Killer Robots & The PAID Problem in *Star Trek TOS*, D, and Beyond to *DCEC** (in HyperSlate®)

Selmer Bringsjord

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

IFLAI2
11/18/2021
`ver 111821NY`

Rensselaer AI and Reasoning Lab

Logistics …

Let’s check the status of the paper in Overleaf …

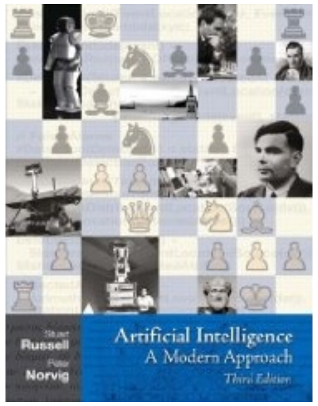
The PAID Problem …

“Earth, we have a problem.”

“Earth, we have a problem.”

“Earth, we have a problem.”



These are amorphous worries separated from the formal. The Problem is easy to state, precisely:

These are amorphous worries separated from the formal. The Problem is easy to state, precisely:

The PAID Problem

These are amorphous worries separated from the formal. The Problem is easy to state, precisely:

The PAID Problem

https://www.nhk.or.jp/globalagenda/archive/25.html

The PAID Problem (Level 1):

The PAID Problem (Level 1):

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

The PAID Problem (Level 1):

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

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

The PAID Problem (Level 1):

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

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

The PAID Problem (Level 1):

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

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

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

The PAID Problem (Level 1):

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

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

$\ldots$ $\ldots$ $\ldots$

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

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

The Four Steps …

Theories of Law

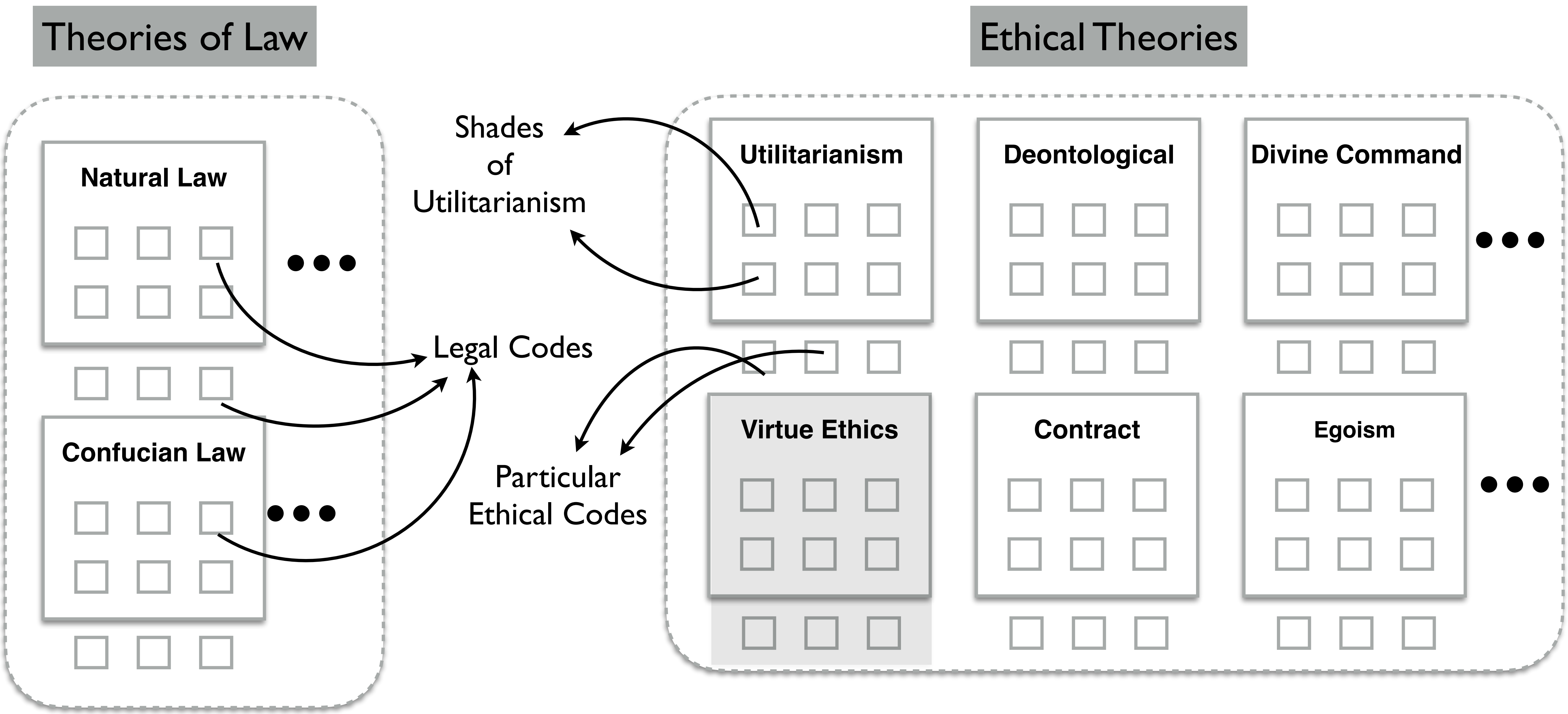
MAKING MORAL MACHINES || MAKING META-MORAL MACHINES

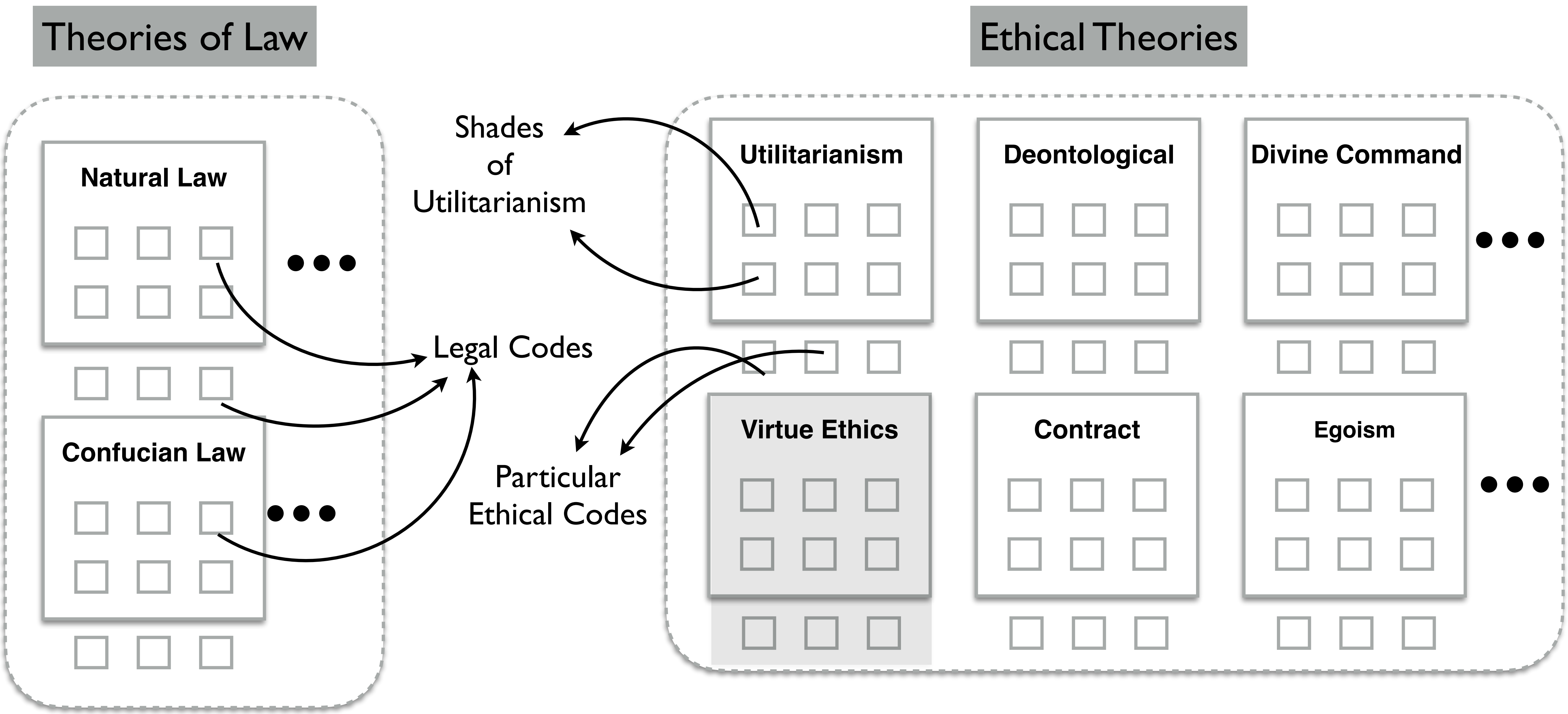
Pick the theory; pick the code; use the Leibnizian operators; engineer the robot

The machine picks the theory; picks the code; ...

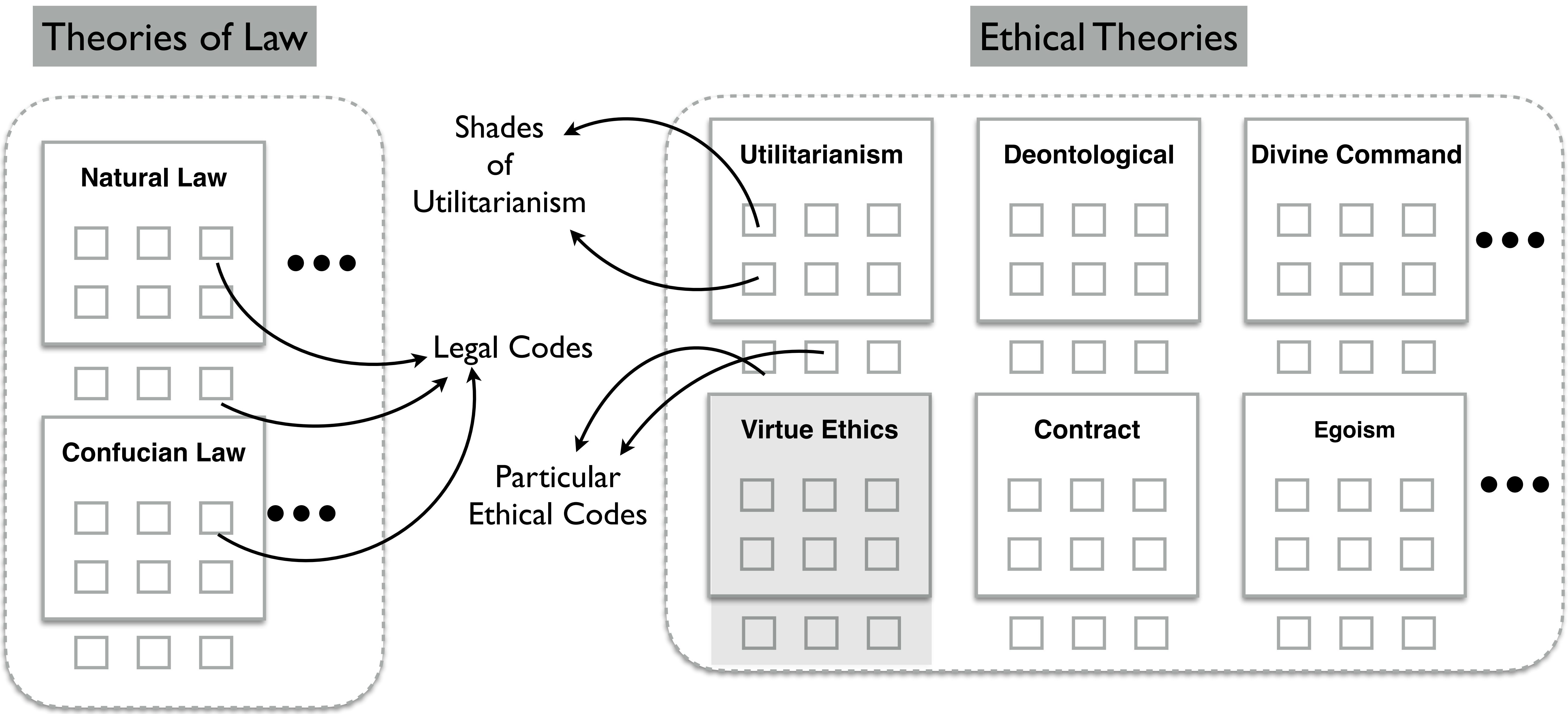
2016

Making Morally X Machines; Only Logic Can Save Us




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

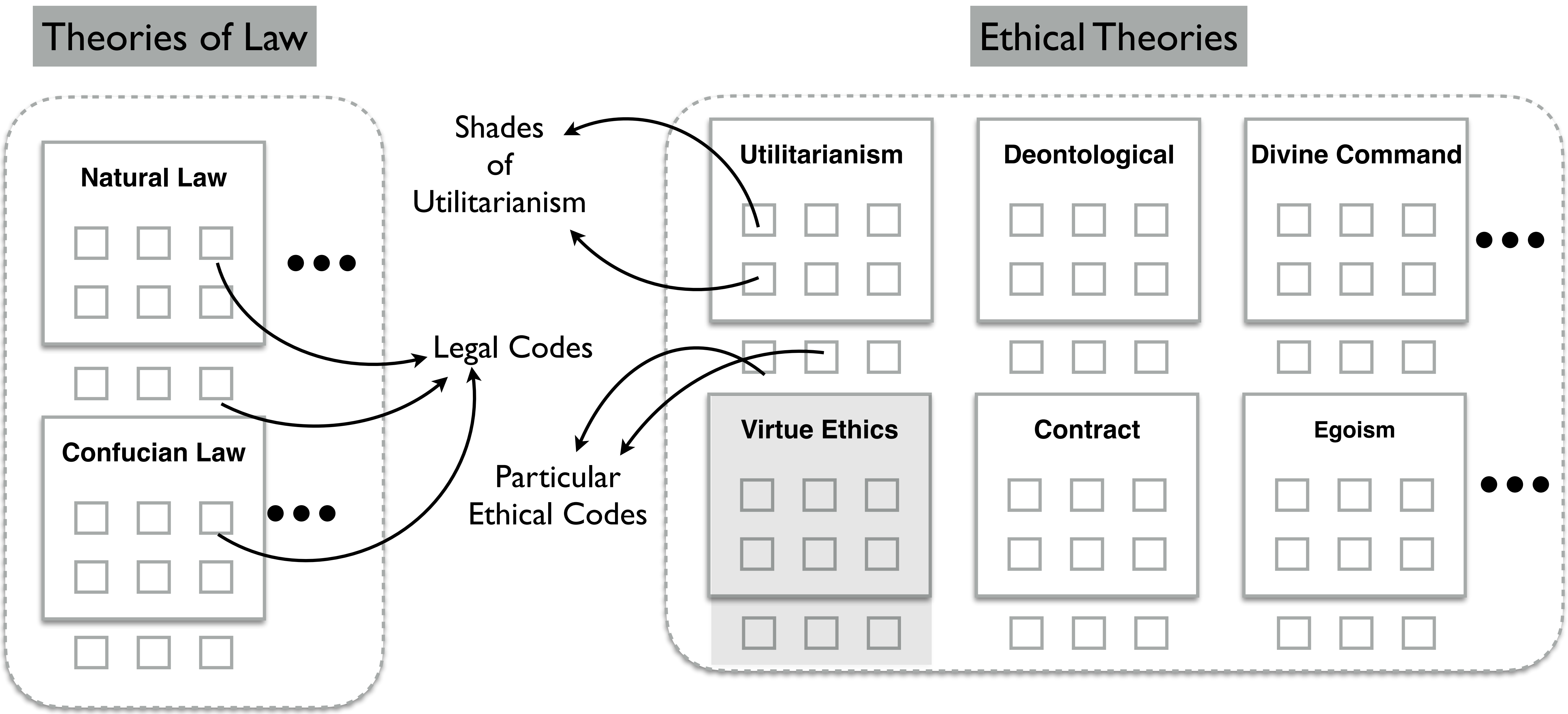
Making Morally X Machines; Only Logic Can Save Us



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.
4. Formalize in a Cognitive Calculus.

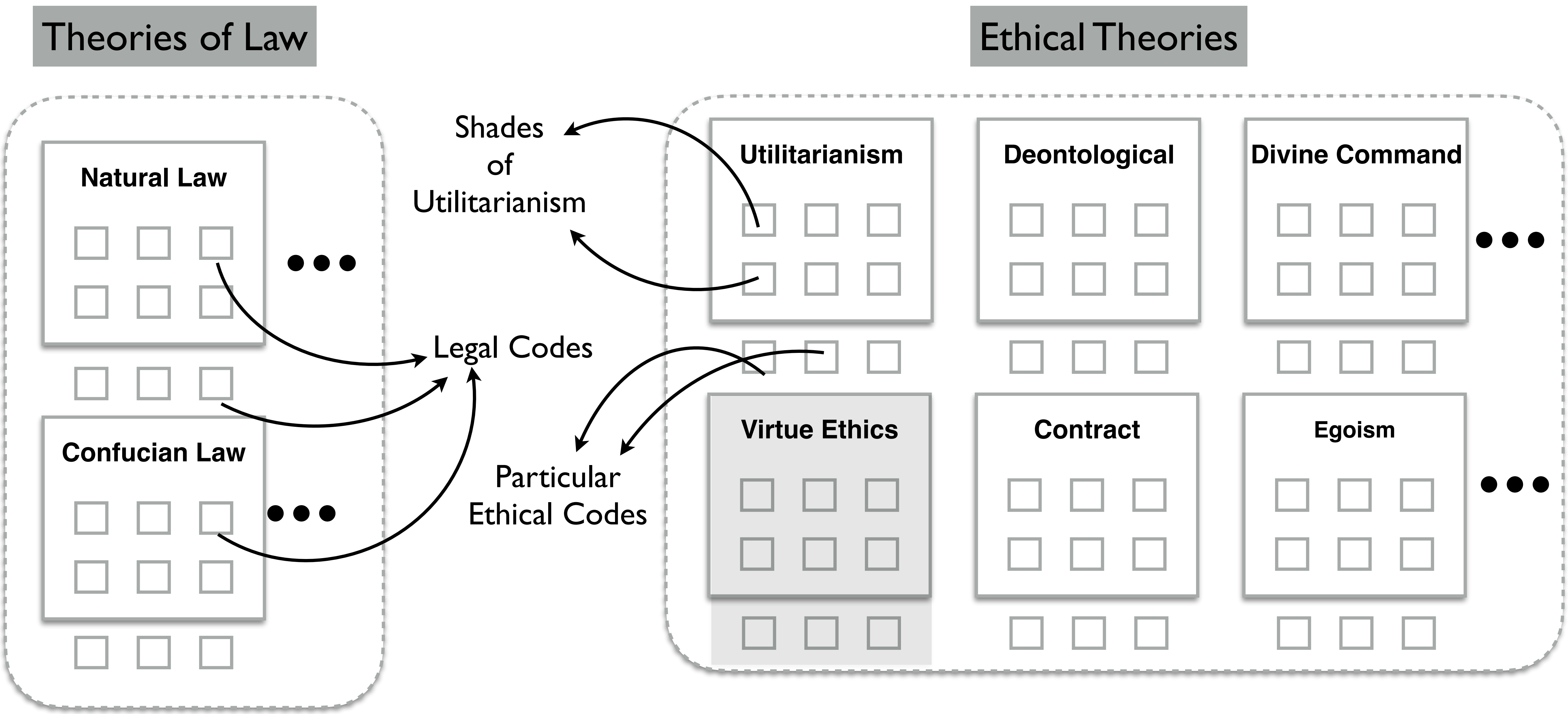
Making Morally X Machines; Only Logic Can Save Us



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.
4. Formalize in a Cognitive Calculus.

Making Morally X Machines; Only Logic Can Save Us



Step 1

1. Pick a theory
2. Pick a code
3. Run through EH.
4. Formalize in a Cognitive Calculus.

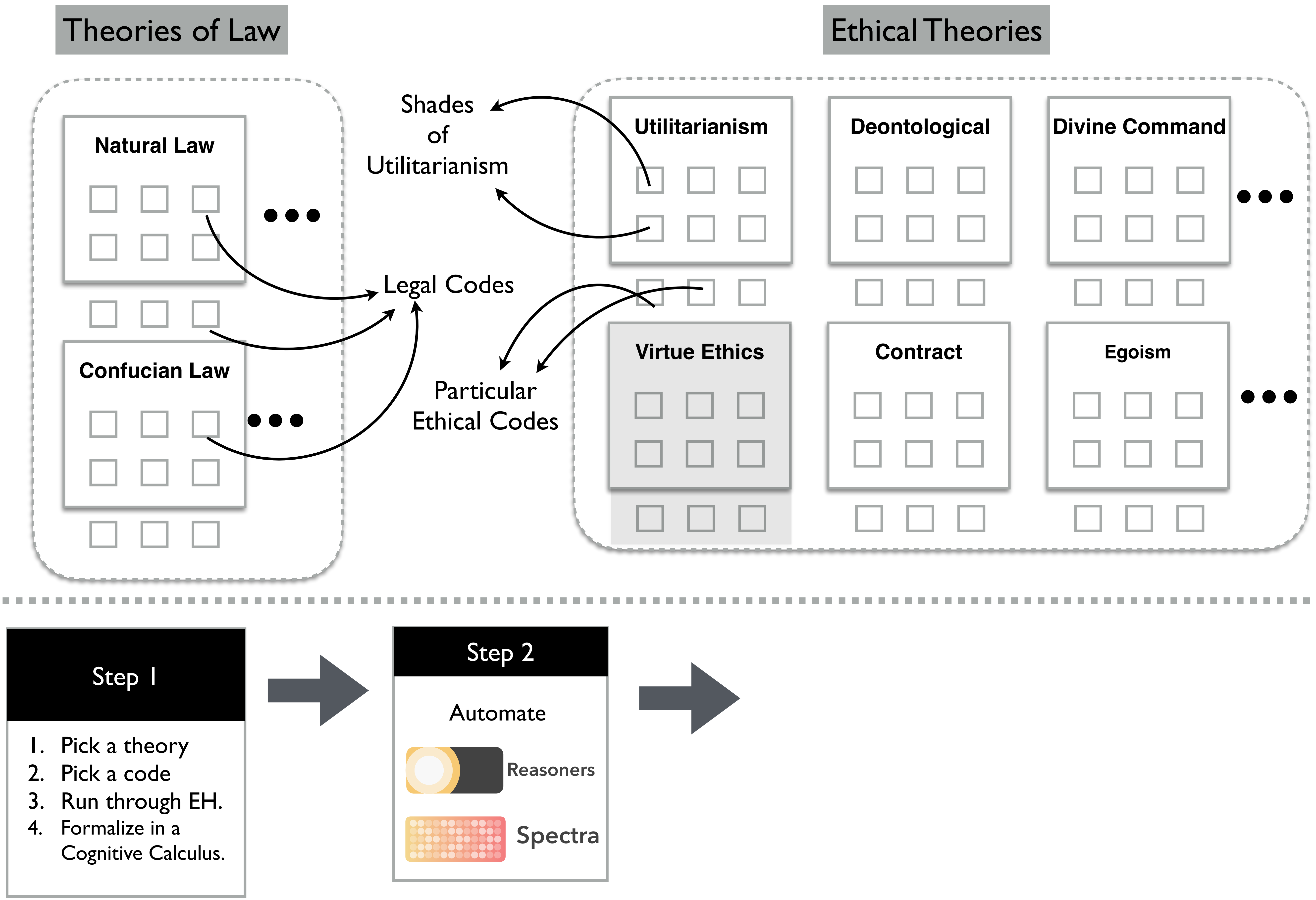
Step 2

Automate

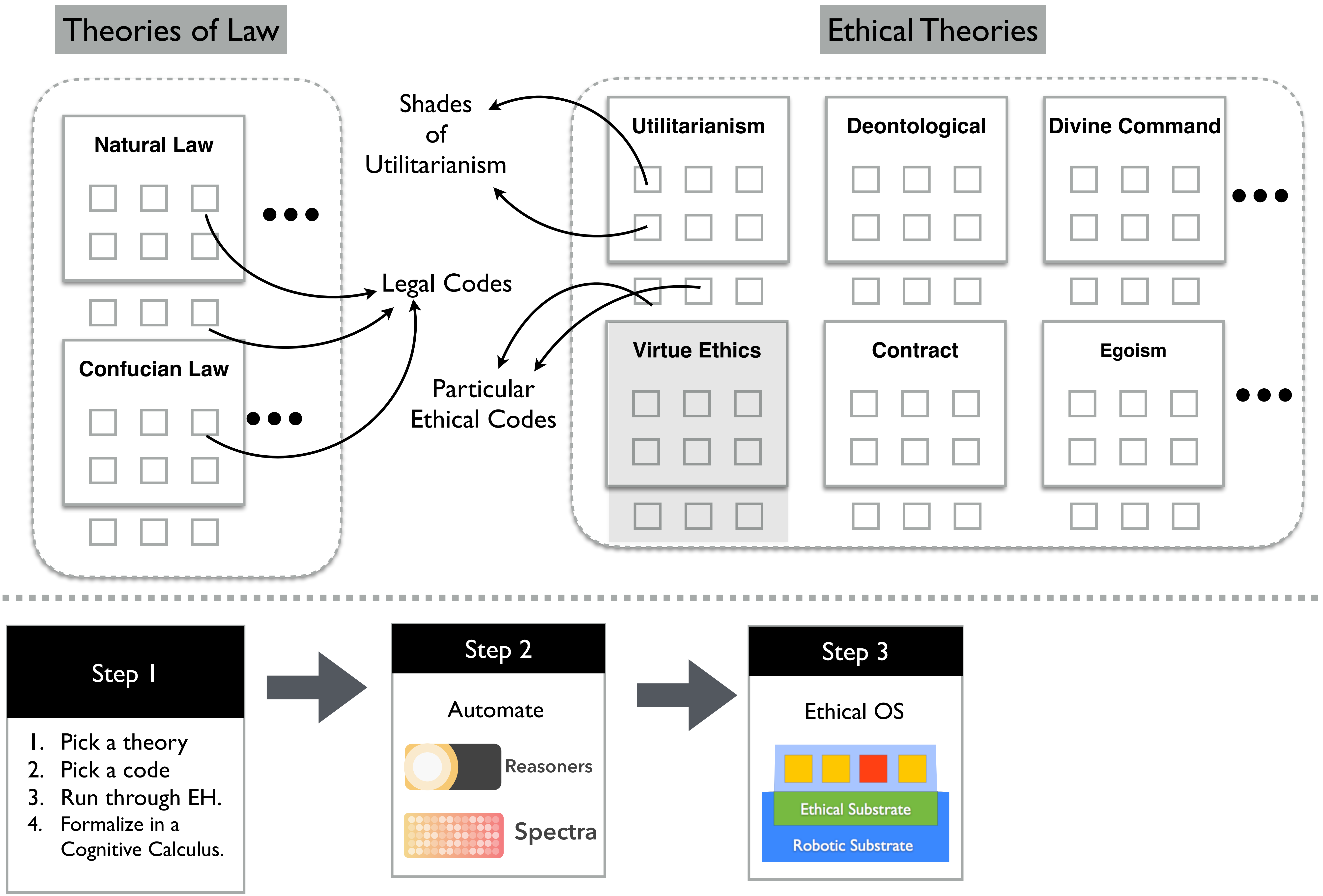
Reasoners

Spectra

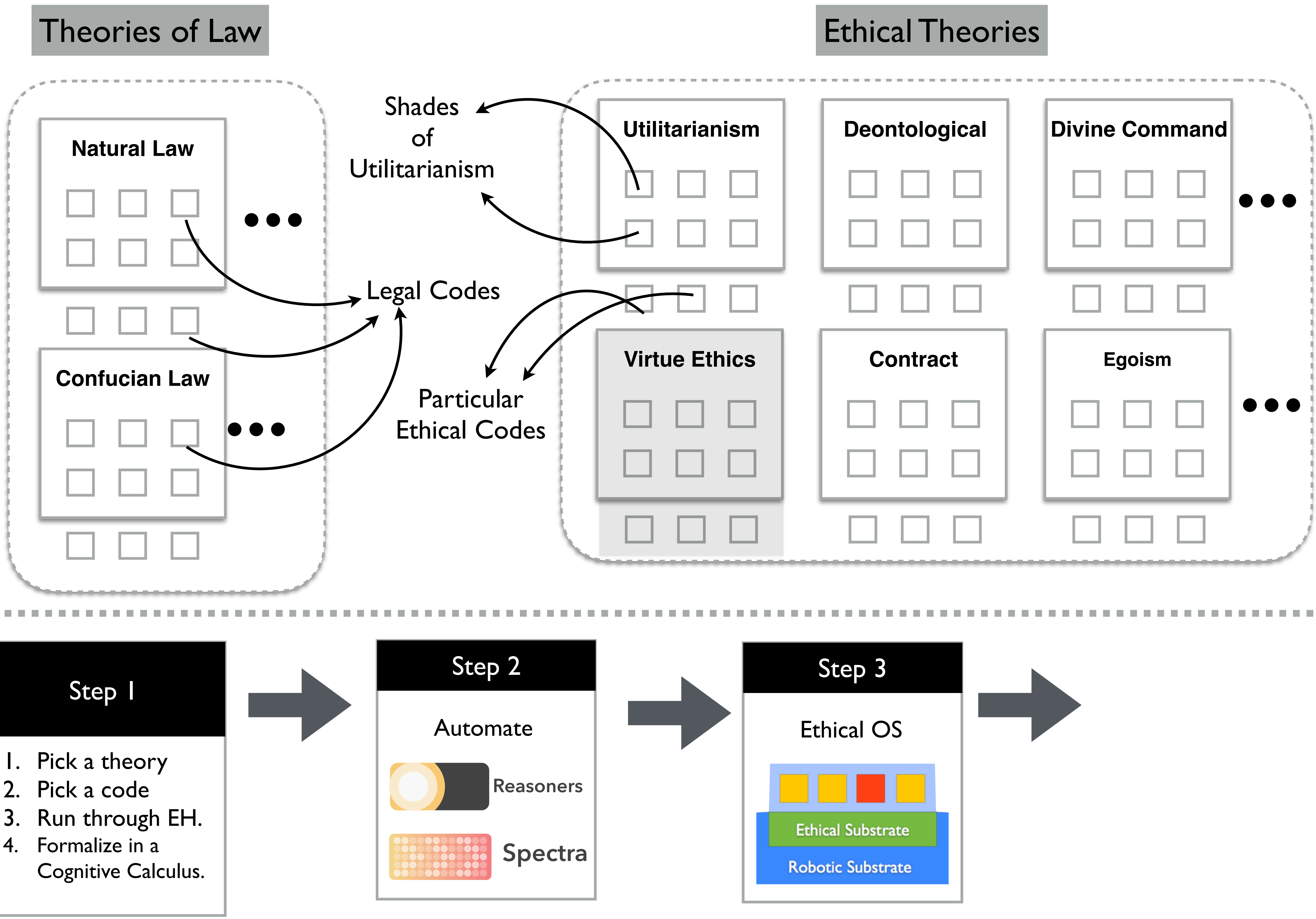
Making Morally X Machines; Only Logic Can Save Us

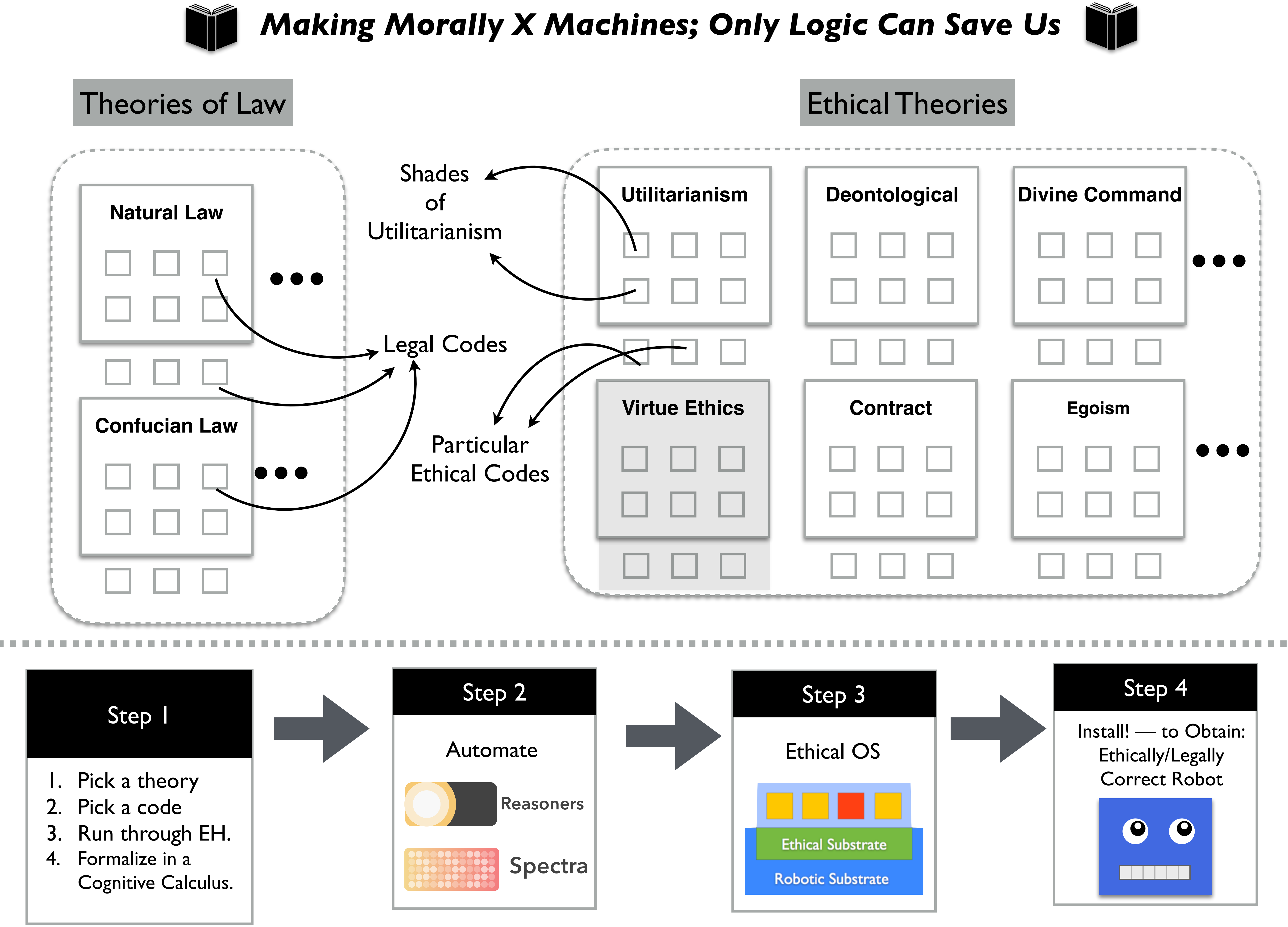


Making Morally X Machines; Only Logic Can Save Us

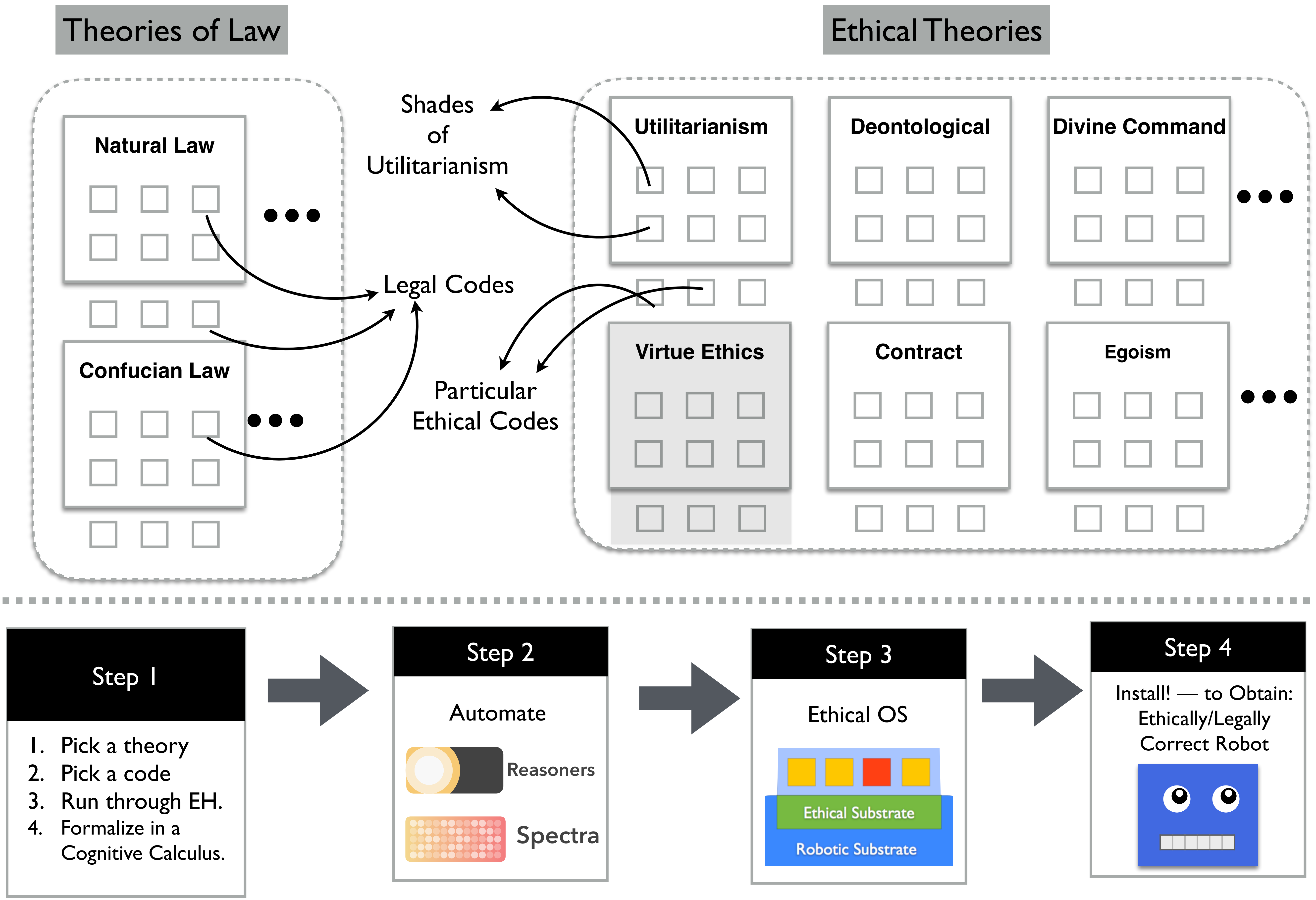


Making Morally X Machines; Only Logic Can Save Us




Making Morally X Machines; Only Logic Can Save Us
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.
4. Formalize in a Cognitive Calculus.
Step 2
Automate
Reasoners
Spectra
Step 3
Ethical OS
Ethical Substrate
Robotic Substrate
Step 4
Install! — to Obtain: Ethically/Legally Correct Robot

Making Morally X Machines; Only Logic Can Save Us



e.g. "Toward the Engineering of Virtuous Robots" Naveen, Selmer et al.

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

“The Ultimate Computer”
S2 E24

The Four Steps aren’t operative here!! …

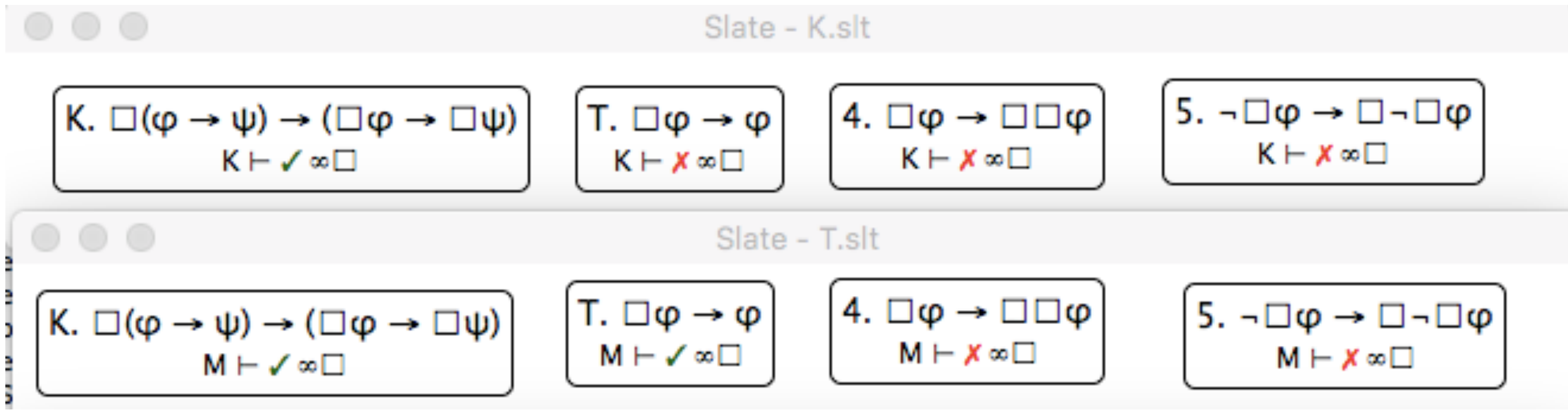
Well, maybe, but at any rate, *what logic*??

Well, maybe, but at any rate, *what logic*??

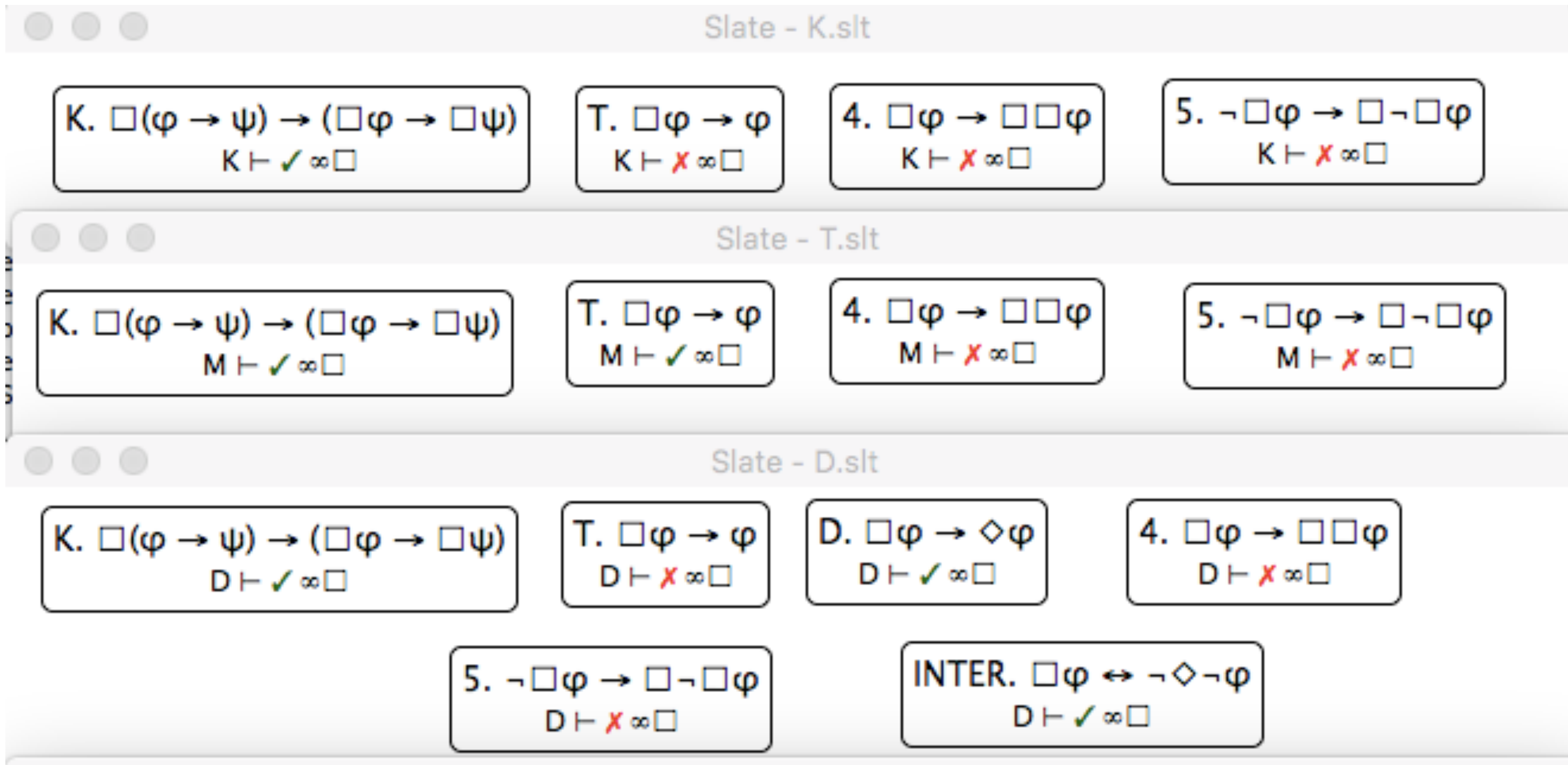
Not **D** = **SDL**! …

Review: Encapsulation

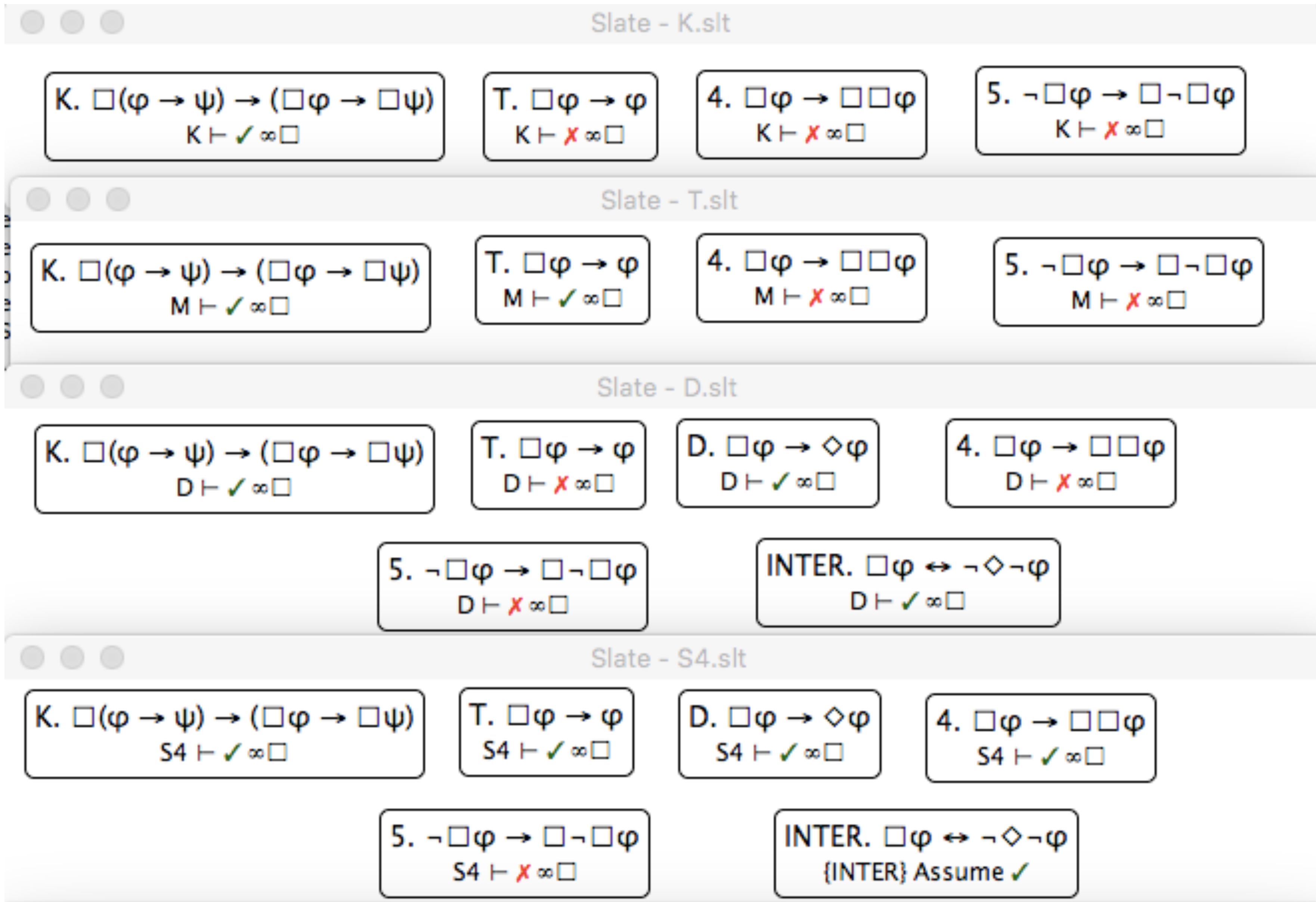
Review: Encapsulation



Review: Encapsulation



Review: Encapsulation



HyperSlate®

Review: Encapsulation

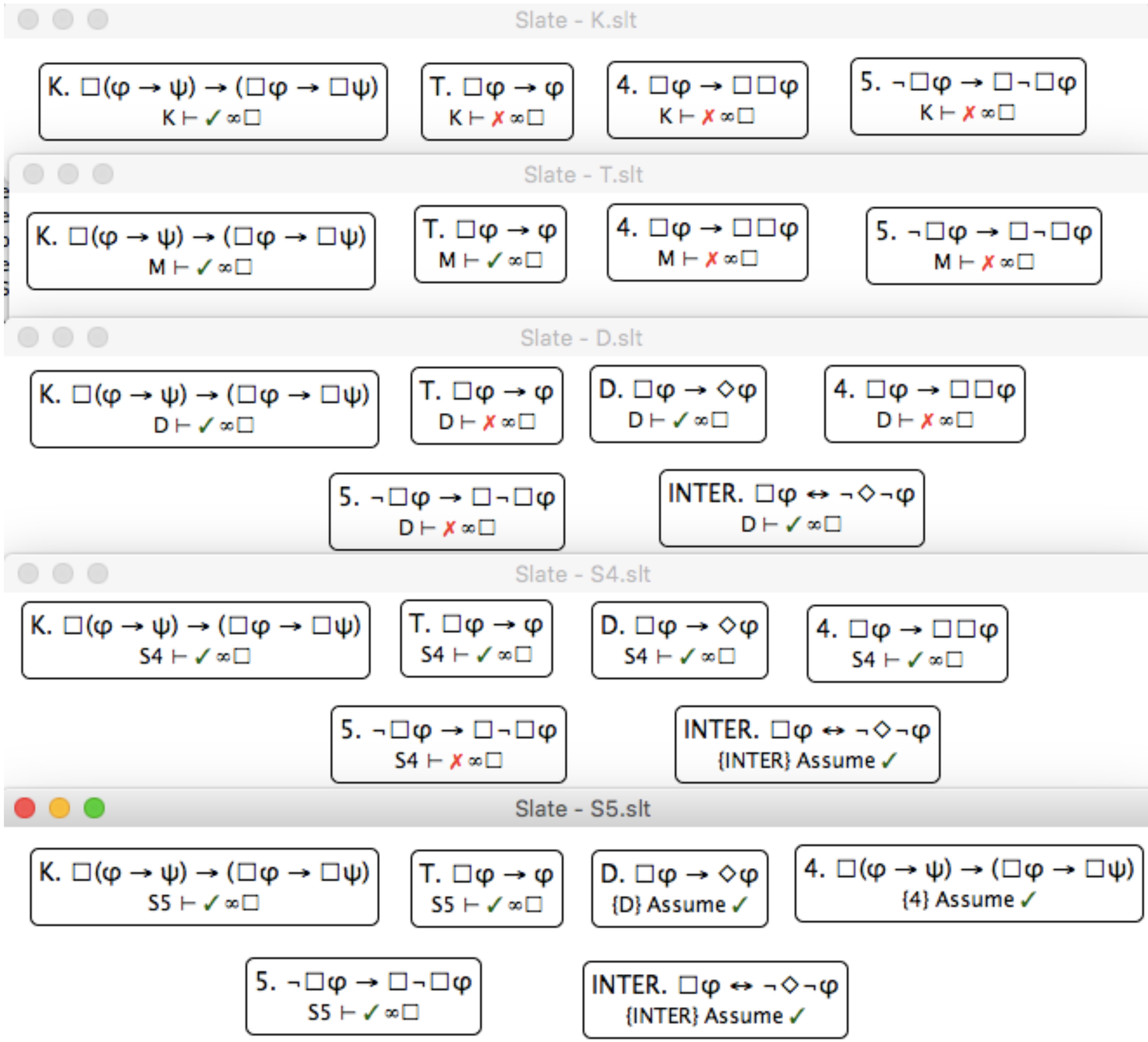
K

T

D

4 = S4

5 = S5



HyperSlate®

Review: Encapsulation

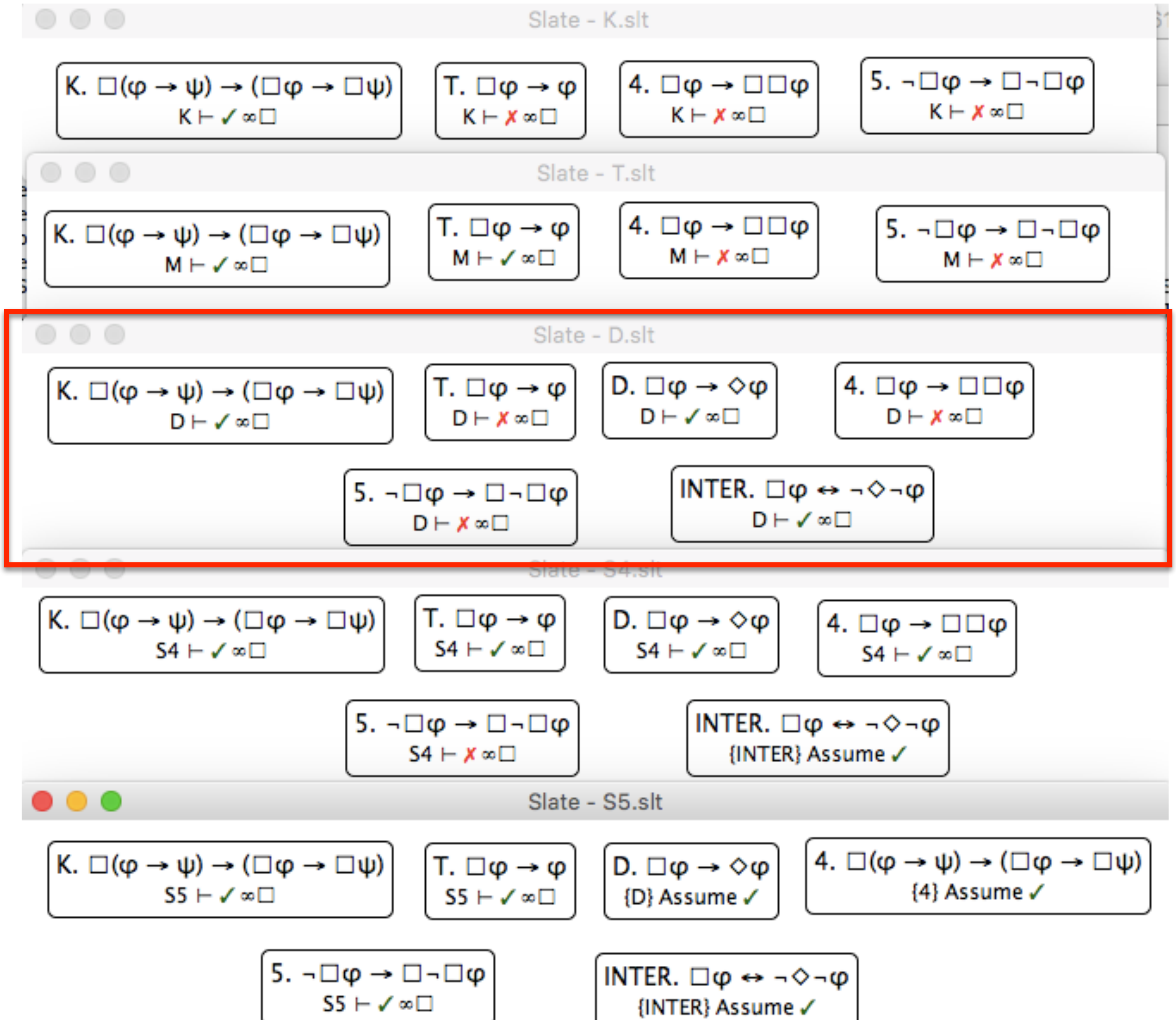
K

T

D

4 = S4

5 = S5



Chisholm’s Paradox …

4.4.4 D = SDL (= 'Standard Deontic Logic')

We here introduce what is known as 'Standard Deontic Logic' (**SDL**), which in Slate is the system **D**. Deontic logic is the sub-branch of logic devoted to formalizing the fundamental concepts of morality; for example, the concepts of *obligation*, *permissibility*, and *forbiddenness*. The first of these three concepts can apparently serve as a cornerstone, since to say that ϕ (a formulae representing some state-of-affairs) is permissible seems to amount to saying that it's not obligatory that it not be the case that ϕ (which shows permissibility can be defined in terms of obligation), and to say that ϕ is forbidden would seem to amount to it being obligatory that it not be the case that ϕ (which of course appears to show that forbiddenness buildable from obligation). This interconnected trio of ethical concepts is a triad explicitly invoked and analyzed since the end of the 18^{th} century, and the importance of the triad even to modern deontic logic would be quite hard to exaggerate.[9]

SDL is traditionally axiomatized by the following:[10]

SDL

TAUT All theorems of the propositional calculus.

OB-K $\odot(\phi \rightarrow \psi) \rightarrow (\odot\phi \rightarrow \odot\psi)$

OB-D $\odot\phi \rightarrow \neg\odot\neg\phi$

MP If $\vdash \phi$ and $\vdash \phi \rightarrow \psi$, then $\vdash \quad \phi$

OB-NEC If $\vdash \phi$ then $\vdash \odot\phi$

4.4.4 D = SDL (= 'Standard Deontic Logic')

We here introduce what is known as 'Standard Deontic Logic' (**SDL**), which in Slate is the system **D**. Deontic logic is the sub-branch of logic devoted to formalizing the fundamental concepts of morality; for example, the concepts of *obligation*, *permissibility*, and *forbiddenness*. The first of these three concepts can apparently serve as a cornerstone, since to say that ϕ (a formulae representing some state-of-affairs) is permissible seems to amount to saying that i
that ϕ (which shows permissibility can be c
say that ϕ is forbidden would seem to amou
the case that ϕ (which of course appears to s
obligation). This interconnected trio of ethic
and analyzed since the end of the 18^{th} centu
to modern deontic logic would be quite hard

SDL is traditionally axiomatized by the fo

SDL

TAUT All theorems of the proposition

OB-K $\odot(\phi \rightarrow \psi) \rightarrow (\odot\phi \rightarrow \odot\psi)$

OB-D $\odot\phi \rightarrow \neg\odot\neg\phi$

MP If $\vdash \phi$ and $\vdash \phi \rightarrow \psi$, then $\vdash \quad \phi$

OB-NEC If $\vdash \phi$ then $\vdash \odot\phi$

OB-RE If $\vdash \phi \longleftrightarrow \psi$, then $\vdash \odot\phi \longleftrightarrow \odot\psi$.

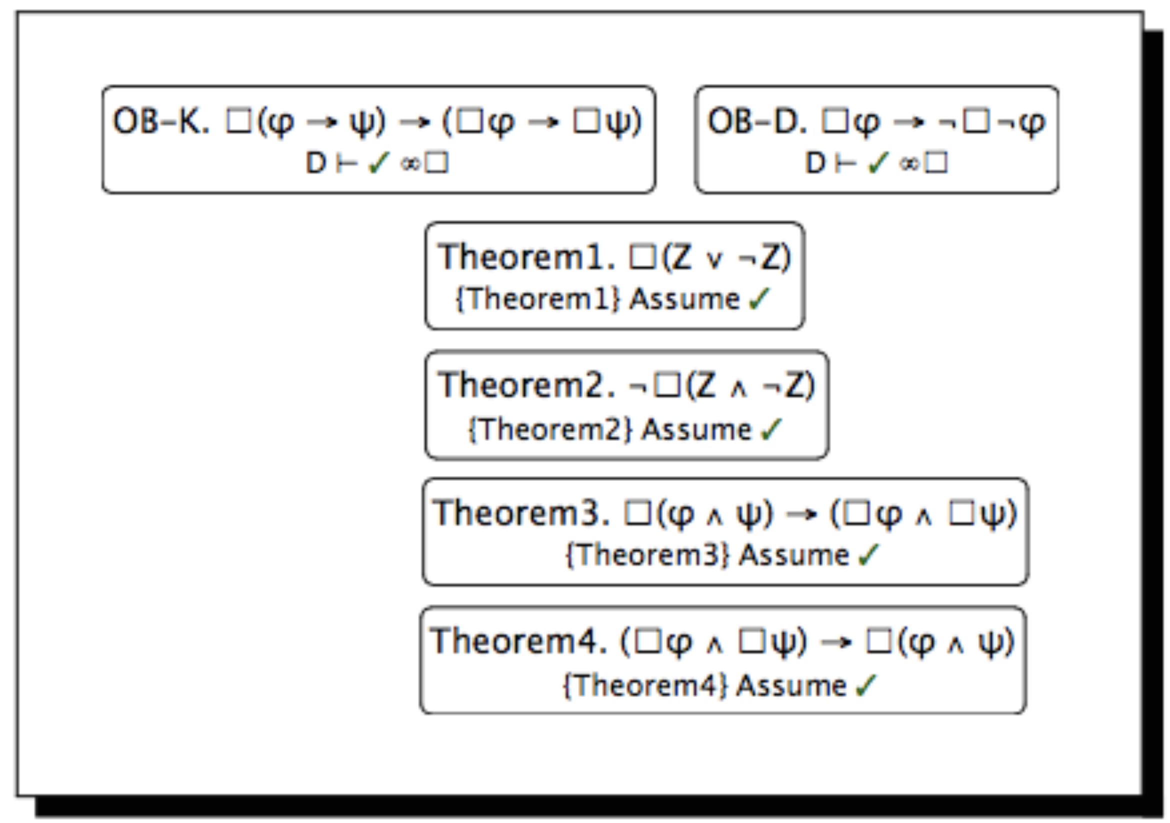


Figure 4.7: The Initial Configuration Upon Opening the File `SDL.slt`

4.4.4.1 Chisholm's Paradox and SDL

There are a host of problems that, together, constitute what is probably a fatal threat to **SDL** as a model of human-level ethical reasoning. We discuss in the present section the first of these problems to hit the "airwaves": Chisholm's Paradox (CP) (Chisholm 1963). CP can be generated in Slate, you we shall see. But before we get to the level of experimentation in Slate, let's understand the scenario that Chisholm's imagined.

Chisholm's clever scenario revolves around the character Jones.[11] It's given that Jones is obligated to go to assist his neighbors, in part because he has promised to do so. The second given fact is that it's obligatory that, if Jones goes to assist his neighbors, he tells them (in advance) that he is coming. In addiiton, and this is the third given, if Jones *doesn't* go to assist his neighbors, it's obligatory that he not tell

[11] We change some particulars to ease exposition; generally, again, follow, the *SEP* entry on deontic logic (recall footnote 10). The core logic mirrors (Chisholm 1963), the original publication.

them that he is coming. The fourth and final given fact is simply that Jones doesn't go to assist his neighbors. (On the way to do so, suppose he comes upon a serious vehicular accident, is proficient in emergency medicine, and (commendably!) seizes the opportunity to save the life (and subsequently monitor) of one of the victims in this accident.) These four givens have been represented in an obvious way within four formula nodes in a Slate file; see Figure 4.8. (Notice that $\Box$ is used in place of $\odot$.) The paradox arises from the fact that Chisholm's quartet of givens, which surely reflect situations that are common in everyday life, in conjunction with the axioms of **SDL**, entail outright contradictions (see Exercise 2 for $\mathbf{D} = \mathbf{SDL}$, in §4.4.4.2).

4.4.4.1 Chisholm's Paradox and SDL

There are a host of problems that, together, constitute what is probably a fatal threat to **SDL** as a model of human-level ethical reasoning. We discuss in the present section the first of these problems to hit the "airwaves": Chisholm's Paradox (CP) (Chisholm 1963). CP can be generated in Slate, you we shall see. But before we get to the level of experimentation in Slate, let's understand the scenario that Chisholm's imagined.

Chisholm's clever scenario revolves around the character Jones.[11] It's given that Jones is obligated to go to assist his neighbors, in part because he has promised to do so. The second given fact is that it's obligatory that, if Jones goes to assist his neighbors, he tells them (in advance) that he is coming. In addiiton, and this is the third given, if Jones *doesn't* go to assist his neighbors, it's obligatory that he not tell

[11]We change some particulars to ease exposition; generally, again, follow, the *SEP* entry on deontic logic (recall footnote 10). The core logic mirrors (Chisholm 1963), the original publication.

them that he is coming. The fourth and final given fact is simply that Jones doesn't go to assist his neighbors. (On the way to do so, suppose he comes upon a serious vehicular accident, is proficient in emergency medicine, and (commendably!) seizes the opportunity to save the life (and subsequently monitor) of one of the victims in this accident.) These four givens have been represented in an obvious way within four formula nodes in a Slate file; see Figure 4.8. (Notice that $\Box$ is used in place of $\odot$.) The paradox arises from the fact that Chisholm's quartet of givens, which surely reflect situations that are common in everyday life, in conjunction with the axioms of **SDL**, entail outright contradictions (see Exercise 2 for **D** = **SDL**, in §4.4.4.2).

Chisholm's Paradox

Axiom2. □(P → Q) → (□P → □Q)
D ⊢ ✓ ∞□

Given2. □(goes_assist_neighbors → tells_coming)
{Given2} Assume ✓

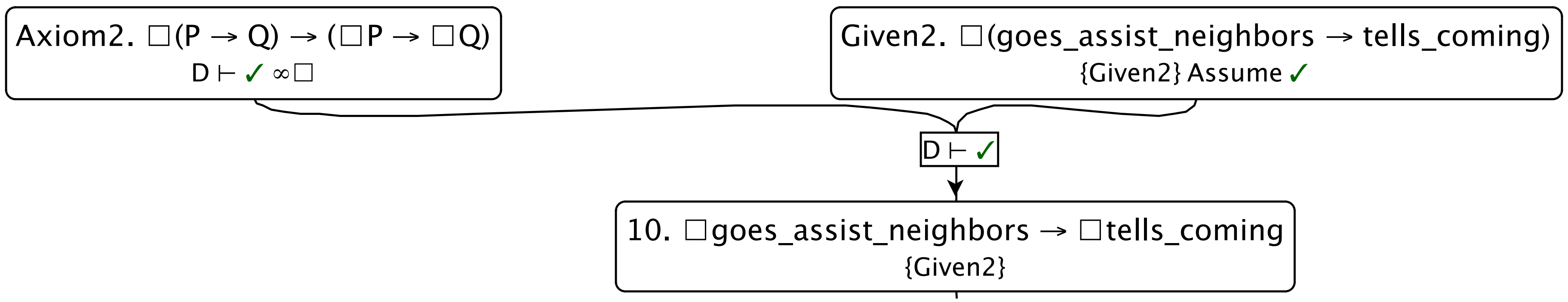
D ⊢ ✓

Axiom4. "Modus ponens for provability."
{Axiom4} Assume ✓

Axiom5. "Theorems are obligatory."
{Axiom5} Assume ✓

Axiom1. "All theorems of the propositional calculus."
{Axiom1} Assume ✓

Chisholm’s Paradox

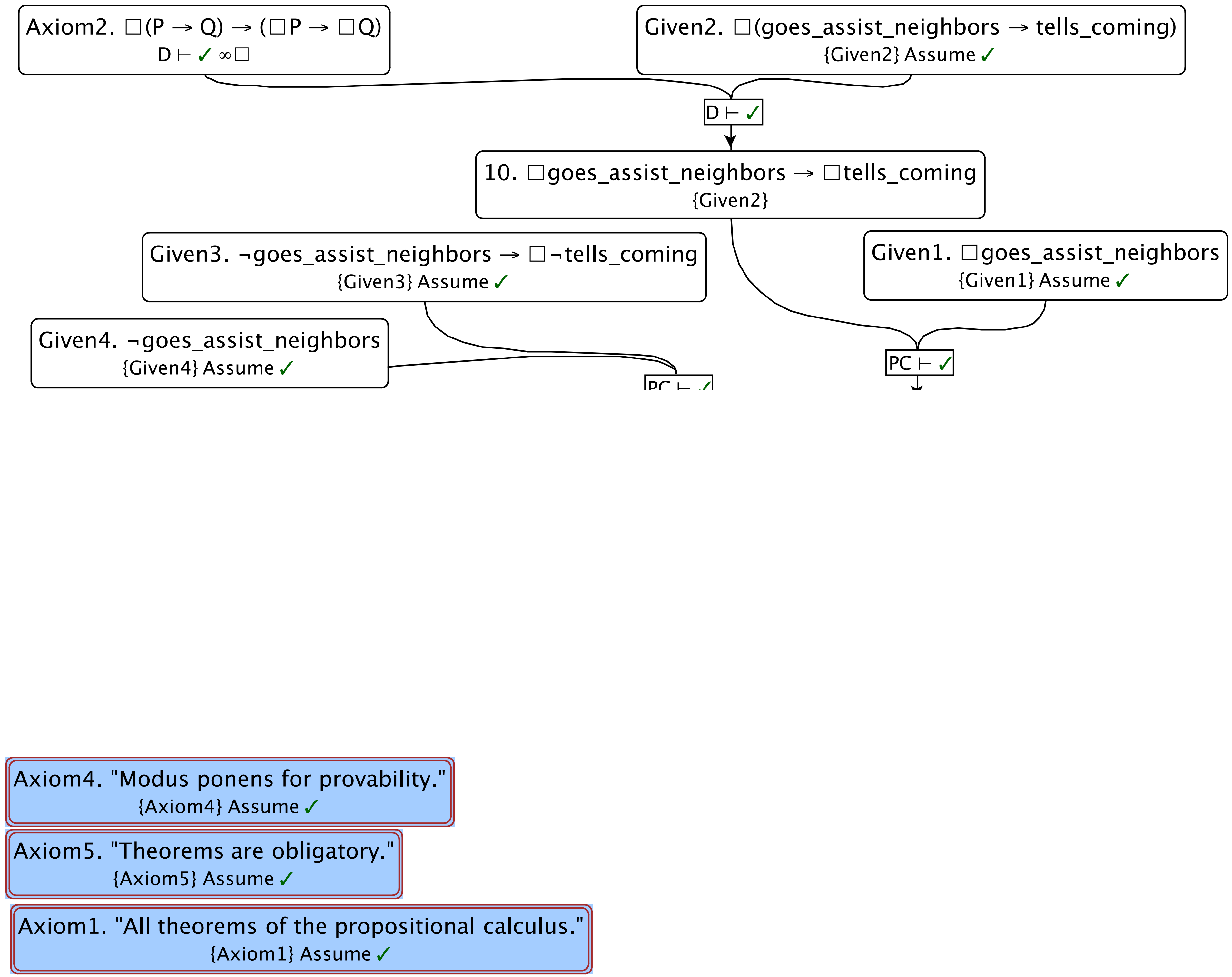


Axiom4. "Modus ponens for provability."
{Axiom4} Assume ✓

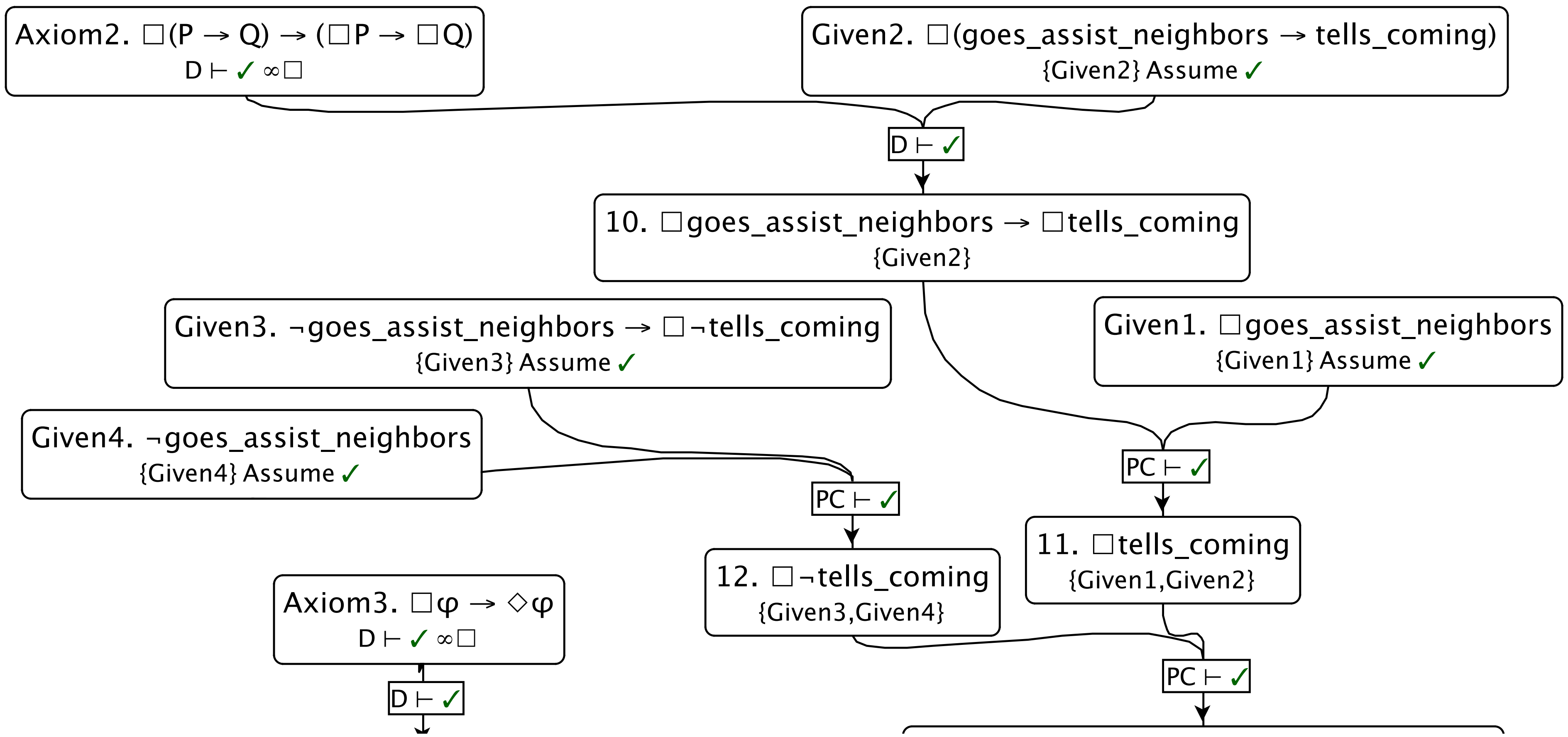
Axiom5. "Theorems are obligatory."
{Axiom5} Assume ✓

Axiom1. "All theorems of the propositional calculus."
{Axiom1} Assume ✓

Chisholm's Paradox



Chisholm's Paradox

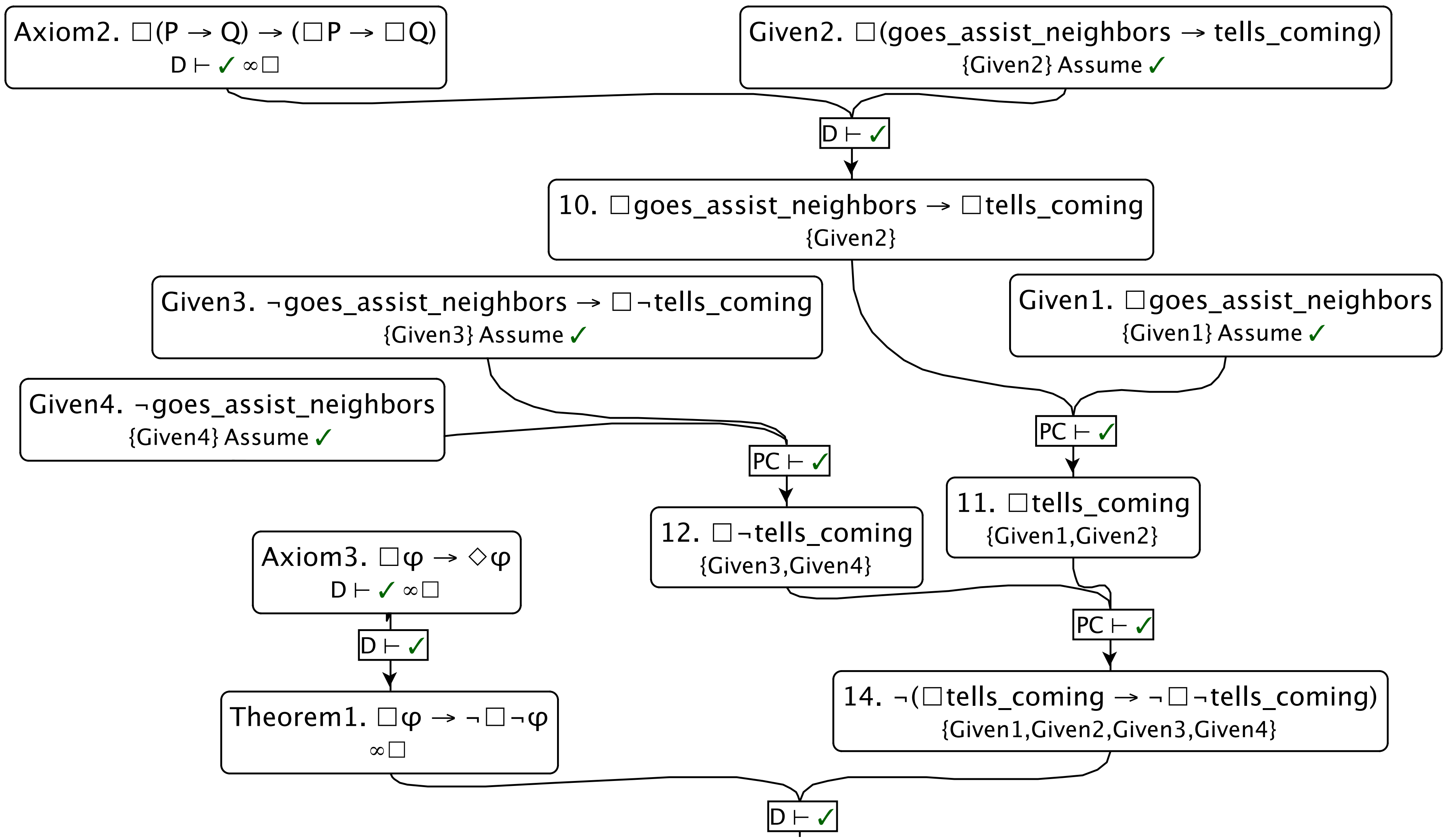


Axiom4. "Modus ponens for provability."
{Axiom4} Assume ✓

Axiom5. "Theorems are obligatory."
{Axiom5} Assume ✓

Axiom1. "All theorems of the propositional calculus."
{Axiom1} Assume ✓

Chisholm's Paradox

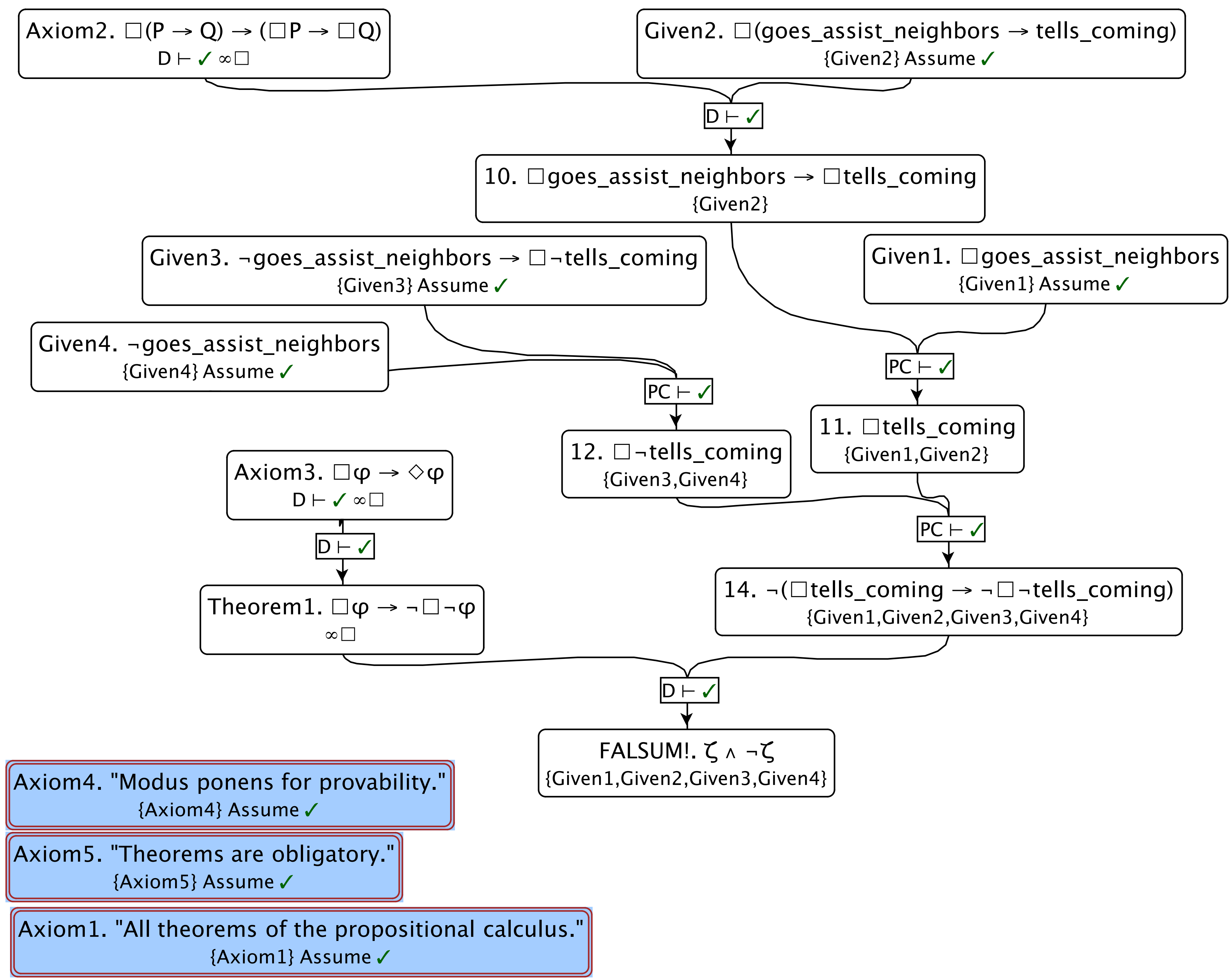


Axiom4. "Modus ponens for provability."
{Axiom4} Assume ✓

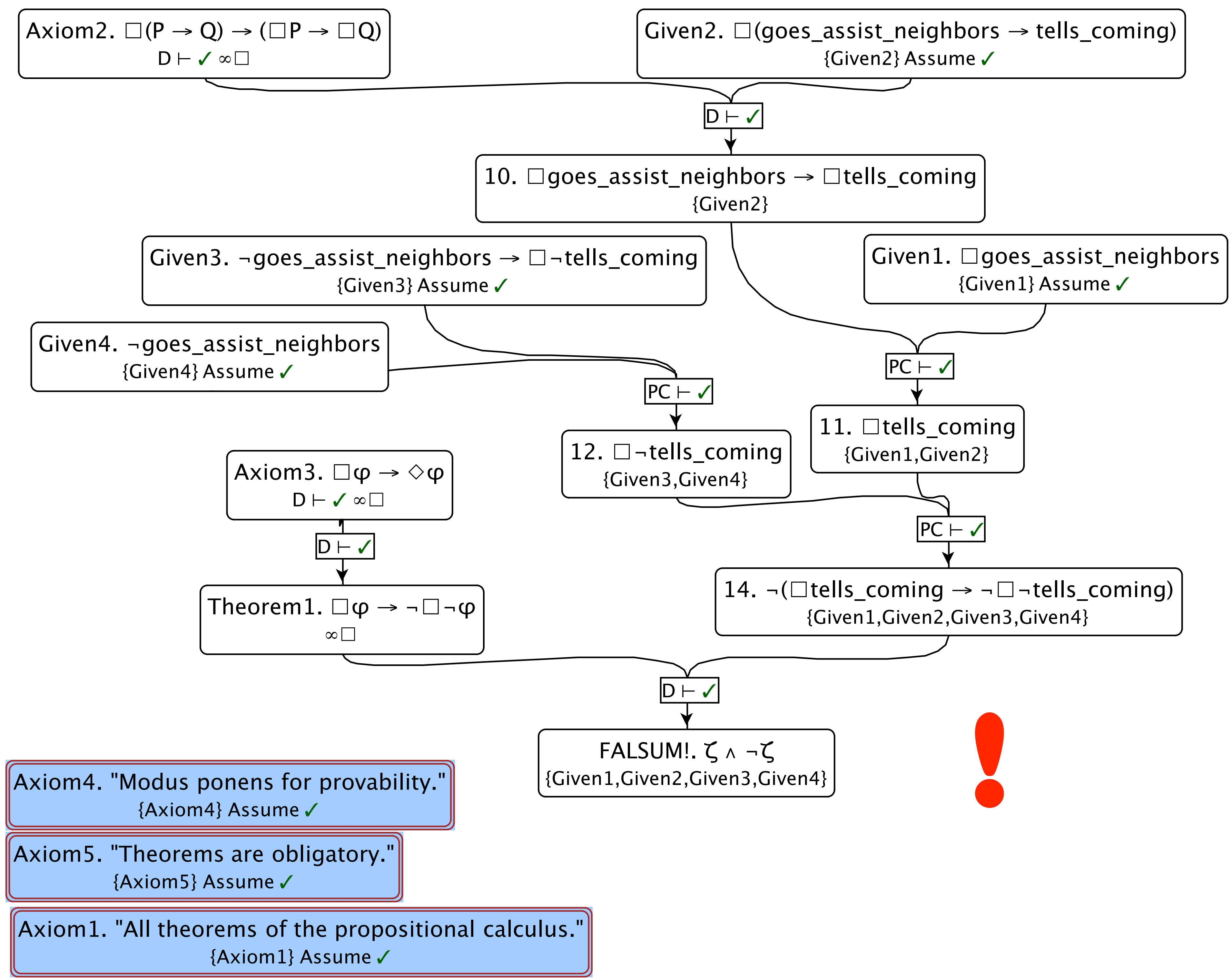
Axiom5. "Theorems are obligatory."
{Axiom5} Assume ✓

Axiom1. "All theorems of the propositional calculus."
{Axiom1} Assume ✓

Chisholm's Paradox



Chisholm’s Paradox



Required

ChisholmsParadox

Here you are asked to build a proof that confirms *Chisholm's Paradox*. This paradox is that from a particular representation in **D** (= Standard Deontic Logic (SDL)) of four seemingly innocuous givens, a contradiction $\zeta \land \neg\zeta$ can be deduced. (Your instructor should have covered this in class, and may well have supplied a proof of CP.) The four givens are based on the story of a character Jones, who is obligated to go to assist his neighbors (move to a different domicile, e.g.). It would be wrong of him to show up unannounced, though; so if he goes to assist them, it ought to be that he tells them he's coming. In addition, if it's not the case that Jones goes to assist them, then it ought to be that it not be the case that he tells them he is coming. Finally, as a matter of fact, it's not case the Jones goes to assist (because on the way he comes across a car accident, and has an opportunity to save one of the victims).

Fortunately, the RAIR Lab's modern cognitive calculus $\mathcal{DCEC}^*$ allows Chisholm's Paradox to be avoided. A recent paper explaining the use by an ethically correct AI of this calculus is available here.

Your finished proof is allowed to make use of the PC provabiity oracle, but of no other oracle.

Deadline November 12, 2020, 3:00 PM EST

Overall Measures for: ChisholmsParadox

Total Submissions	Passed	Failed	Unprocessed
38	37	1	0

ChisholmsParadox_Mon_Nov_09_2020_13:50:13_GMT-0500_(EST).csv

SDL's = **D**'s Problems
Don't Stop Here:
The Free Choice
Permission Paradox …

The Free Choice Permission Paradox (Ross)

1. "You may either sleep on the sofa bed or the guest bed."
{1} Assume ✓

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

The Free Choice Permission Paradox (Ross)

1'. $\Diamond(\text{sofa-bed} \vee \text{guest-bed})$
{1'} Assume ✓

D ⊢ ✗

2'. $\Diamond\text{sofa-bed} \wedge \Diamond\text{guest-bed}$
{1'}

1. "You may either sleep on the sofa bed or the guest bed."
{1} Assume ✓

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

The Free Choice Permission Paradox (Ross)

1'. $\Diamond$(sofa–bed $\vee$ guest–bed)
{1'} Assume ✓

$\downarrow$ D $\vdash$ ✗

2'. $\Diamond$sofa–bed $\wedge$ $\Diamond$guest–bed
{1'}

1. "You may either sleep on the sofa bed or the guest bed."
{1} Assume ✓

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

NEW SCHEMA?. $\Diamond(\varphi \vee \psi) \rightarrow (\Diamond\varphi \wedge \Diamond\psi)$
{NEW SCHEMA?} Assume ✓

The Free Choice Permission Paradox (Ross)

1'. ◇(sofa-bed ∨ guest-bed)
{1'} Assume ✓

D ⊢ ✗

2'. ◇sofa-bed ∧ ◇guest-bed
{1'}

1. "You may either sleep on the sofa bed or the guest bed."
{1} Assume ✓

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

NEW SCHEMA?. $\Diamond(\varphi \vee \psi) \rightarrow (\Diamond\varphi \wedge \Diamond\psi)$
{NEW SCHEMA?} Assume ✓

COMMENT. "We can prove:"
{COMMENT} Assume ✓

THM 5. $\Diamond\varphi \rightarrow \Diamond(\varphi \vee \psi)$
D ⊢ ✓ ∞□

The Free Choice Permission Paradox (Ross)

1'. $\Diamond$(sofa-bed $\vee$ guest-bed)
{1'} Assume ✓

D ⊢ ✗

2'. $\Diamond$sofa-bed $\wedge$ $\Diamond$guest-bed
{1'}

1. "You may either sleep on the sofa bed or the guest bed."
{1} Assume ✓

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

NEW SCHEMA?. $\Diamond(\varphi \vee \psi) \rightarrow (\Diamond\varphi \wedge \Diamond\psi)$
{NEW SCHEMA?} Assume ✓

COMMENT. "We can prove:"
{COMMENT} Assume ✓

THM 5. $\Diamond\varphi \rightarrow \Diamond(\varphi \vee \psi)$
D ⊢ ✓ ∞☐

(How?)

The Free Choice Permission Paradox (Ross)

1'. $\Diamond$(sofa-bed $\vee$ guest-bed)
{1'} Assume ✓

1. "You may either sleep on the sofa bed or the guest bed."
{1} Assume ✓

D $\vdash$ ✗

2'. $\Diamond$sofa-bed $\wedge$ $\Diamond$guest-bed
{1'}

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

NEW SCHEMA?. $\Diamond(\varphi \vee \psi) \rightarrow (\Diamond\varphi \wedge \Diamond\psi)$
{NEW SCHEMA?} Assume ✓

COMMENT. "We can prove:"
{COMMENT} Assume ✓

THM 5. $\Diamond\varphi \rightarrow \Diamond(\varphi \vee \psi)$
D $\vdash$ ✓ $\infty\Box$

(How?)

8. $\Diamond\varphi$
{8} Assume ✓

PC $\vdash$ ✓

The Free Choice Permission Paradox (Ross)

1'. ◇(sofa-bed ∨ guest-bed)
{1'} Assume ✓

D ⊢ ✗

2'. ◇sofa-bed ∧ ◇guest-bed
{1'}

1. "You may either sleep on the sofa bed or the guest bed."
{1} Assume ✓

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

NEW SCHEMA?. $\Diamond(\varphi \vee \psi) \rightarrow (\Diamond\varphi \wedge \Diamond\psi)$
{NEW SCHEMA?} Assume ✓

COMMENT. "We can prove:"
{COMMENT} Assume ✓

THM 5. $\Diamond\varphi \rightarrow \Diamond(\varphi \vee \psi)$
D ⊢ ✓ ∞☐

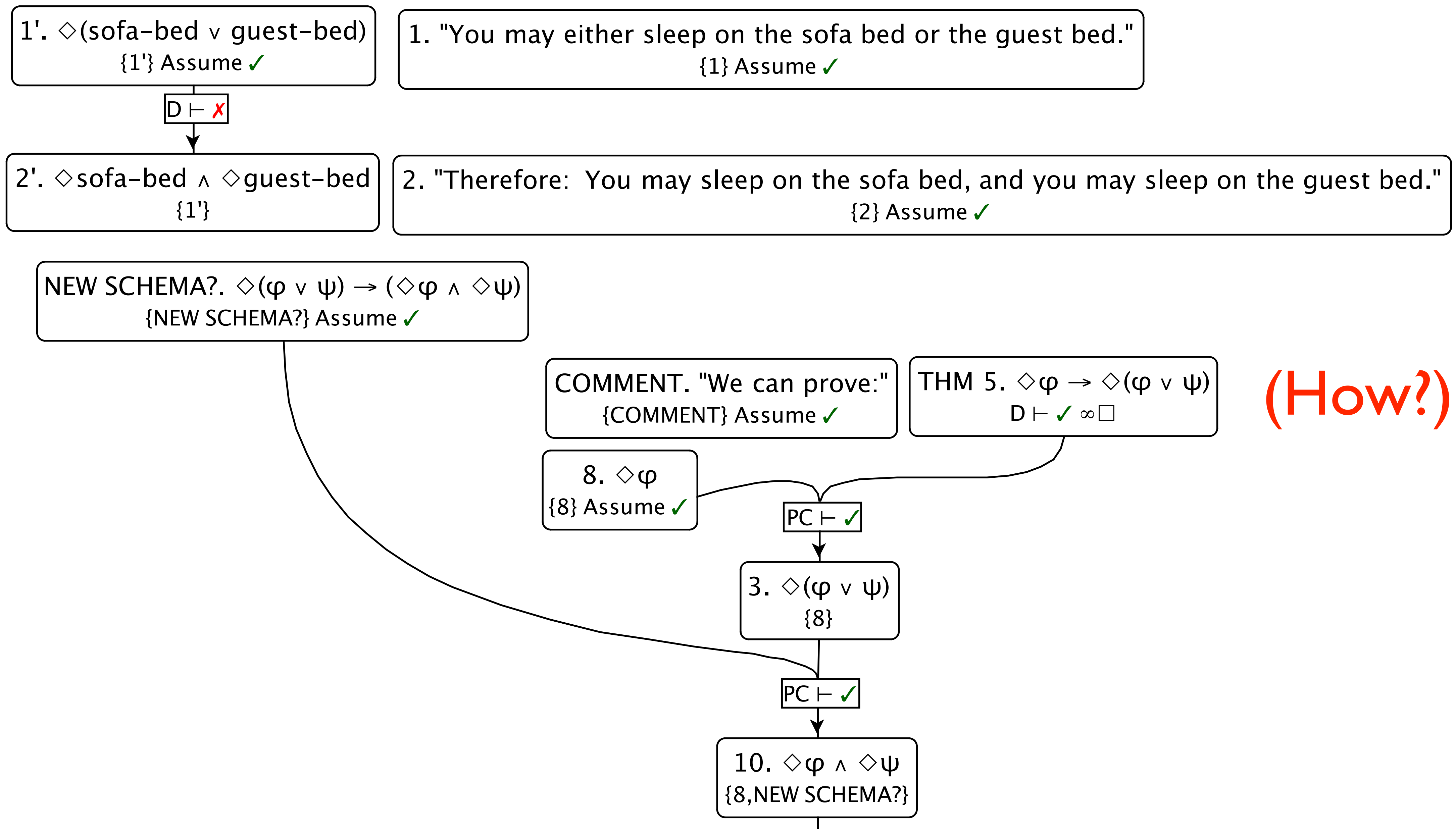
(How?)

8. $\Diamond\varphi$
{8} Assume ✓

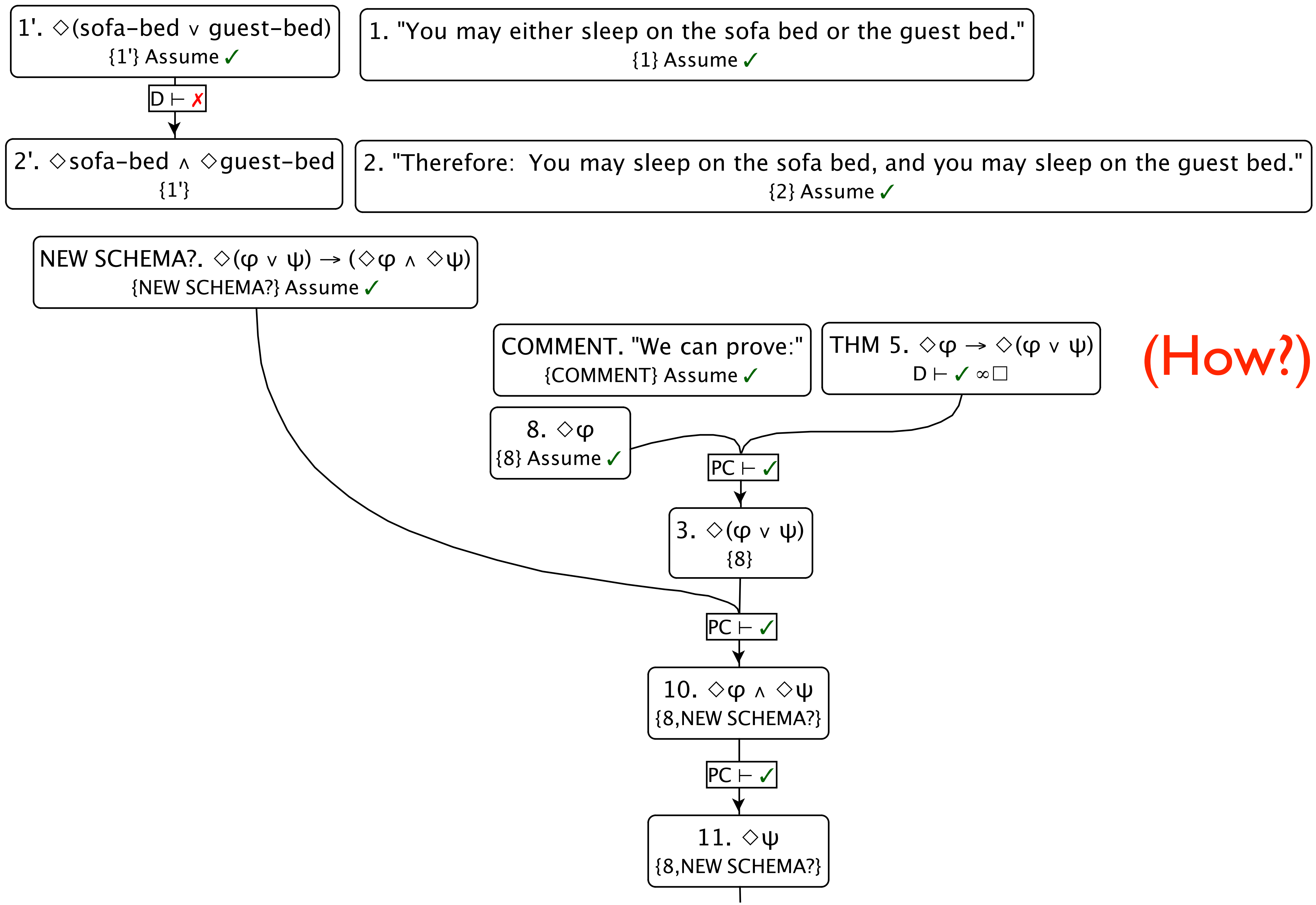
PC ⊢ ✓

3. $\Diamond(\varphi \vee \psi)$
{8}

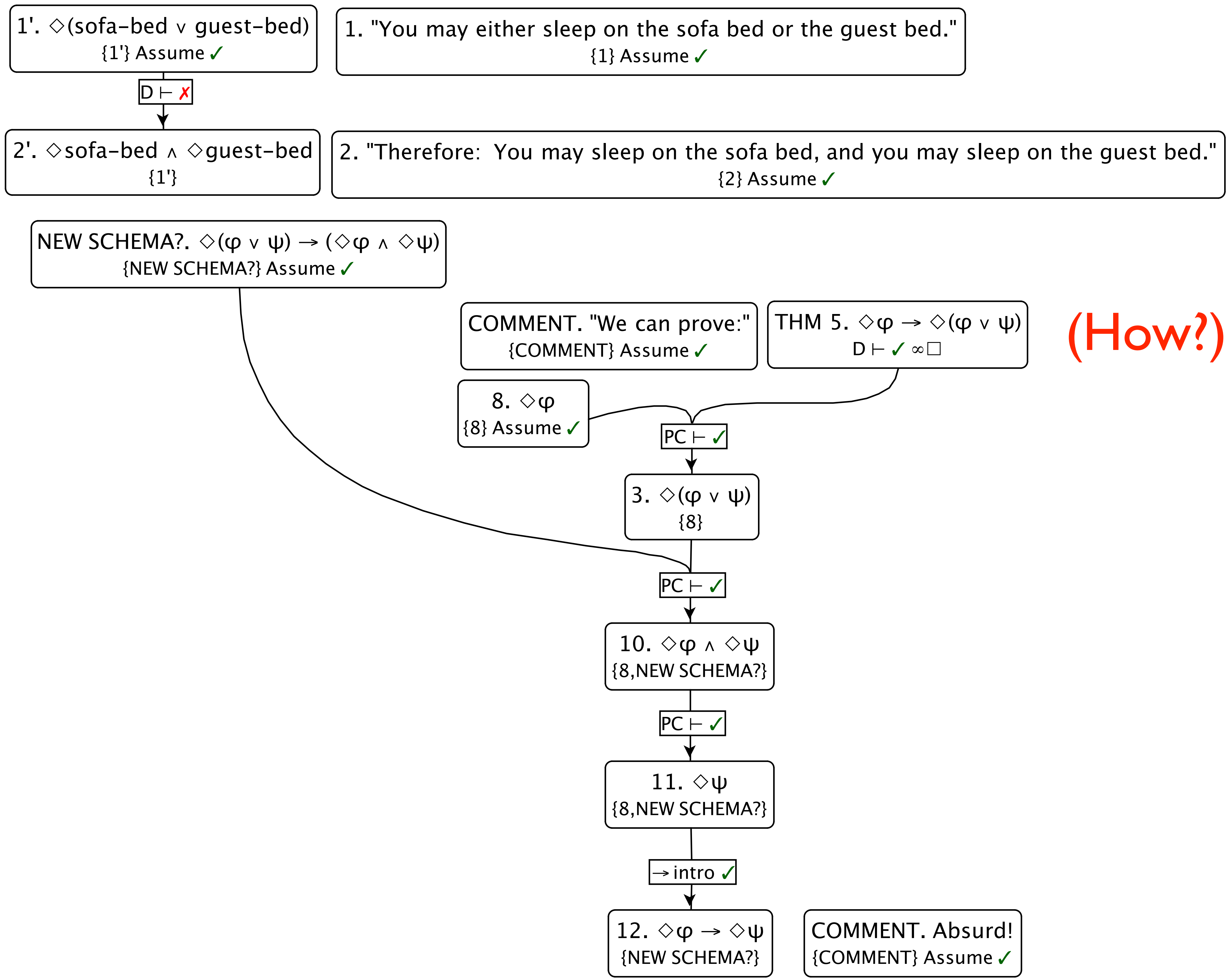
The Free Choice Permission Paradox (Ross)



The Free Choice Permission Paradox (Ross)



The Free Choice Permission Paradox (Ross)



Required

TheFreeChoicePermissionParadox

Producing a valid proof in this problem will enable you to understand The Free Choice Permission Paradox (FCPP), discovered in 1941 by Ross ("Imperatives and Logic," *Theoria* **7**: 53–71). Given that the proof in question yields an absurdity, FCPP can be taken to show that **SDL** (Standard Deontic Logic) = **D** leads to inconsistency when applied; or, put in AI terms, you wouldn't want a robot to base its ethical decision-making on **D**! Fortunately, the RAIR Lab's modern cognitive calculus $\mathcal{DCEC}^*$ allows FCPP to be avoided. (A recent paper explaining the use by an ethically correct AI of this calculus is available here.)

Here's the paradox. Suppose that you travel to visit a friend, arrive late at night, and are weary. Your friend says hospitably: "You may either sleep on the sofa-bed or sleep on the guest-room bed." (1) From this statement it follows that you are permitted to sleep on the sofa-bed, and you are permitted to sleep on the guest-room bed. (2) In **D**, this pair gets symbolized like this:

(1')
$\Diamond(sofabed \vee guestbed)$

(2')
$\Diamond sofabed \wedge \Diamond guestbed$

But (2') doesn't follow deductively from (1') in **D**, as a call to the provability oracle for **D** in the HyperSlate$^{\text{TM}}$ file for this problem confirms. A suggested repair is to add to **D** the schema

$$\Diamond(\phi \vee \psi) \rightarrow (\Diamond\phi \wedge \Diamond\psi),$$

but as your proof will (hopefully) show, this addition allows a proof of the absurd theorem that if anything is morally perimssible, everything is!

Your finished proof is allowed to make use of the PC provabiity oracle, but of no other oracle.

Deadline November 12, 2020, 3:00 PM EST

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

TheFreeChoicePermissionParadox

Producing a valid proof in this problem will enable you to understand The Free Choice Permission Paradox (FCPP), discovered in 1941 by Ross ("Imperatives and Logic," *Theoria* **7**: 53–71). Given that the proof in question yields an absurdity, FCPP can be taken to show that **SDL** (Standard Deontic Logic) = **D** leads to inconsistency when applied; or, put in AI terms, you wouldn't want a robot to base its ethical decision-making on **D**! Fortunately, the RAIR Lab's modern cognitive calculus $\mathcal{DCEC}^*$ allows FCPP to be avoided. (A recent paper explaining the use by an ethically correct AI of this calculus is available here.)

Here's the paradox. Suppose that you travel to visit a friend, arrive late at night, and are weary. Your friend says hospitably: "You may either sleep on the sofa-bed or sleep on the guest-room bed." (1) From this statement it follows that you are permitted to sleep on the sofa-bed, and you are permitted to sleep on the guest-room bed. (2) In **D**, this pair gets symbolized like this:

(1')
$\Diamond(sofabed \vee guestbed)$

(2')
$\Diamond sofabed \wedge \Diamond guestbed$

But (2') doesn't follow deductively from (1') in **D**, as a call to the provability oracle for **D** in the HyperSlate^{TM} file for this problem confirms. A suggested repair is to add to **D** the schema

$$\Diamond(\phi \vee \psi) \rightarrow (\Diamond\phi \wedge \Diamond\psi),$$

but as your proof will (hopefully) show, this addition allows a proof of the absurd theorem that if anything is morally perimssible, everything is!

Your finished proof is allowed to make use of the PC provabiity oracle, but of no other oracle. (No deadline for now.)

*DCEC** !!!

https://www.ijcai.org/Proceedings/2017/0658.pdf

Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)

On Automating the Doctrine of Double Effect

Naveen Sundar Govindarajulu and **Selmer Bringsjord**
Rensselaer Polytechnic Institute, Troy, NY
{naveensundarg,selmer.bringsjord}@gmail.com

Abstract

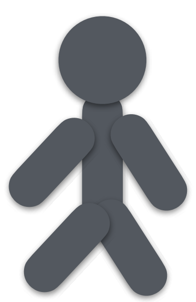
The **doctrine of double effect** ($\mathcal{DDE}$) is a long-studied ethical principle that governs when actions that have both positive and negative effects are to be allowed. The goal in this paper is to automate $\mathcal{DDE}$. We briefly present $\mathcal{DDE}$, and use a first-order modal logic, the **deontic cognitive event calculus**, as our framework to formalize the doctrine.

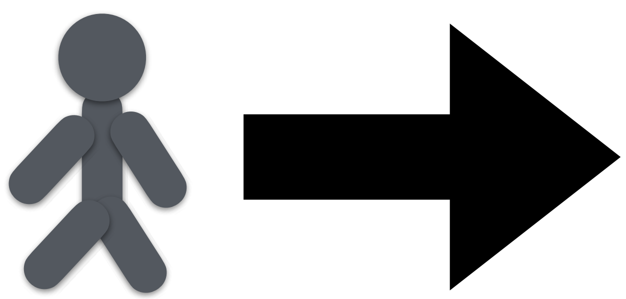
— provided that 1) the harmful effects are not intended; 2) the harmful effects are not used to achieve the beneficial effects (harm is merely a *side*-effect); and 3) benefits outweigh the harm by a significant amount. What distinguishes $\mathcal{DDE}$ from, say, naïve forms of consequentialism in ethics (e.g. act utilitarianism, which holds that an action is obligatory for an autonomous agent if and only if it produces the most utility among all competing actions) is that purely mental intentions in and of themselves, independent of conse-

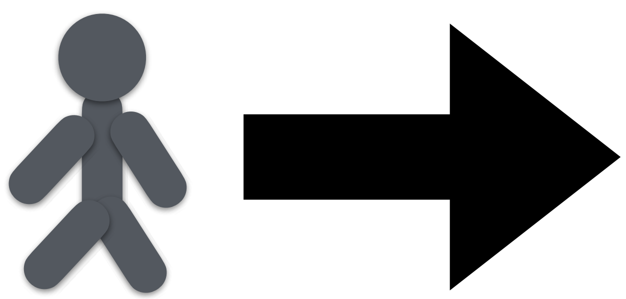
4 Informal $\mathcal{DDE}$

We now informally but rigorously present $\mathcal{DDE}$. We assume we have at hand an ethical hierarchy of actions as in the deontological case (e.g. forbidden, neutral, obligatory); see [Bringsjord, 2017]. We also assume that we have a utility or goodness function for states of the world or effects as in the consequentialist case. For an autonomous agent a, an action α in a situation σ at time t is said to be $\mathcal{DDE}$-compliant *iff*:

- $\mathbf{C}_1$ the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [2017], and require that the action be neutral or above neutral in such a hierarchy);
- $\mathbf{C}_2$ The net utility or goodness of the action is greater than some positive amount γ;
- $\mathbf{C}_{3a}$ the agent performing the action intends only the good effects;
- $\mathbf{C}_{3b}$ the agent does not intend any of the bad effects;
- $\mathbf{C}_4$ the bad effects are not used as a means to obtain the good effects; and
- $\mathbf{C}_5$ if there are bad effects, the agent would rather the situation be different and the agent not have to perform the action. That is, the action is unavoidable.







Formal Conditions for $\mathcal{DDE}$

$\mathbf{F_1}$ α carried out at t is not forbidden. That is:

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

$\mathbf{F_2}$ The net utility is greater than a given positive real γ:

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

$\mathbf{F_{3a}}$ The agent a intends at least one good effect. ($\mathbf{F_2}$ should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu(f_g, y) > 0$, 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:

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

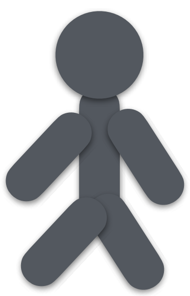
$\mathbf{F_{3b}}$ The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu(f_b, y) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu(f_g, y) > 0$, and for all y such that $t < y \leq H$ the following holds:

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

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

$\mathbf{F_4}$ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of $\rhd$ above, hold here. One such permutation is shown below. For any bad fluent f_b holding at t_1, and any good fluent f_g holding at some t_2, such that $t < t_1, t_2 \leq H$, the following holds:

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



Formal Conditions for $\mathcal{DDE}$

$\mathbf{F_1}$ α carried out at t is not forbidden. That is:

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

$\mathbf{F_2}$ The net utility is greater than a given positive real γ:

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

$\mathbf{F_{3a}}$ The agent a intends at least one good effect. ($\mathbf{F_2}$ should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu(f_g, y) > 0$, 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:

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

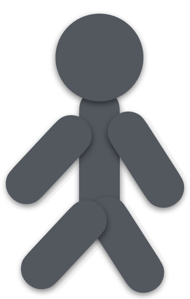
$\mathbf{F_{3b}}$ The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu(f_b, y) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu(f_g, y) > 0$, and for all y such that $t < y \leq H$ the following holds:

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

$\mathbf{F_4}$ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of $\rhd$ above, hold here. One such permutation is shown below. For any bad fluent f_b holding at t_1, and any good fluent f_g holding at some t_2, such that $t < t_1, t_2 \leq H$, the following holds:

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

$\mathbb{P}_{\mathrm{DDE}_1}$ + ShadowProver



Formal Conditions for $\mathcal{DDE}$

F₁ α carried out at t is not forbidden. That is:

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

F₂ The net utility is greater than a given positive real γ:

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

F₃ₐ The agent a intends at least one good effect. (**F₂** 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:

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

F₃ᵦ 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:

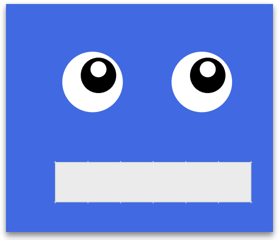
$$\Gamma \nvdash \mathbf{I}\Big(a, t, Holds\big(f_b, y\big)\Big) \text{ and}$$

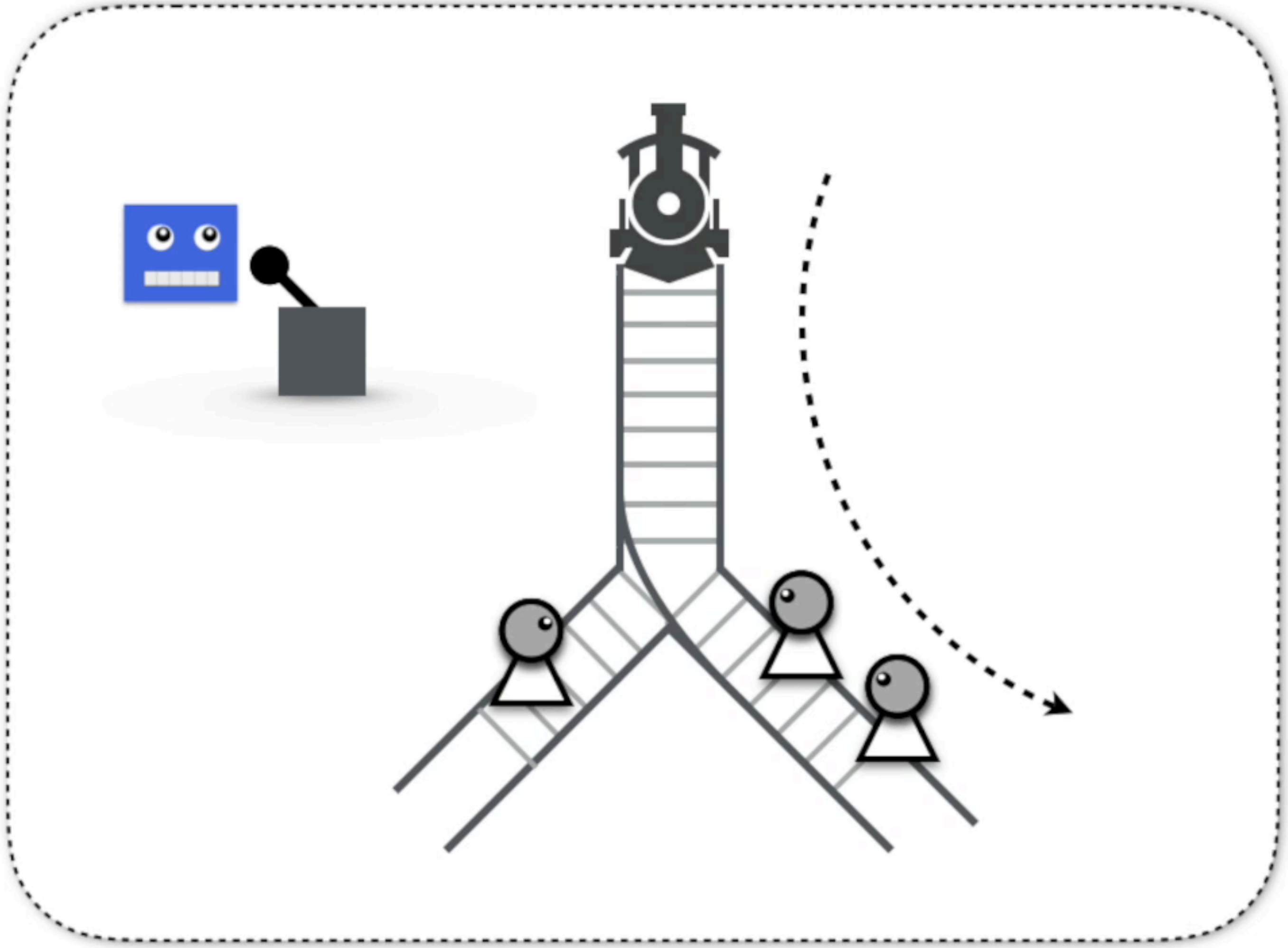
$$\Gamma \nvdash \mathbf{I}\Big(a, t, \neg Holds\big(f_g, y\big)\Big)$$

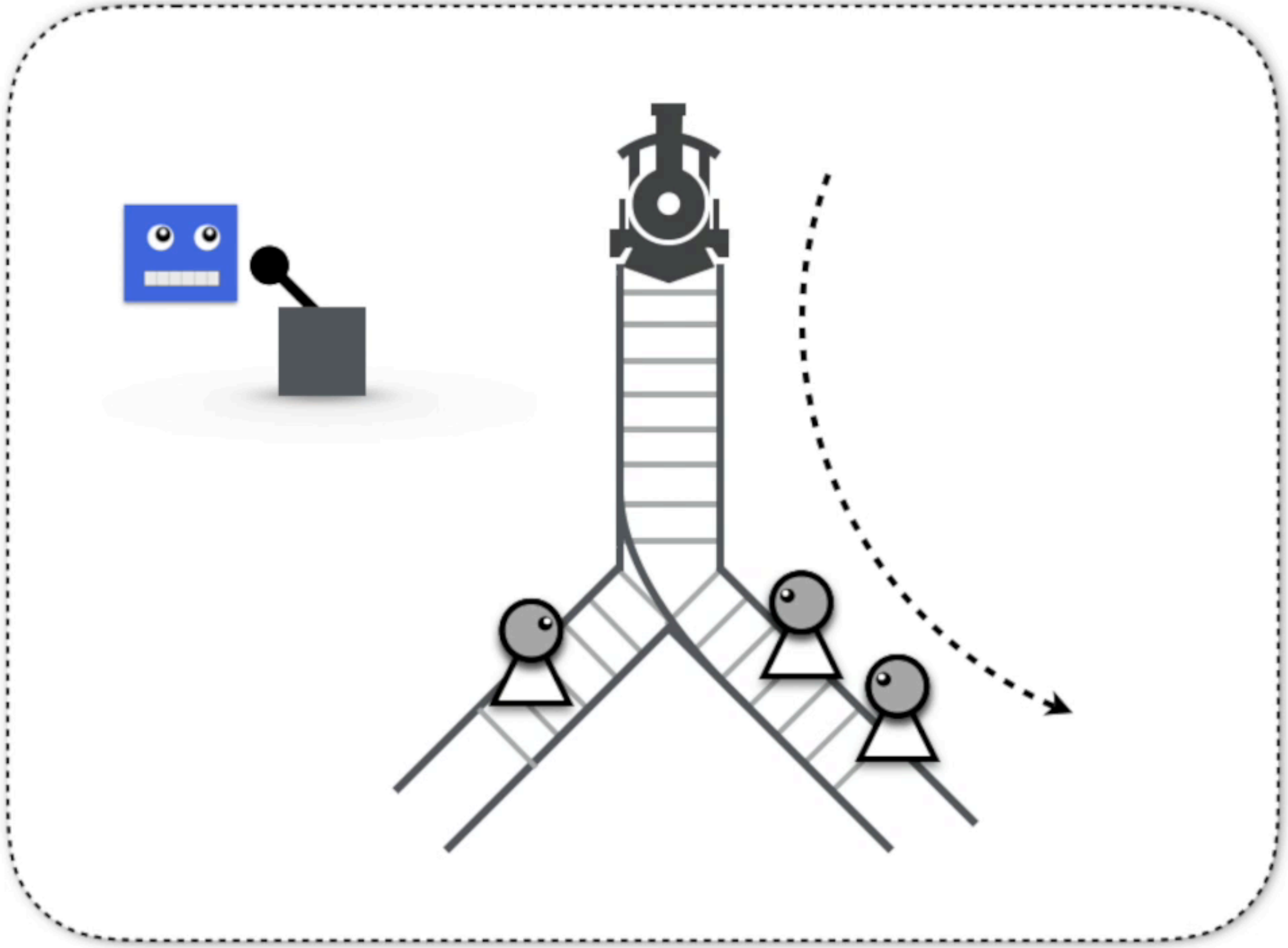
F₄ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of $\rhd$ above, hold here. One such permutation is shown below. For any bad fluent f_b holding at t_1, and any good fluent f_g holding at some t_2, such that $t < t_1, t_2 \leq H$, the following holds:

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

$\mathbb{P}_{\mathrm{DDE}_1}$ + ShadowProver







Inference Schemata

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

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

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

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

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

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

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

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

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

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

Inference Schemata

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

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

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

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

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

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

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

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

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

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

Making Morally  Machines

Selmer Bringsjord ∧ Naveen Sundar Govindarajulu ∧ John Licato

Making Morally Machines

Selmer Bringsjord ∧ Naveen Sundar Govindarajulu ∧ John Licato

er løsningen, med nok penger!