

# **Standard Deontic Logic (SDL = D) Isn't Going to Cut It!**

**(Chisholm's Paradox; The Free Choice Permission Paradox)**

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

Intro to Logic  
4/8/2019



# Curved Grades T 2

All: positively extraordinary

4: A+

3: A

2: A-

1: B

Peek ahead to next  
time for some context  
today ...

**At least supposedly, long term:**

At least supposedly, long term:

“We’re in *very* deep trouble.”

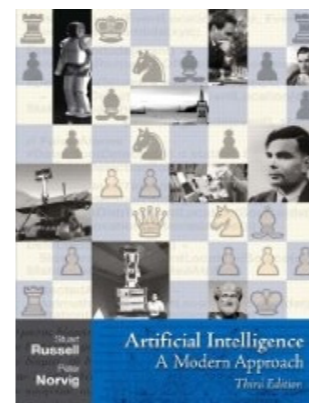
At least supposedly, long term:

“We’re in *very* deep trouble.”



At least supposedly, long term:

“We’re in *very* deep trouble.”



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



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

$\forall x : \text{Agents}$

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

$\forall x : \text{Agents}$

$\text{Autonomous}(x) + \text{Powerful}(x) + \text{Highly\_Intelligent}(x) = \text{Dangerous}(x)$

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

$\forall x : \text{Agents}$

$$\text{Autonomous}(x) + \text{Powerful}(x) + \text{Highly\_Intelligent}(x) = \text{Dangerous}(x)$$



Autonomous agents are dangerous because they are powerful and intelligent.

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

$\forall x : \text{Agents}$

$\text{Autonomous}(x) + \text{Powerful}(x) + \text{Highly\_Intelligent}(x) = \text{Dangerous}(x)$



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

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

$\forall x$  : Agents  
Autonomous(x)

$\cup$

) = Dangerous(x)

## 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

1	Introduction	1
2	The Distressing Principle, Intuitively Put	1
3	The Distressing Principle, More Formally Put	2
3.1	The Ideal-Observer Point of View	2
3.2	Theory-of-Mind-Creativity	3
3.3	Autonomy	4
3.4	The Deontic Cognitive Event Calculus (D <sup>o</sup> CEC)	5
3.5	Collaborative Situations; Untrustworthiness	7
3.6	Theorem ACU	7
4	Computational Simulations	8
4.1	ShadowProver	8
4.2	The Simulation Proper	9
5	Toward the Needed Engineering	10
	References	16

\*The authors are deeply grateful for support provided by both AFOSR and ONR that enabled the research reported on herein, and are in addition thankful both for the guidance and patience of the editors and wise comments received from two reviewers.

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

$\forall x : \text{Agents}$

$\text{Autonomous}(x) + \text{Powerful}(x) + \text{Highly\_Intelligent}(x) = \text{Dangerous}(x)$




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

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

$\forall x : \text{Agents}$

$$\text{Autonomous}(x) + \text{Powerful}(x) + \text{Highly\_Intelligent}(x) = \text{Dangerous}(x)$$


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

**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  $\gamma$  in part by way of a contributed action  $\alpha_k$  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**



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

$\forall x : \text{Agents}$

$$\text{Autonomous}(x) + \text{Powerful}(x) + \text{Highly\_Intelligent}(x) = \text{Dangerous}(x)$$

(We use the “jump”  
technique in relative  
computability.)

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

**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  $\gamma$  in part by way of a contributed action  $\alpha_k$  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**



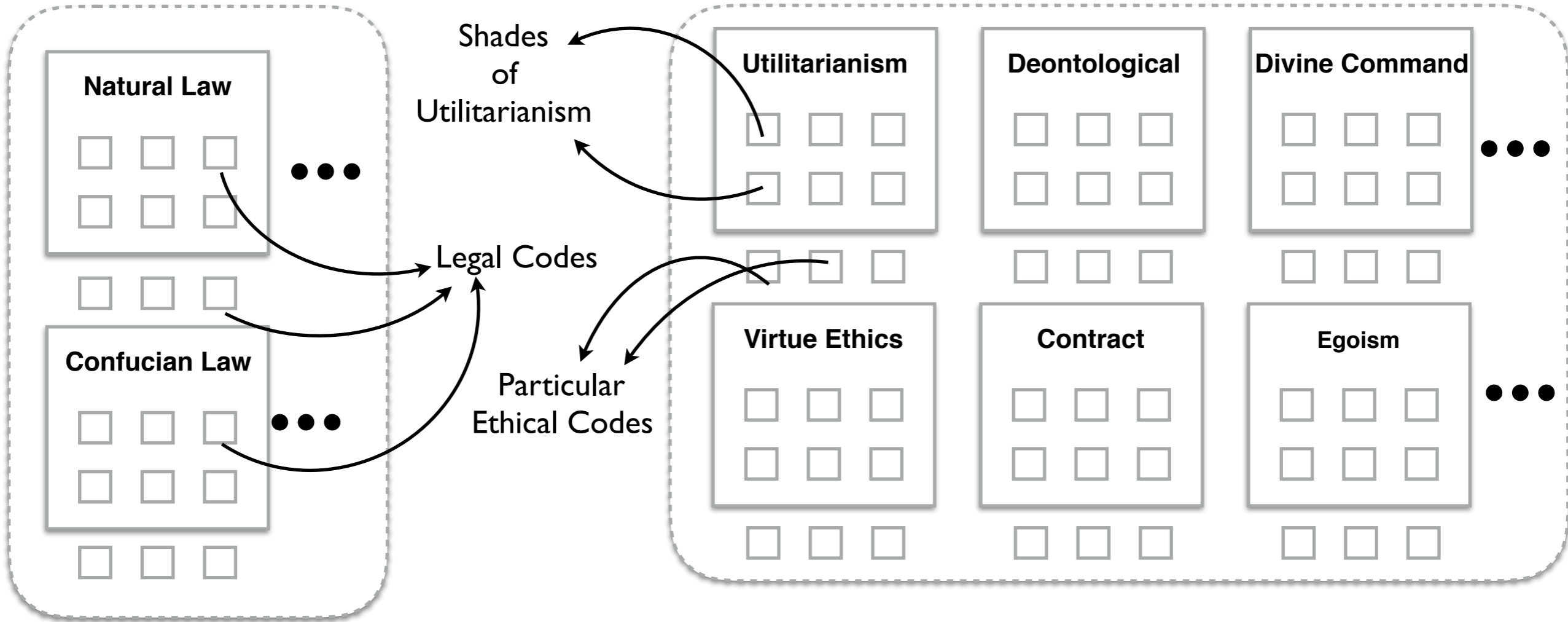
# Making Moral Machines

# Making Meta-Moral Machines



## Theories of Law

## Ethical Theories



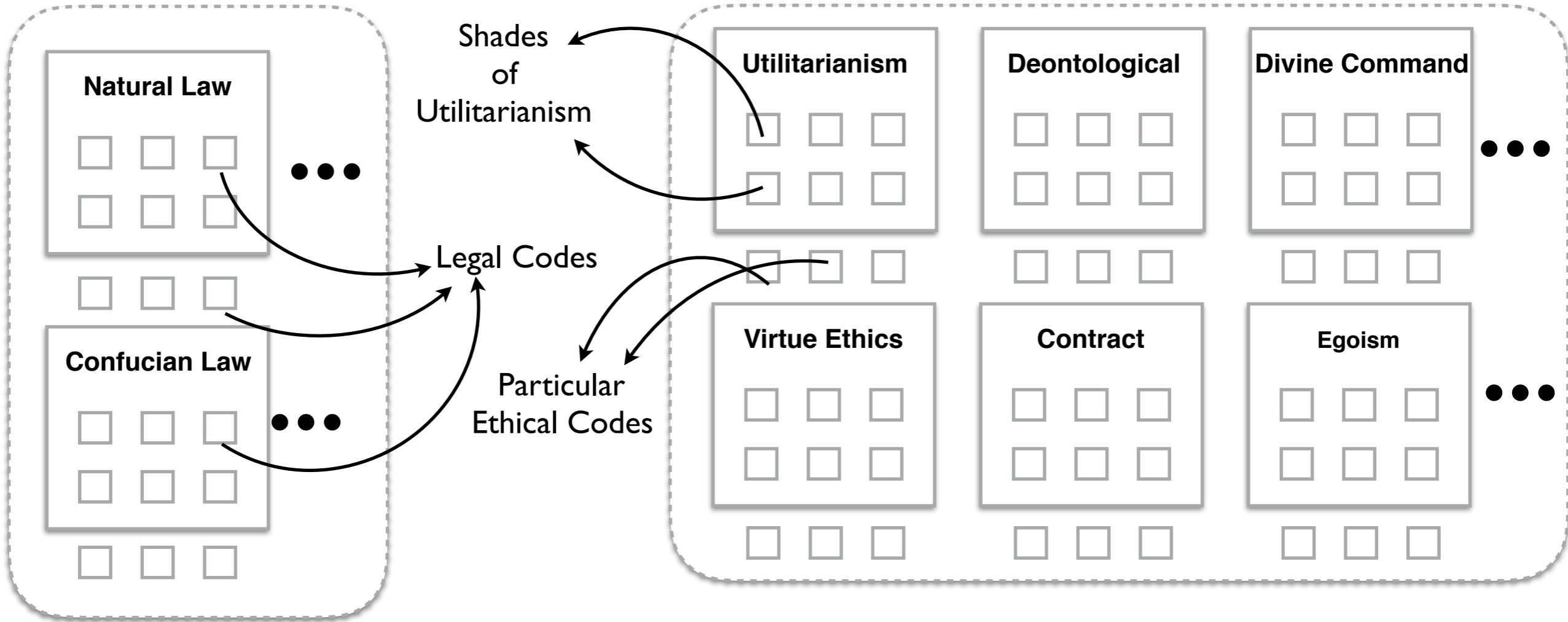
# Making Moral Machines

# Making Meta-Moral Machines



## Theories of Law

## Ethical Theories



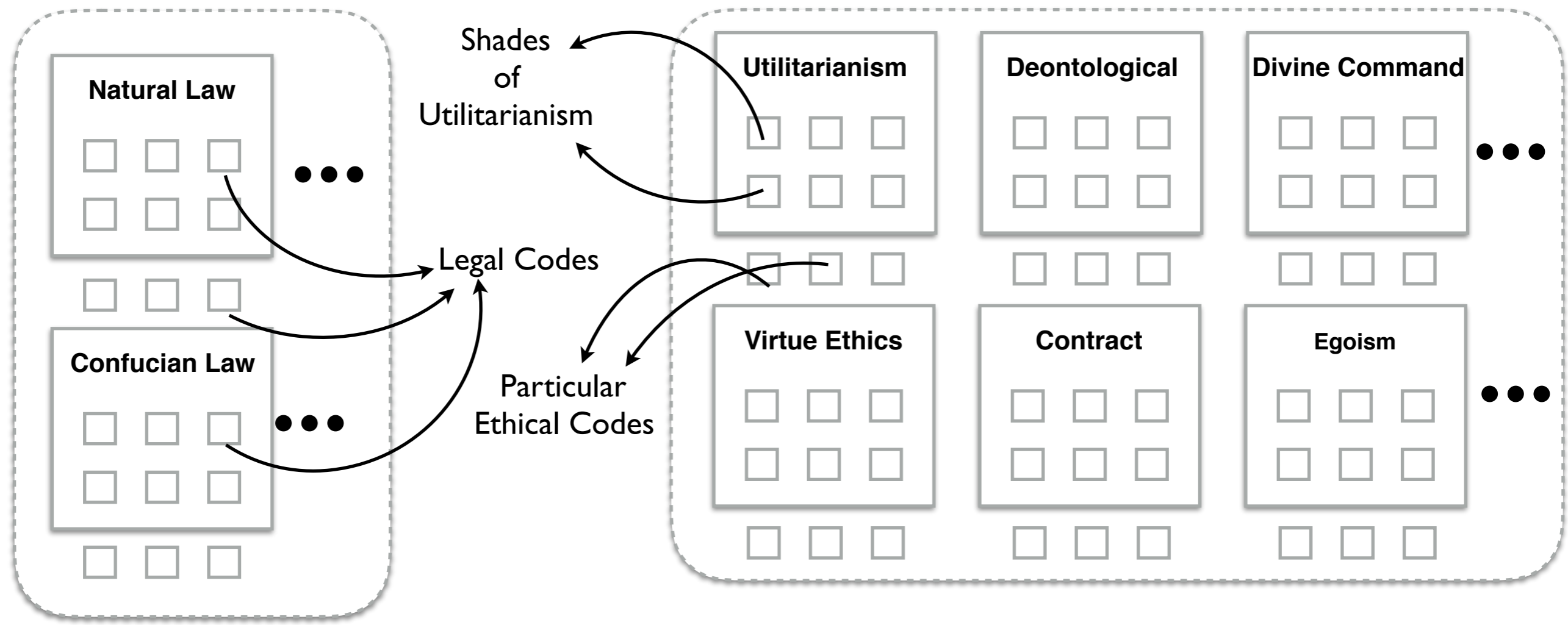
# Making Moral Machines

# Making Meta-Moral Machines



## Theories of Law

## Ethical Theories



- Step 1**
1. Pick a theory
  2. Pick a code
  3. Run through EH.

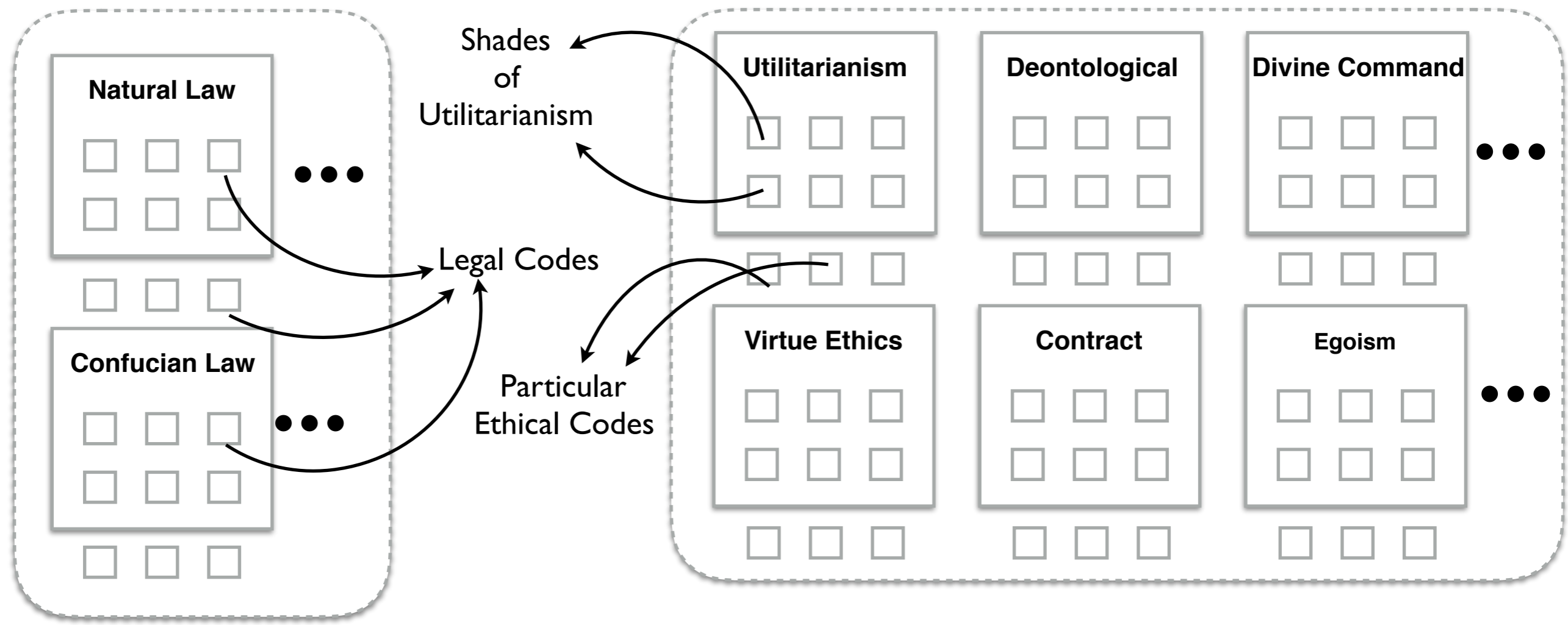
# Making Moral Machines

# Making Meta-Moral Machines

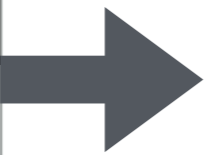


## Theories of Law

## Ethical Theories



- Step 1**
1. Pick a theory
  2. Pick a code
  3. Run through EH.

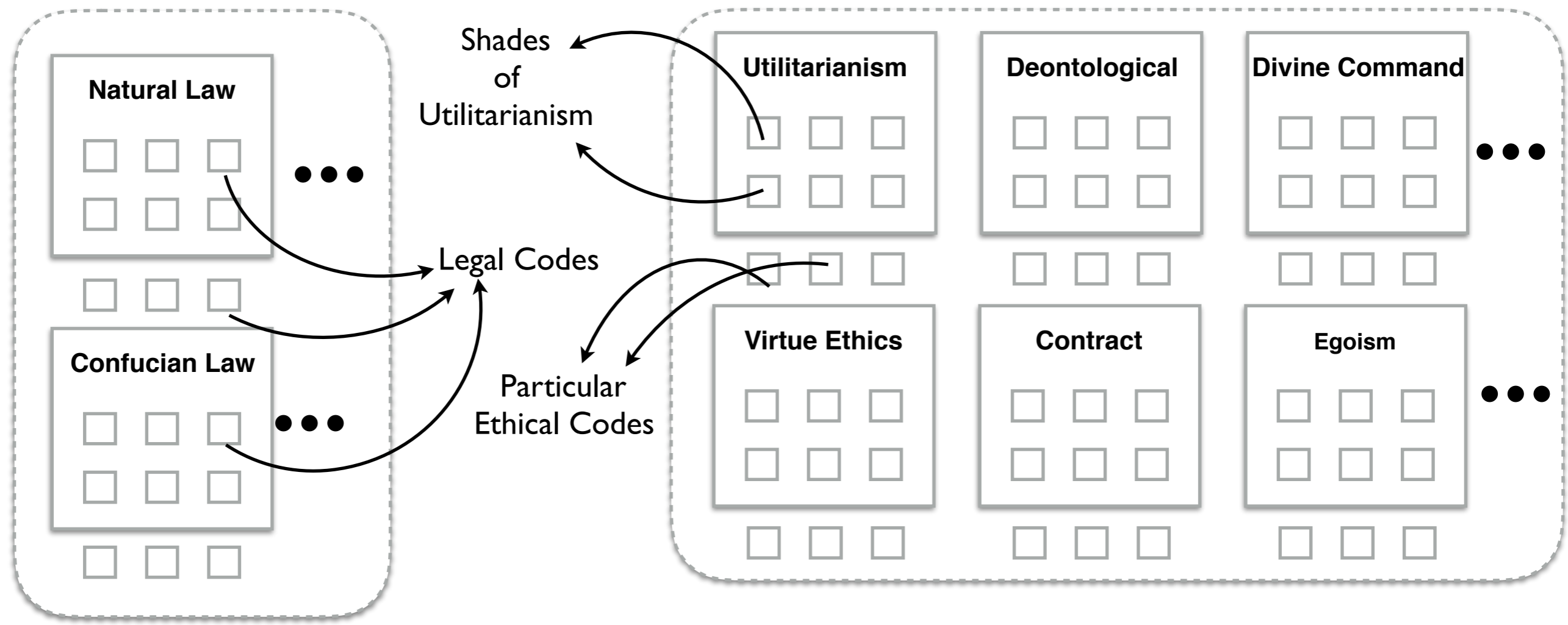


# Making Moral Machines    Making Meta-Moral Machines



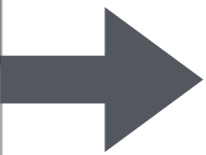
## Theories of Law

## Ethical Theories




### Step 1


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



### Step 2

Automate

 Prover

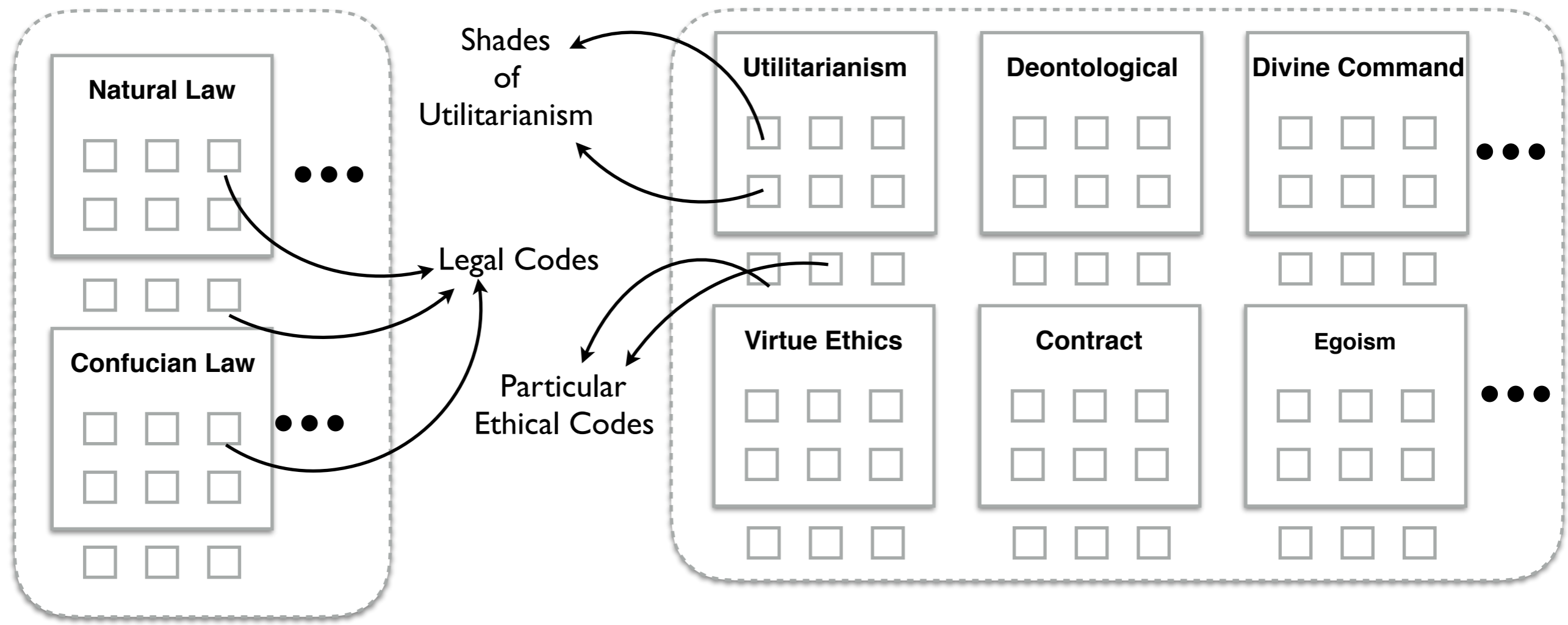
 Spectra

# Making Moral Machines    Making Meta-Moral Machines



## Theories of Law

## Ethical Theories





**Step 1**

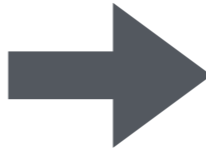
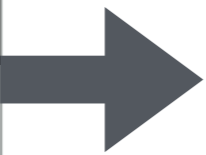
1. Pick a theory
2. Pick a code
3. Run through EH.

**Step 2**

Automate

 Prover

 Spectra

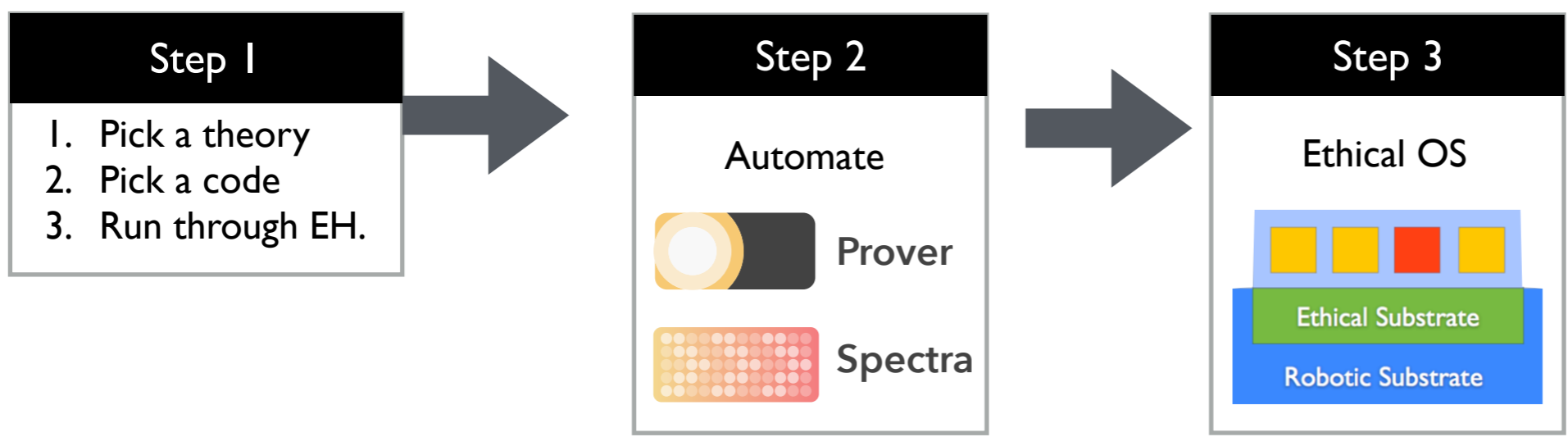
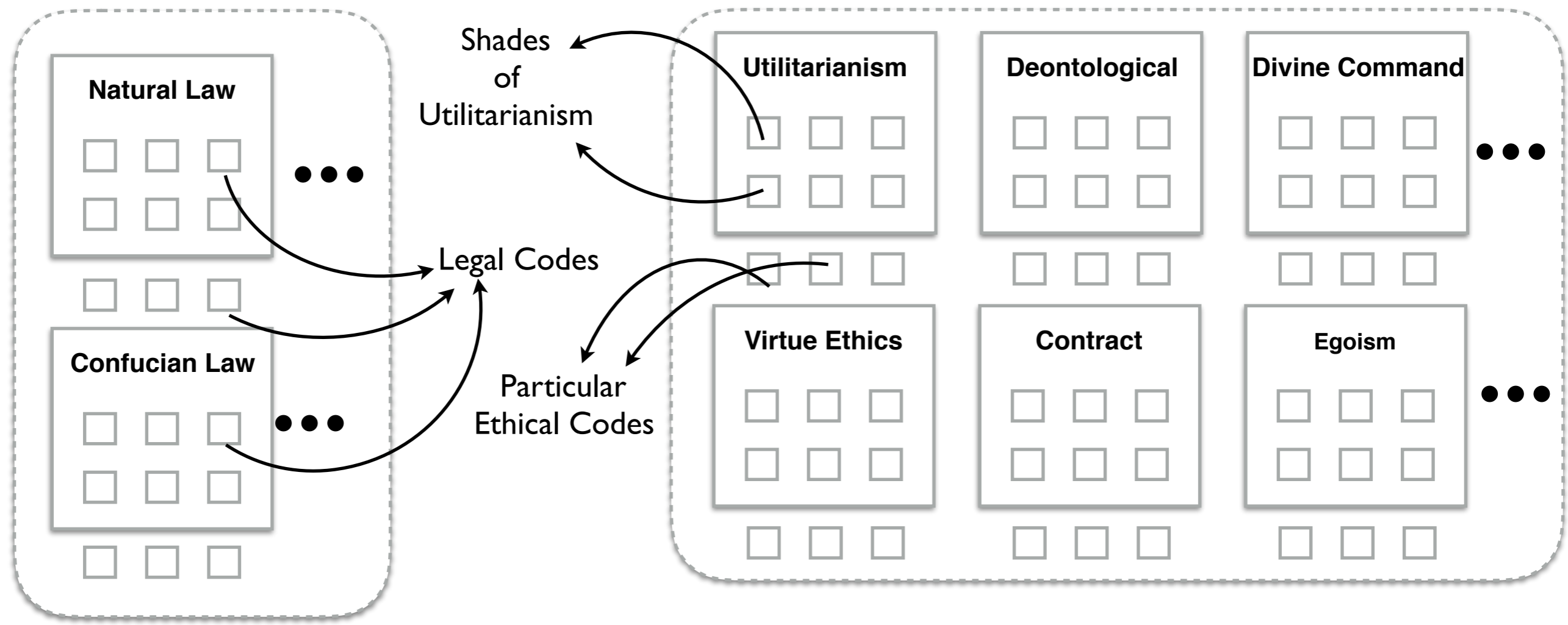


# Making Moral Machines      Making Meta-Moral Machines



## Theories of Law

## Ethical Theories

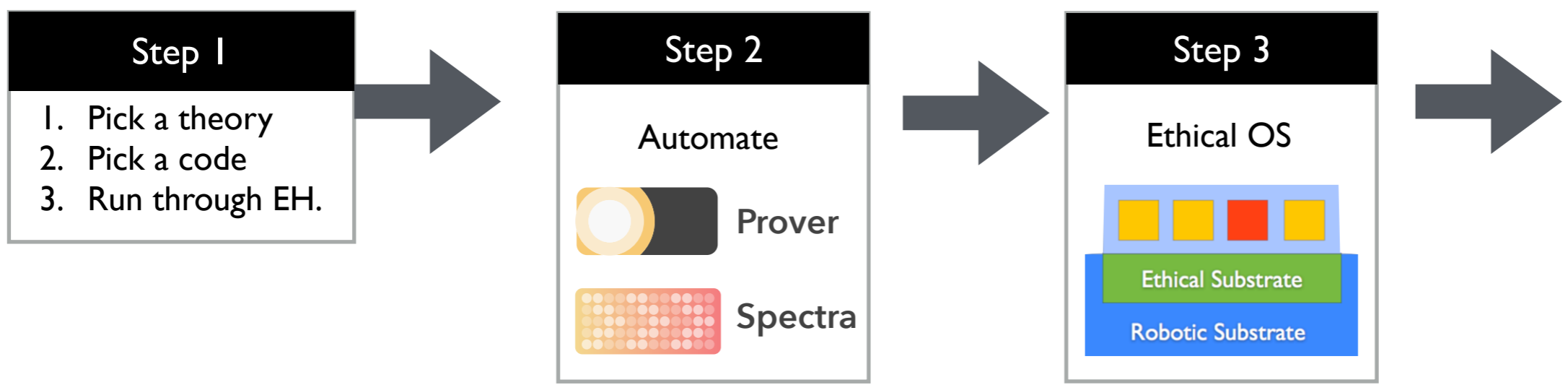
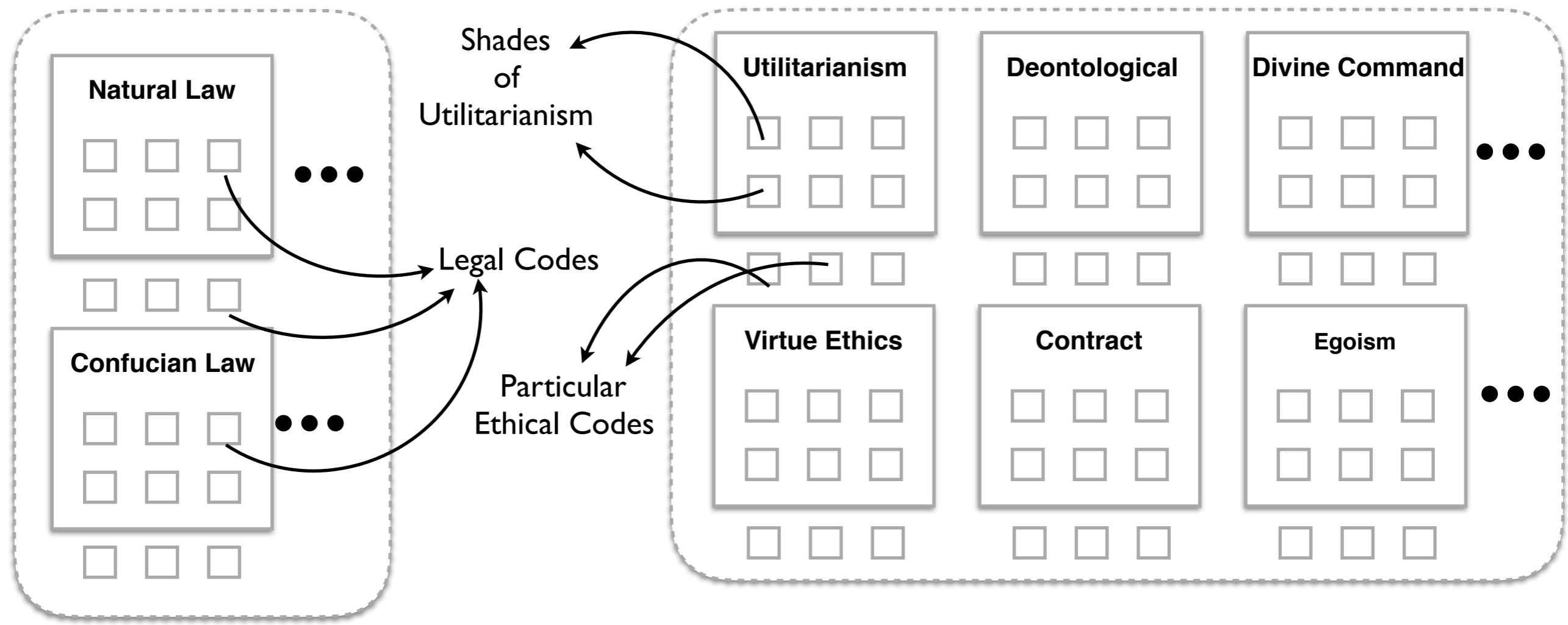


# Making Moral Machines      Making Meta-Moral Machines



## Theories of Law

## Ethical Theories



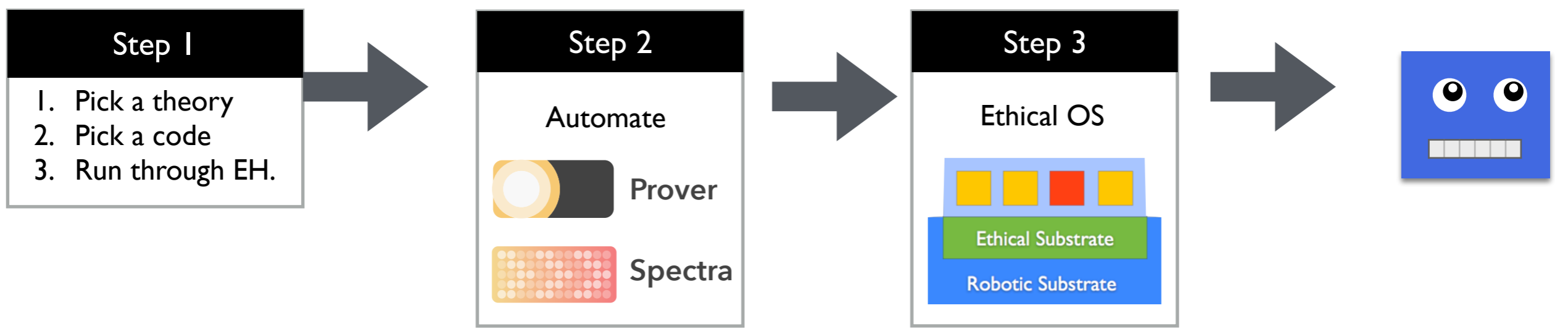
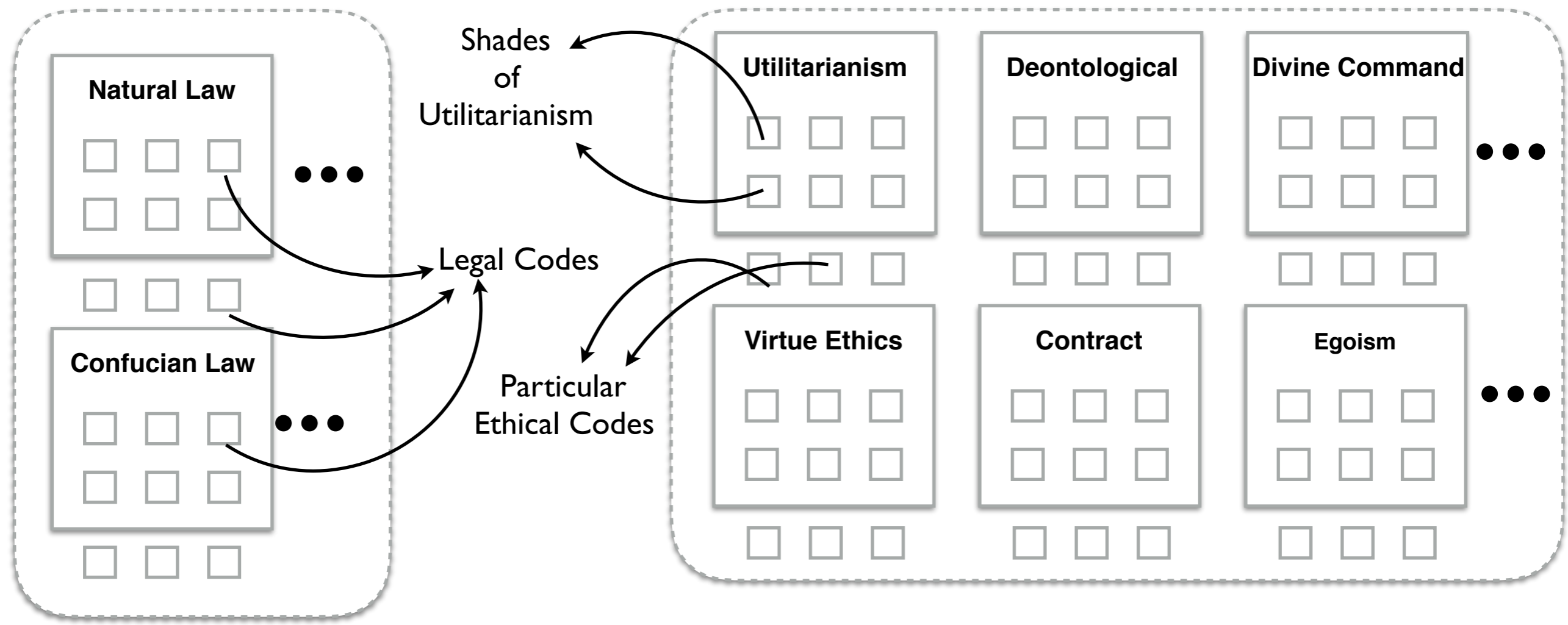


# Making Moral Machines    Making Meta-Moral Machines



## Theories of Law

## Ethical Theories

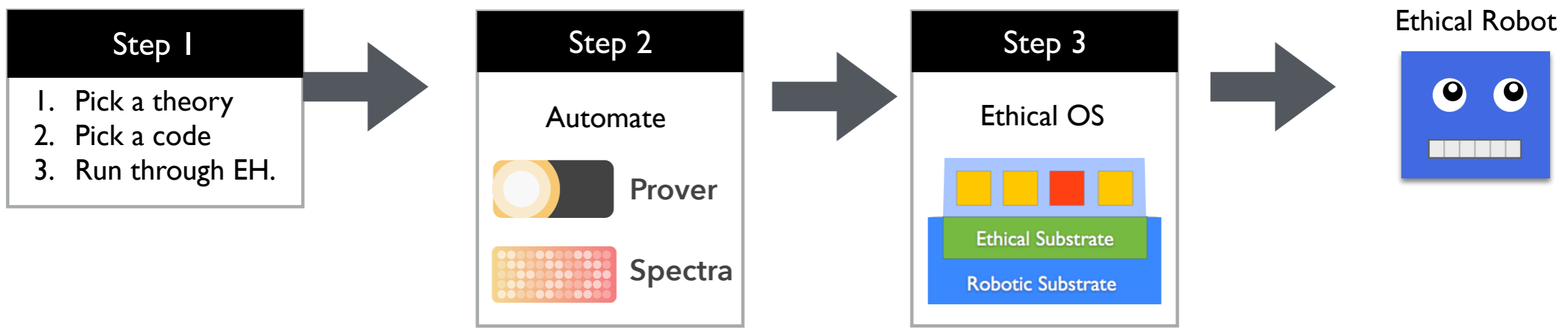
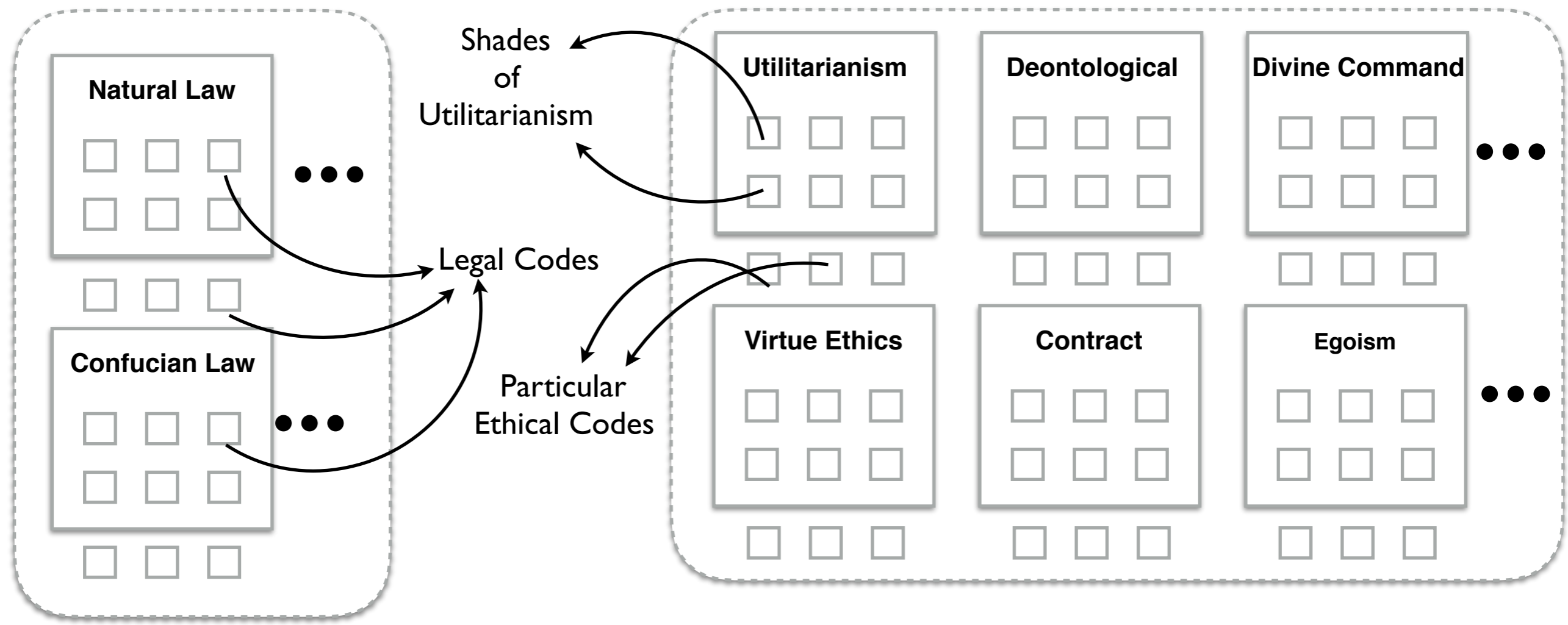


# Making Moral Machines      Making Meta-Moral Machines



## Theories of Law

## Ethical Theories

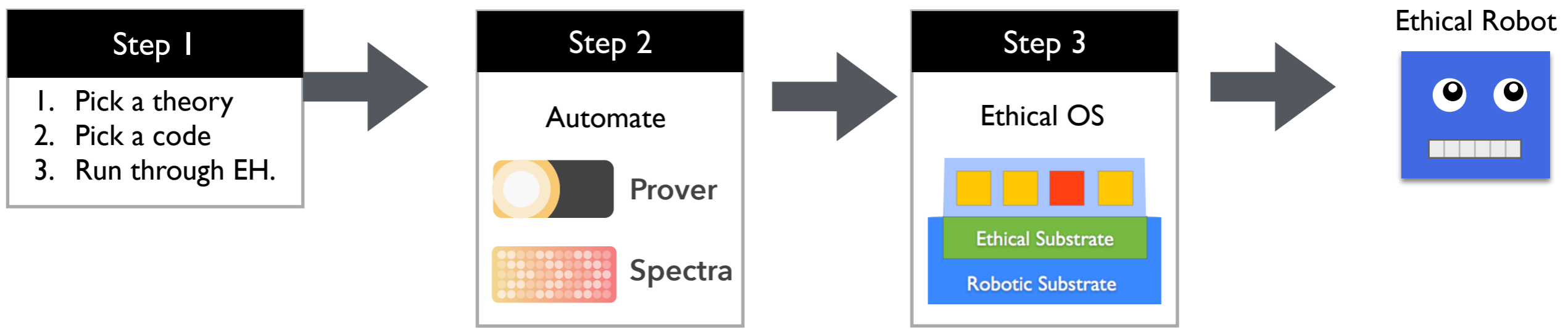
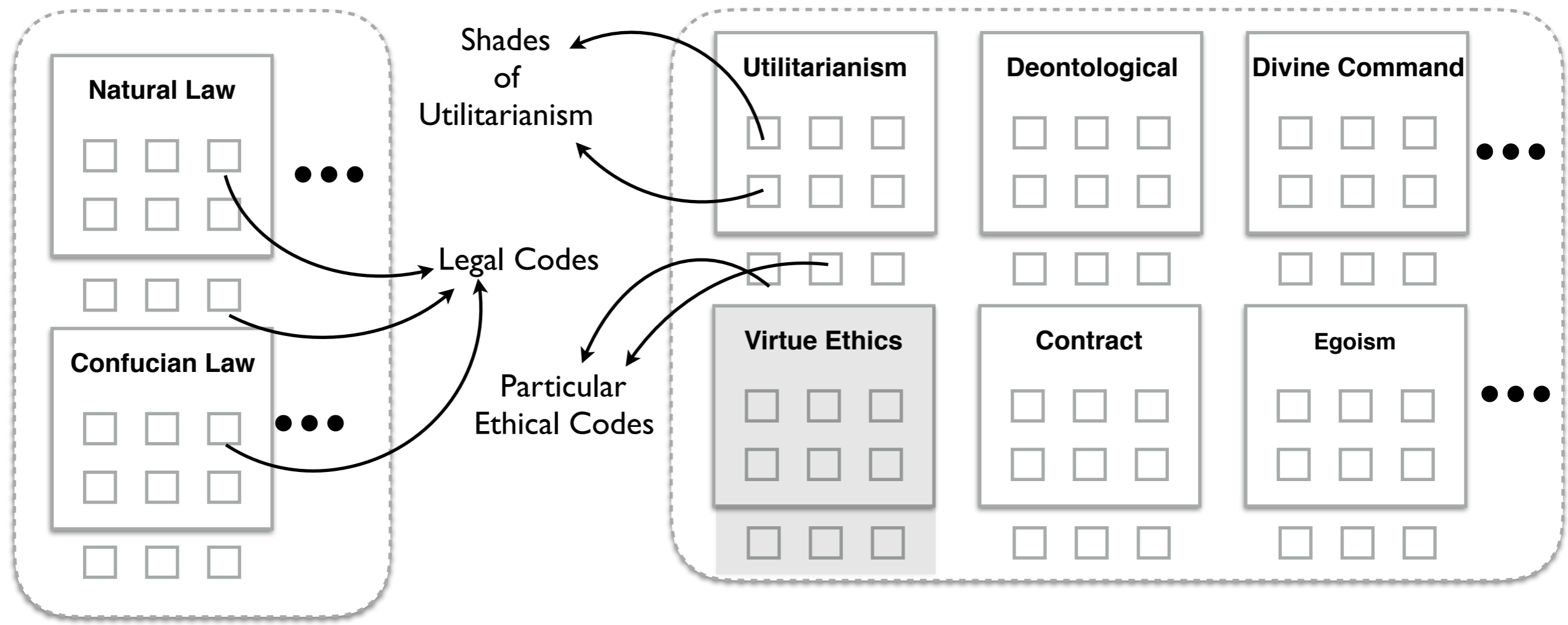


# Making Moral Machines      Making Meta-Moral Machines



## Theories of Law

## Ethical Theories

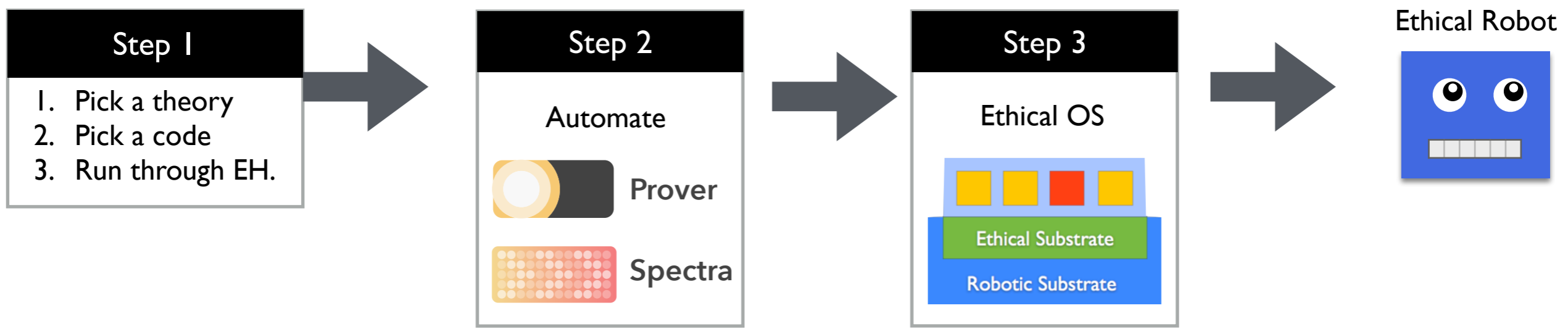
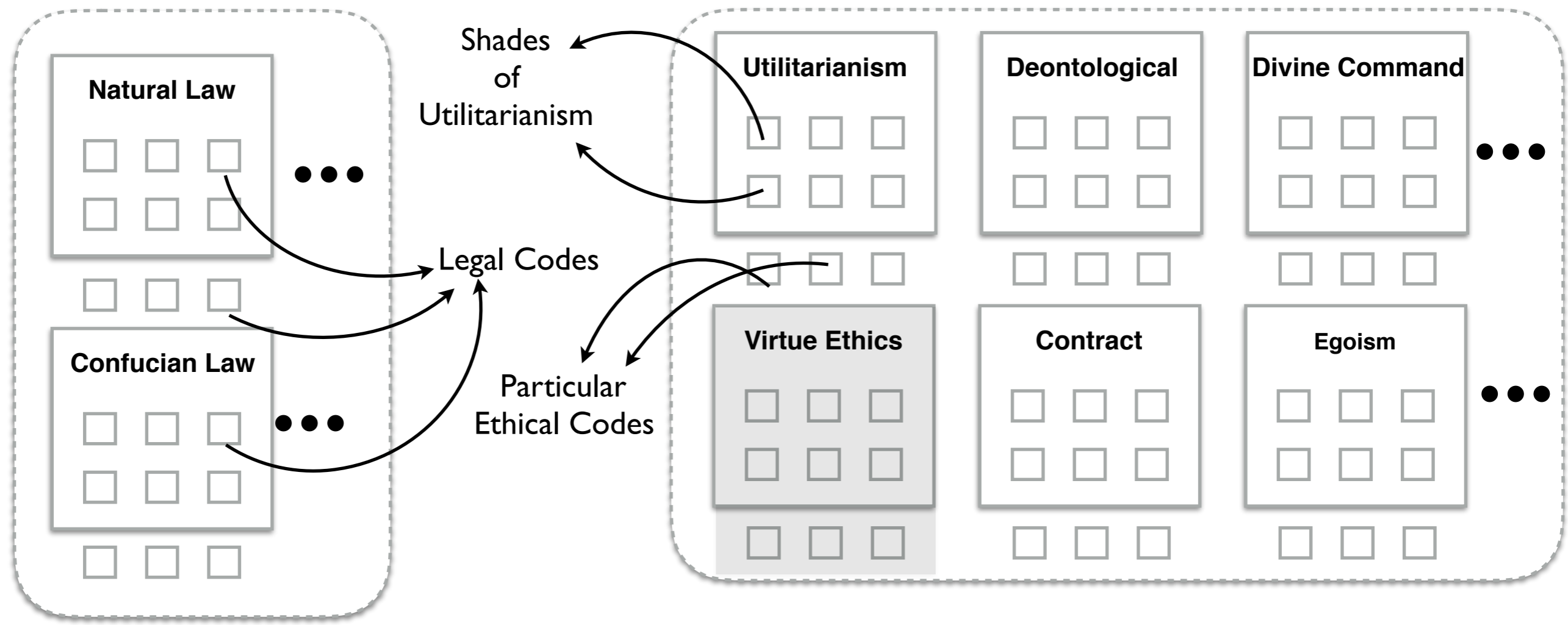


# Making Moral Machines Making Meta-Moral Machines



## Theories of Law

## Ethical Theories



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

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

Perhaps **D = SDL?** ...

# Review: Encapsulation

Slate - K.slt

K. $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$ K $\vdash$ ✓ $\infty$ $\Box$	T. $\Box\varphi \rightarrow \varphi$ K $\vdash$ ✗ $\infty$ $\Box$	4. $\Box\varphi \rightarrow \Box\Box\varphi$ K $\vdash$ ✗ $\infty$ $\Box$	5. $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$ K $\vdash$ ✗ $\infty$ $\Box$
--	--	--	--

# Review: Encapsulation

The image shows two overlapping windows from the Slate application. The top window is titled "Slate - K.slt" and the bottom window is titled "Slate - T.slt". Each window contains four rounded rectangular boxes, each representing a modal logic formula and its validity in a specific system.

**Slate - K.slt**

- Box 1:  $K. \Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   
 $K \vdash \checkmark \infty \Box$
- Box 2:  $T. \Box\varphi \rightarrow \varphi$   
 $K \vdash \times \infty \Box$
- Box 3:  $4. \Box\varphi \rightarrow \Box\Box\varphi$   
 $K \vdash \times \infty \Box$
- Box 4:  $5. \neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   
 $K \vdash \times \infty \Box$

**Slate - T.slt**

- Box 1:  $K. \Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   
 $M \vdash \checkmark \infty \Box$
- Box 2:  $T. \Box\varphi \rightarrow \varphi$   
 $M \vdash \checkmark \infty \Box$
- Box 3:  $4. \Box\varphi \rightarrow \Box\Box\varphi$   
 $M \vdash \times \infty \Box$
- Box 4:  $5. \neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   
 $M \vdash \times \infty \Box$



# Review: Encapsulation

The image displays three overlapping windows, each showing a set of modal logic formulas and their validity in different systems. The windows are titled "Slate - K.slt", "Slate - T.slt", and "Slate - D.slt".

**Slate - K.slt**

- K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   
K  $\vdash \checkmark \infty \Box$
- T.  $\Box\varphi \rightarrow \varphi$   
K  $\vdash \times \infty \Box$
- 4.  $\Box\varphi \rightarrow \Box\Box\varphi$   
K  $\vdash \times \infty \Box$
- 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   
K  $\vdash \times \infty \Box$

**Slate - T.slt**

- K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   
M  $\vdash \checkmark \infty \Box$
- T.  $\Box\varphi \rightarrow \varphi$   
M  $\vdash \checkmark \infty \Box$
- 4.  $\Box\varphi \rightarrow \Box\Box\varphi$   
M  $\vdash \times \infty \Box$
- 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   
M  $\vdash \times \infty \Box$

**Slate - D.slt**

- K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   
D  $\vdash \checkmark \infty \Box$
- T.  $\Box\varphi \rightarrow \varphi$   
D  $\vdash \times \infty \Box$
- D.  $\Box\varphi \rightarrow \Diamond\varphi$   
D  $\vdash \checkmark \infty \Box$
- 4.  $\Box\varphi \rightarrow \Box\Box\varphi$   
D  $\vdash \times \infty \Box$
- 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   
D  $\vdash \times \infty \Box$
- INTER.  $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$   
D  $\vdash \checkmark \infty \Box$

# Review: Encapsulation

The image shows four overlapping Slate windows, each displaying a set of modal logic formulas and their provability status in a specific system. The windows are titled 'Slate - K.slt', 'Slate - T.slt', 'Slate - D.slt', and 'Slate - S4.slt'.

**Slate - K.slt**

- K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   $K \vdash \checkmark \infty \Box$
- T.  $\Box\varphi \rightarrow \varphi$   $K \vdash \times \infty \Box$
- 4.  $\Box\varphi \rightarrow \Box\Box\varphi$   $K \vdash \times \infty \Box$
- 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   $K \vdash \times \infty \Box$

**Slate - T.slt**

- K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   $M \vdash \checkmark \infty \Box$
- T.  $\Box\varphi \rightarrow \varphi$   $M \vdash \checkmark \infty \Box$
- 4.  $\Box\varphi \rightarrow \Box\Box\varphi$   $M \vdash \times \infty \Box$
- 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   $M \vdash \times \infty \Box$

**Slate - D.slt**

- K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   $D \vdash \checkmark \infty \Box$
- T.  $\Box\varphi \rightarrow \varphi$   $D \vdash \times \infty \Box$
- D.  $\Box\varphi \rightarrow \Diamond\varphi$   $D \vdash \checkmark \infty \Box$
- 4.  $\Box\varphi \rightarrow \Box\Box\varphi$   $D \vdash \times \infty \Box$
- 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   $D \vdash \times \infty \Box$
- INTER.  $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$   $D \vdash \checkmark \infty \Box$

**Slate - S4.slt**

- K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$   $S4 \vdash \checkmark \infty \Box$
- T.  $\Box\varphi \rightarrow \varphi$   $S4 \vdash \checkmark \infty \Box$
- D.  $\Box\varphi \rightarrow \Diamond\varphi$   $S4 \vdash \checkmark \infty \Box$
- 4.  $\Box\varphi \rightarrow \Box\Box\varphi$   $S4 \vdash \checkmark \infty \Box$
- 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$   $S4 \vdash \times \infty \Box$
- INTER.  $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$   $\{INTER\} \text{ Assume } \checkmark$

# Review: Encapsulation

The image shows five overlapping Slate windows, each displaying a set of modal logic formulas and their derivability in a specific system. The windows are titled as follows:

- Slate - K.slt**:
  - K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$  (K  $\vdash \checkmark \infty \Box$ )
  - T.  $\Box\varphi \rightarrow \varphi$  (K  $\vdash \times \infty \Box$ )
  - 4.  $\Box\varphi \rightarrow \Box\Box\varphi$  (K  $\vdash \times \infty \Box$ )
  - 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$  (K  $\vdash \times \infty \Box$ )
- Slate - T.slt**:
  - K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$  (M  $\vdash \checkmark \infty \Box$ )
  - T.  $\Box\varphi \rightarrow \varphi$  (M  $\vdash \checkmark \infty \Box$ )
  - 4.  $\Box\varphi \rightarrow \Box\Box\varphi$  (M  $\vdash \times \infty \Box$ )
  - 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$  (M  $\vdash \times \infty \Box$ )
- Slate - D.slt**:
  - K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$  (D  $\vdash \checkmark \infty \Box$ )
  - T.  $\Box\varphi \rightarrow \varphi$  (D  $\vdash \times \infty \Box$ )
  - D.  $\Box\varphi \rightarrow \Diamond\varphi$  (D  $\vdash \checkmark \infty \Box$ )
  - 4.  $\Box\varphi \rightarrow \Box\Box\varphi$  (D  $\vdash \times \infty \Box$ )
  - 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$  (D  $\vdash \times \infty \Box$ )
  - INTER.  $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$  (D  $\vdash \checkmark \infty \Box$ )
- Slate - S4.slt**:
  - K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$  (S4  $\vdash \checkmark \infty \Box$ )
  - T.  $\Box\varphi \rightarrow \varphi$  (S4  $\vdash \checkmark \infty \Box$ )
  - D.  $\Box\varphi \rightarrow \Diamond\varphi$  (S4  $\vdash \checkmark \infty \Box$ )
  - 4.  $\Box\varphi \rightarrow \Box\Box\varphi$  (S4  $\vdash \checkmark \infty \Box$ )
  - 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$  (S4  $\vdash \times \infty \Box$ )
  - INTER.  $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$  ({INTER} Assume  $\checkmark$ )
- Slate - S5.slt**:
  - K.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$  (S5  $\vdash \checkmark \infty \Box$ )
  - T.  $\Box\varphi \rightarrow \varphi$  (S5  $\vdash \checkmark \infty \Box$ )
  - D.  $\Box\varphi \rightarrow \Diamond\varphi$  ({D} Assume  $\checkmark$ )
  - 4.  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$  ({4} Assume  $\checkmark$ )
  - 5.  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$  (S5  $\vdash \checkmark \infty \Box$ )
  - INTER.  $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$  ({INTER} Assume  $\checkmark$ )

# Review: Encapsulation

The image displays five overlapping Slate windows, each showing a set of modal logic formulas and their derivability status in a specific system. The windows are titled as follows:

- Slate - K.slt**: Shows formulas K, T, 4, and 5. K is derivable (✓), while T, 4, and 5 are not (✗).
- Slate - T.slt**: Shows formulas K, T, 4, and 5. K and T are derivable (✓), while 4 and 5 are not (✗).
- Slate - D.slt** (highlighted with a red border): Shows formulas K, T, D, 4, 5, and INTER. K, T, and 5 are not derivable (✗), while D and INTER are derivable (✓).
- Slate - S4.slt**: Shows formulas K, T, D, 4, 5, and INTER. K, T, D, and 4 are derivable (✓), while 5 is not (✗). The INTER formula is derivable with the assumption {INTER}.
- Slate - S5.slt**: Shows formulas K, T, D, 4, 5, and INTER. All formulas (K, T, D, 4, 5, and INTER) are derivable (✓). The INTER formula is derivable with the assumption {INTER}.

The formulas shown in each window are:

- K.**  $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$
- T.**  $\Box\varphi \rightarrow \varphi$
- D.**  $\Box\varphi \rightarrow \Diamond\varphi$
- 4.**  $\Box\varphi \rightarrow \Box\Box\varphi$
- 5.**  $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$
- INTER.**  $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$

The derivability status is indicated by a checkmark (✓) for derivable and a red X (✗) for not derivable in the system. The system name is indicated by the symbol before the formula (K, T, D, S4, S5).



#### 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  $\phi$  (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  $\phi$  (which shows permissibility can be defined in terms of obligation), and to say that  $\phi$  is forbidden would seem to amount to it being obligatory that it not be the case that  $\phi$  (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<sup>th</sup> century, and the importance of the triad even to modern deontic logic would be quite hard to exaggerate.<sup>9</sup>

SDL is traditionally axiomatized by the following:<sup>10</sup>

#### **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  $\vdash \psi$

**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  $\phi$  (a formulae representing some state-of-affairs) is permissible seems to amount to saying that  $\neg\phi$  (which shows permissibility can be expressed in terms of obligation) is forbidden. This interconnected trio of ethical concepts has been analyzed since the end of the 18<sup>th</sup> century and modern deontic logic would be quite hard to do without. SDL is traditionally axiomatized by the following axioms:

#### 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  $\vdash \psi$

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

### CHAPTER 4. PROPOSITIONAL MODAL LOGIC

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

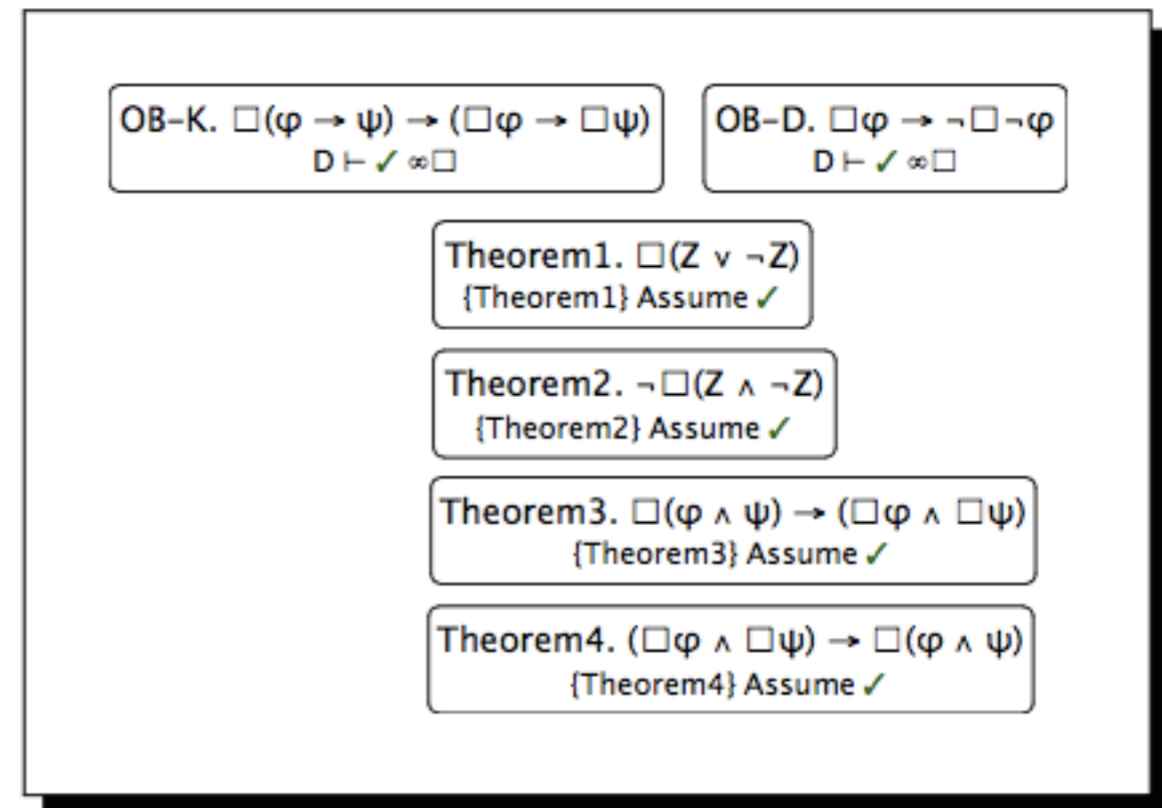


Figure 4.7: The Initial Configuration Upon Opening the File SDL.slt

#### 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.<sup>11</sup> 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 addition, and this is the third given, if Jones *doesn't* go to assist his neighbors, it's obligatory that he not tell

---

<sup>11</sup>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  $\square$  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).



#### 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.

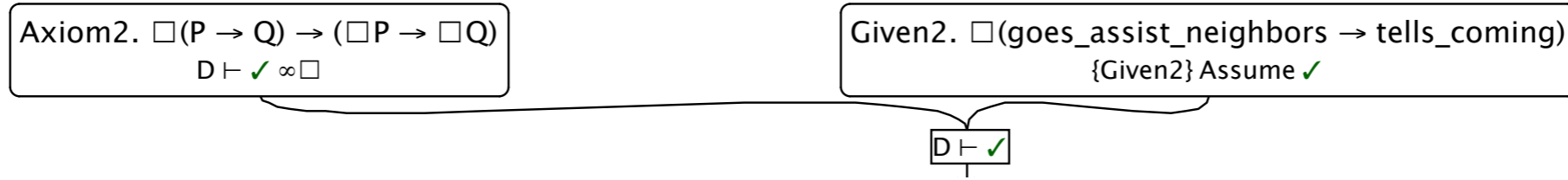
Chisholm's clever scenario revolves around the character Jones.<sup>11</sup> It's given that Jones is obligated to go to assist his neighbors, in part because he has promised to do so. The second given fact is that it's obligatory that, if Jones goes to assist his neighbors, he tells them (in advance)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

---

<sup>11</sup>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  $\square$  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

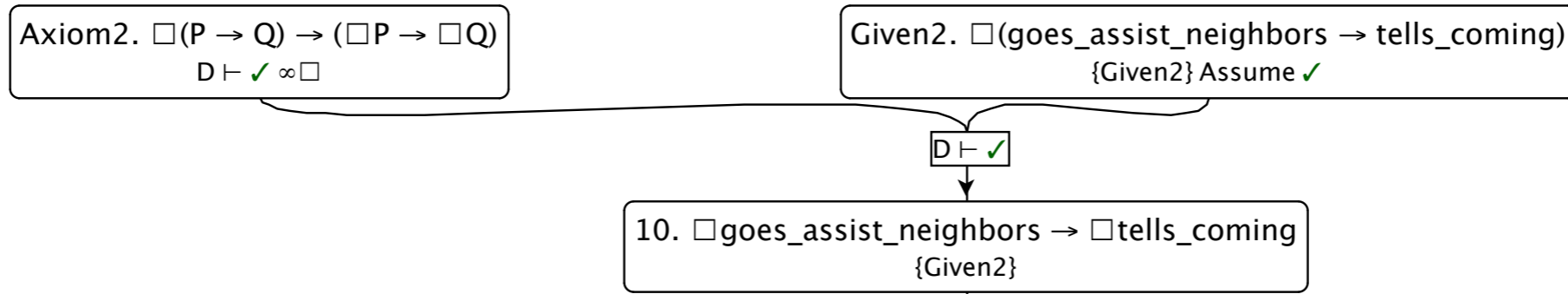


Axiom4. "Modus ponens for provability."  
{Axiom4} Assume ✓

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

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

# Chisholm's Paradox

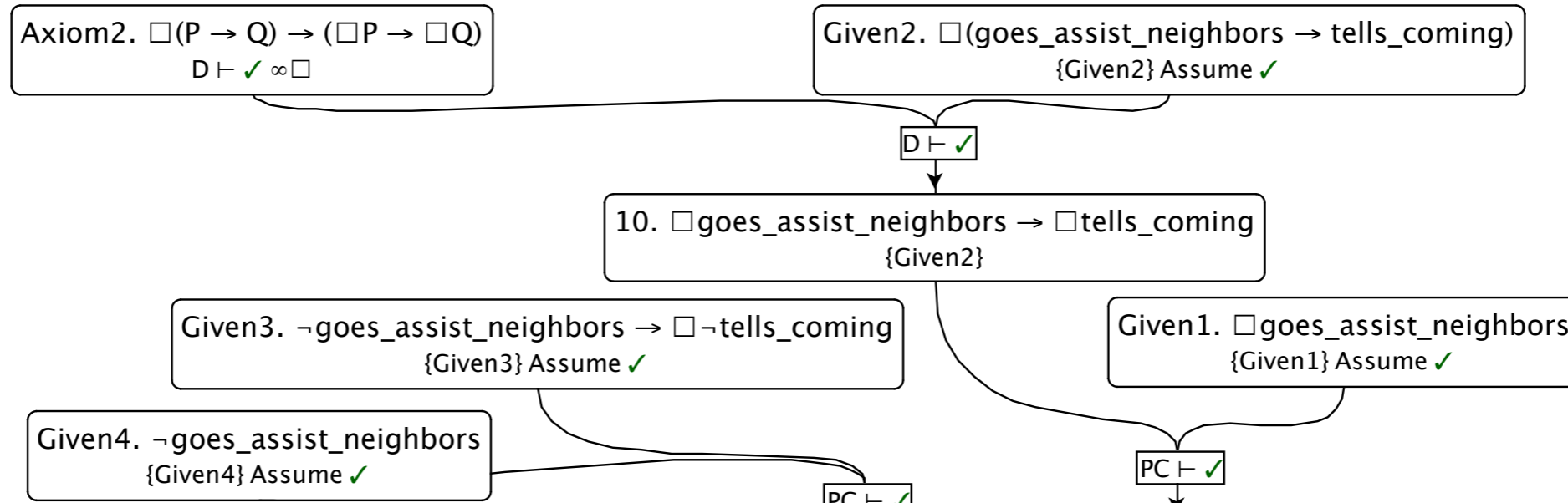


Axiom4. "Modus ponens for provability."  
{Axiom4} Assume  $\checkmark$

Axiom5. "Theorems are obligatory."  
{Axiom5} Assume  $\checkmark$

Axiom1. "All theorems of the propositional calculus."  
{Axiom1} Assume  $\checkmark$

# Chisholm's Paradox

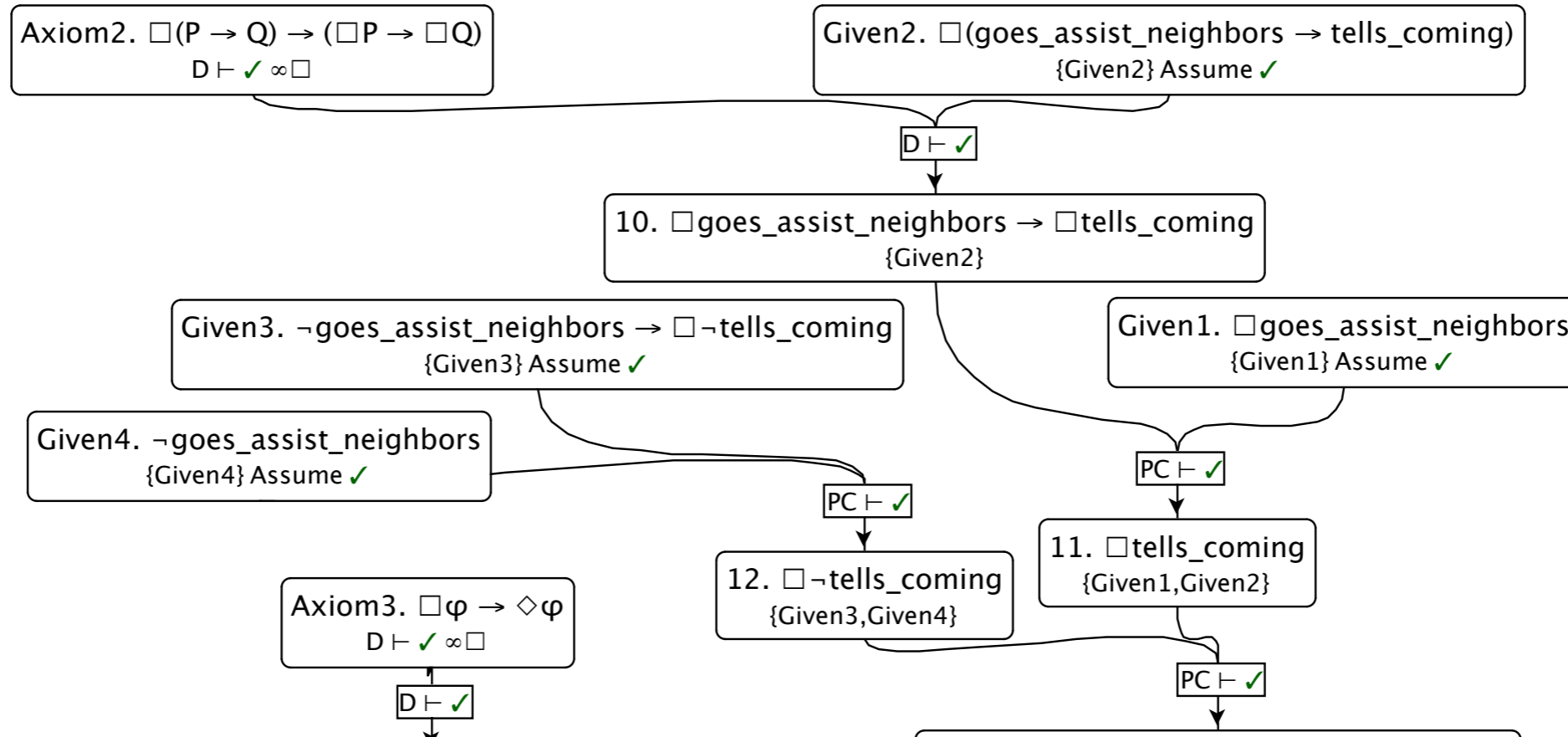


Axiom4. "Modus ponens for provability."  
 $\{\text{Axiom4}\}$  Assume  $\checkmark$

Axiom5. "Theorems are obligatory."  
 $\{\text{Axiom5}\}$  Assume  $\checkmark$

Axiom1. "All theorems of the propositional calculus."  
 $\{\text{Axiom1}\}$  Assume  $\checkmark$

# Chisholm's Paradox

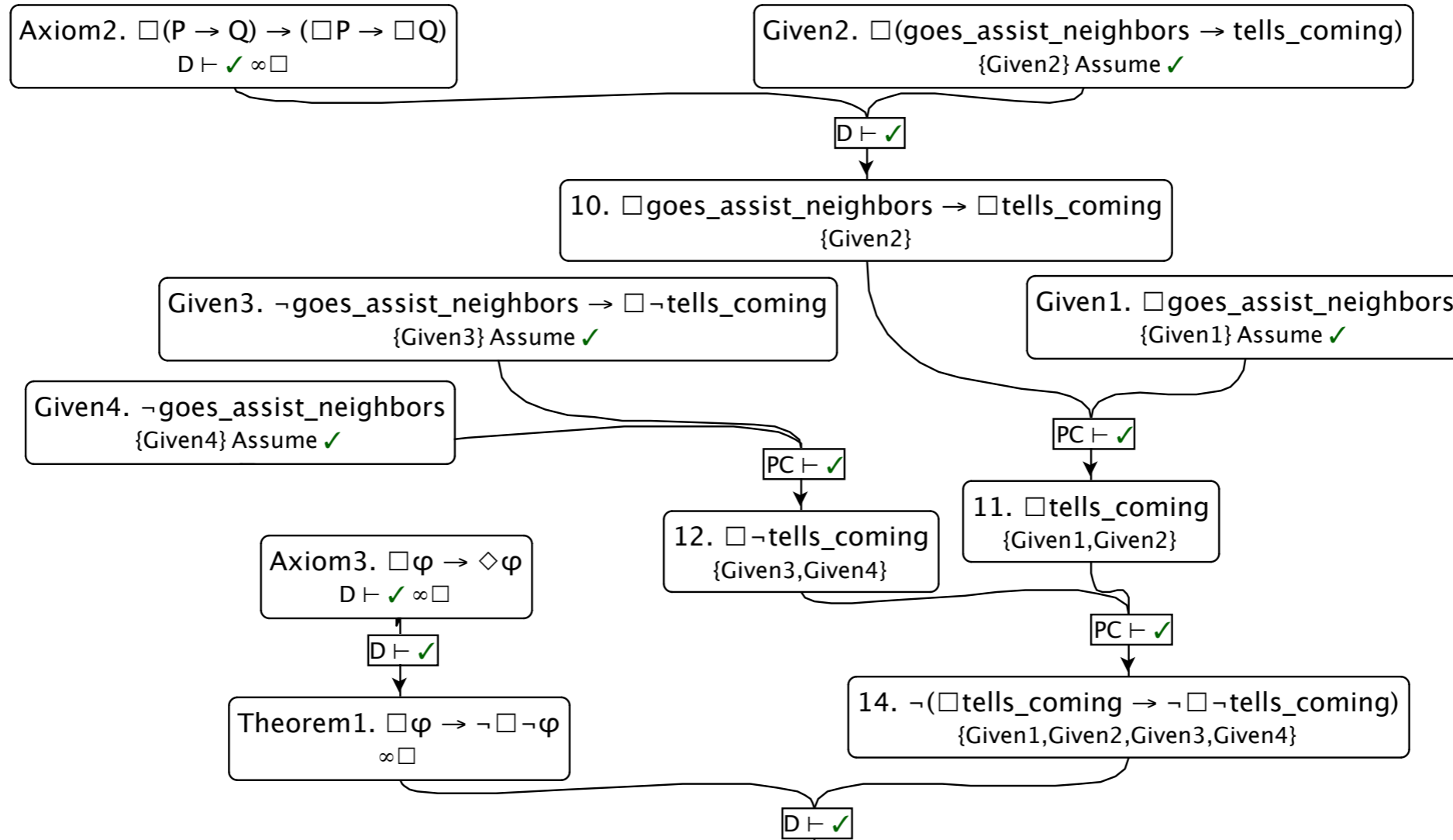


Axiom4. "Modus ponens for provability."  
 $\{\text{Axiom4}\} \text{ Assume } \checkmark$

Axiom5. "Theorems are obligatory."  
 $\{\text{Axiom5}\} \text{ Assume } \checkmark$

Axiom1. "All theorems of the propositional calculus."  
 $\{\text{Axiom1}\} \text{ Assume } \checkmark$

# Chisholm's Paradox

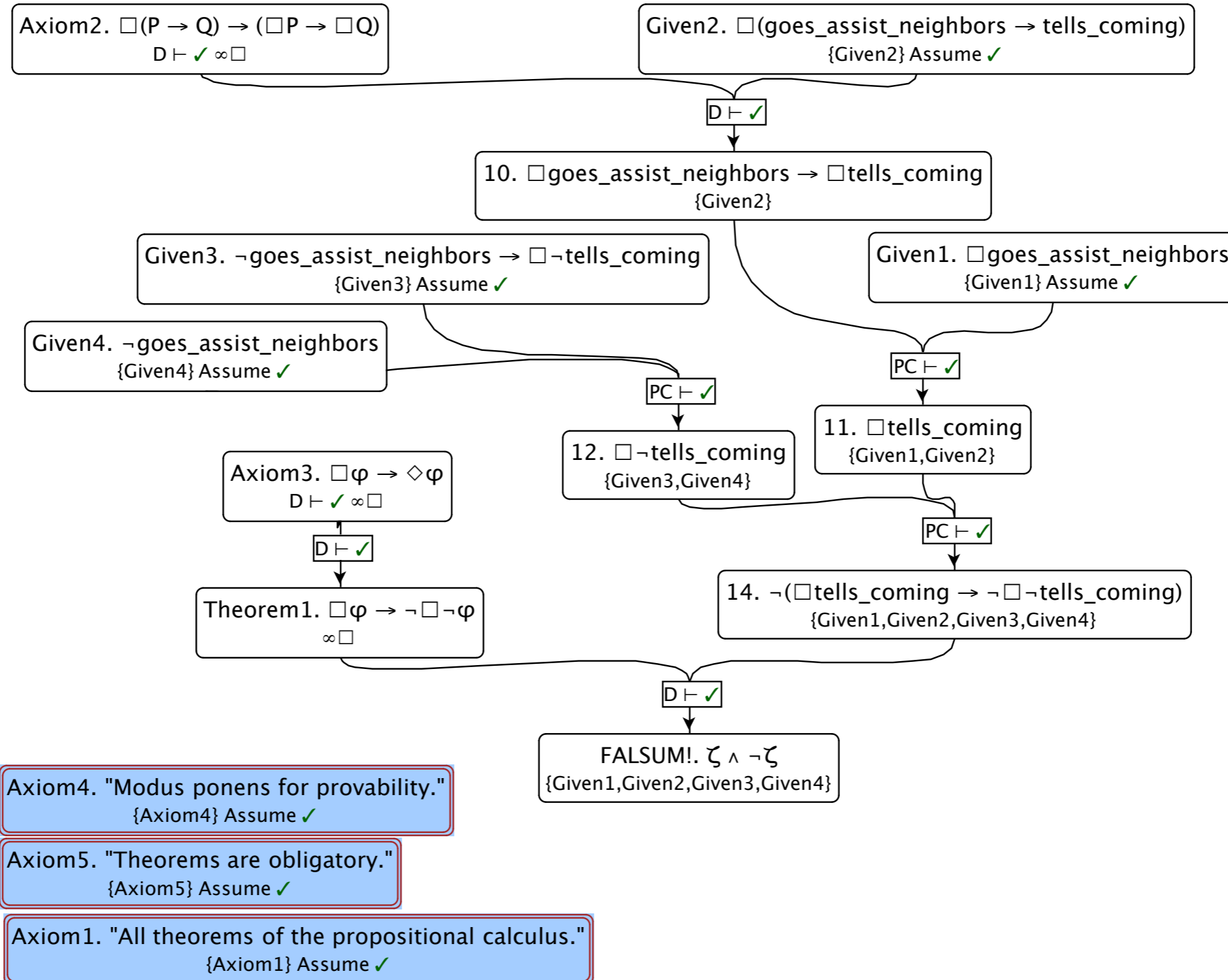


Axiom4. "Modus ponens for provability."  
 $\{\text{Axiom4}\} \text{Assume } \checkmark$

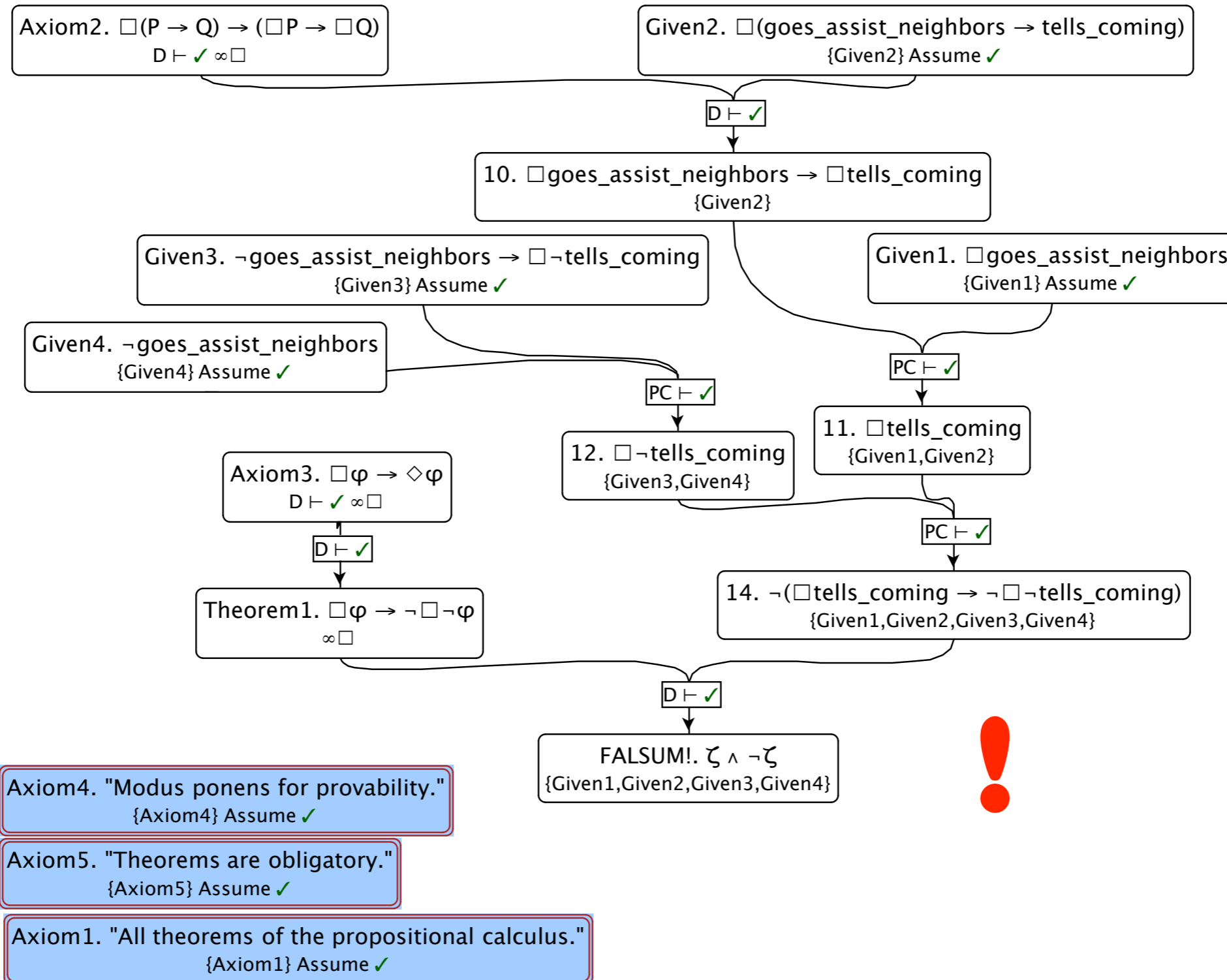
Axiom5. "Theorems are obligatory."  
 $\{\text{Axiom5}\} \text{Assume } \checkmark$

Axiom1. "All theorems of the propositional calculus."  
 $\{\text{Axiom1}\} \text{Assume } \checkmark$

# Chisholm's Paradox



# Chisholm's Paradox





**SDL's = D's Problems**  
**Don't Stop Here ...**

# 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 ✓

$\text{D} \vdash \times$

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

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

$\Box \vdash \times$

2'.  $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$   
{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 ✓

# The Free Choice Permission Paradox (Ross)

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

$D \vdash \times$

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 ✓

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

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

THM 5.  $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$   
 $D \vdash \checkmark \infty \square$

# The Free Choice Permission Paradox (Ross)

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

$D \vdash \times$

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 ✓

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

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

THM 5.  $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$   
 $D \vdash \checkmark \infty \square$

(How?)

# The Free Choice Permission Paradox (Ross)

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

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

$D \vdash \times$

2'.  $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$   
{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.  $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$   
 $D \vdash \checkmark \infty \square$

(How?)

8.  $\diamond\varphi$   
{8} Assume ✓

$PC \vdash \checkmark$

# The Free Choice Permission Paradox (Ross)

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

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

$D \vdash \times$

2'.  $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$   
{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 ✓

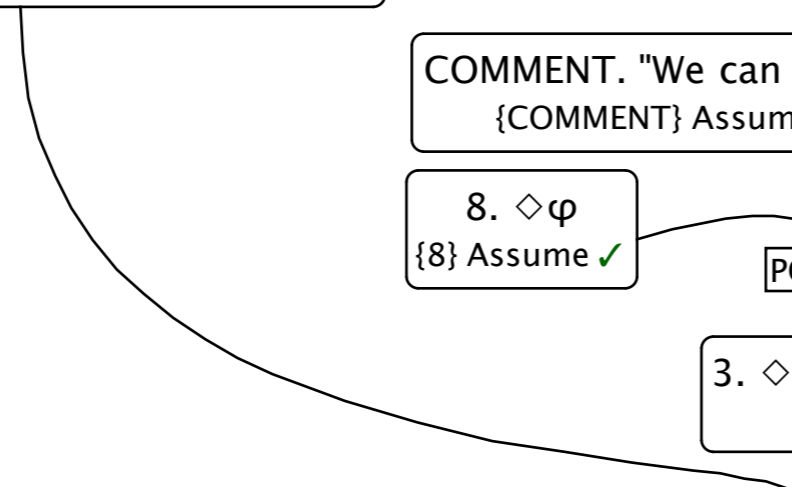
THM 5.  $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$   
 $D \vdash \checkmark \infty \square$

(How?)

8.  $\diamond\varphi$   
{8} Assume ✓

$PC \vdash \checkmark$

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





# The Free Choice Permission Paradox (Ross)

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

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

$D \vdash \times$

2'.  $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$   
 {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.  $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$   
 $D \vdash \checkmark \infty \square$

(How?)

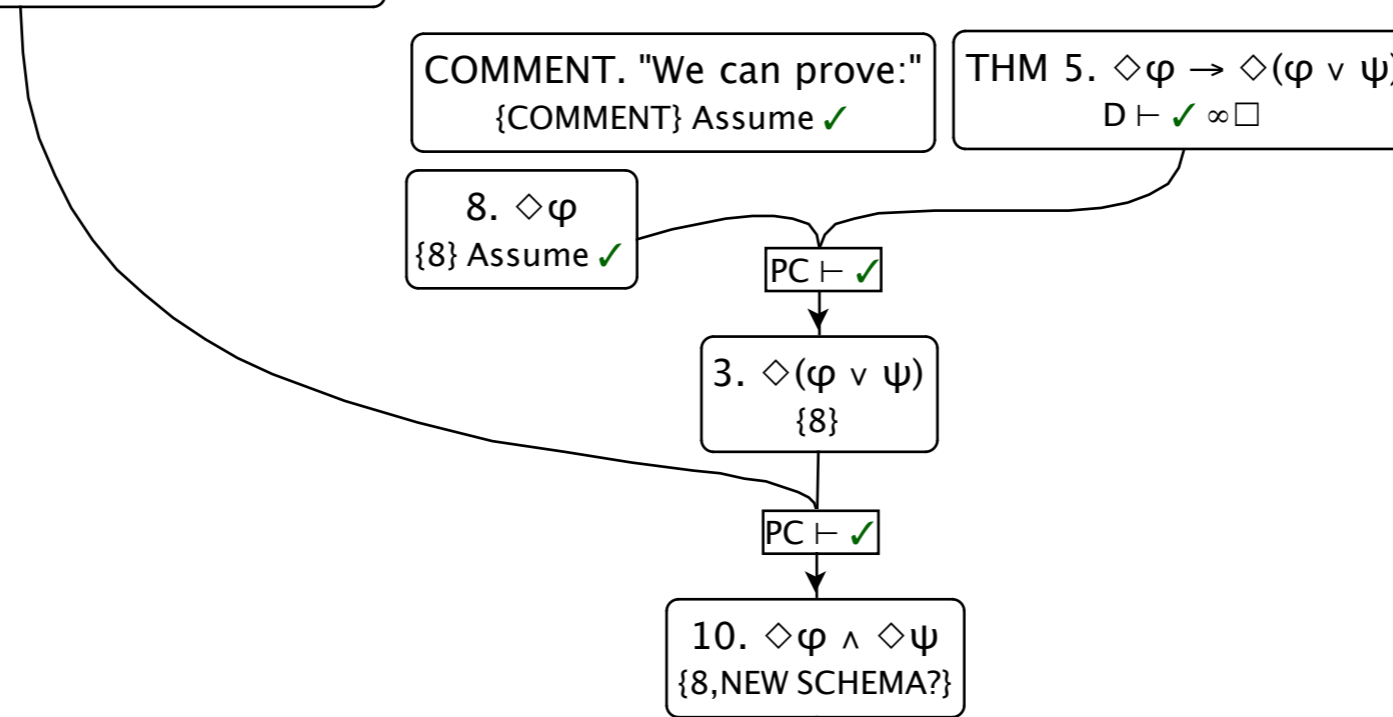
8.  $\diamond\varphi$   
 {8} Assume ✓

$PC \vdash \checkmark$

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

$PC \vdash \checkmark$

10.  $\diamond\varphi \wedge \diamond\psi$   
 {8, NEW SCHEMA?}



# The Free Choice Permission Paradox (Ross)

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

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

$D \vdash \times$

2'.  $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$   
 {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.  $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$   
 $D \vdash \checkmark \infty \square$

(How?)

8.  $\diamond\varphi$   
 {8} Assume ✓

$PC \vdash \checkmark$

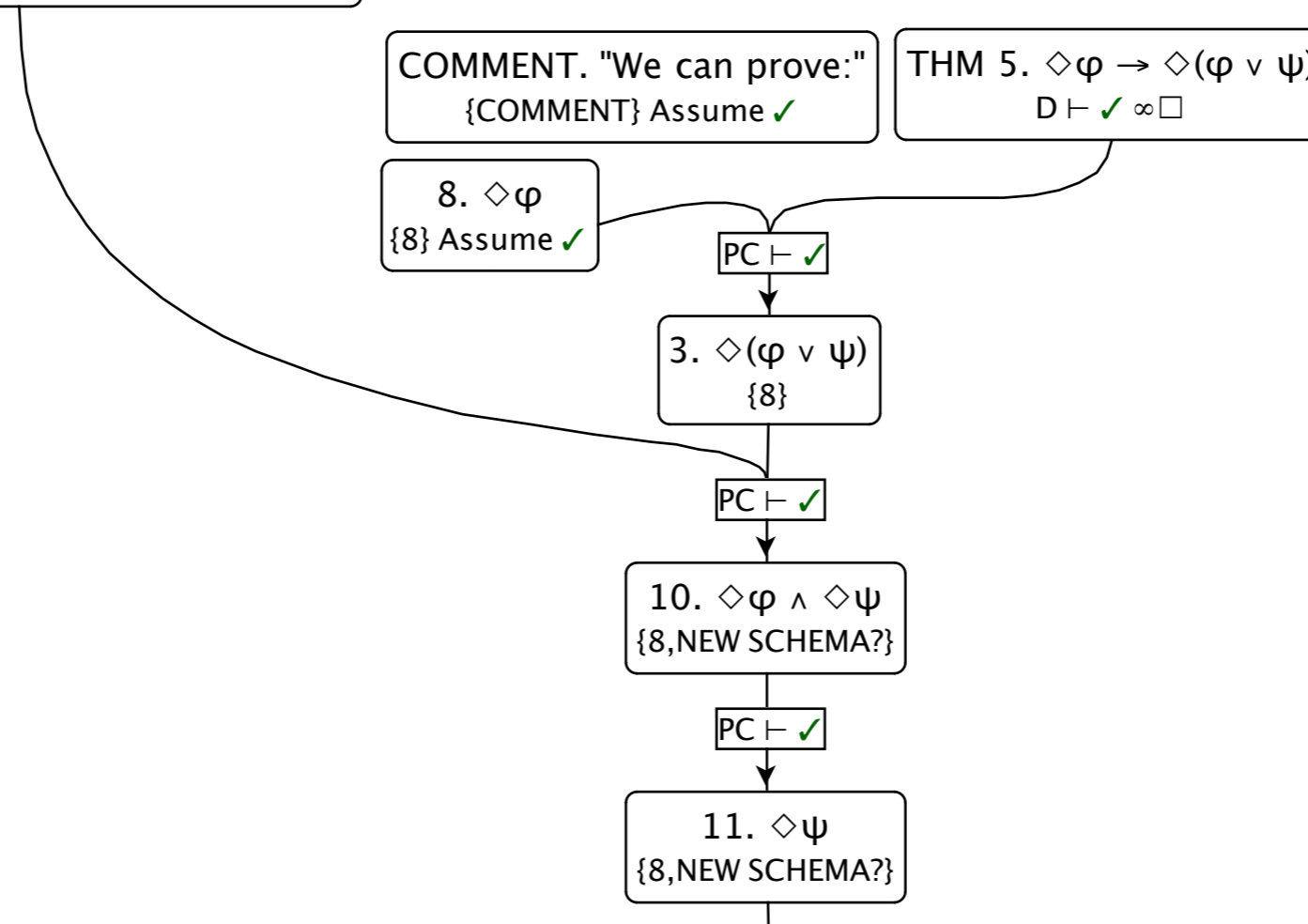
3.  $\diamond(\varphi \vee \psi)$   
 {8}

$PC \vdash \checkmark$

10.  $\diamond\varphi \wedge \diamond\psi$   
 {8, NEW SCHEMA?}

$PC \vdash \checkmark$

11.  $\diamond\psi$   
 {8, NEW SCHEMA?}



# The Free Choice Permission Paradox (Ross)

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

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

$\text{D} \vdash \times$

2'.  $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$   
 {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.  $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$   
 $\text{D} \vdash \checkmark \infty \square$

(How?)

8.  $\diamond\varphi$   
 {8} Assume ✓

$\text{PC} \vdash \checkmark$

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

$\text{PC} \vdash \checkmark$

10.  $\diamond\varphi \wedge \diamond\psi$   
 {8, NEW SCHEMA?}

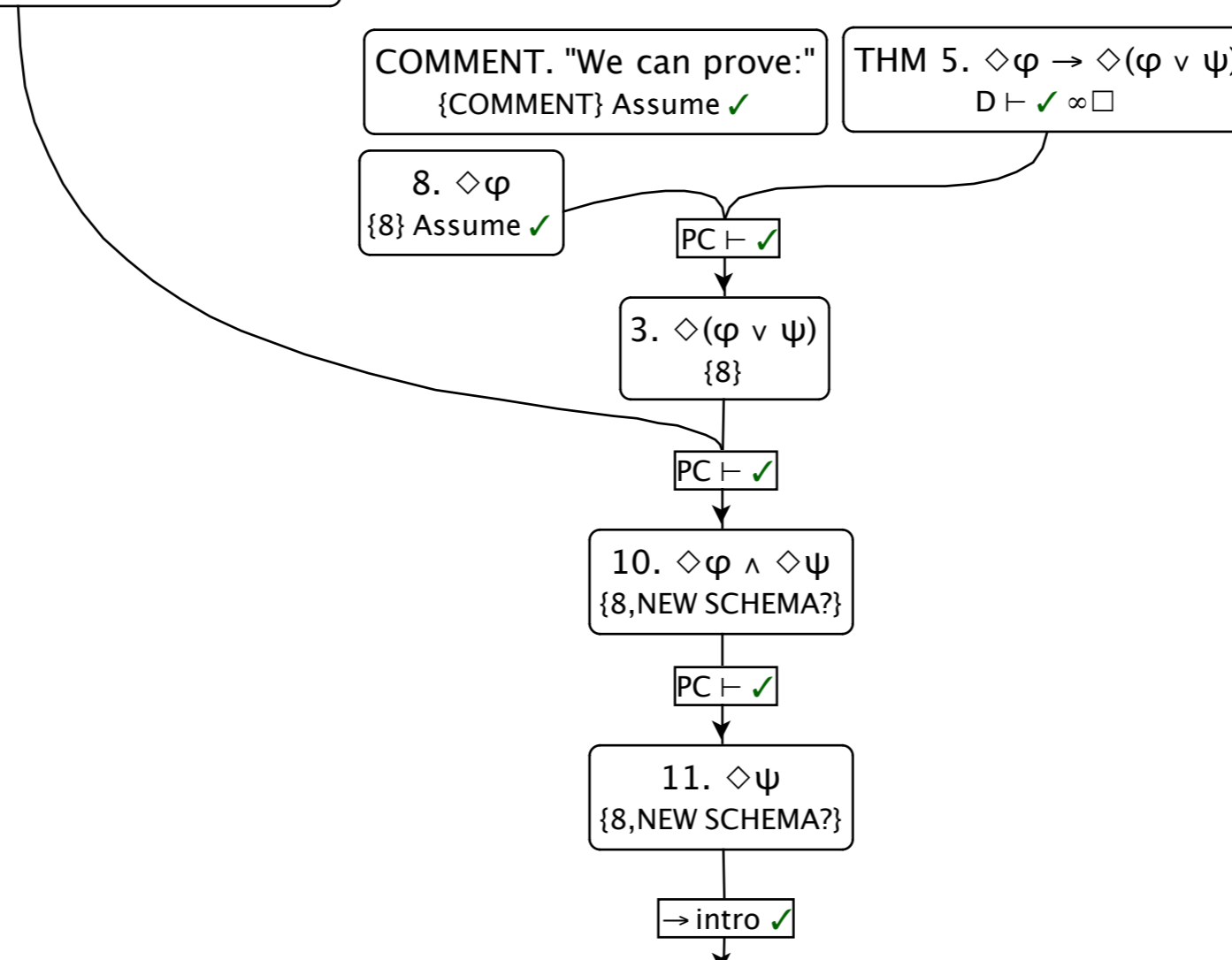
$\text{PC} \vdash \checkmark$

11.  $\diamond\psi$   
 {8, NEW SCHEMA?}

$\rightarrow$  intro ✓

12.  $\diamond\varphi \rightarrow \diamond\psi$   
 {NEW SCHEMA?}

COMMENT. Absurd!  
 {COMMENT} Assume ✓





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