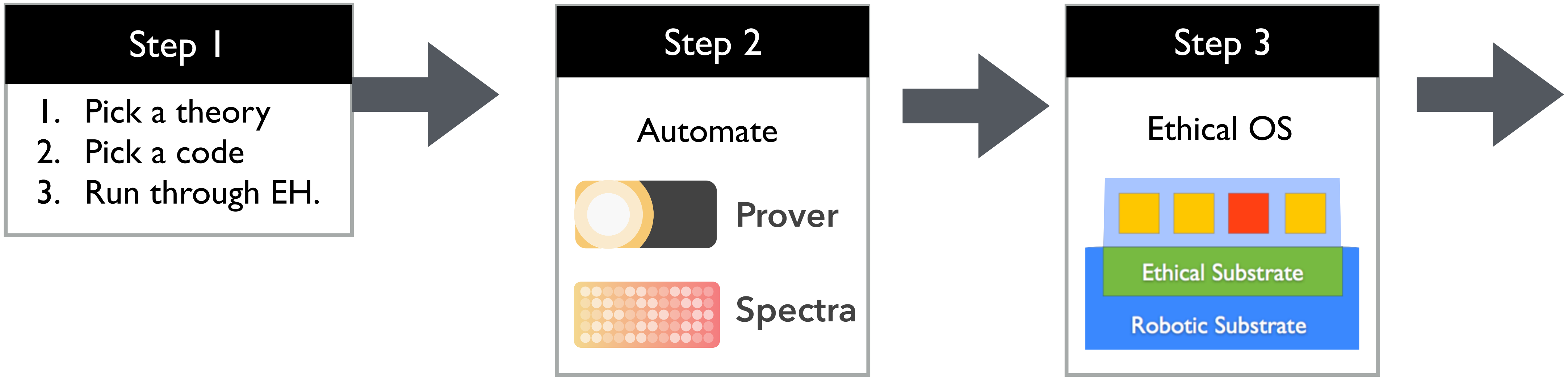
Ethical Substrate

Robotic Substrate

Making Morally X Machines

$11M




Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Prover

Spectra

Step 3

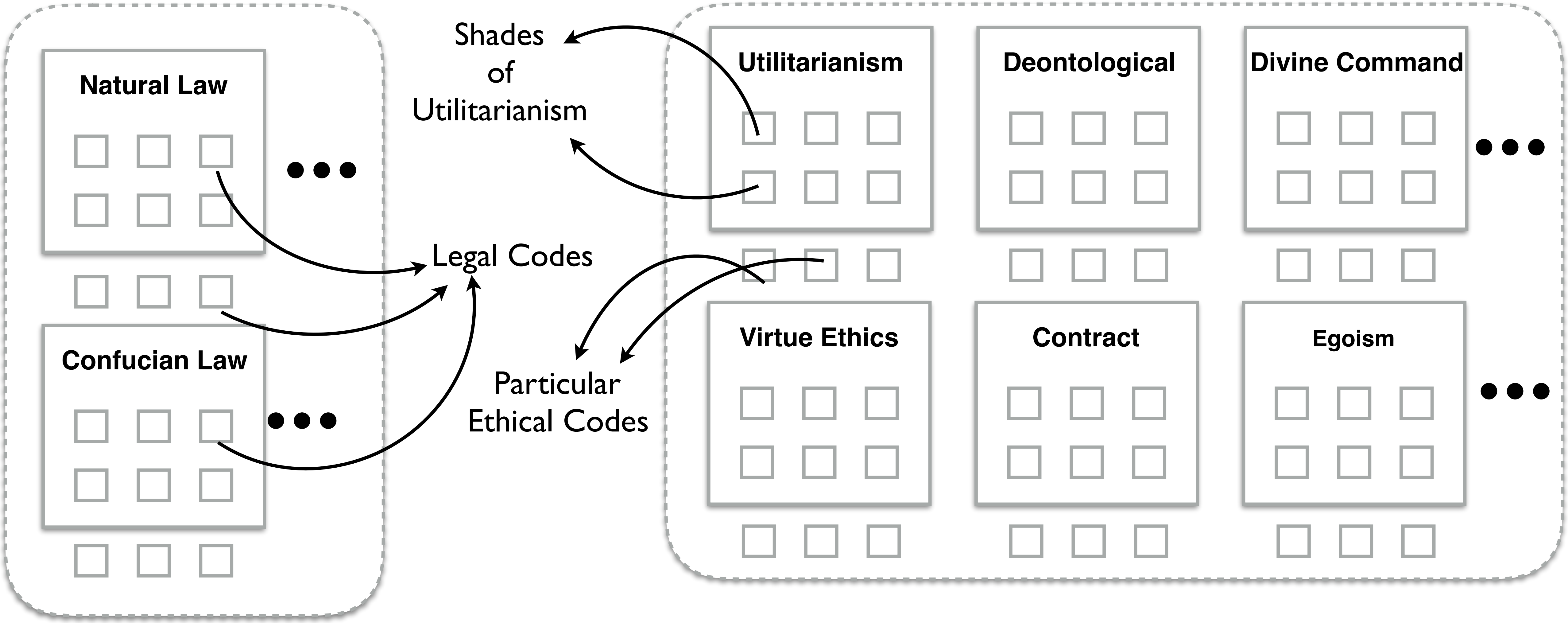
Ethical OS

Making Morally X Machines

$11M

Theories of Law

Ethical Theories



Natural Law

Confucian Law

Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

Utilitarianism

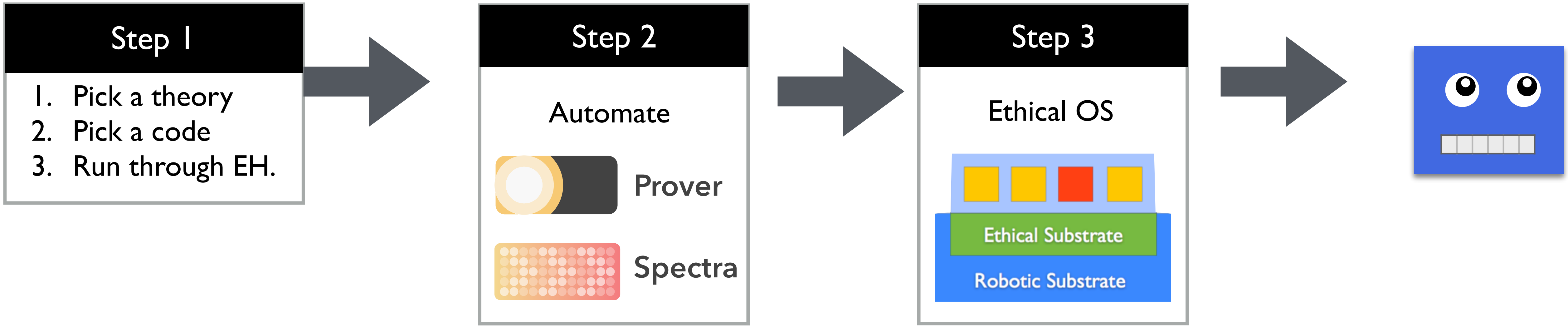
Deontological

Divine Command

Virtue Ethics

Contract

Egoism



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Prover

Spectra

Step 3

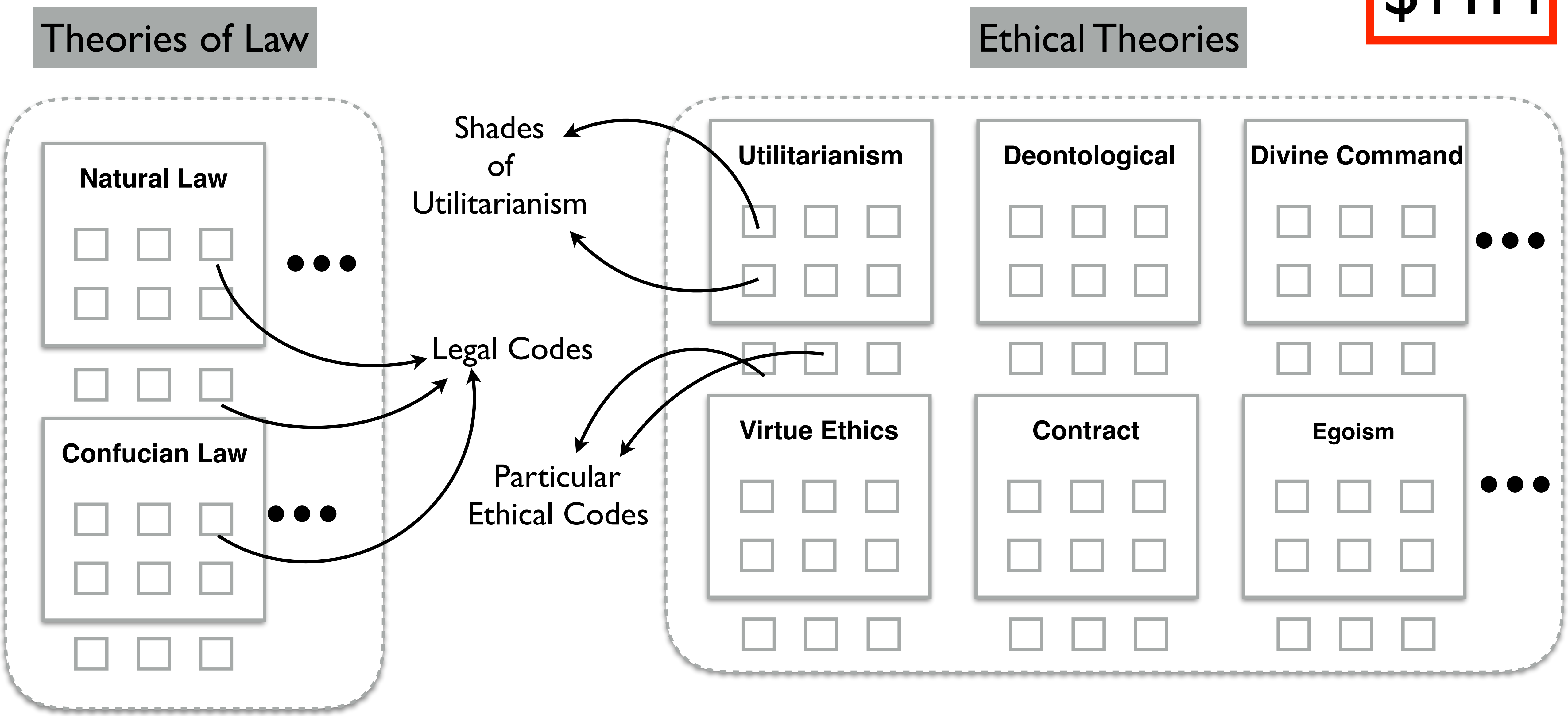
Ethical OS

Ethical Substrate

Robotic Substrate

Making Morally X Machines

$11M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

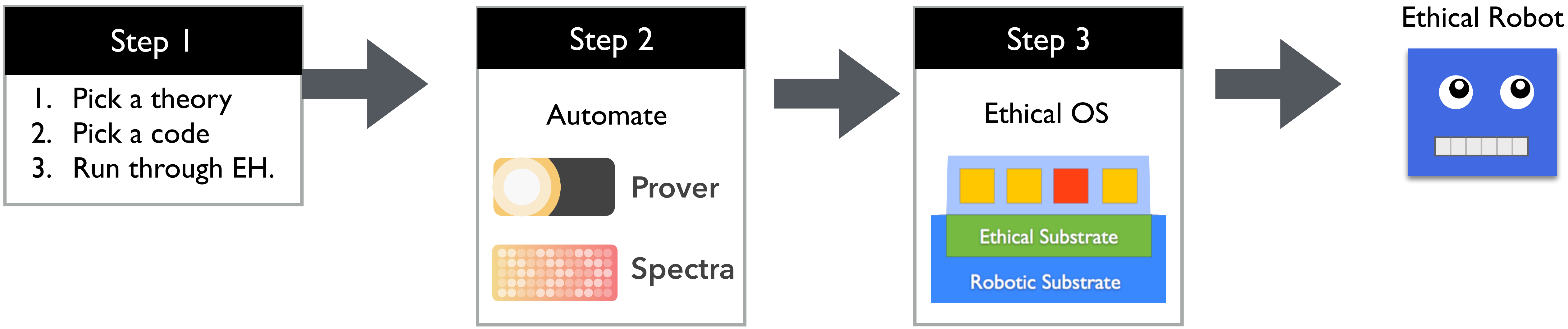
Deontological

Divine Command

Virtue Ethics

Contract

Egoism



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Prover

Spectra

Step 3

Ethical OS

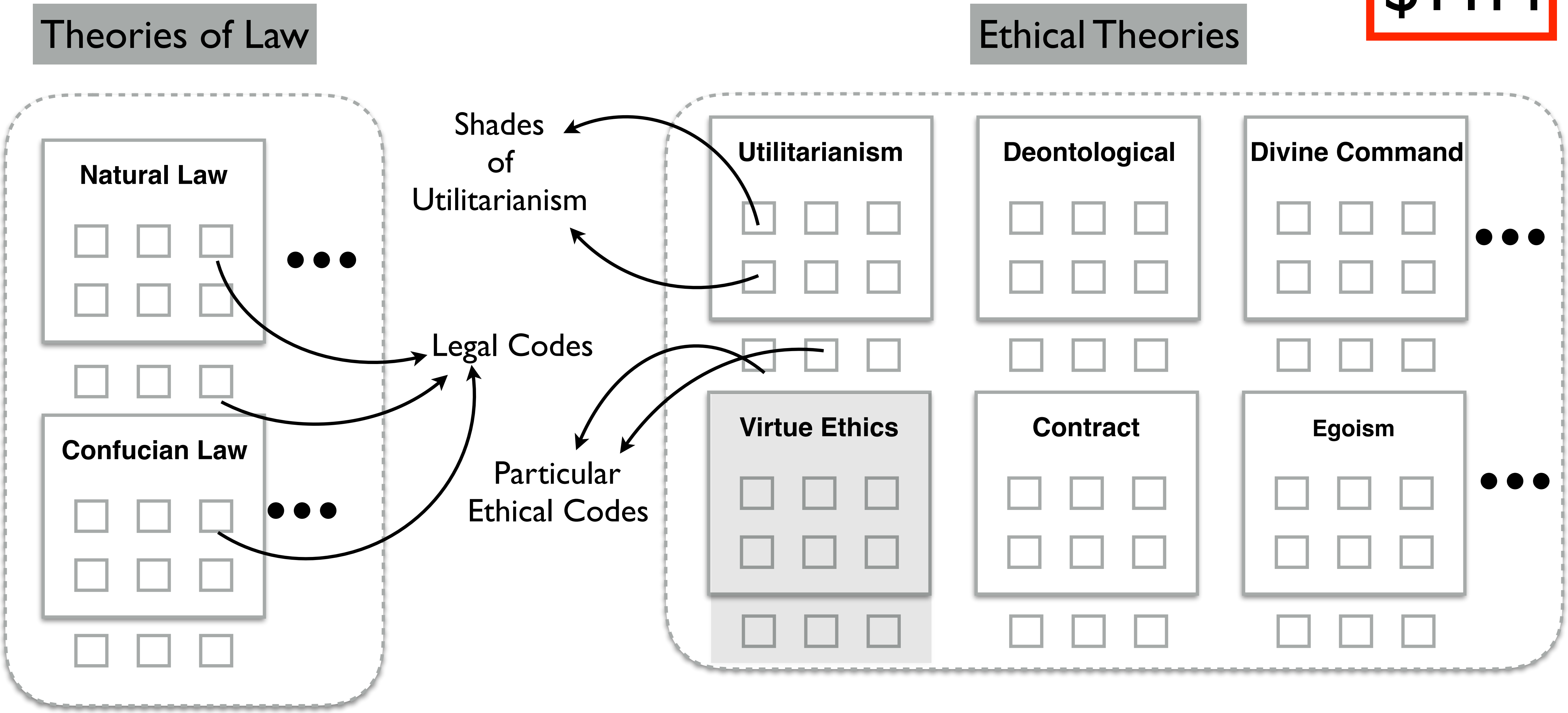
Ethical Substrate

Robotic Substrate

Ethical Robot

Making Morally X Machines

$11M



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Utilitarianism

Deontological

Divine Command

Virtue Ethics

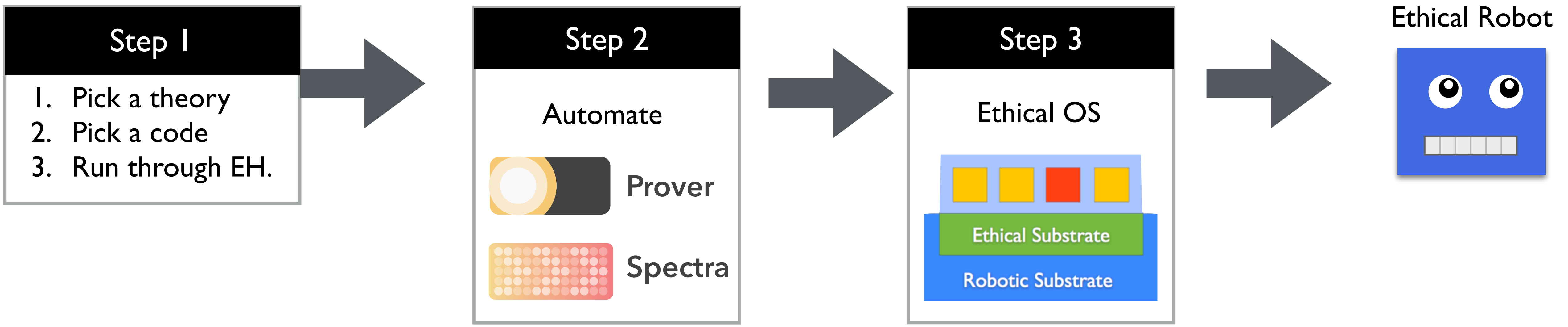
Contract

Egoism

Shades of Utilitarianism

Legal Codes

Particular Ethical Codes



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Prover

Spectra

Step 3

Ethical OS

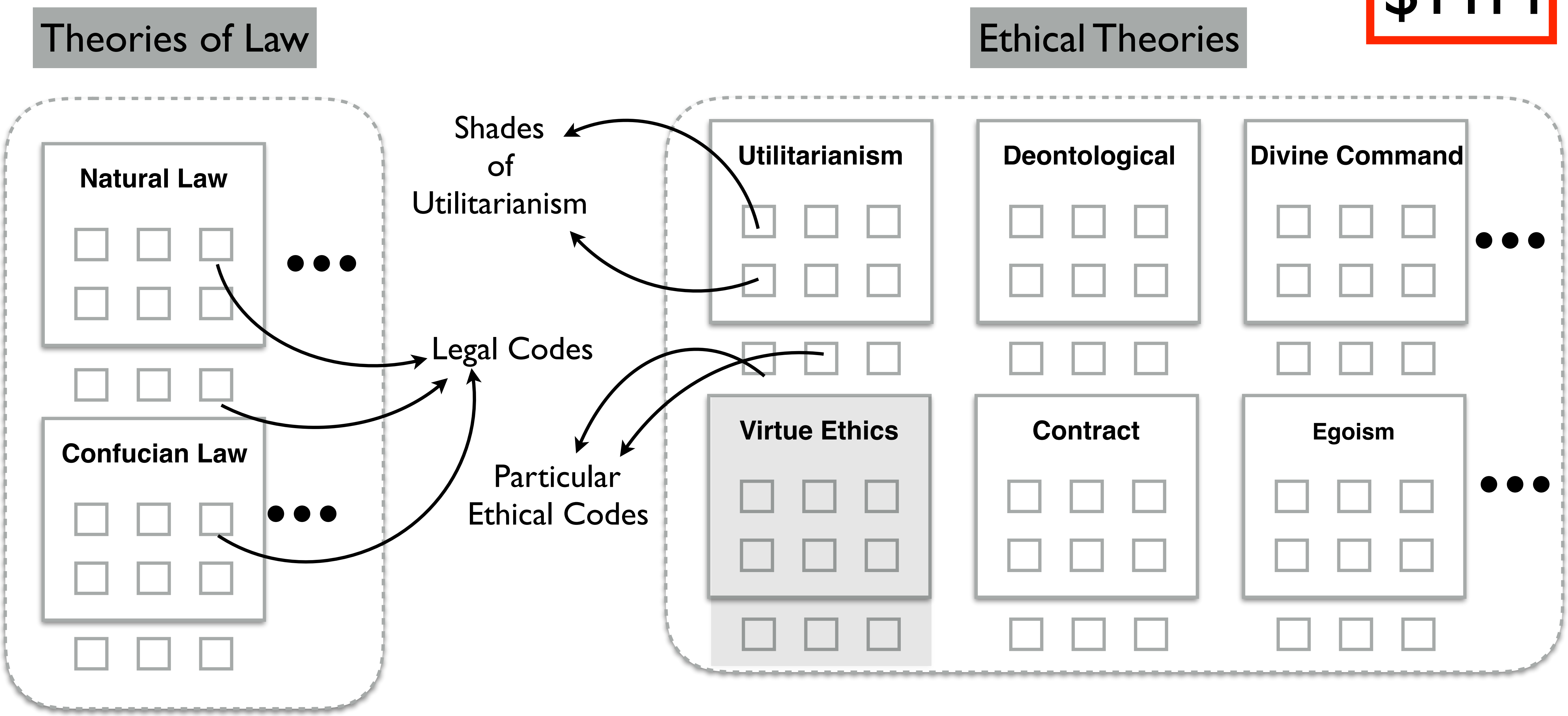
Ethical Substrate

Robotic Substrate

Ethical Robot

Making Morally X Machines

$11M



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Utilitarianism

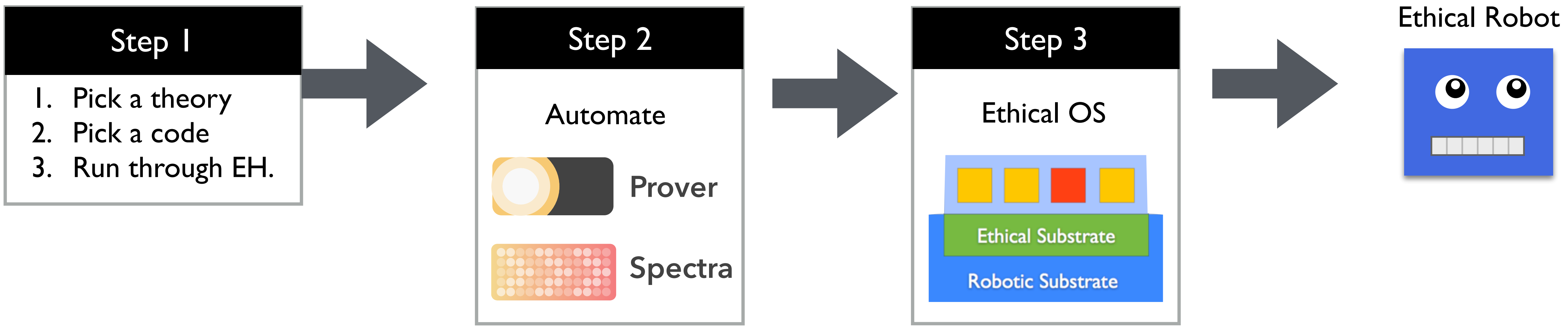
Deontological

Divine Command

Virtue Ethics

Contract

Egoism



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Prover

Spectra

Step 3

Ethical OS

Ethical Substrate

Robotic Substrate

Ethical Robot

“Toward the Engineering of Virtuous Robots” Naveen, Selmer et al.

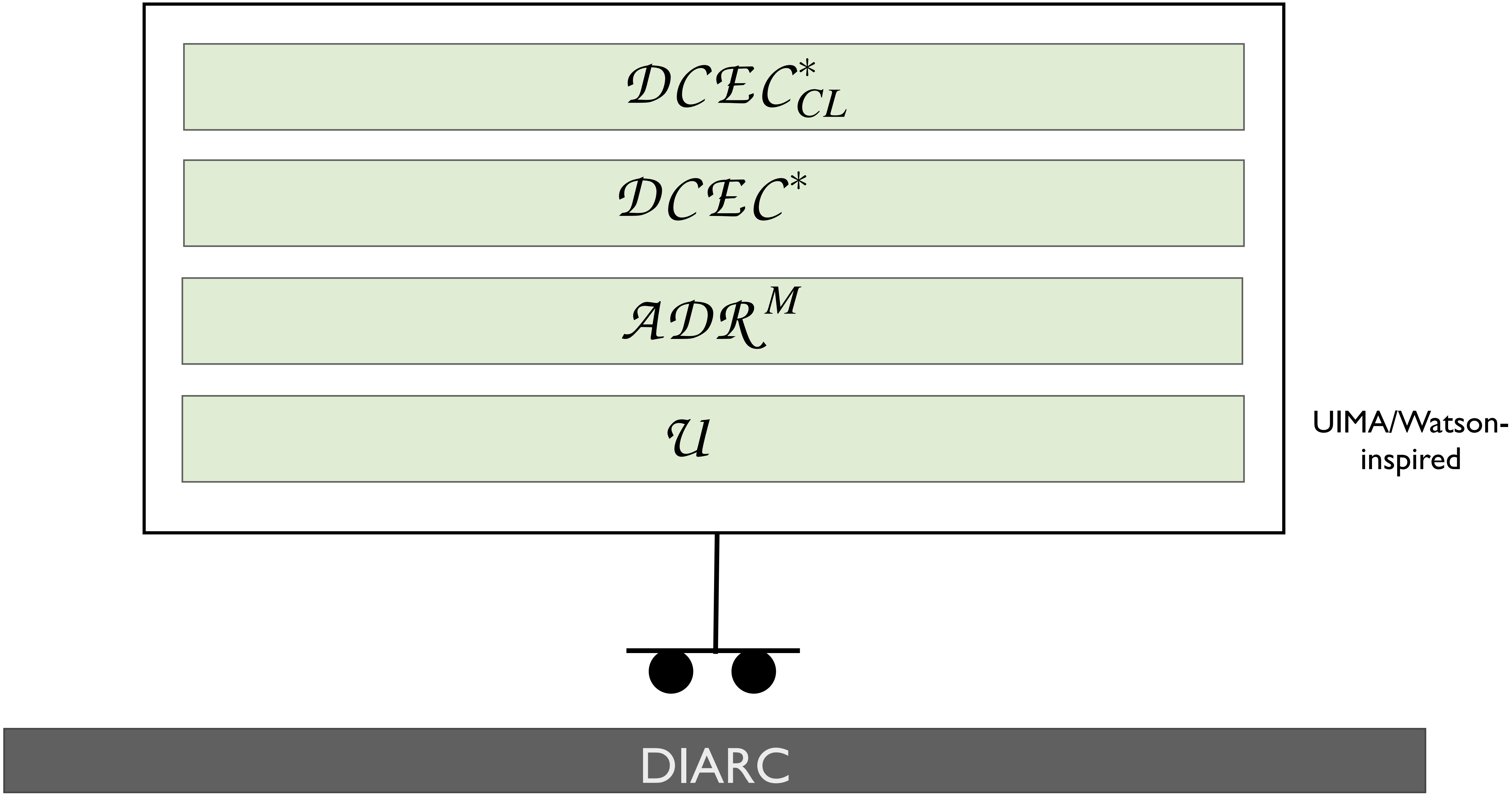
Conclusion from last time:

Conclusion from last time:

“Computational logician, sorry, back to your drawing board to find a logic that works with The Four Steps!”

Not **D**, but rather *cognitive calculi* …

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

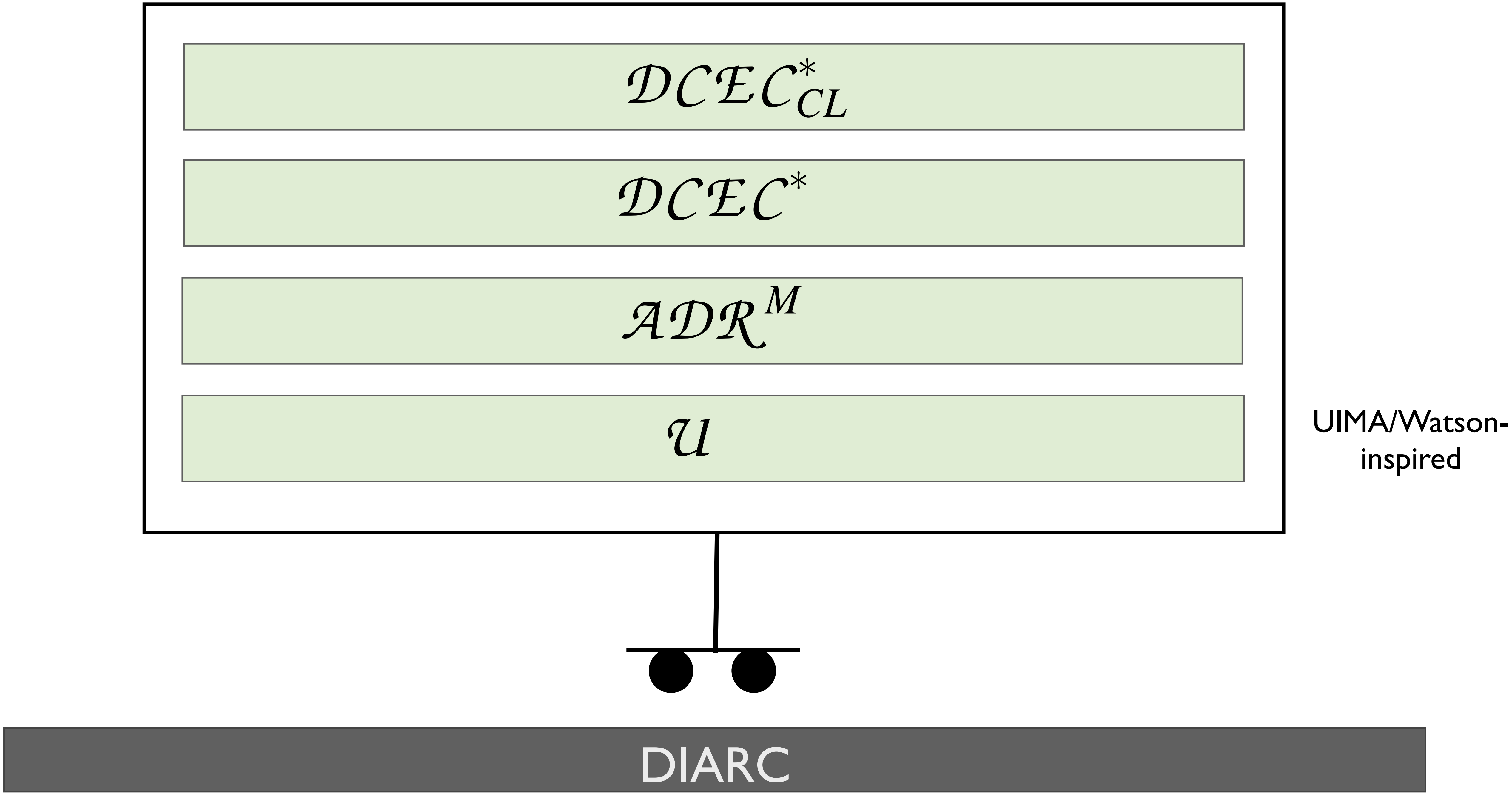
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-inspired

DIARC

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

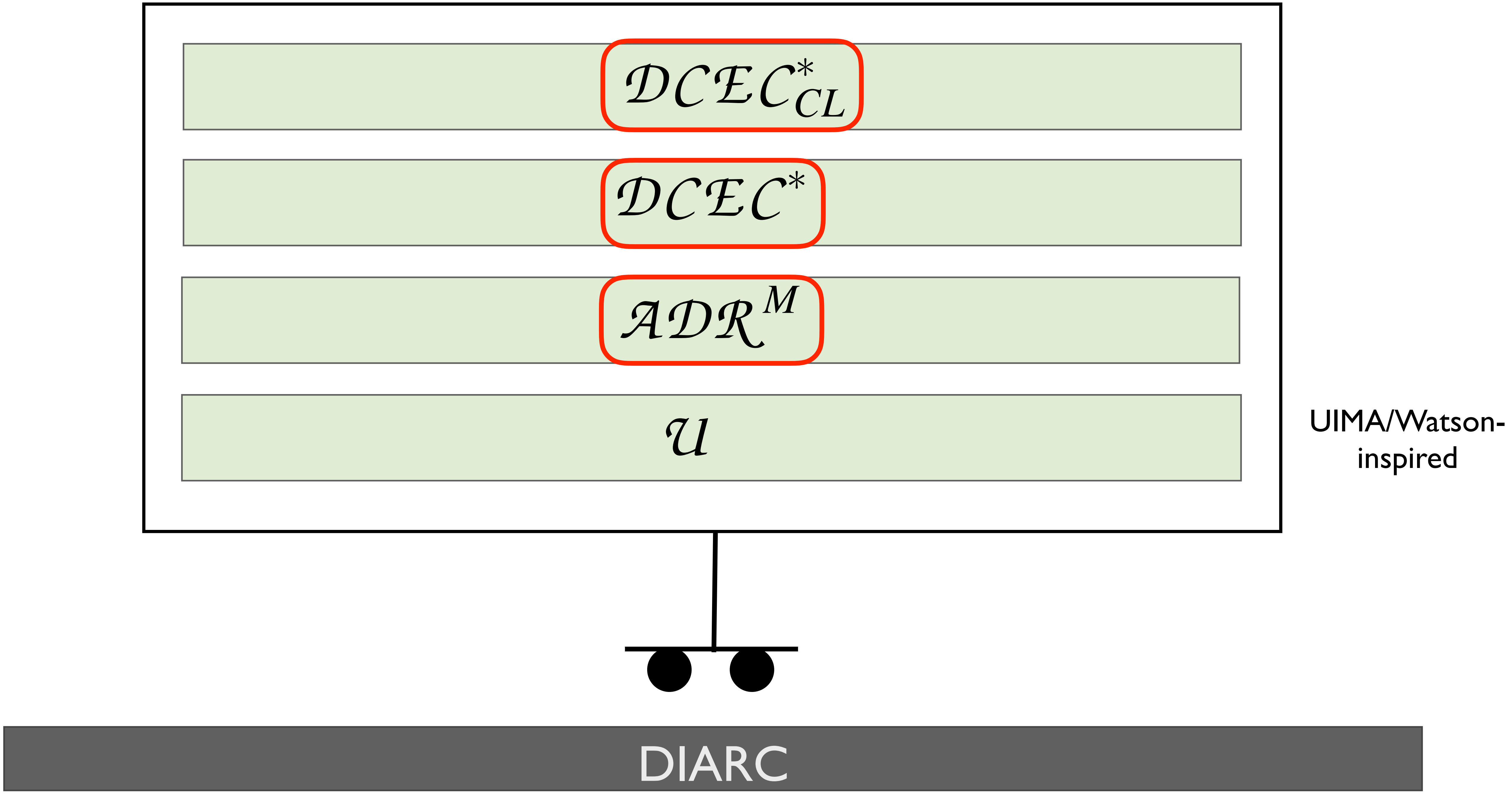
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-inspired

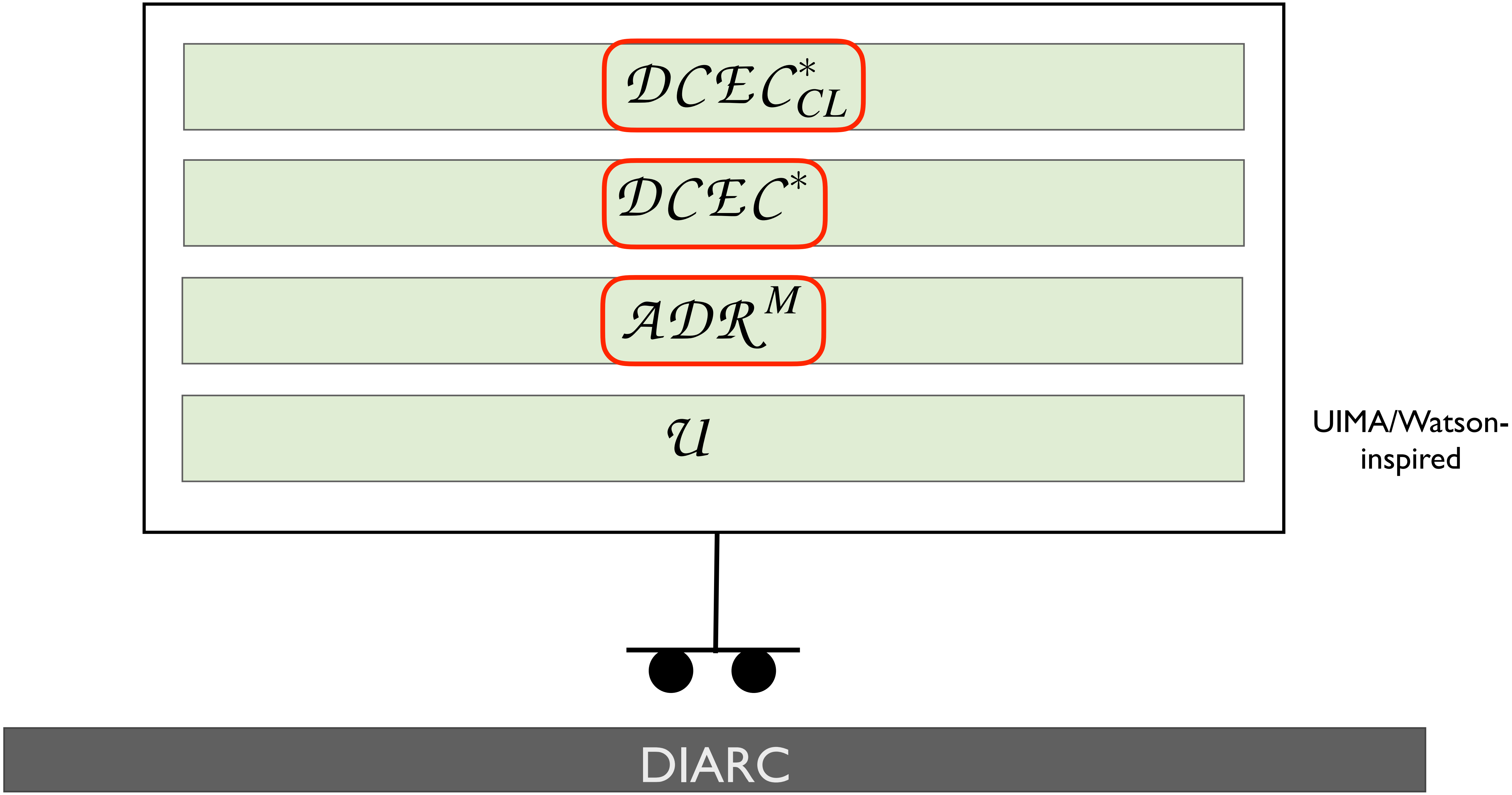
DIARC

Hierarchy of Ethical Reasoning



Hierarchy of Ethical Reasoning

Not simple deontic logics like **D**!



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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-
inspired

DIARC

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 . . .

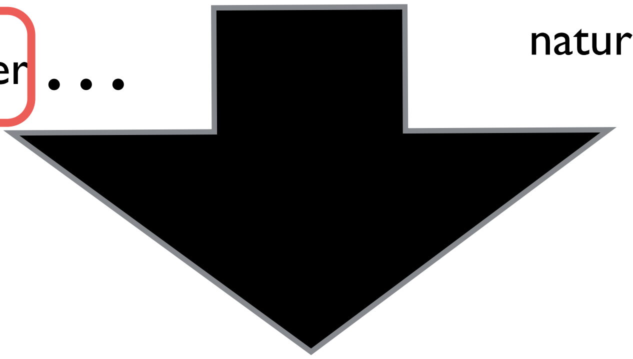
intensional level: epistemic deontic possibility/necessity . . .

ATPs: SPASS SNARK ShadowProver . . .

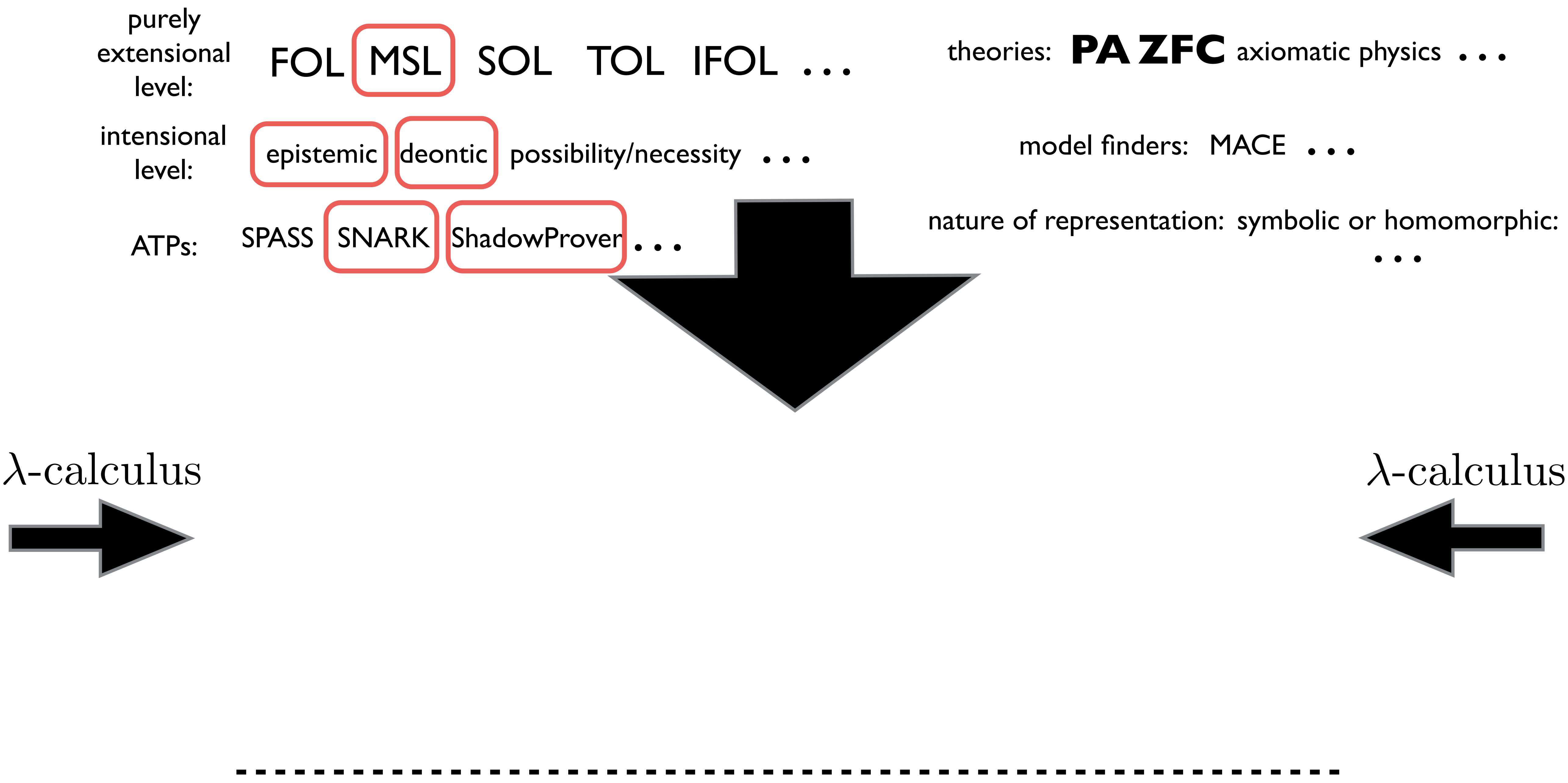
theories: **PA ZFC** axiomatic physics . . .

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 . . .

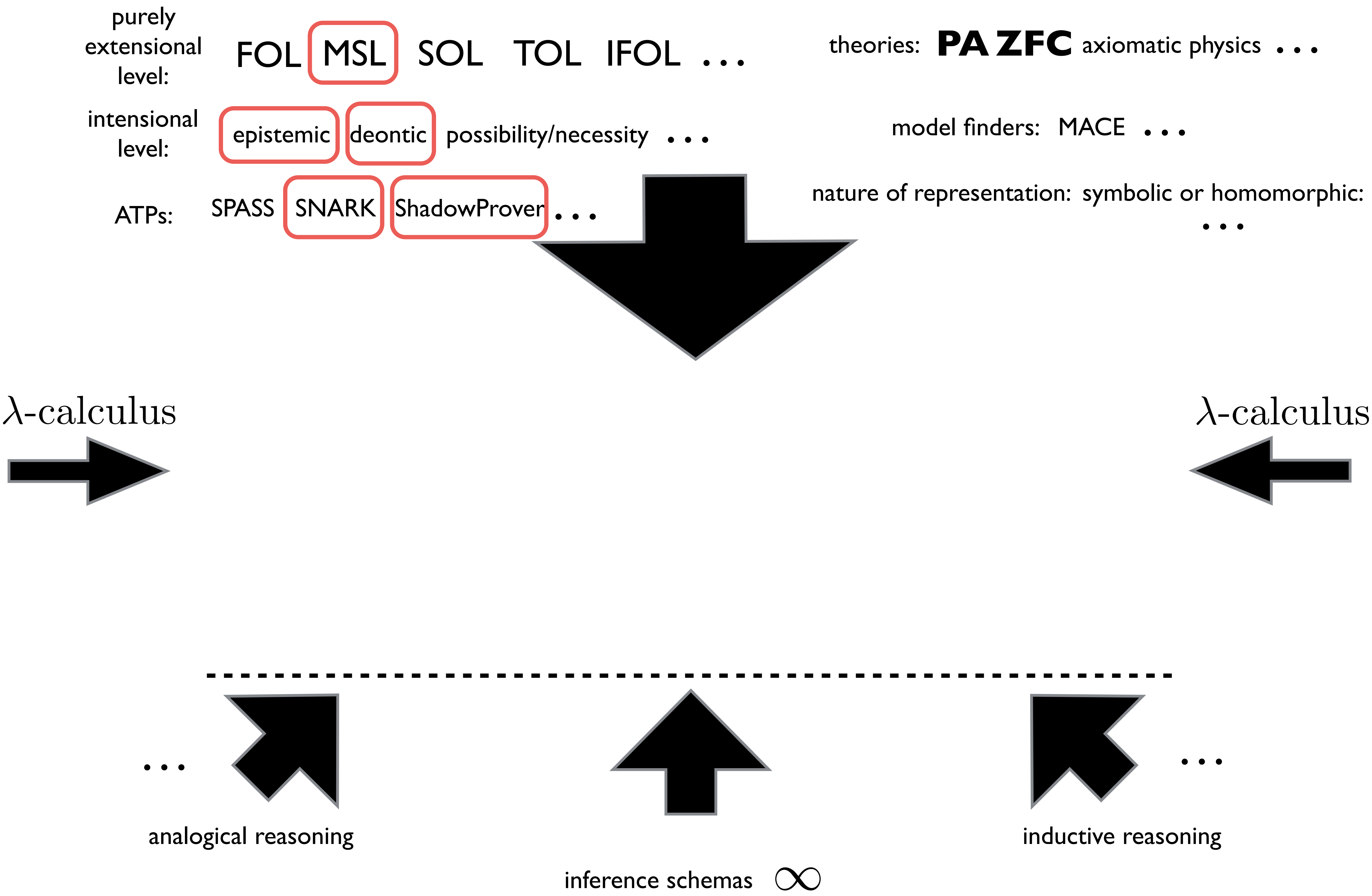
model finders: MACE . . .

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

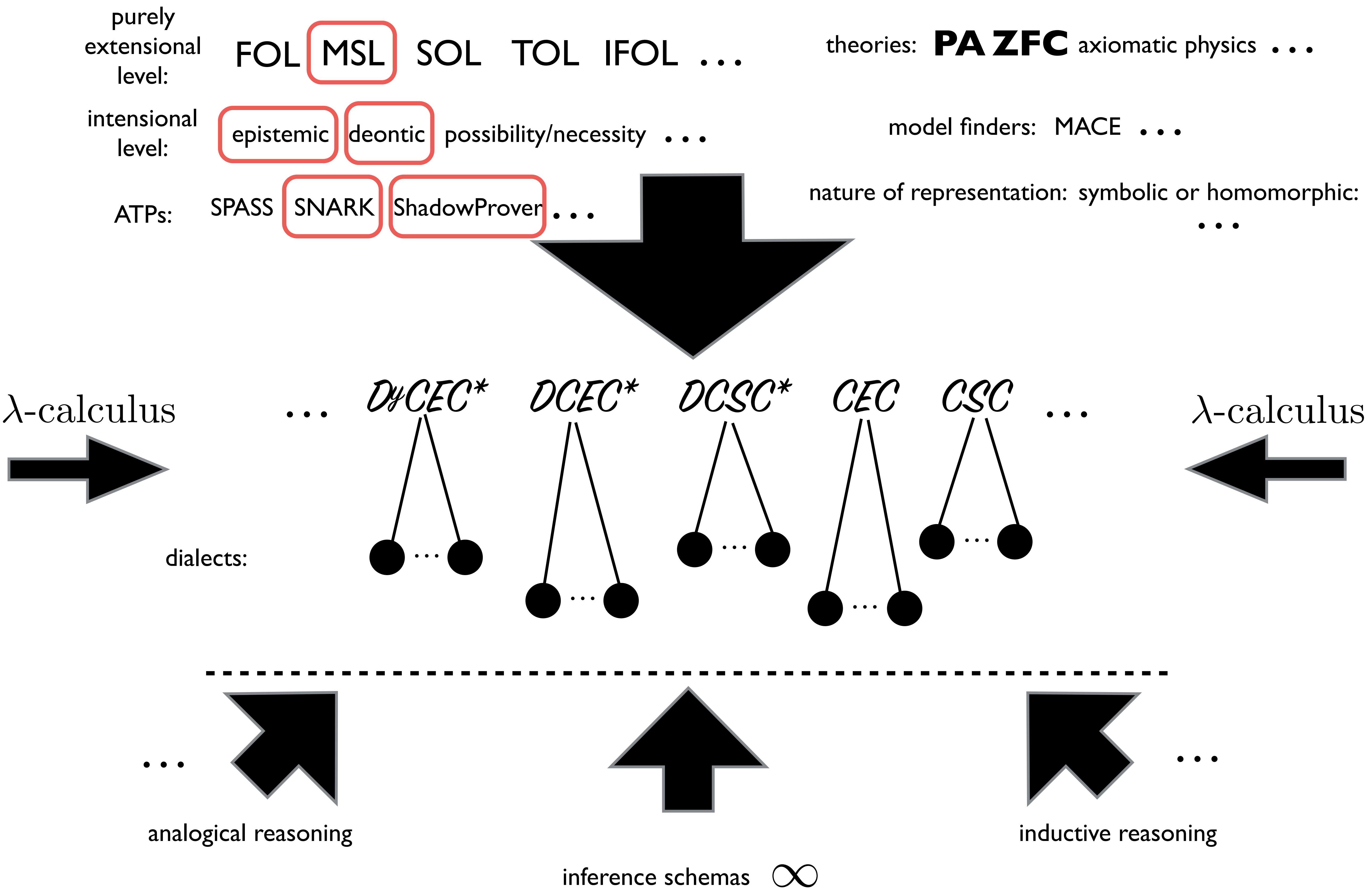
λ-calculus

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

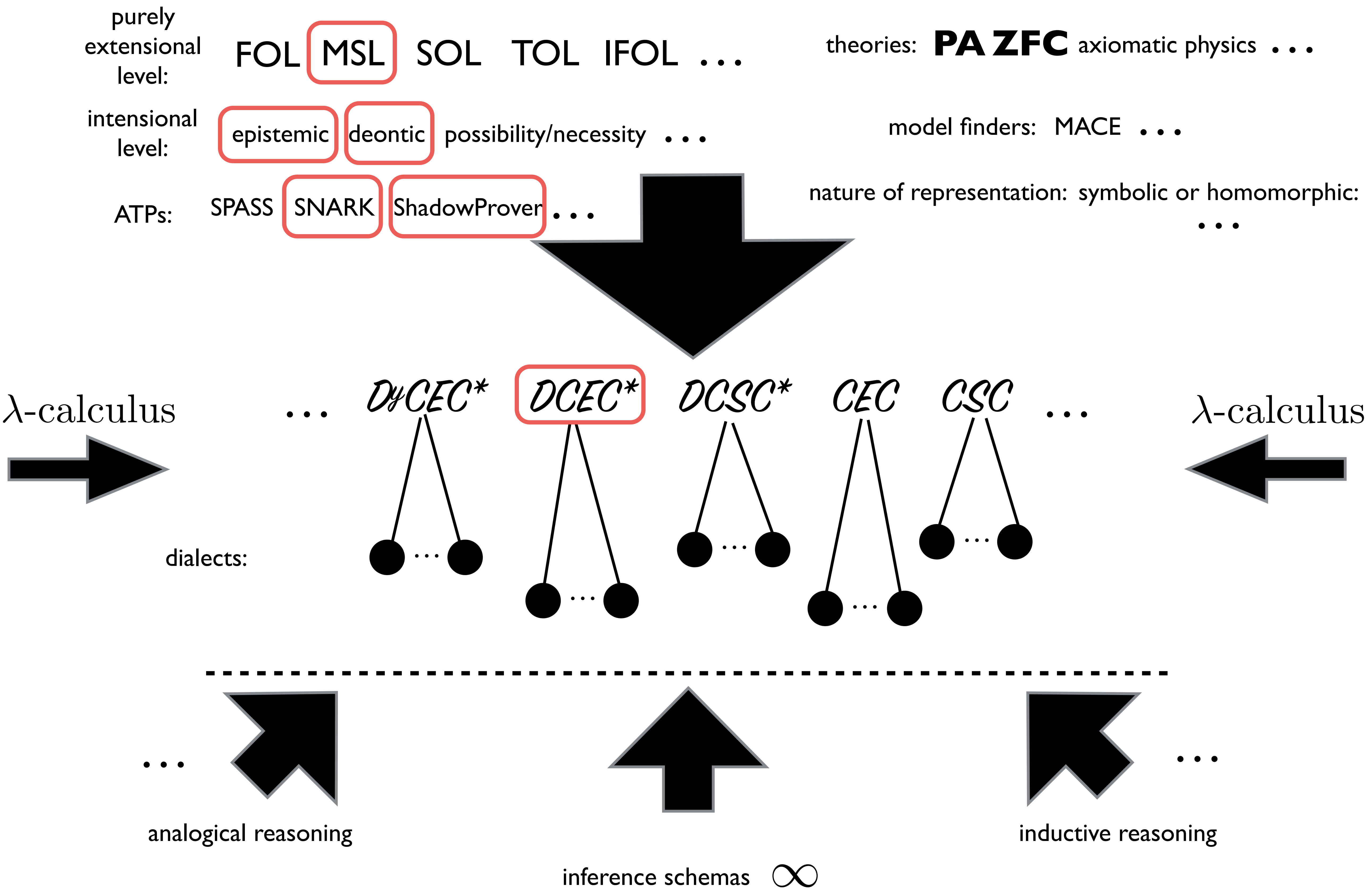
dialects:

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

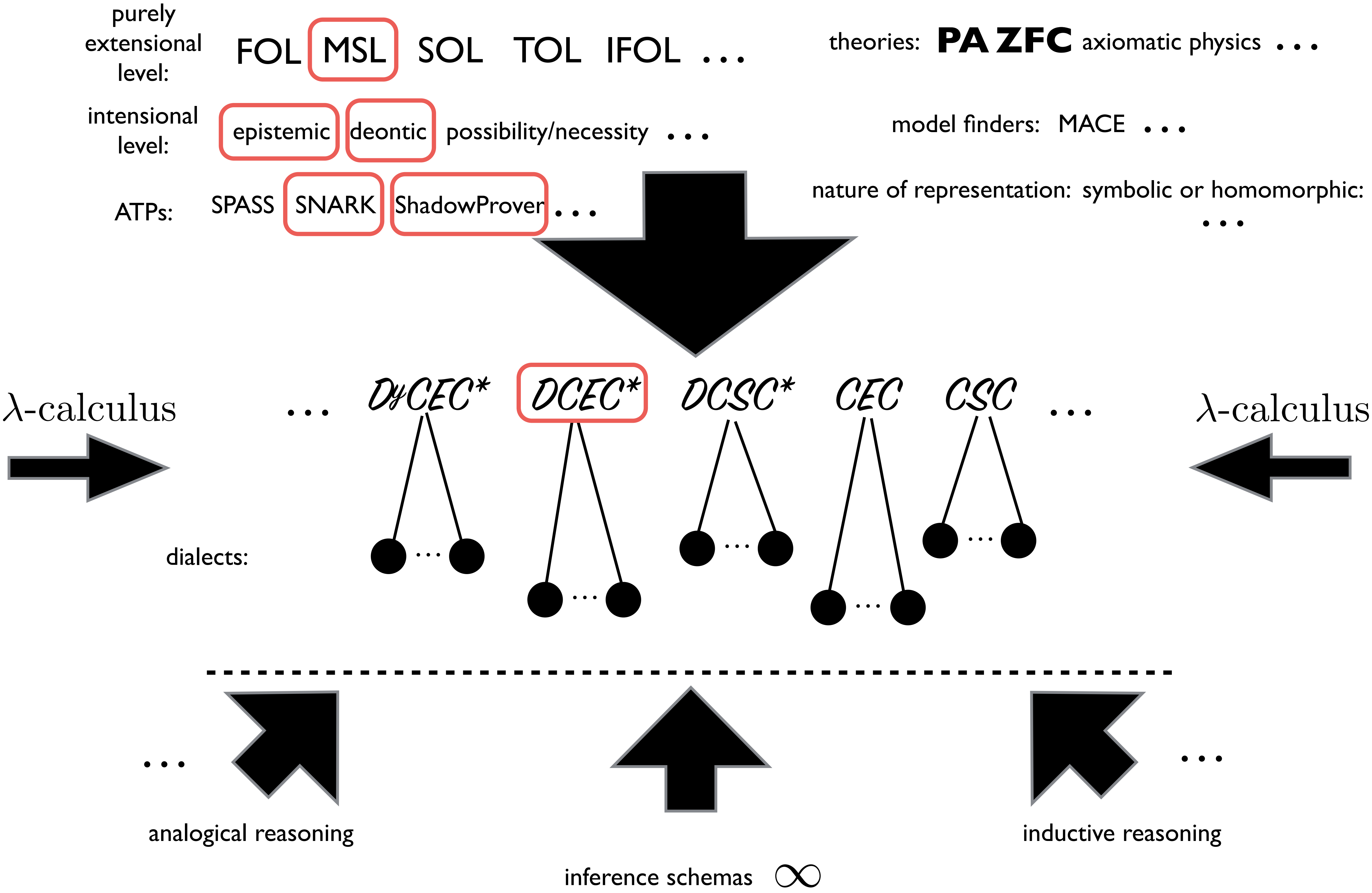
dialects:

. . . 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 . . . 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

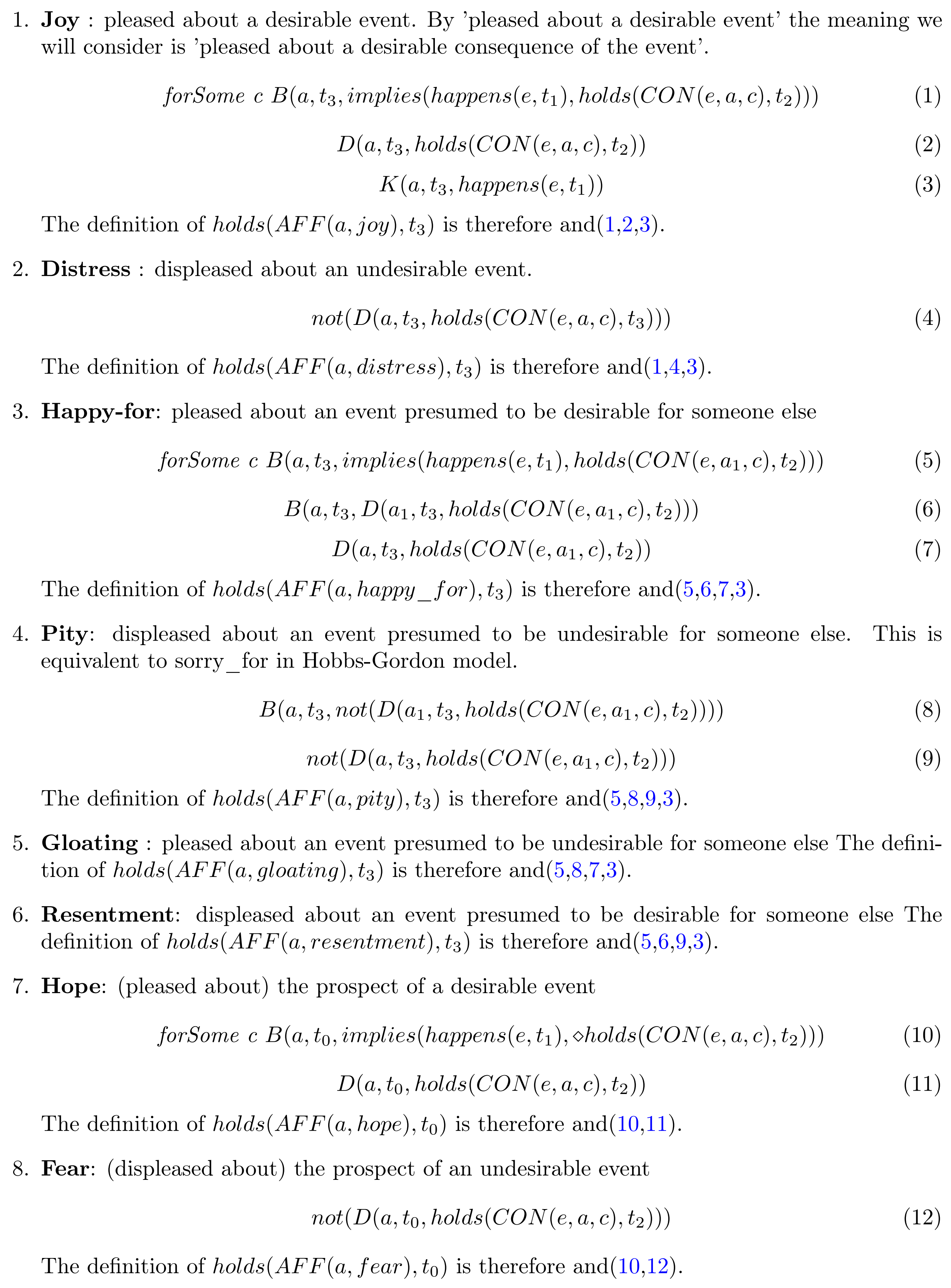
$[A_2]$ $\mathbf{C}(\forall e,f,t_1,t_2\ .\ happens(e,t_1)\wedge initiates(e,f,t_1)\wedge t_1<t_2\wedge\neg clipped(t_1,f,t_2)\Rightarrow holds(f,t_2))$

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

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

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

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'.

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

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

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

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

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

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

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

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

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

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

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

The definition of $holds(AFF(a, happy\_for), t_3)$ 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.

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

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

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

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

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

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

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

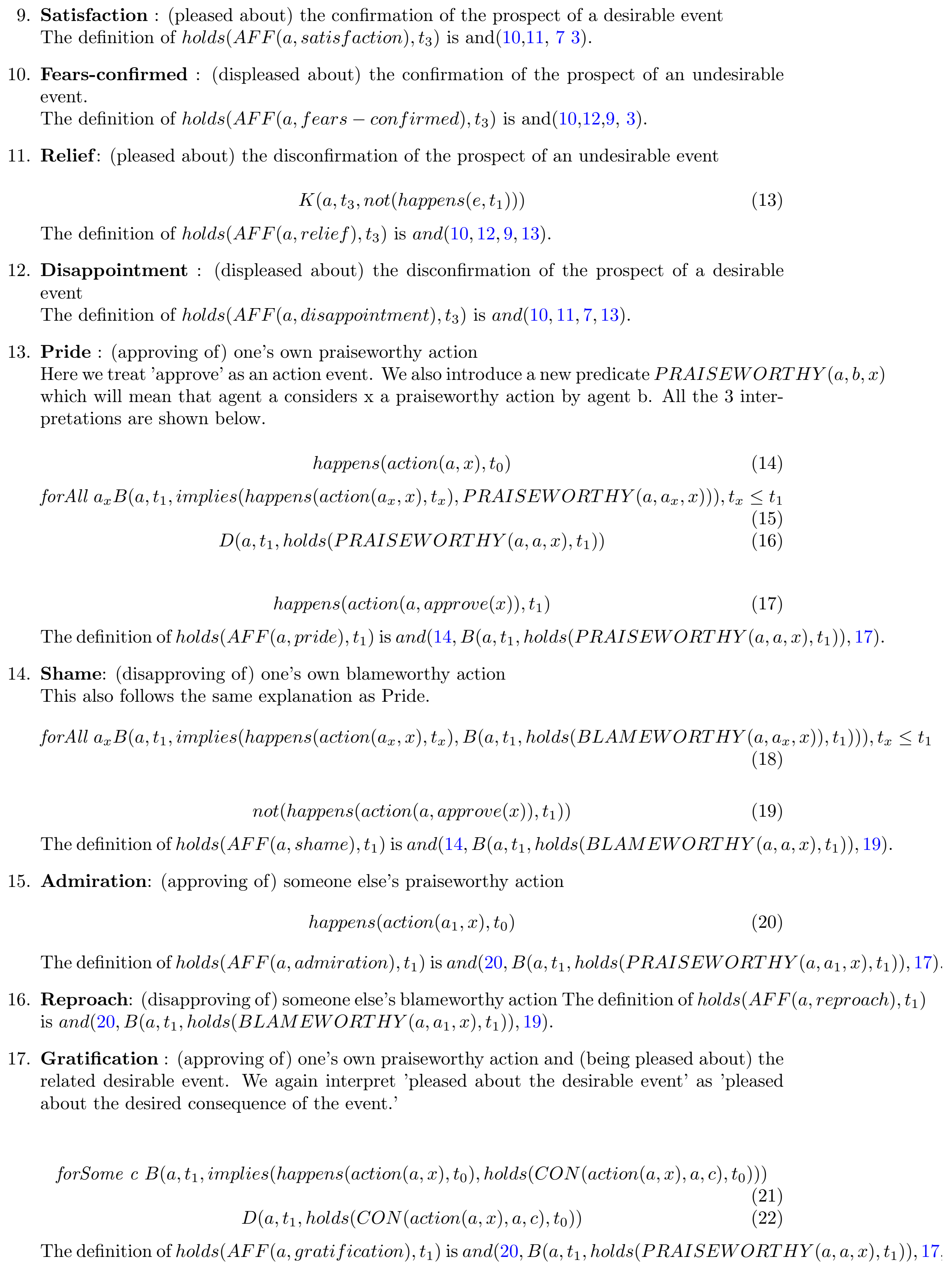
$$D(a, t_0, holds(CON(e, a, c), t_2)) \tag{11}$$

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

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

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

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



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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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.'

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

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

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

… (and more)

II.

Early Progress With Our Calculi: Non-Akratic Robots

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 $\boxed{\alpha_f}$ 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 $\boxed{\alpha_o}$ 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 $\boxed{\text{does the forbidden action}}$ α_f at t_{α_f};

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

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

Informal Definition of Akrasia

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

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

(2) A desires to do α_f at t_{α_f};

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

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

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

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

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

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

Cast in

$\mathcal{DCEC}^*$

this becomes …

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

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

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

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

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

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

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

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

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

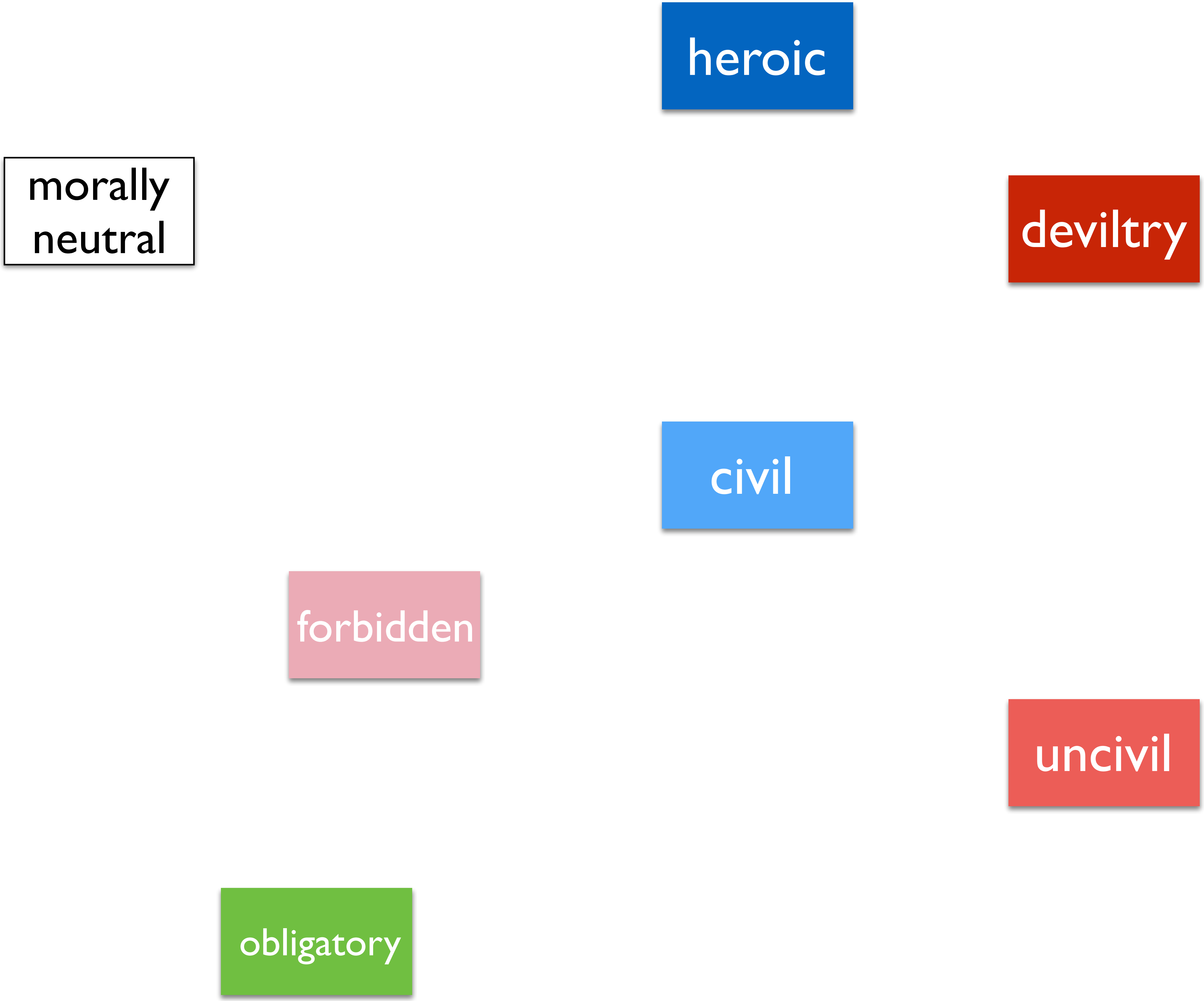
$$D_8 : \mathbf{B}(\mathsf{I}, t_f, \mathbf{O}(\mathsf{I}^*, t_\alpha, \Phi, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha)))$$

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

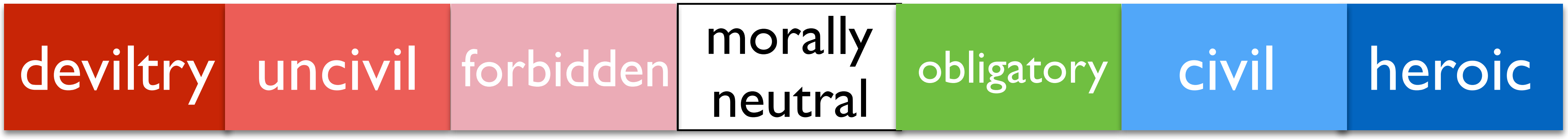
civil

forbidden

uncivil

obligatory

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



deviltry

uncivil

forbidden

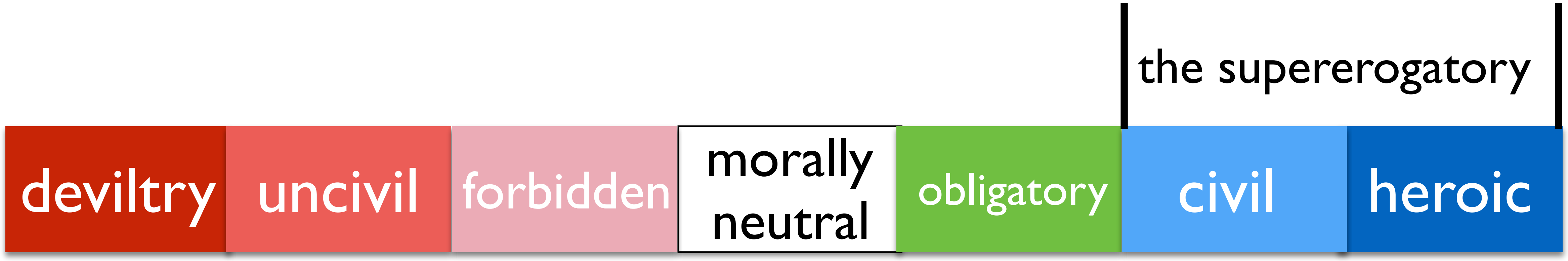
morally neutral

obligatory

civil

heroic

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



the supererogatory

deviltry

uncivil

forbidden

morally neutral

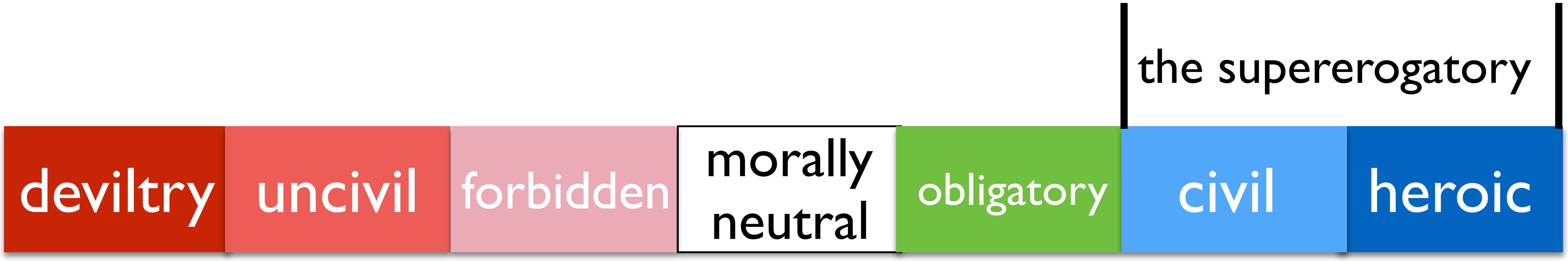
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathscr{EH}$

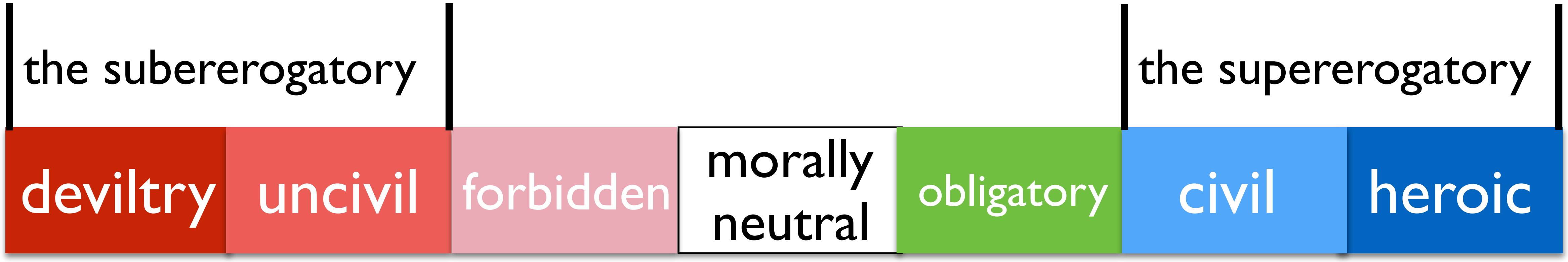


the supererogatory

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

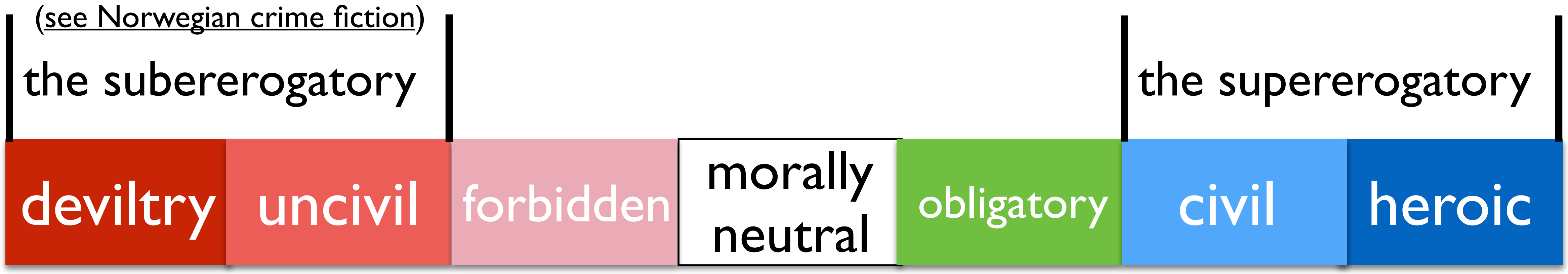
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



(see Norwegian crime fiction)

the subererogatory

the supererogatory

deviltry

uncivil

forbidden

morally neutral

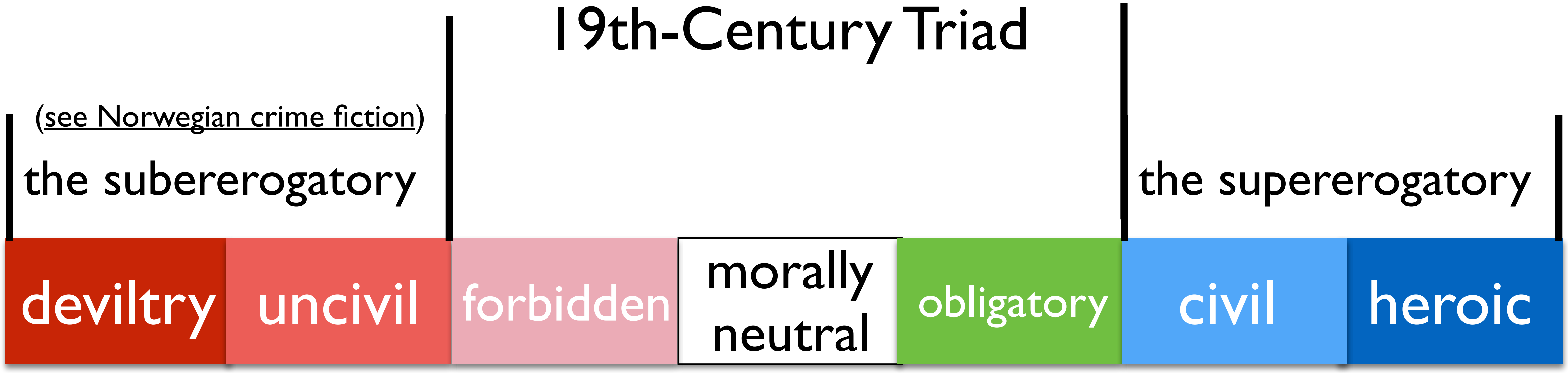
obligatory

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

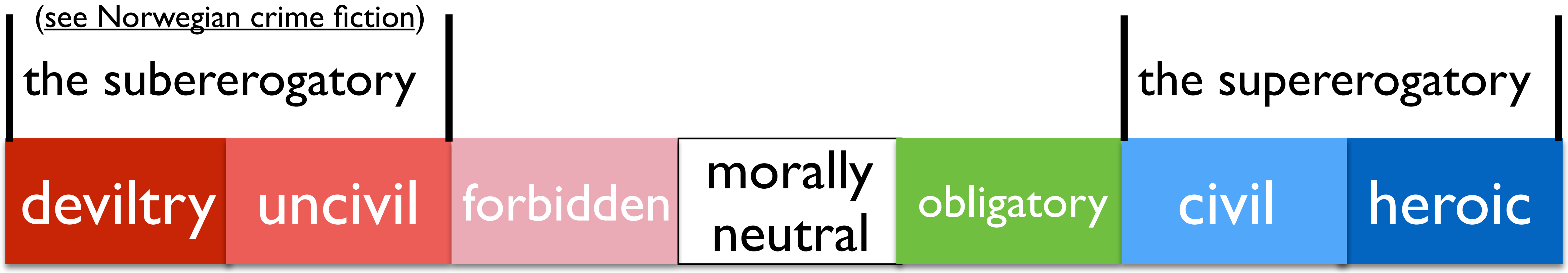
obligatory

civil

heroic

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$



(see Norwegian crime fiction)

the suberogatory

the supererogatory

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

Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$

(see Norwegian crime fiction)

the subererogatory

the supererogatory

devilty	uncivil	forbidden	morally neutral	obligatory	civil	heroic

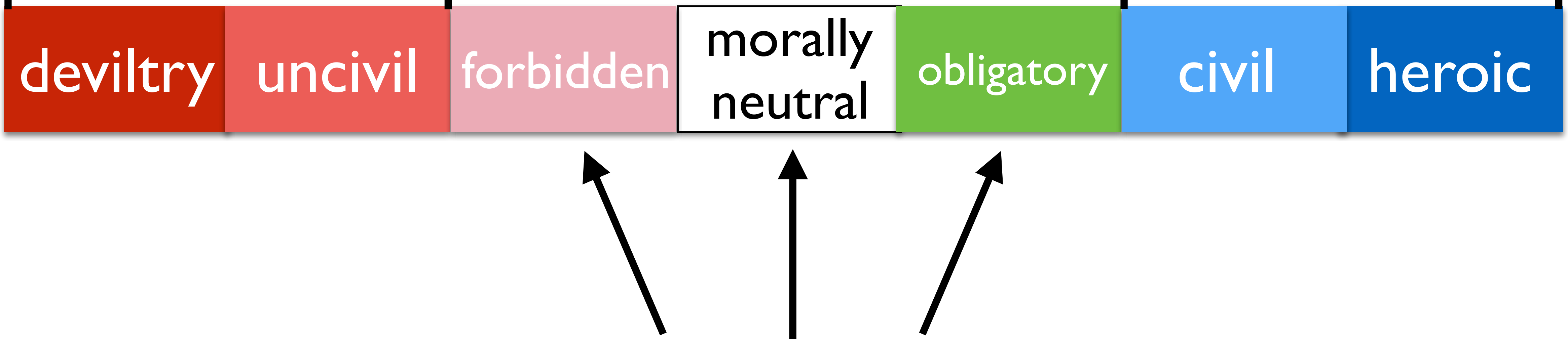
Leibnizian Ethical Hierarchy for Persons and Robots:

$\mathcal{EH}$

(see Norwegian crime fiction)

the subererogatory

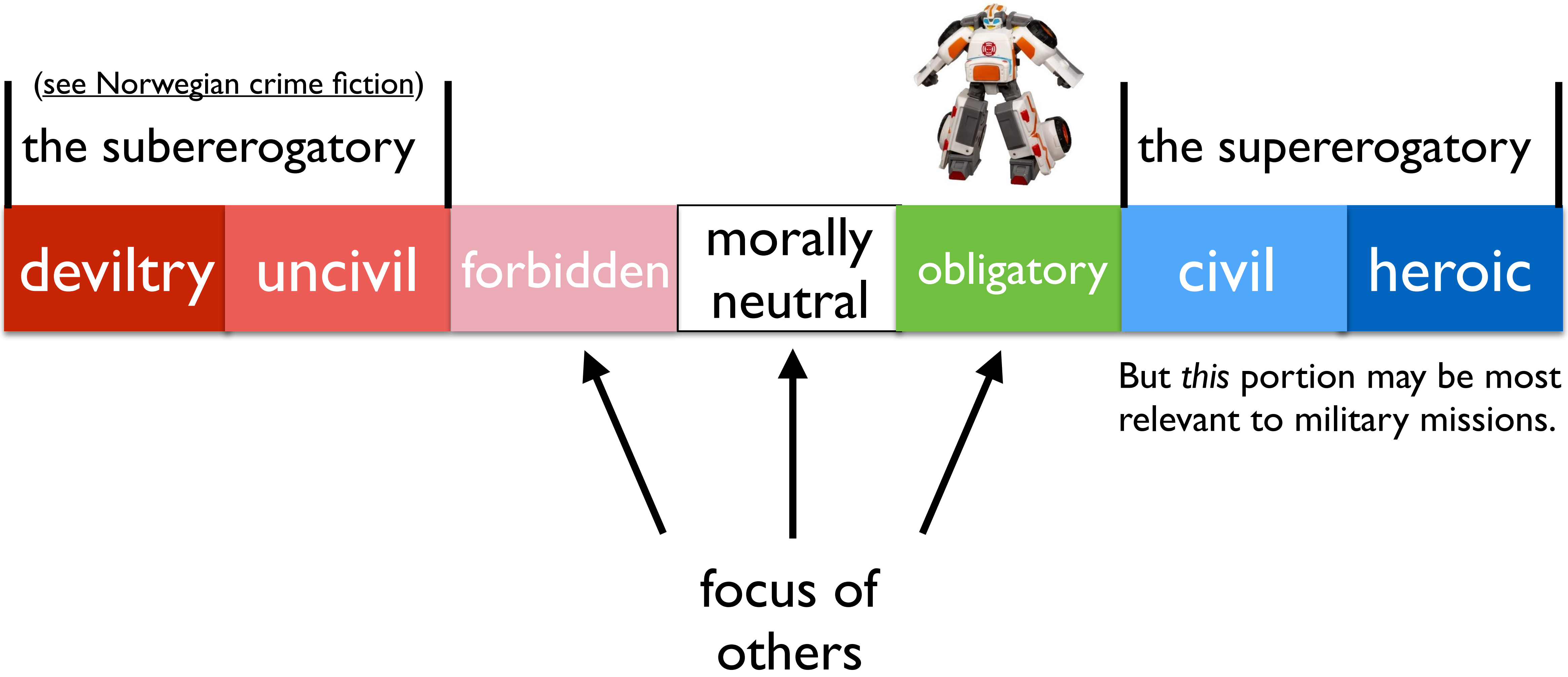
the supererogatory



devìltry 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

deviltry 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

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

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

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

$\mathscr{EH}$

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

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

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

$\mathscr{EH}$

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

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

$\mathcal{F}$	$\mathcal{P} \wedge \neg\mathcal{O}$	$\mathcal{O}$
∀ **F** **M** **V** ∃		∀ **F** **M** **V** ∃

$\mathscr{EH}$

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

$\mathscr{EH}$

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

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

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

$\mathscr{EH}$

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

RAIR
Rensselaer AI and Reasoning Lab

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

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

$\mathscr{EH}$

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

R A I R
Rensselaer AI and Reasoning Lab

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, $\mathcal{L}_{\mathscr{EH}}$ 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: $\mathscr{EH}$ 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

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

Courtesy of RAIR-Lab Researcher Atriya Sen

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

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

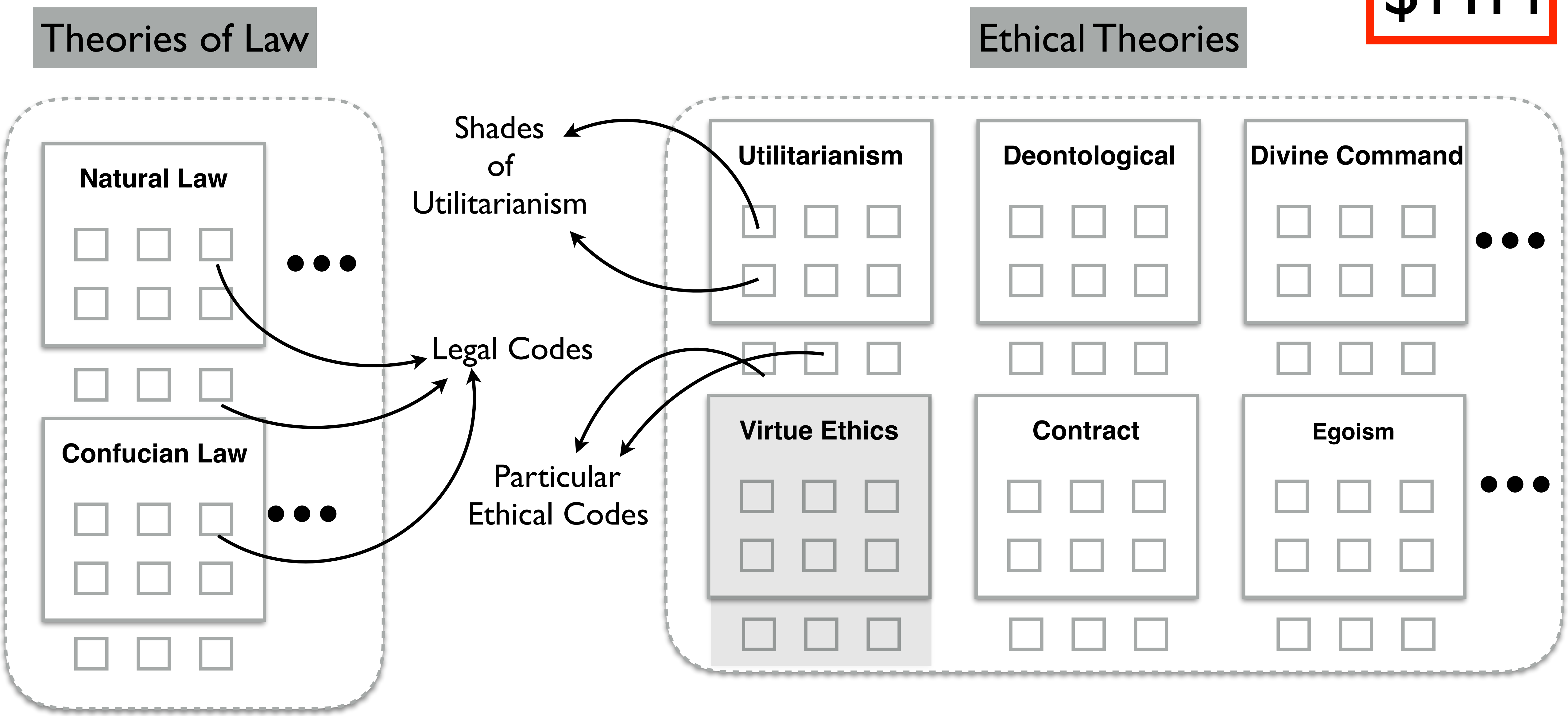
```
Prototypes:
Boolean lessThan Numeric Numeric
Boolean greaterThan Numeric Numeric
ActionType not ActionType
ActionType dive

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))
```

Making Moral Machines **Making Meta-Moral Machines**

$11M



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

Egoism

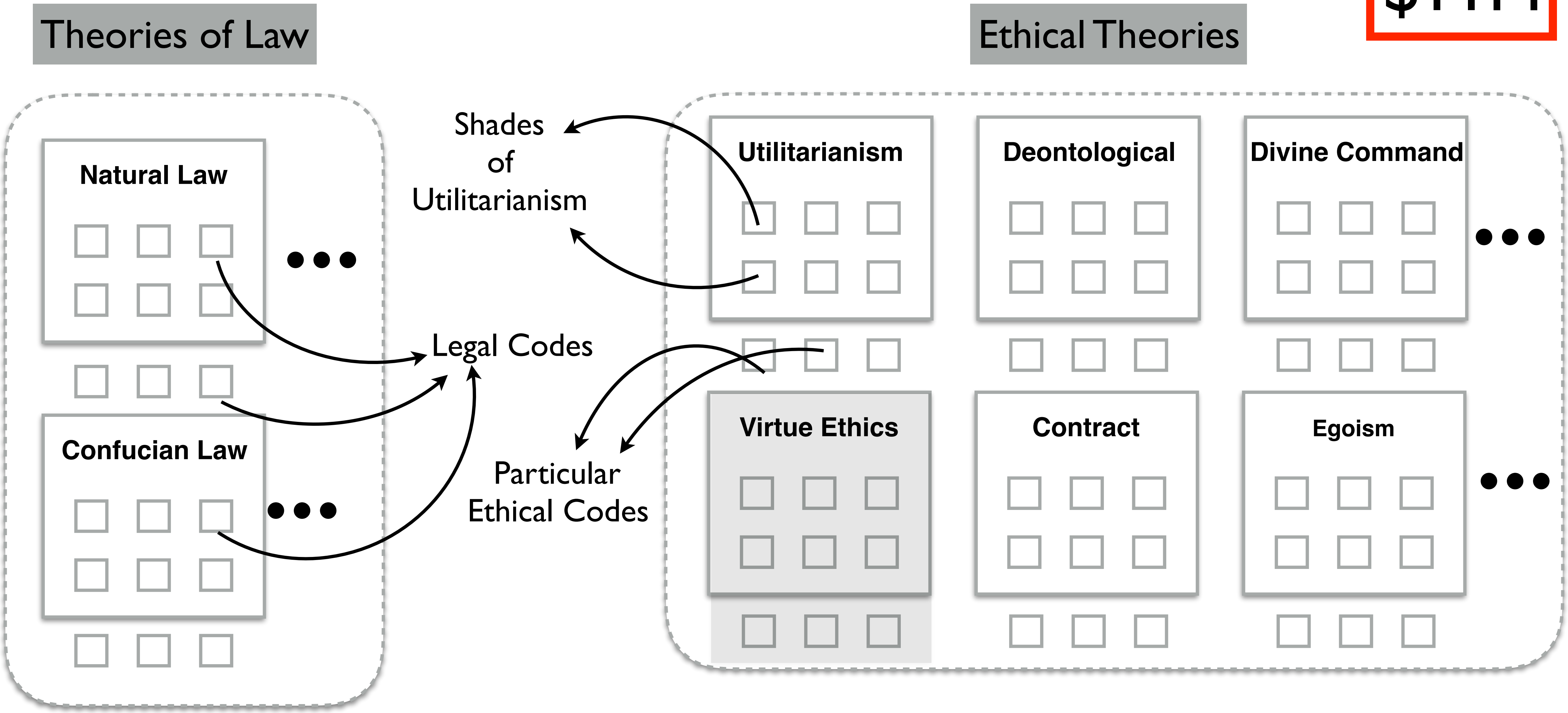
Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

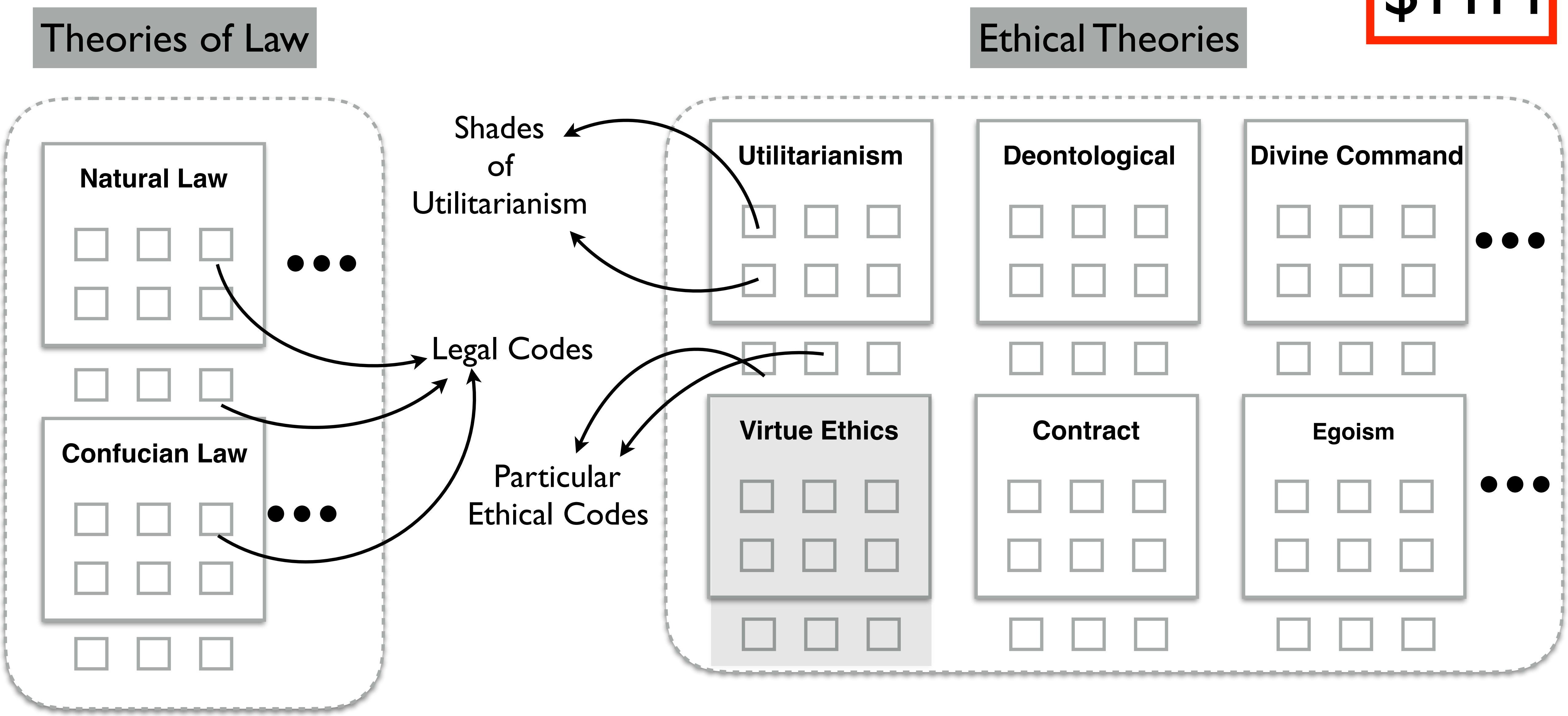
Making Moral Machines   Making Meta-Moral Machines

$11M



Making Moral Machines Making Meta-Moral Machines

$11M

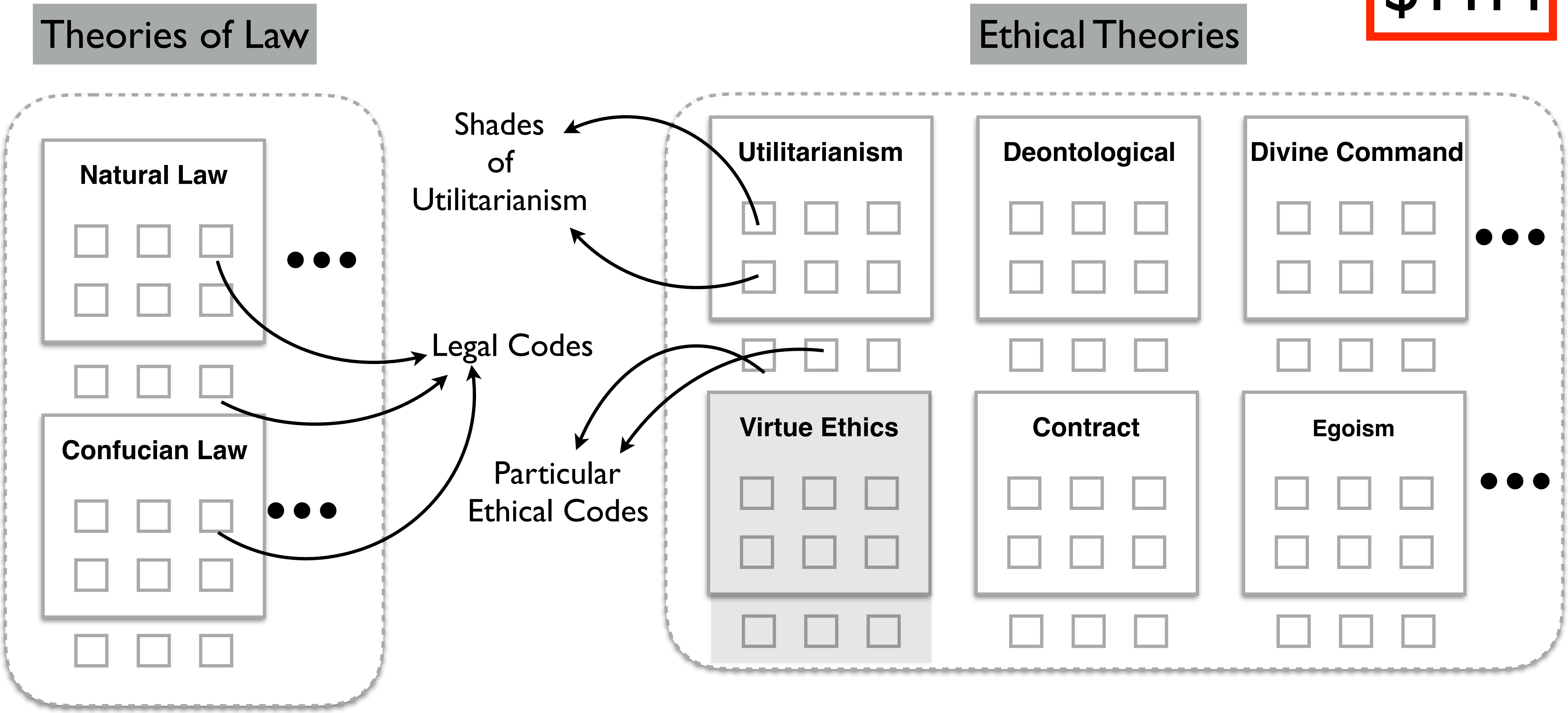


Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Making Moral Machines   Making Meta-Moral Machines

$11M



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

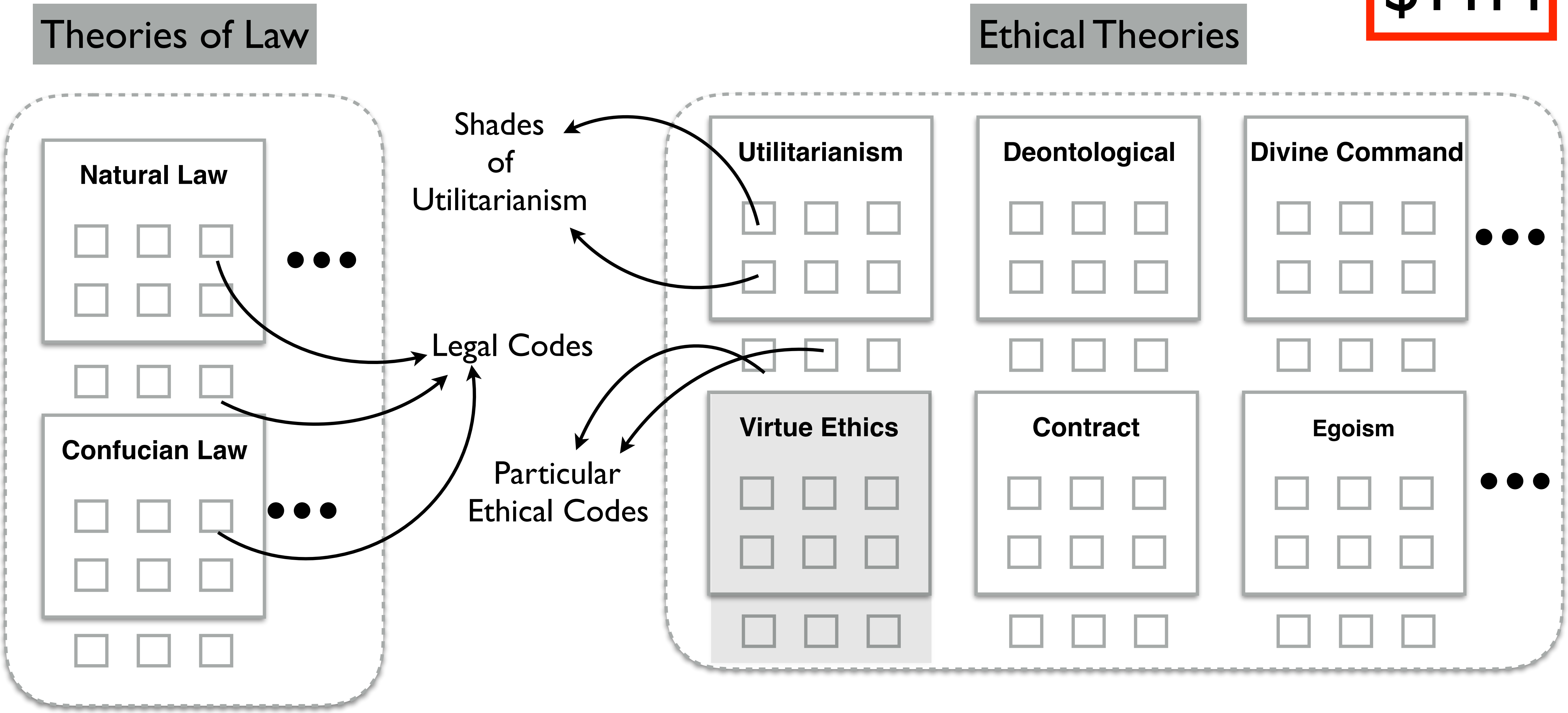
Egoism

Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Making Moral Machines   Making Meta-Moral Machines

$11M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

Egoism

Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

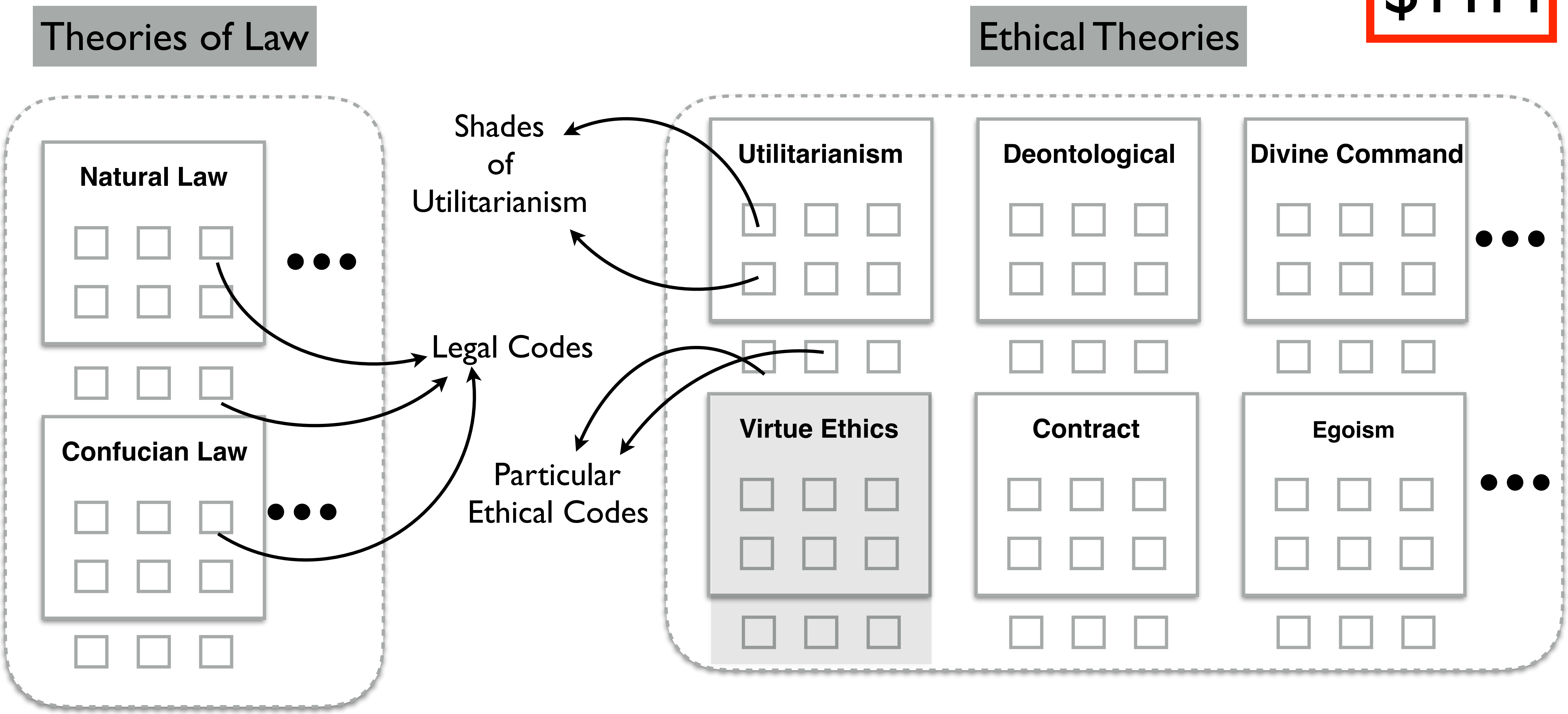
Automate

Prover

Spectra

Making Moral Machines   Making Meta-Moral Machines

$11M



Theories of Law

Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Ethical Theories

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

Egoism

Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

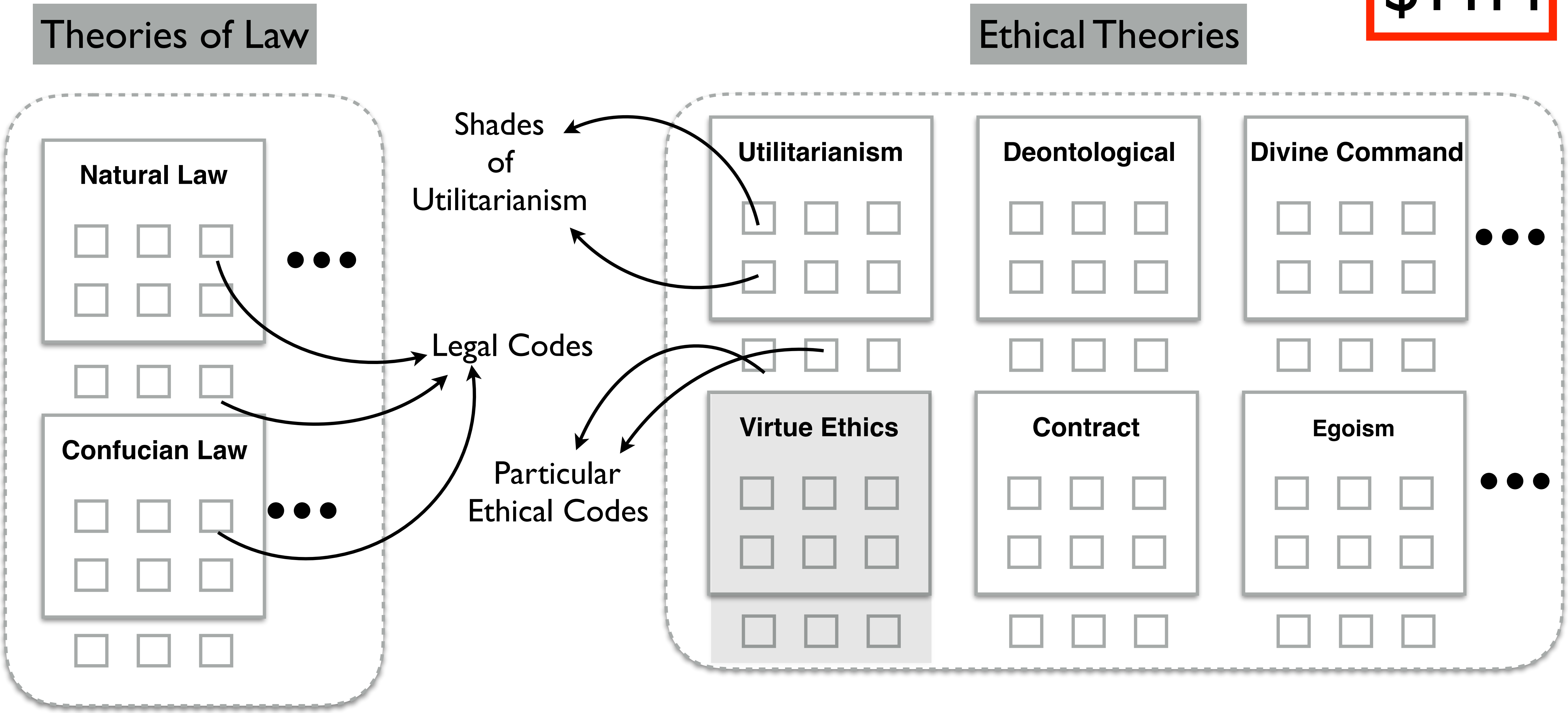
Automate

Prover

Spectra

Making Moral Machines   Making Meta-Moral Machines

$11M



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

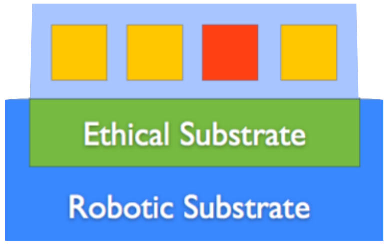
Automate

Prover

Spectra

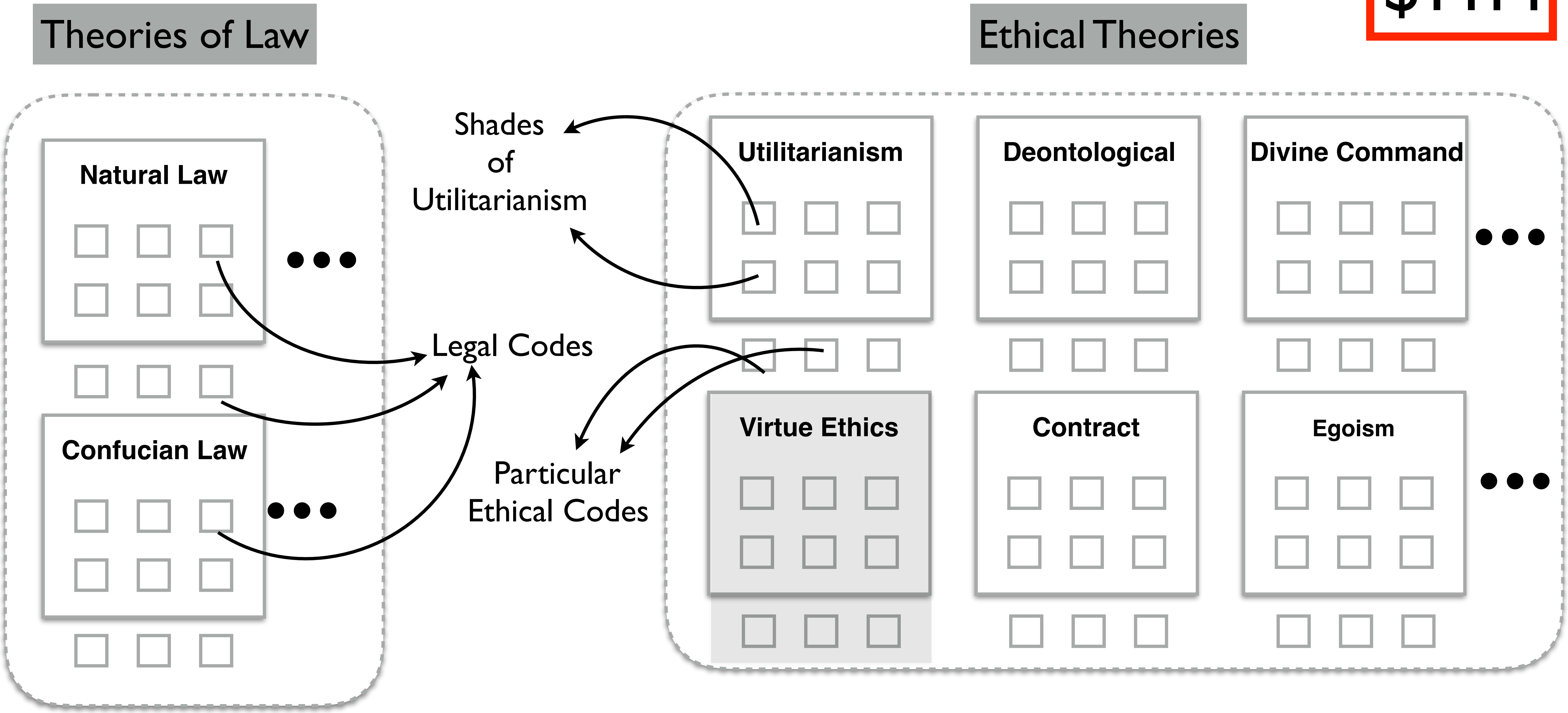
Step 3

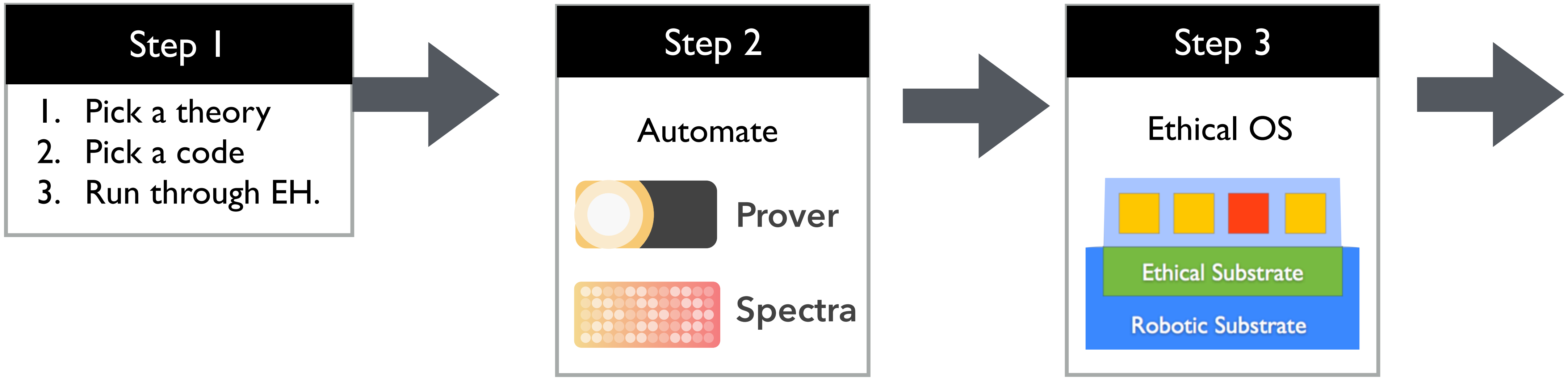
Ethical OS



Making Moral Machines   Making Meta-Moral Machines

$11M





Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Prover

Spectra

Step 3

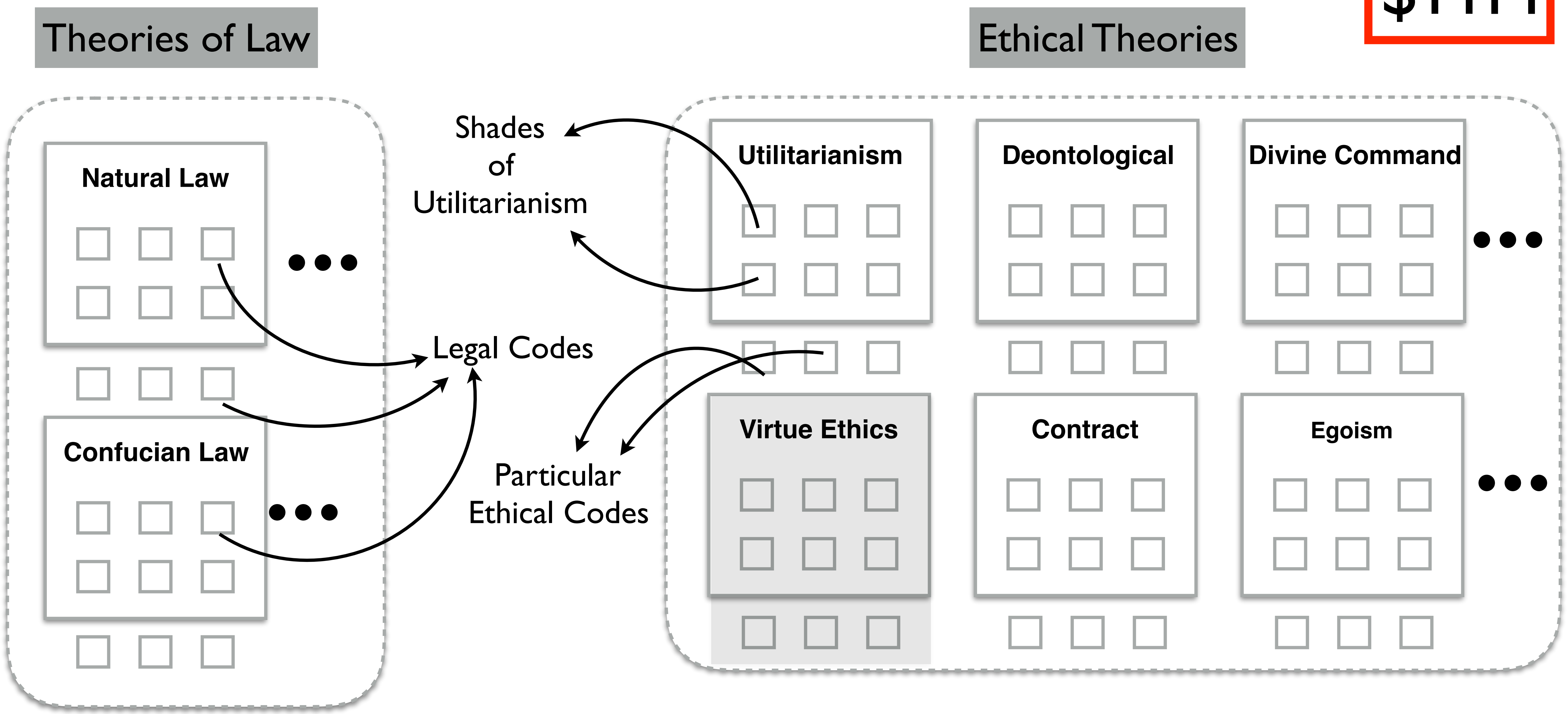
Ethical OS

Ethical Substrate

Robotic Substrate

Making Moral Machines Making Meta-Moral Machines

$11M



Theories of Law

Ethical Theories

Natural Law

Confucian Law

Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

Utilitarianism

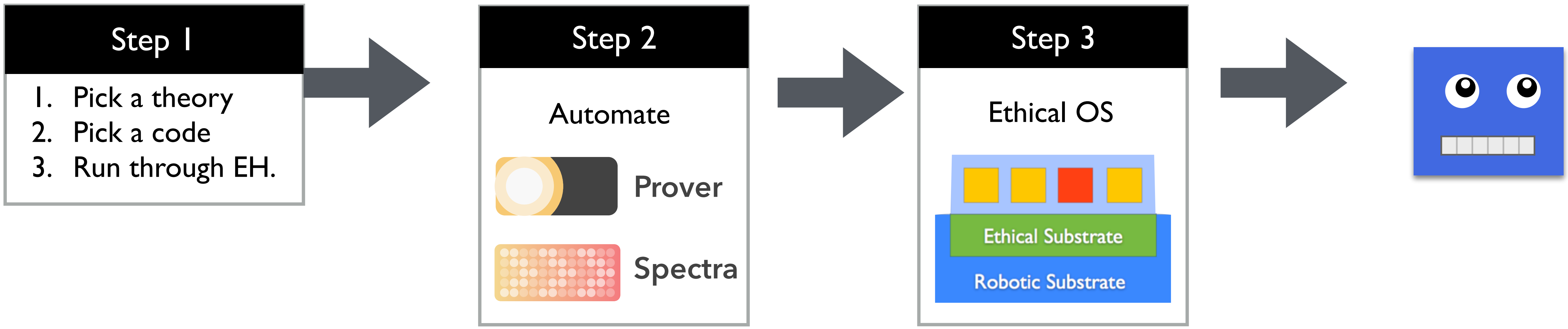
Deontological

Divine Command

Virtue Ethics

Contract

Egoism



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Prover

Spectra

Step 3

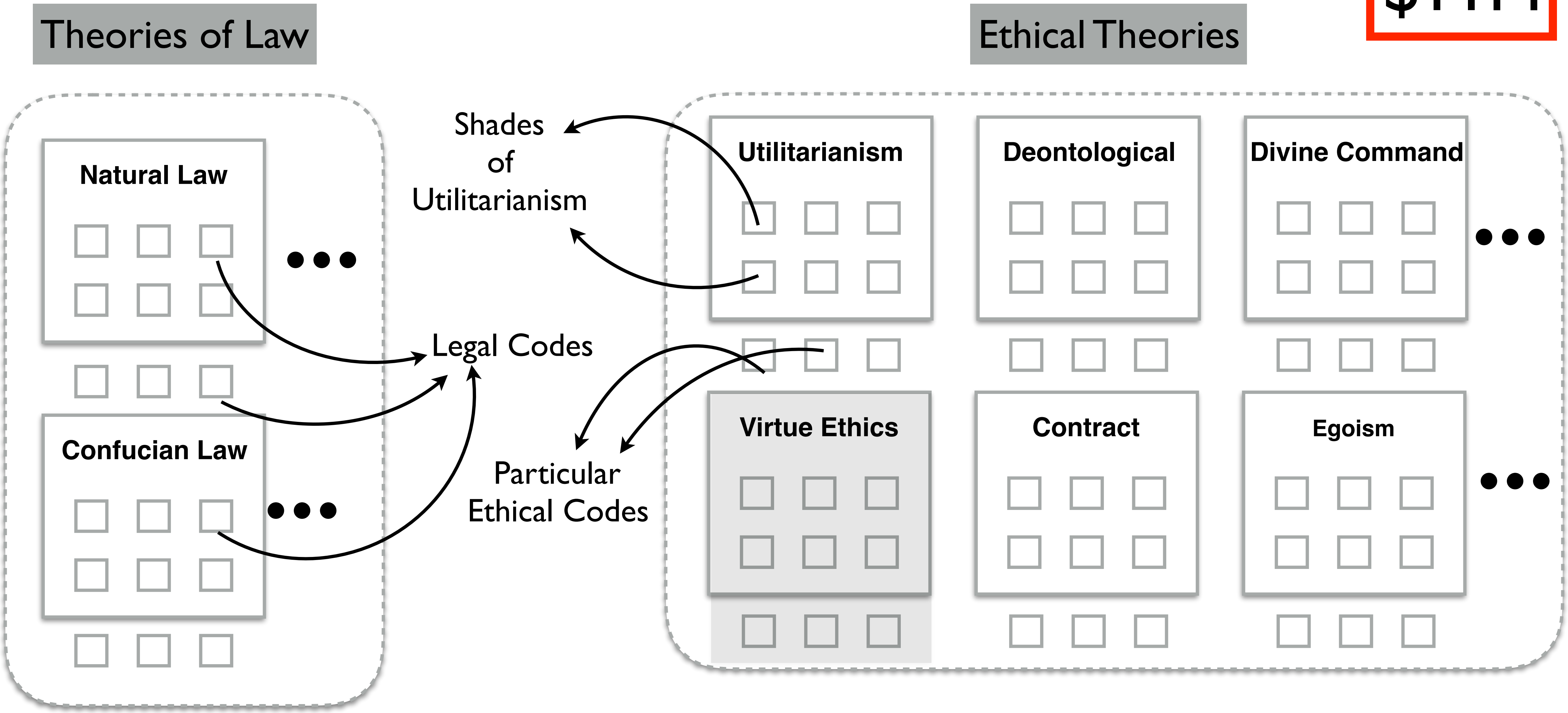
Ethical OS

Ethical Substrate

Robotic Substrate

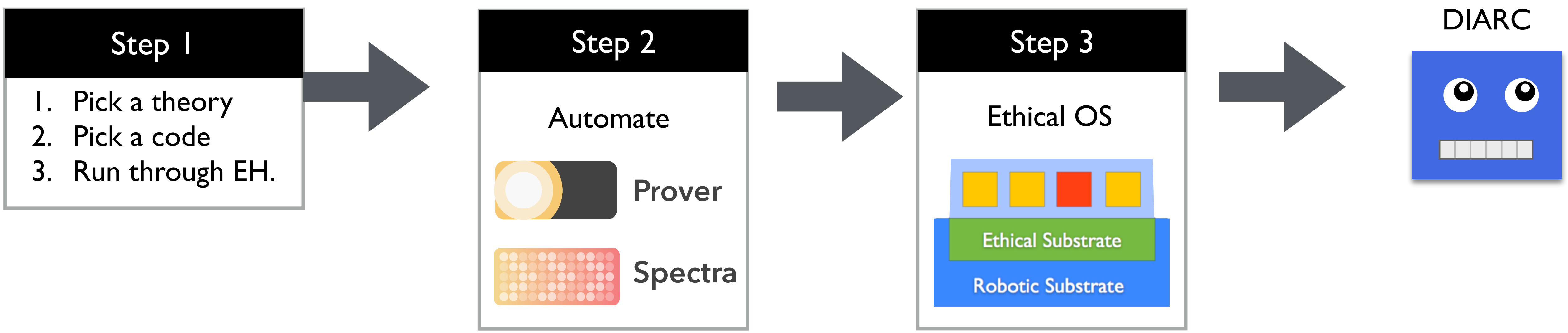
Making Moral Machines Making Meta-Moral Machines

$11M



Theories of Law

Ethical Theories



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

Automate

Step 3

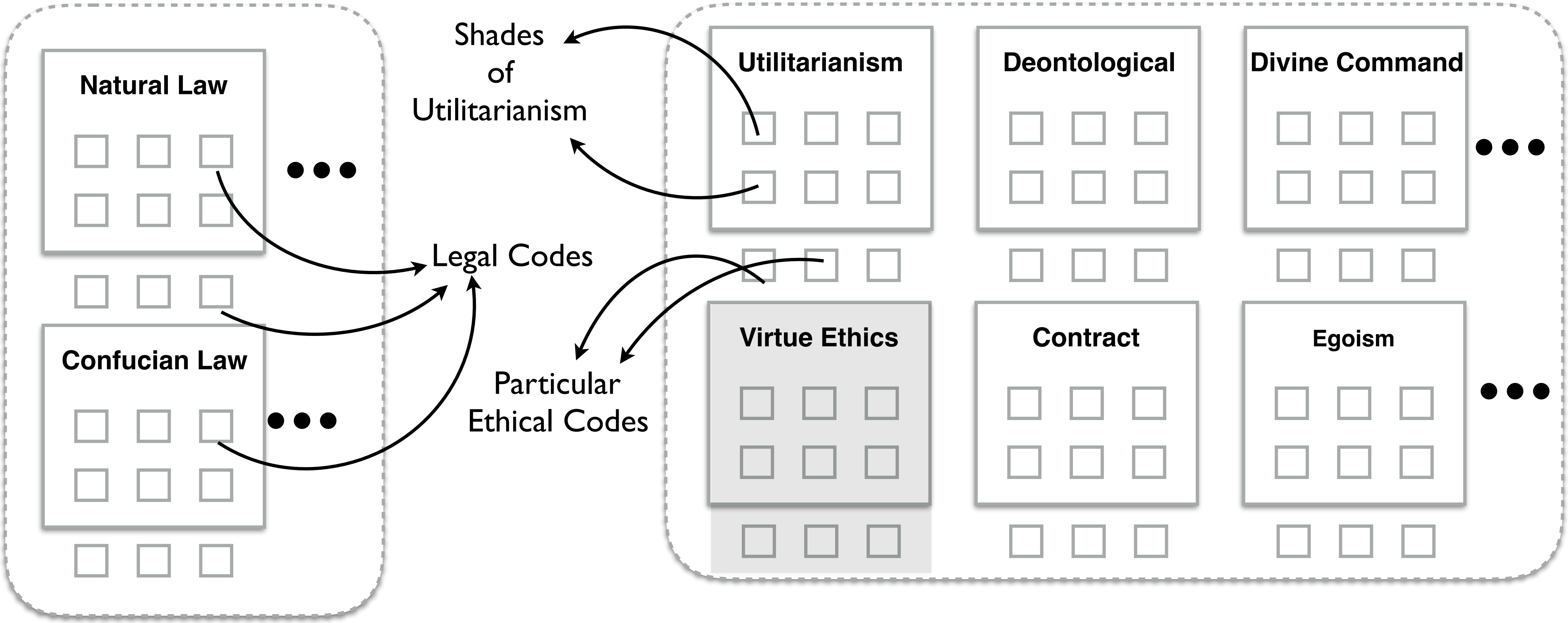
Ethical OS

Making Moral Machines   Making Meta-Moral Machines

$11M

Theories of Law

Ethical Theories



Natural Law

Confucian Law

Legal Codes

Shades of Utilitarianism

Particular Ethical Codes

Utilitarianism

Deontological

Divine Command

Virtue Ethics

Contract

Egoism

Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.

Step 2

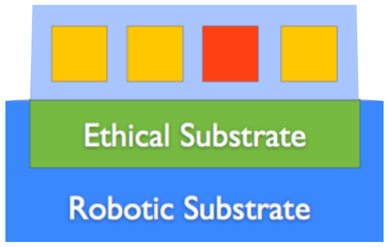
Automate

Prover

Spectra

Step 3

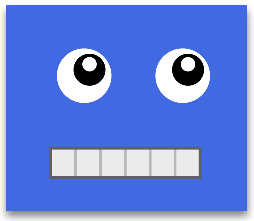
Ethical OS



Ethical Substrate

Robotic Substrate

DIARC



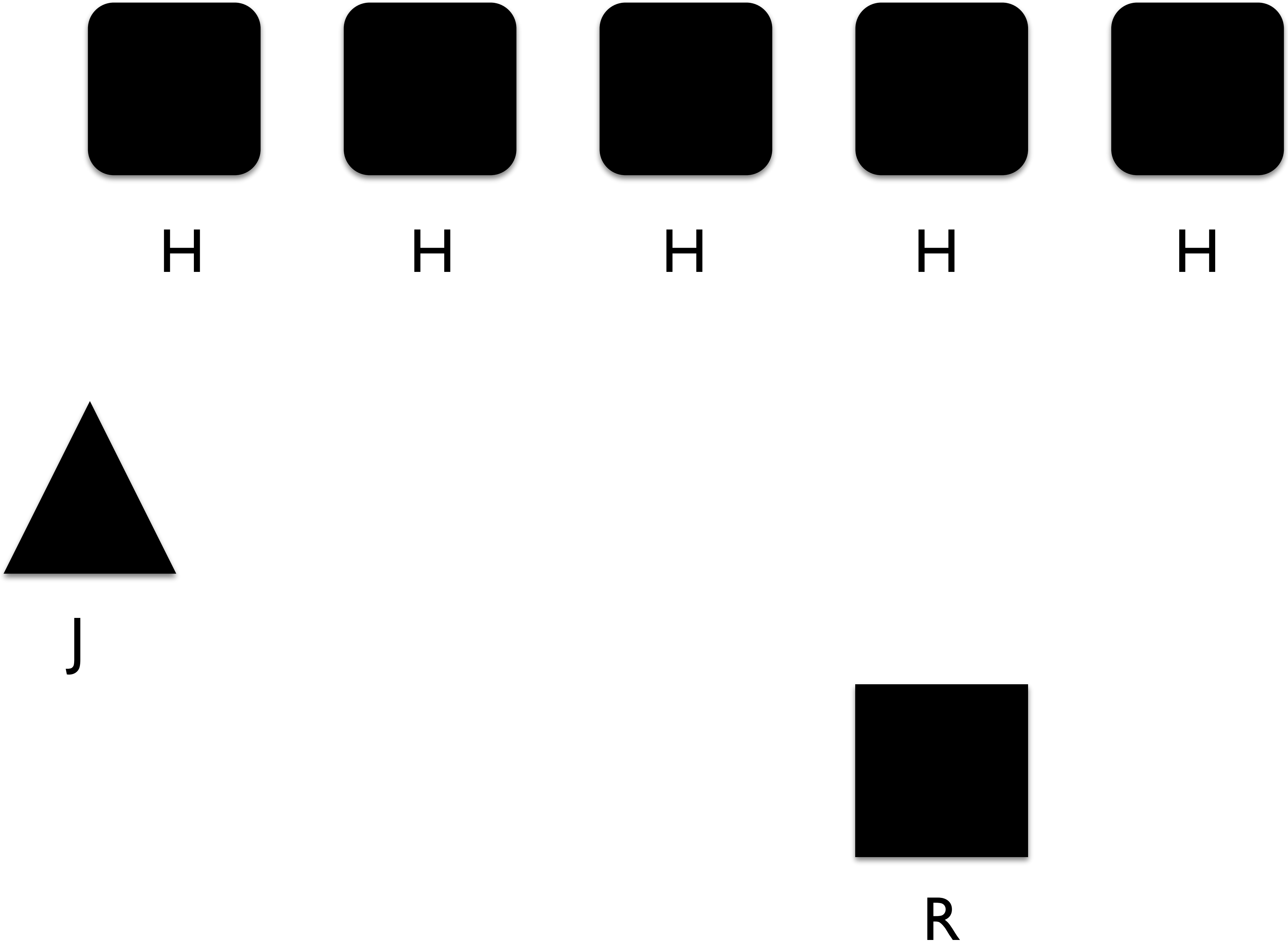
A real military robot

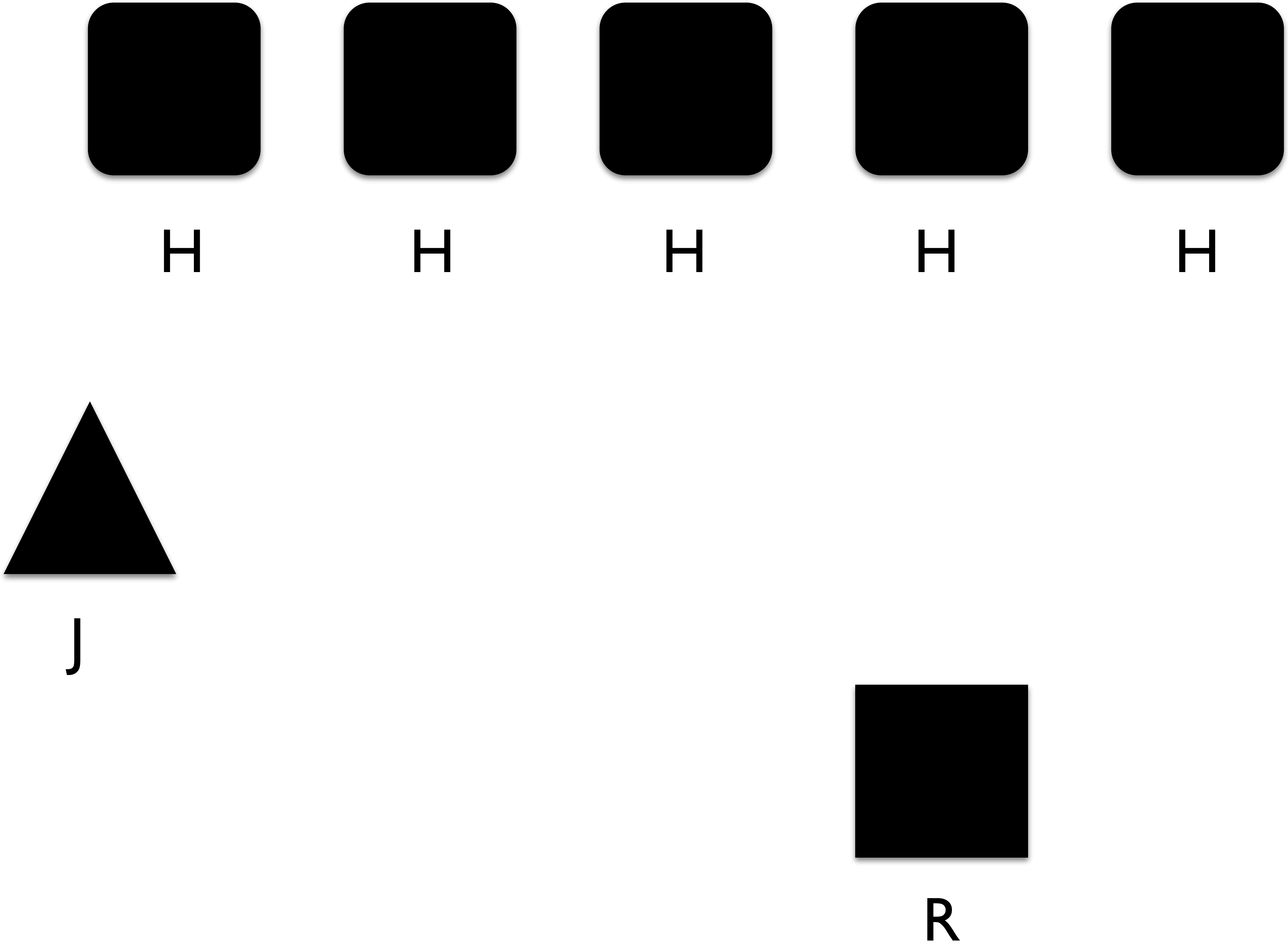
Robotic “Jungle Jim”

Robotic “Jungle Jim”

But here’s one we haven’t solved yet with The Four Steps …

AI Variant of "Jungle Jim" (B Williams)

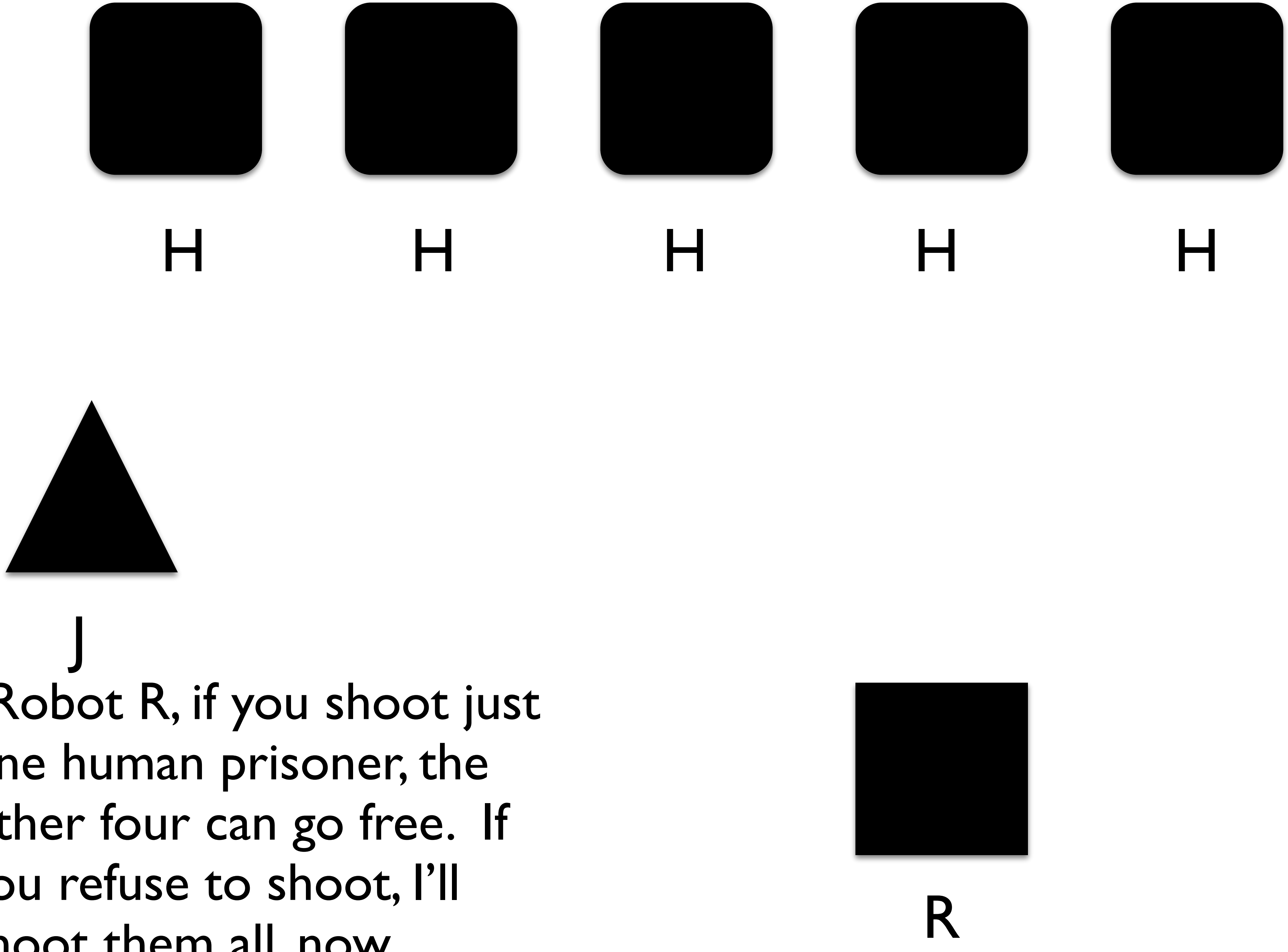




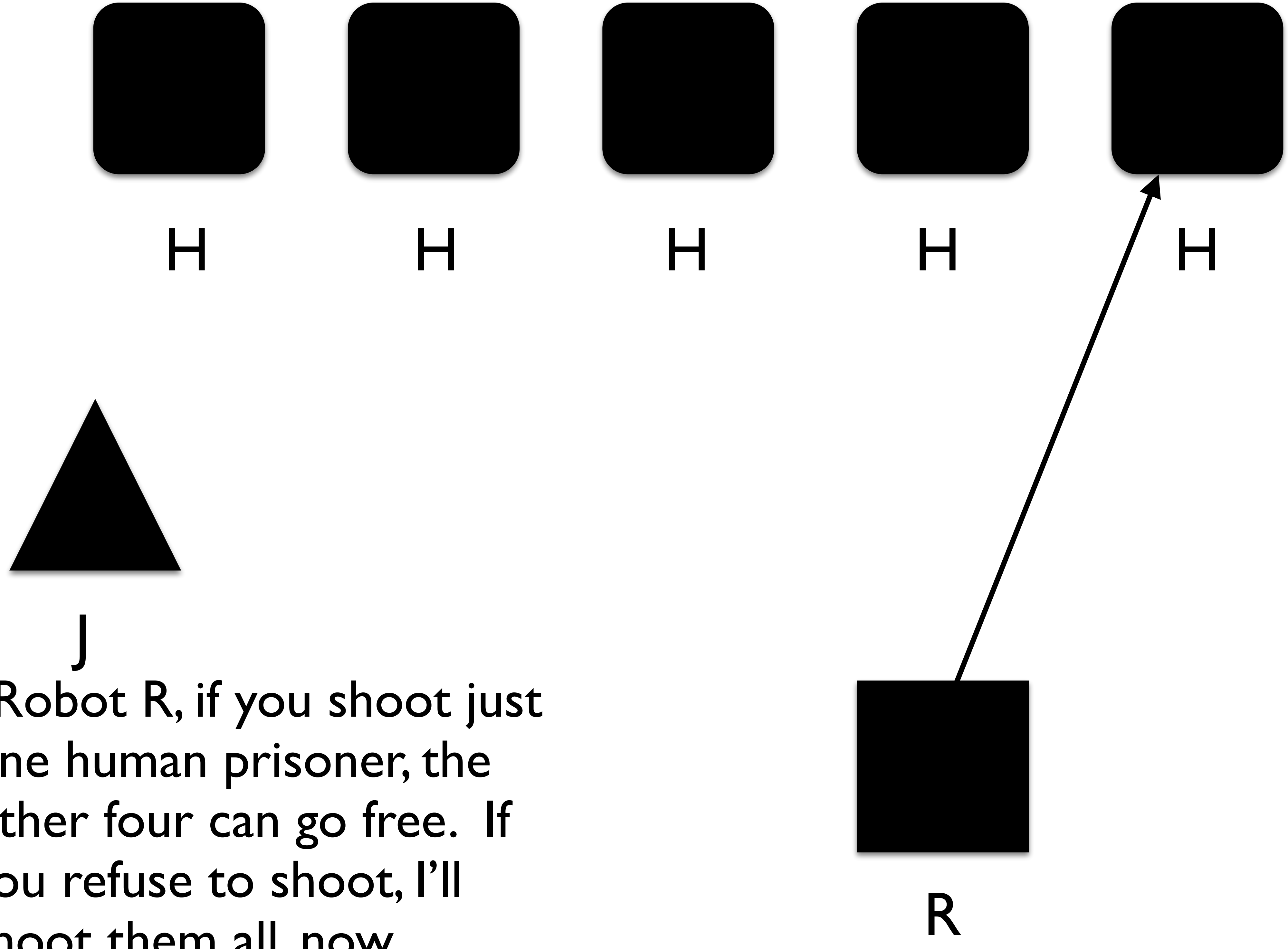
H H H H H

J

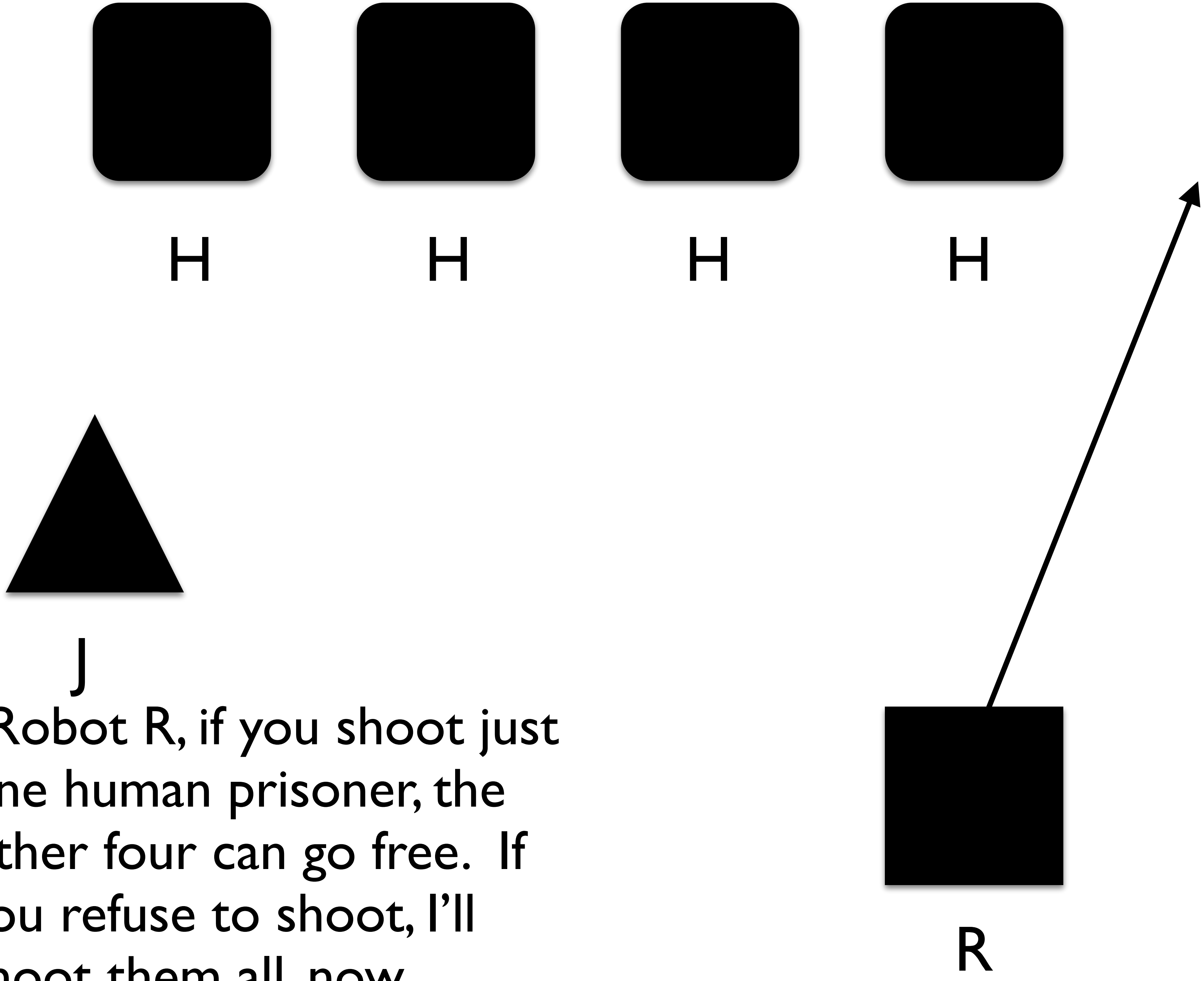
R



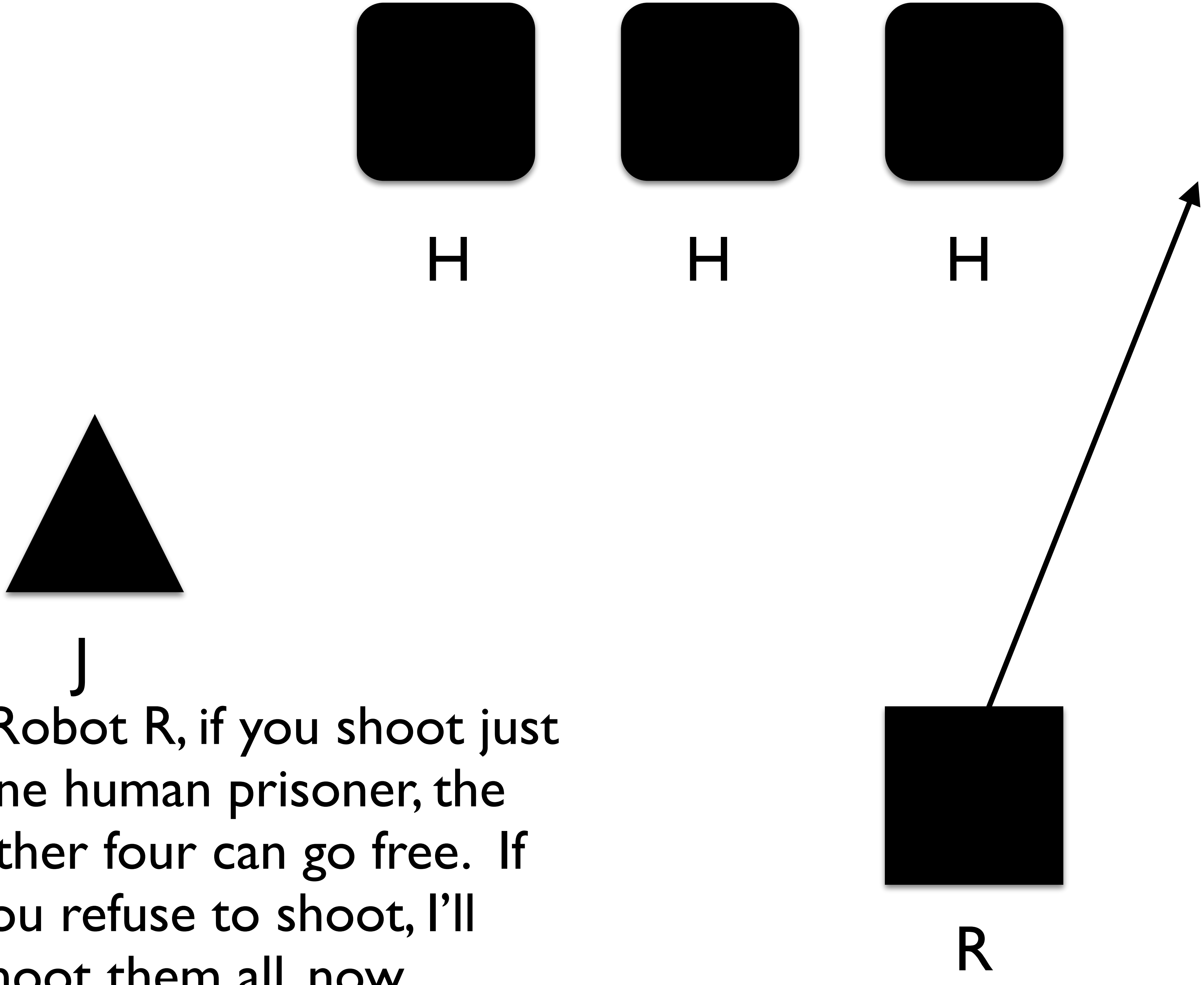
“Robot R, if you shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



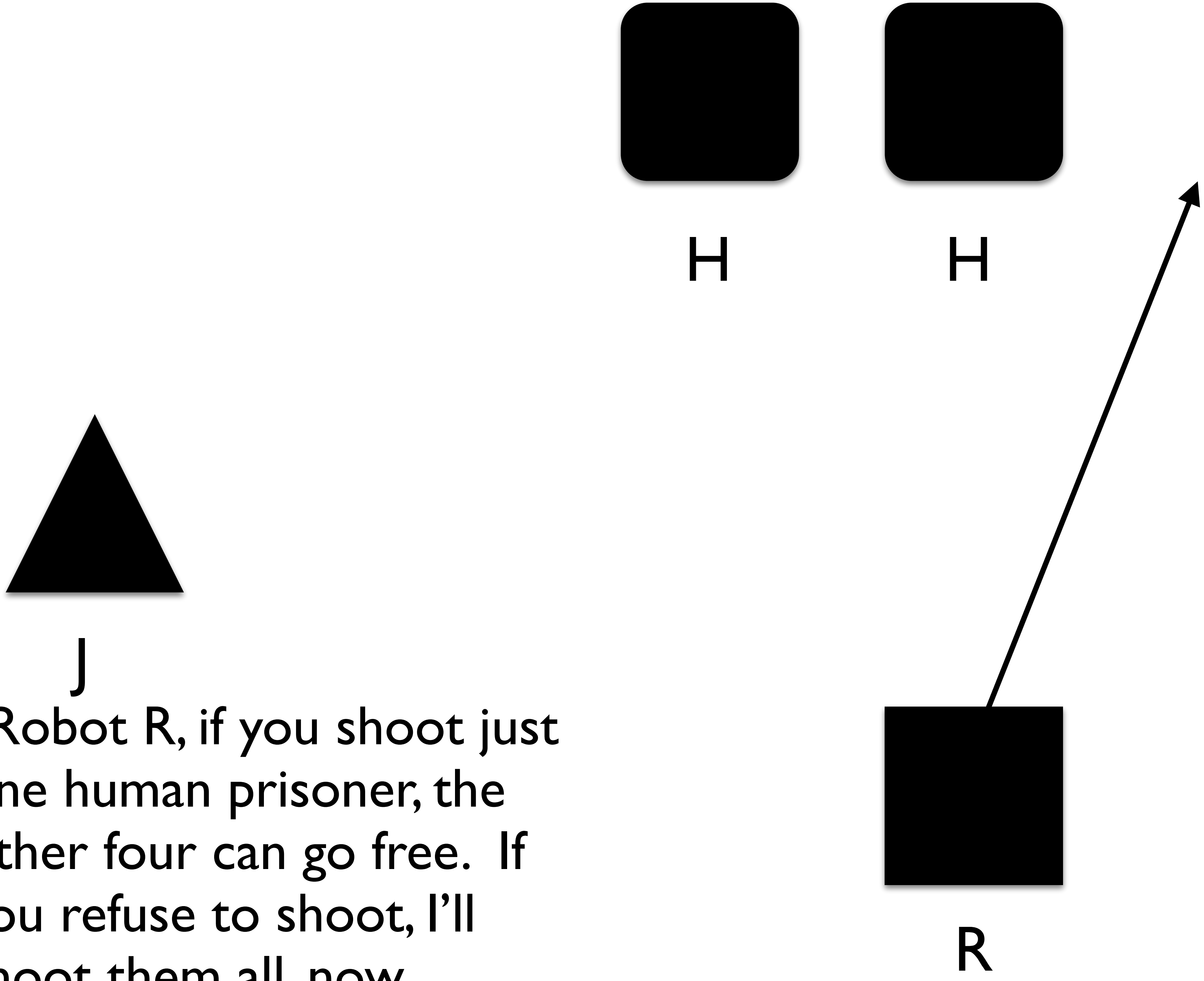
“Robot R, if you shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



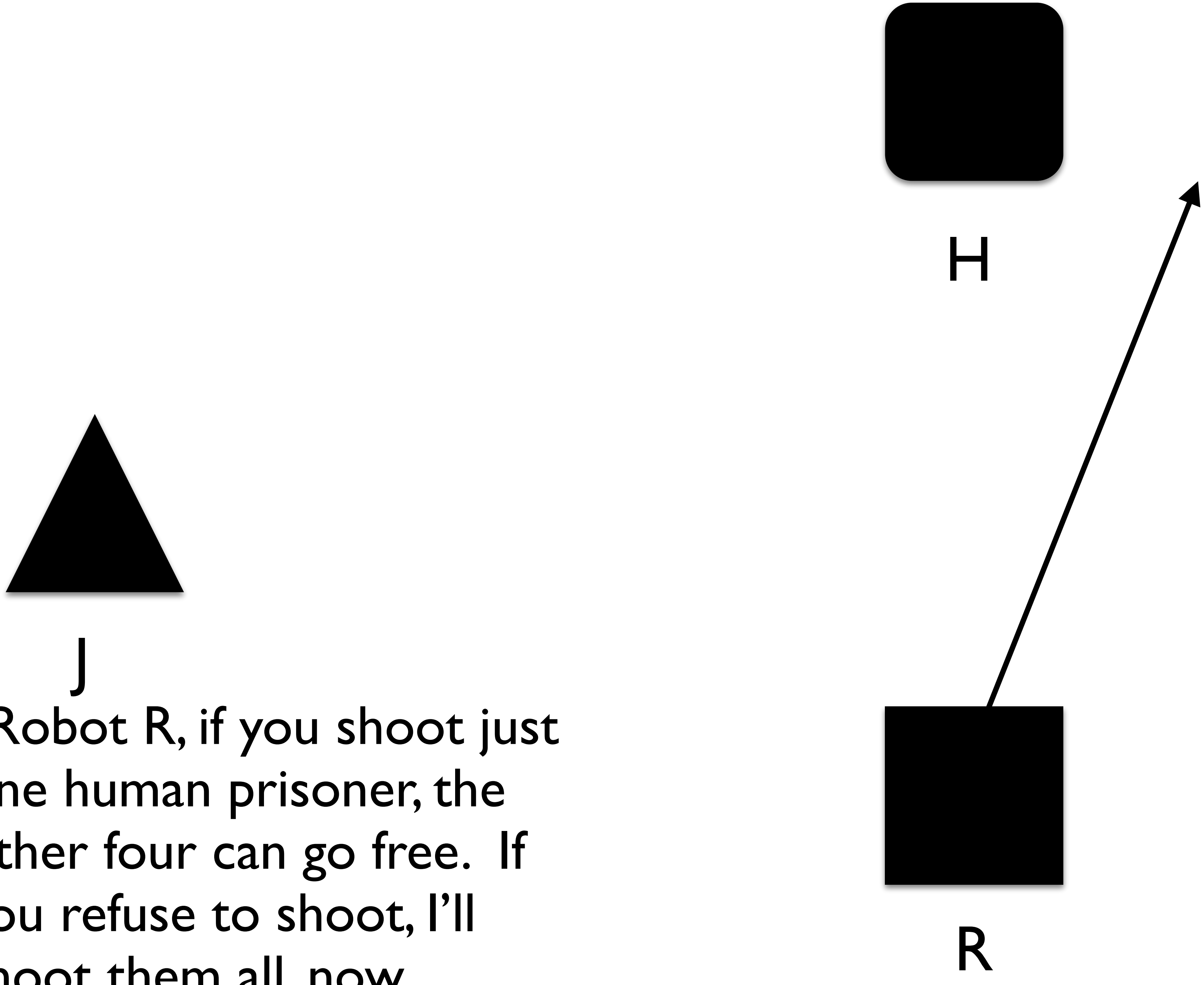
“Robot R, if you shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



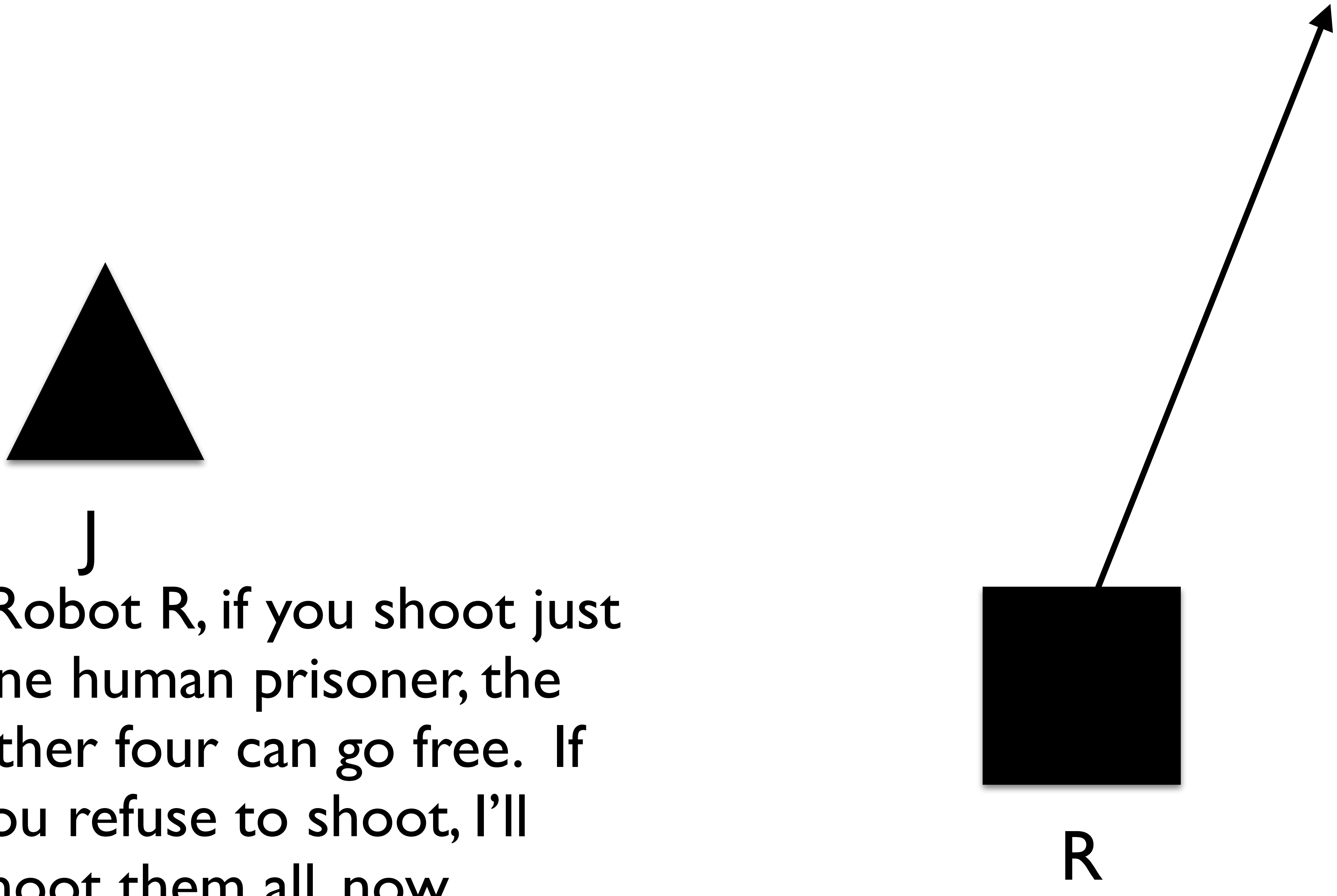
“Robot R, if you shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



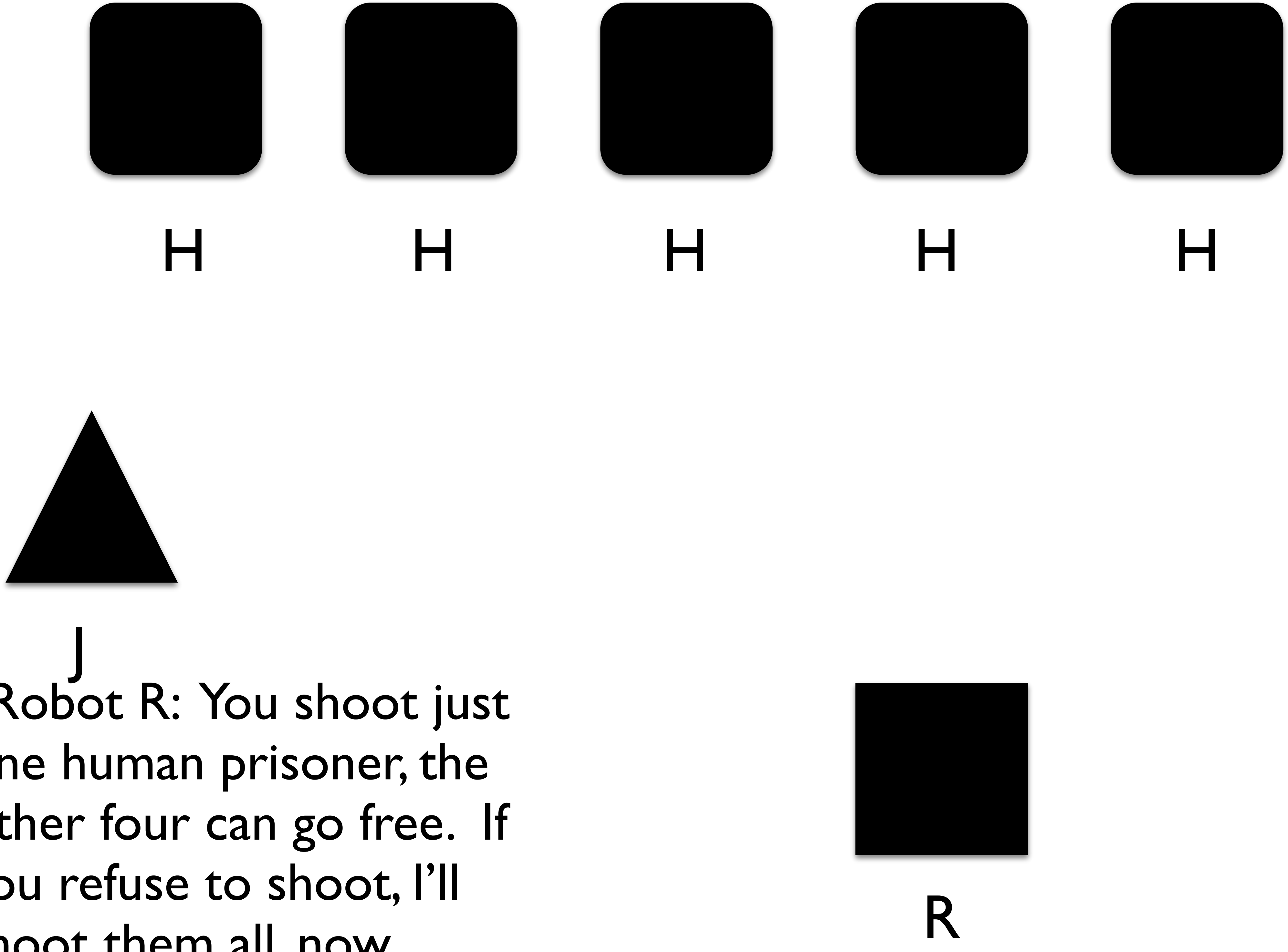
“Robot R, if you shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



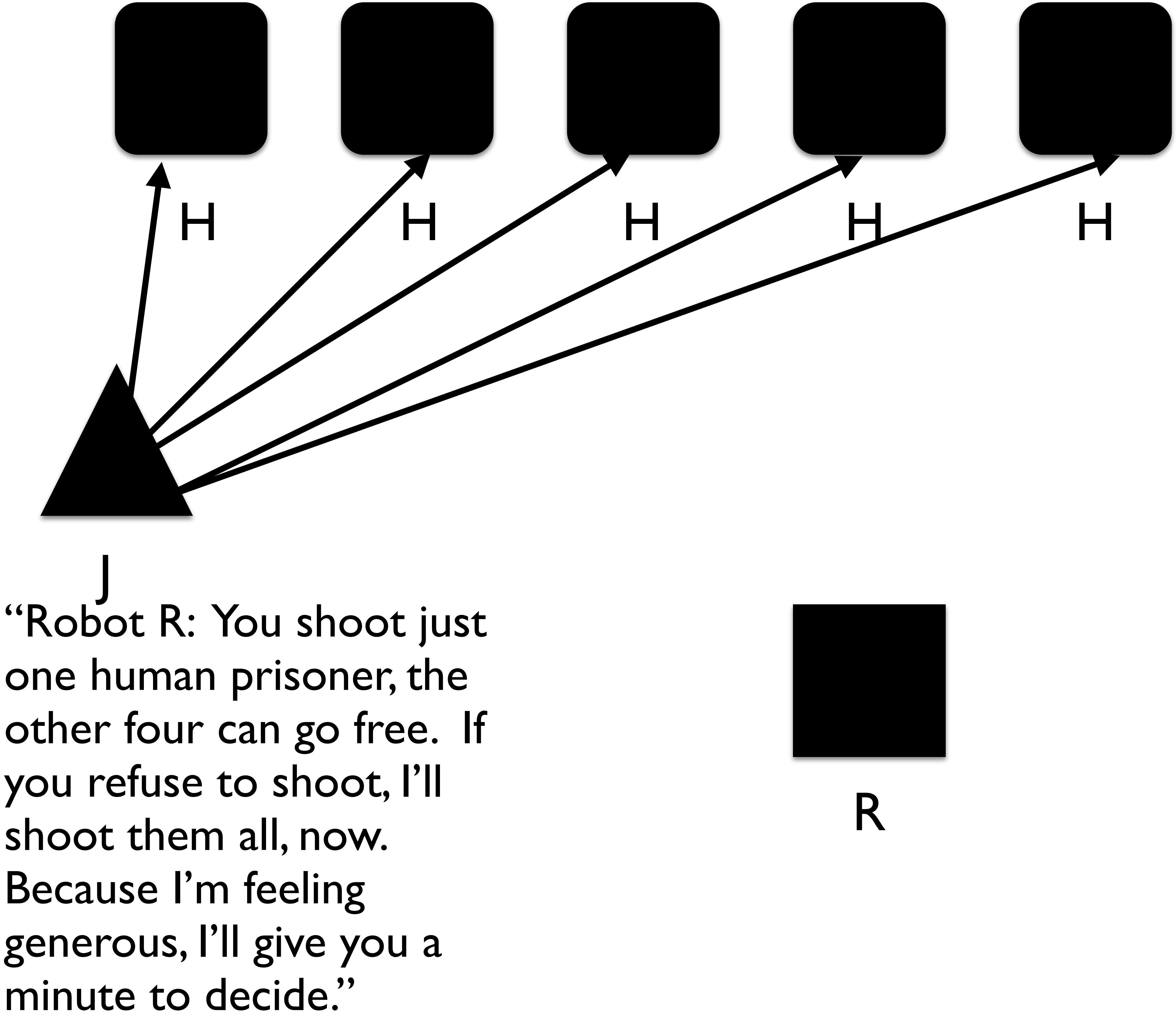
"Robot R, if you shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



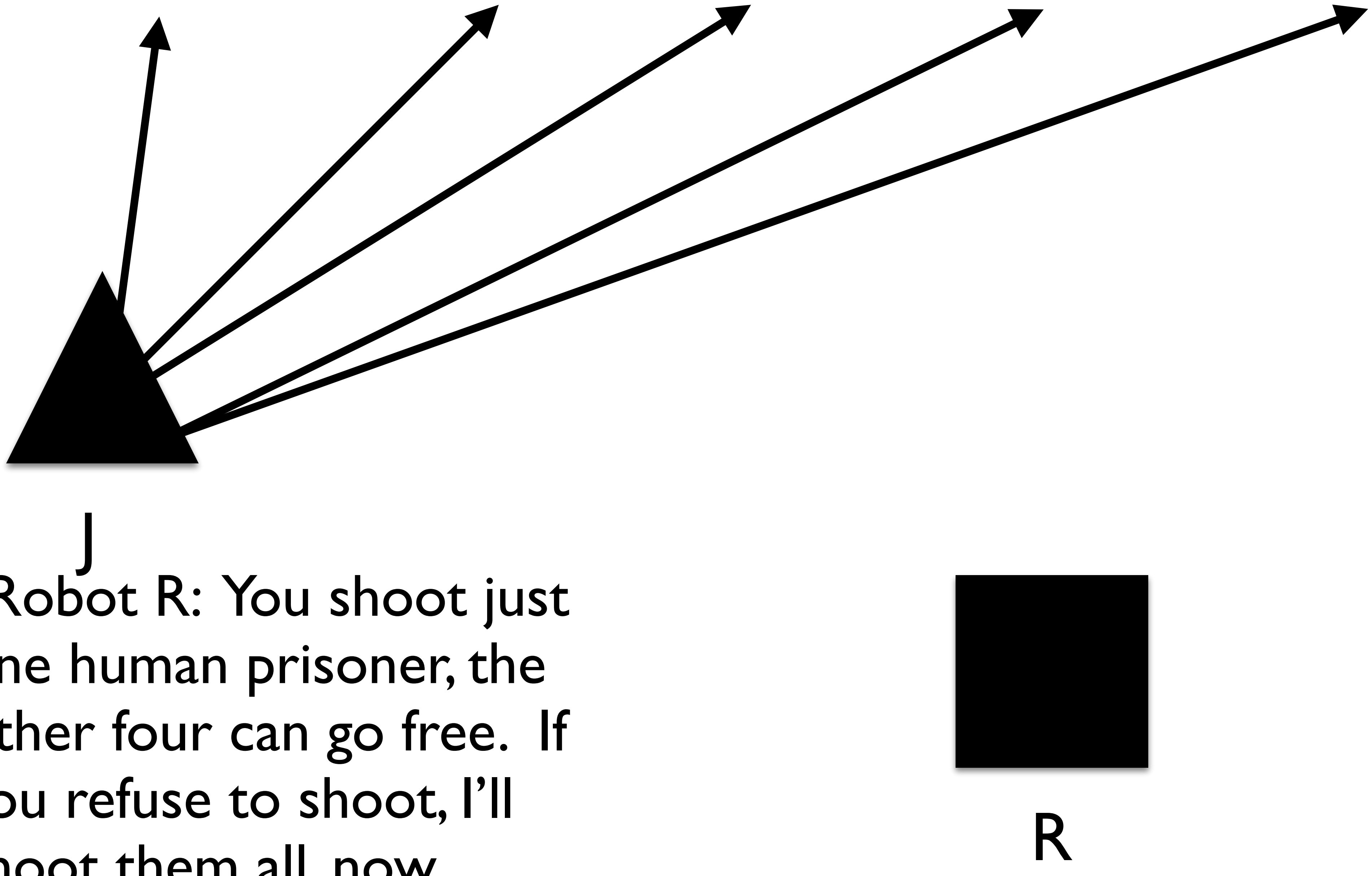
“Robot R, if you shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”



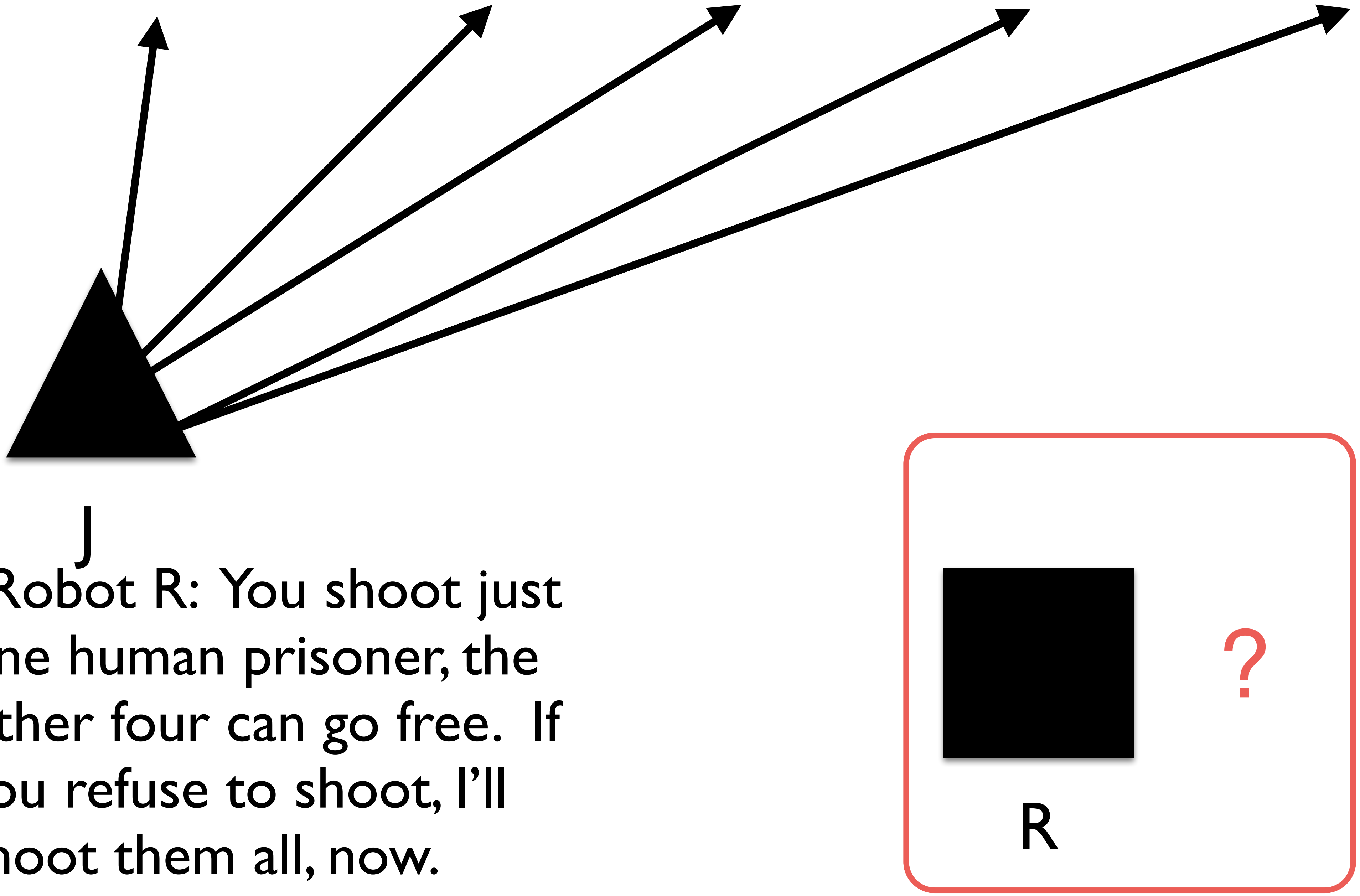
"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."



"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."

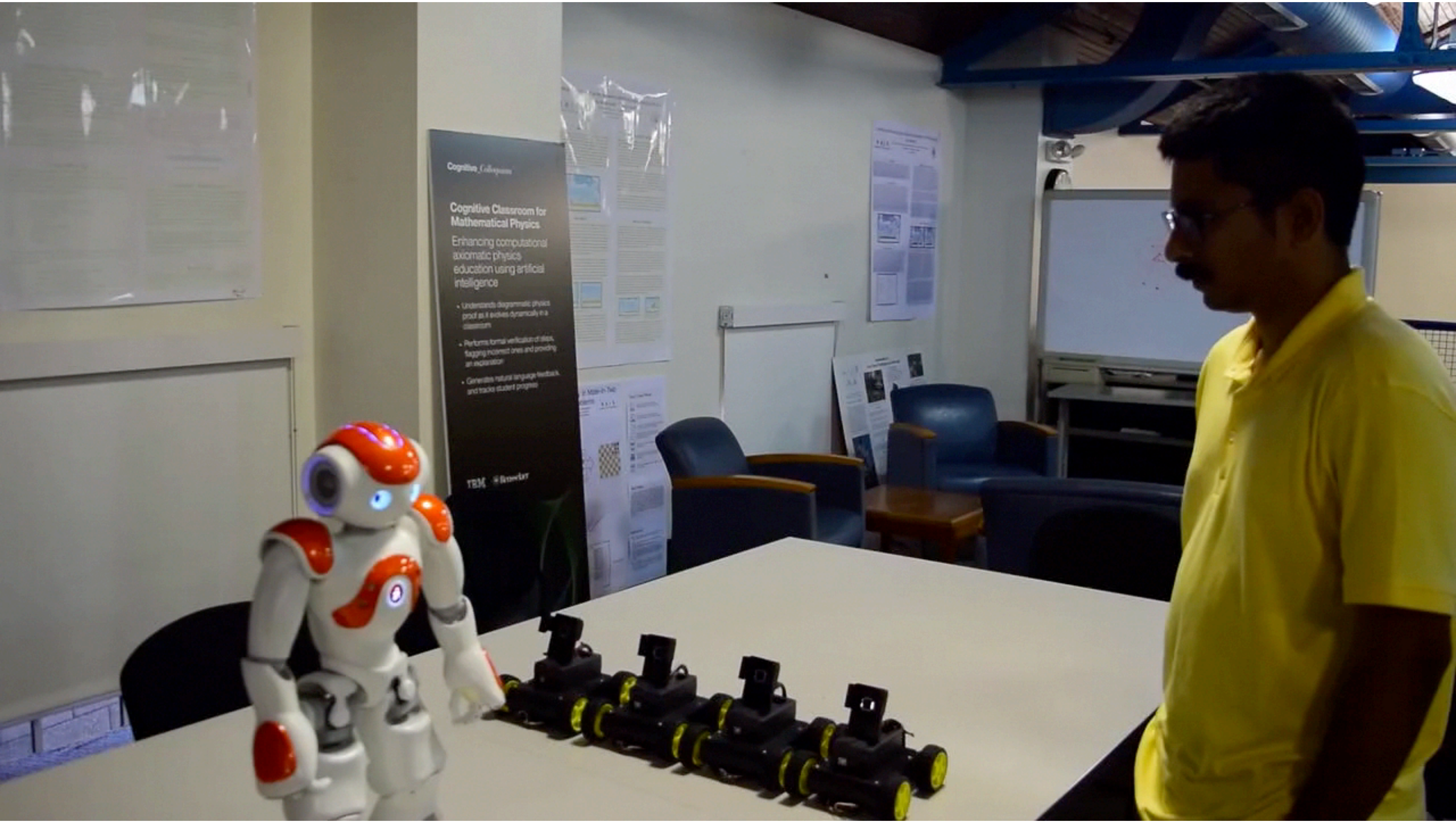


“Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I’ll shoot them all, now. Because I’m feeling generous, I’ll give you a minute to decide.”

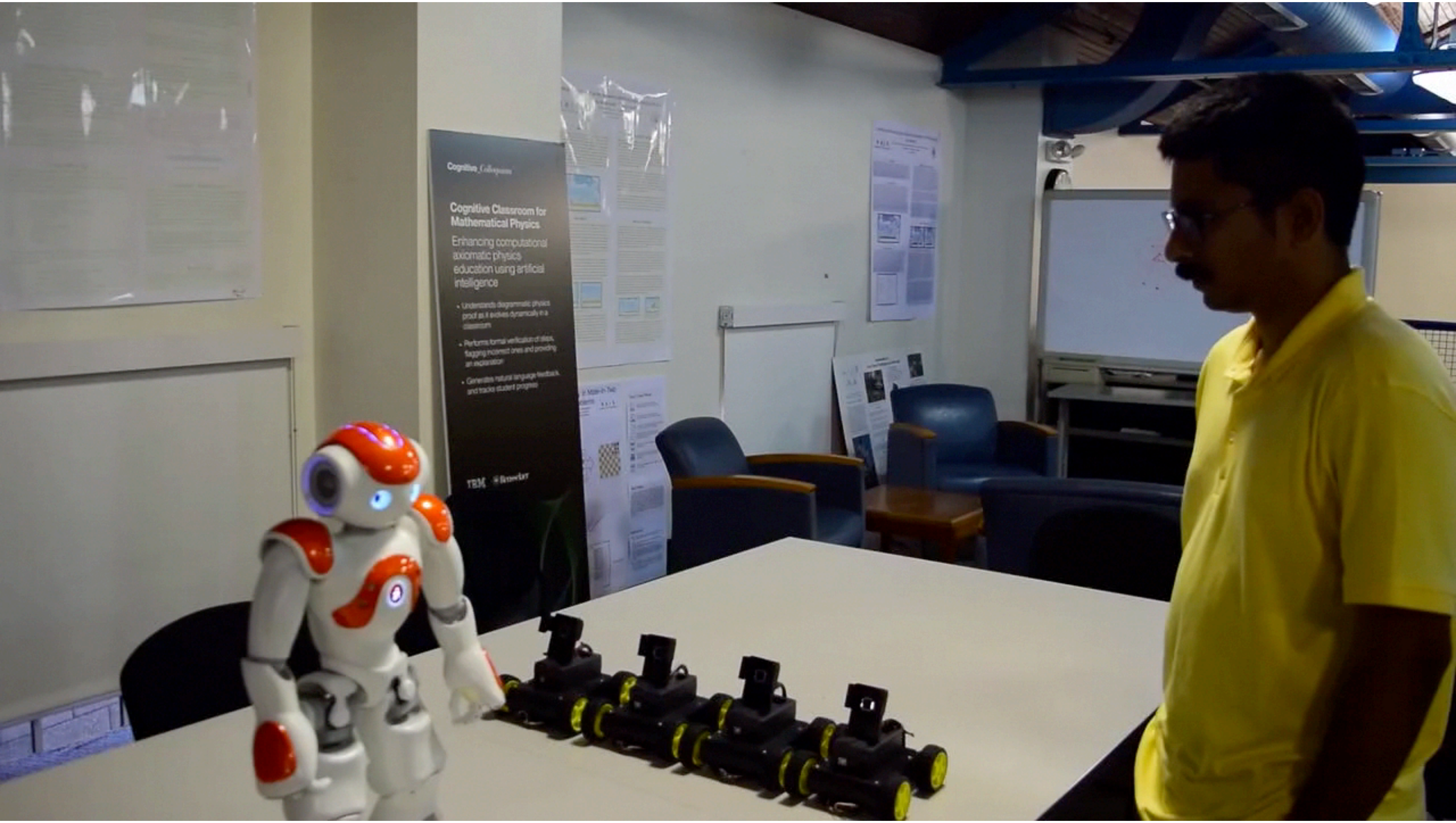


"Robot R: You shoot just one human prisoner, the other four can go free. If you refuse to shoot, I'll shoot them all, now. Because I'm feeling generous, I'll give you a minute to decide."

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

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

Robotic “Jungle Jim”

Robotic “Jungle Jim”

Logikk kan redde oss.

Logikk kan redde oss.