

# 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

IFLAI  
3/27/2023



**At least supposedly, long term:**

At least supposedly, long term:

“We’re in *very* deep trouble.”

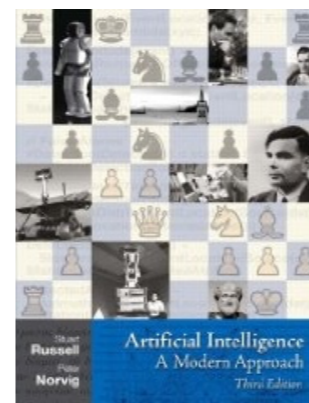
At least supposedly, long term:

“We’re in *very* deep trouble.”



At least supposedly, long term:

“We’re in *very* deep trouble.”



**(Russell himself?) ...**

# Can “Provably Beneficial AI” Save Us?

Selmer Bringsjord <i>Rensselaer AI &amp; Reasoning Lab</i> RPI Troy, USA selmer.bringsjord@gmail.com	Naveen Sundar Govindarajulu <i>Rensselaer AI &amp; Reasoning Lab</i> RPI Troy, USA naveen.sundar.g@gmail.com	John Licato <i>Advancing Machine &amp; Human Reasoning Lab</i> University of South Florida Tampa, USA john.licato@gmail.com
--	--	---

**Abstract**—AI-polymath Stuart Russell, in the face of fear about superhuman AI arriving within 80 years and doing the human race in, commendably offers a recipe (based upon inductive reinforcement learning) for salvation quite different than our own (the sharing of which is beyond the current scope of the present paper). He does this in his recent book *Human Compatible*. Unfortunately, as we explain, Russell’s recipe is afflicted by four fatal defects.

**Index Terms**—machine ethics, robot ethics, inductive reinforcement learning

## I. INTRODUCTION: THE PROBLEM

AI-polymath<sup>1</sup> Stuart Russell, in the face of fear about superhuman AI arriving within 80 years and doing the human race in, offers a recipe for salvation quite different than our own (the sharing of which is beyond the current scope of the present short paper, but see e.g. [8]). He does this in his book *Human Compatible* [11]. Russell does not rely upon The Singularity (or any other such speculative thing) to justify his belief that superintelligent machines will arrive.<sup>2</sup> On the other hand, Russell is of the opinion that the arrival of superintelligent AI could very well be quite sudden. He writes:

My timeline of, say, eighty years is considerably more conservative than that of the typical AI researcher. Recent surveys suggest that most active researchers expect human-level AI to arrive around the middle of this century. Our experience with nuclear physics suggests that it would be prudent to assume that progress could occur quite quickly and to prepare accordingly. If just one conceptual breakthrough were needed, analogous to Szilard’s idea for a neutron-induced nuclear chain reaction, superintelligent AI in some form could arrive quite suddenly. The chances are that we would be unprepared: if we built superintelligent machines with any degree of autonomy, we would soon find ourselves unable to control them. I am, however, fairly confident that we have some breathing space because there are several major breakthroughs needed between here and superintelligence, not just one. [11, Chap. 3, § 7]

The remainder shall unfold straightforwardly as follows. In the next section we summarize what Russell offers as a

<sup>1</sup>Lead author of the encyclopedic, leading introduction and overview of AI, now out in its fourth edition: [12].

<sup>2</sup>The fact is, he does not really tell us in his book why he is so sure superintelligent AI will arrive — but he certainly is sure it will. Our educated guess is that Russell is content with his observing in his book the failure of numerous arguments against the proposition that superintelligent AI will arrive.

solution to the threat to humanity from superintelligent AI. The section after that presents in sequence four problems that plague his proposal. Finally, the paper concludes with a brief discussion of the next steps to be taken in our assessment of Russell’s approach, and in our consideration of competing approaches.

## II. RUSSELL’S PROPOSED SOLUTION

What is the solution Russell proposes? We cannot cover the ins and outs of his solution, as doing so would require a detailed explanation of *reinforcement learning* (RL), including *inverse RL* (IRL), upon which his proposal rests. While these forms of learning are mathematically simple frameworks in which agents gradually get better at reaching toward a goal, we nonetheless have not the time and space here to burn in exposition — and besides which RL and IRL are well-known to AI researchers. (Russell’s [11] *Human Compatible* is in fact itself an excellent non-technical introduction to these forms of learning.) Fortunately, the core of Russell’s proposed solution, what he calls “Provably Beneficial AI” (PBAI), can be quite efficiently conveyed here. The core of PBAI is that we take care to engineer robots driven solely by a “desire” to reach goals that accord with the goals of humanity. Of course, desire in the human case entails that the human doing the desiring has some states of “phenomenal” or “subjective” consciousness (what Block [1] calls ‘p-consciousness’). This is so because, as we humans all know, when one desires something, one *feels* things, inevitably. For example, if one intensely desires to get some reward, and works ferociously toward it, but keeps failing to even get close to obtaining it, one is likely to e.g. feel frustrated, angry, despondent, and so on. Thus, we use scare quotes around ‘desire’ so as not to assume any such thing as that the robots Russell seeks will have p-consciousness.<sup>3</sup>

Encapsulated, what then in Russell’s PBAI is the reward “desired” by the machines? He maintains that that reward will be none other than our own collective maximal well-being. Since we can safely assume that such goals in our case include that our species survives, and indeed overall thrives, if such a “desire” can be counted upon to really and truly drive our future robots, we should as a species be in good shape. In addition, we must be able to comfortably *prove*

<sup>3</sup>According to the first author, they will have no such thing, and in fact no one at present has the slightest clue as to how to proceed with engineering that can be rationally regarded to move a nanometer closer to p-conscious AIs, as explained in [2].

that the robots are beneficial to humanity. Here is how Russell expresses overall his rather rosy take on things:

[M]y proposal for beneficial machines: machines whose actions can be expected to achieve *our* objectives. Because these objectives are in us, and not in them, the machines will need [via IRL] to learn more about *what we really want* from observations of the choices we make and how we make them. Machines designed in this way will defer to humans: they will ask permission; they will act cautiously when guidance is unclear; and they will allow themselves to be switched off. [11, ¶ 2, § “Beneficial Machines” in Chap. 10 “Problem Solved?”; emphasis ours]

Unfortunately, while we have deep respect for the formality of Russell’s approach (unsurprising since any real formality is rooted in formal logic and proofs therein: there is no other way to achieve a proof by to employ formal logic) there are four each-fatal-in-their-own-right problems plaguing Russell’s proposal, as we now explain. Here now are these problems.

### III. FOUR PROBLEMS AFFLICTING RUSSELL’S PBAI

As promised, we now proceed to explain, in turn, four defects (among others) that afflict PBAI.

#### A. Problem 1: *Sola Utilitarianism?*

The first problem is simple to grasp, and simply devastating; it is that Russell’s proposal to save our race is based upon *only* the family of consequentialist ethical theories. This family includes the familiar ethical theory known as *act utilitarianism*, according to which what is obligatory are actions that maximize overall happiness; a precise account can be found in the classic [7]. But surely this particular family is only an *option* from among many families of ethical theories; and, these families are pairwise inconsistent. That is, pick any two families, and the definitions they include for the central operators of any ethical theory, for instance for *obligatory*, and one will arrive at contradictions, by elementary deductive reasoning over these definitions in garden-variety contexts. To see this, let us pick for consideration another ethical-theory family. Specifically, let us pick for expository purposes the family of *divine-command* ethical theories. Divine-command ethical theories are based upon the core notion that what is obligatory, permissible, forbidden, and so on is wholly determined by God’s commands. A seminal presentation of a divine-command ethical theory is given by [10]. Exploration of divine-command ethical theories in a manner that conforms to what is needed in attempts to engineer morally correct machines is carried out in [4]. Note that when one considers the entire population of planet Earth, and subscription among its members to a dominant family of ethical theories, it is probably the divine-command family that has the largest number of adherents, by far.<sup>4</sup>

<sup>4</sup>There are currently e.g. about 2.2 billion Christians on Earth, and about 2 billion Muslims. For both groups, by definition, it is first and foremost what God commands that determines what is obligatory. Orthodox and conservative Jews would of course be in precisely the same category. (This is of course not at all to say that the three religions here each perfectly agree on every attribute ascribed to God. The main ones, though — e.g. omnipotence, omniscience, omnipresence, omnibenevolence, creator of all contingent things — are indeed ascribed to God in the case of each of the trio of religions we cite here.

Now, given the setup supplied in the previous paragraph, here is a pair of relevant biconditionals, one from each of the two families we have just cited.<sup>5</sup> The first is part of act utilitarianism; the second is from all divine-command ethical theories.

Ob<sub>U</sub> An agent (a category that includes human persons) is obligated at time  $t$ , given (context)  $\Phi$ , to do action  $a$  at later time  $t'$  if and only if  $a$ , from among all viable alternative actions available to this agent, brings about the most happiness for the most people.

Ob<sub>DC</sub> A human person is obligated at time  $t$ , given (context)  $\Phi$ , to do action  $a$  at later time  $t'$  if and only if the performance of  $a$  has been commanded by God (or is deductively entailed by what has been commanded by God).

We are quite sure the reader can see the problem. By ‘context’ here, represented by ‘ $\Phi$ ,’ is meant simply a collection of declarative formulae, or for our somewhat informal exposition here, declarative propositions, that sets the situation. We can consider a hypothetical to make this more concrete: Molycarp is a devout Christian living under a brutal dictatorship whose key tenets include those of rabid and unrelenting atheism, and Molycarp is imprisoned, tortured, and asked to explicitly utter blasphemous and profane denial of his orthodox conception of Jesus as sinless and divine.<sup>6</sup> *Ex hypothesi*, Molycarp’s agreeing to do this will save his life, ensure the well-being of his family (for which he is the breadwinner), and bring about many, many other happiness-bearing states-of-affairs through an endless array of chains of weal catalyzed by his subsequent actions. However, if he accepts death, only two terrestrial people will ever know what happened to him (the dictator and the executioner), as he will be incinerated, and in fact soon after his death everybody else will thoroughly forget about him. By a suitable instantiation of Ob<sub>DC</sub>, Molycarp is obligated to proclaim his belief in Jesus and his divinity, and die a martyr; but in stark contrast, by a suitable instantiation of Ob<sub>U</sub>, he is obligated to go through the motions of quickly spouting out a few words that will secure his freedom, and a lot of happiness that cannot otherwise be secured. Assuming that no one can be obligated to perform two actions that are impossible to both perform,<sup>7</sup> we have a contradiction.

There is more general, history-centric way to sum up Problem 1 for Russell, and for those inclined to follow him; it is to simply report that the discipline of systematic, theoretical ethics has been in progress since at least Aristotle, three centuries before the birth Christ, and if we know anything at all about the history of the discipline from that ancient timepoint we know that the human race has on hand myriad families of ethical theories, each none other than, as we have noted above, pairwise incompatible. It is thus rather doubtful that the solution to the problem posed by future superintelligent

<sup>5</sup>For easing exposition, let us not worry about which particular ethical theory is in play here from each of the two families we have called out.

<sup>6</sup>The sinlessness and divinity of Jesus is a credal doctrine of orthodox Christianity. See e.g. [13]. Many readers will see in our use of ‘Molycarp’ a thinly disguised reference to the real martyr Polycarp, executed in 155 AD.

<sup>7</sup>This, that “ought implies can,” is known as *Kant’s Law*, and is a staple in deontic logic, the branch of logic devoted to logicizing ethical theories.



machines is to be found in the Russellian engineering of robots whose *modus operandi* is the following the dictates of only one family, consequentialism.

#### B. Problem 2: Mental States Not Inferred from Behavior

The second problem afflicting Russell's approach to the threat to humanity is that this approach at its heart relies upon the ability of present and future AIs to infer a human's interior mentality from that human's exterior, readily observable behavior. After all, what Russell (admirably and rationally) wants is for the machines in question to place our happiness first among the goals they seek — but what is happiness if not a mental state, and as such an *invisible* state? (This is why we emphasized the phrase 'what we really want' in the quote of Russell just above.) This particular sentence is being written (at least in its first version) by author Bringsjord, who is thus simply staring at a screen and typing as characters appear on said screen for this eyes to take in. Okay, so suppose you walk up now to Bringsjord, who is seated, and look at his face, standing above him; and suppose that he stops typing and looks at your face. Can you tell if Bringsjord is happy? You may of course be able to rationally *assert* that he is happy, because you may have empirical data regarding his recent past (e.g., that he had a gourmet lunch featuring arctic char at Manhattan's Aquavit restaurant, a particular favorite of his, before his the current work session you just interrupted), and you may even happen to have a live feed from Selmer's iPhone somehow, giving you his vitals and perhaps all sorts of information about this bodily state, including its over internal condition in many regards, but — again, we assume here that Selmer is staring at you, expressionless — you will only be guessing. And in fact you would be wrong. Reason? Selmer happens to be thinking about an event in his childhood, a rather sad one: the death of his dog King, caused by a car; and his current state is far from a happy one, mixed as it is with some rather dreadful mental movies of what happened that fateful day just outside New York City.

Now, just replace you with a robot (or with an AI using sensors in the relevant room) looking at Selmer, and you will see the problem facing Russell. AIs cannot toil on our behalf by using inductive reinforcement learning because they cannot learn the nature of what they need to reduce or increase: namely, our mental states.

#### C. Problem 3: Cognition Ranges Beyond The Turing Limit

The next problem is quite simple to state. The robots that will be toiling in our favor are explicitly asserted by Russell to be boxed in by what a Turing machine can do. This is easy to confirm, because when he offers a theorem-schema that, when proved, will provide the ultimate assurance he seeks in the face of impending doom from superintelligent machines, that theorem-schema employs 'machine,' and this term means *Turing-level* machine. (We look at Russell's theorem-sketch below, in the final section.) Put another way, the robots with which Russell is concerned are all constrained by the Turing Limit, the level of computational power beyond which Turing

machines (and lesser machines, e.g. linear-bounded Turing machines). But that means that if our cognition, our intellectual power, extends *beyond* this limit, the robots will not be able to grasp and abide by our cognition. But according to Bringsjord, human cognition is indeed of this nature; see for example assertions and defenses of this claim in [3, 5, 6].

It is important to grasp that the problem here for Russell's PBAI paradigm is not weak, vague, or haphazard; in fact the problem is logico-mathematical in nature. Suppose one computing machine  $m_1$  is not capable of computing functions beyond some *bona fide* level  $L_1$ , and that some other computing machine  $m_2$  is capable of computing functions at some level  $L_2$  above  $L_1$ .<sup>8</sup> It then is an easy theorem that  $m_1$ , by observing the operation of the more powerful  $m_2$ , cannot compute functions at  $L_2$ , or for that matter one iota above  $L_1$ . Yet, Russell pins his hopes on robots that will observe us, and figure out how to work to our benefit. But what if our benefit requires doing things that demand as much cognitive power as we have? In that case it is mathematically impossible for his salvific recipe to work.

#### D. Problem 4: Humans Do Not Agree on Weighty Propositions

Let us suppose for the sake of argument that the Russellian beneficial-to-us robots can indeed somehow be magically engineered, so that at every moment of their existence, and perpetually so, they toil for *the* benefit of humanity: their sought-after reward is that very benefit. Notice our emphasis on the word 'the' in the previous sentence. That tiny little word, a so-called "determiner," creates a fatal problem for Russell. The problem is that there is no *the* thing that is humankind's benefit. What would this thing be, after all? Masochists seek their own harm and pain; sadists the harm and pain of others; criminals their own material benefit at the expense and pain of others; Christians perpetual bliss in an afterlife, this earthly life being no more than — quoting David — a vapor and — quoting Solomon — at its best filled with soul-making suffering; "brave" existentialists like Camus expend what they admit is pointless effort to stay alive even though this life is evanescent and absurd; and so on seemingly *ad infinitum* into never-ending heterogeneity. So, there is no *the* benefit, alas. The bottom line for Russell's PBAI explodes it; that bottom line is that each relevant group of humans, with enough wealth, is going to purchase a robot or robots in order to facilitate *their* priorities. If anything, this will just make the world as contentious and chaotic as it is now — maybe more so.

#### IV. NEXT STEPS

The alert reader will recall that there is a 'P' for 'provably' in Russell's 'PBAI.' What is it that Russell says we need to prove in his approach? He gives the general shape of the theorems which, if proved, will constitute assurance. We read:

<sup>8</sup>We spare the reader technical bases beneath this imagined state-of-affairs, but mention here that this means that the levels must be ones in the Arithmetic Hierarchy or Analytic Hierarchy, and genuinely distinct ones therein. We cannot be referring to levels in the Polynomial Hierarchy, because all problems in that hierarchy are Turing-solvable.

Let's look at the kind of theorem we would like eventually to prove about machines that are beneficial to humans. One type might go something like this:

Suppose a machine has components  $A$ ,  $B$ ,  $C$ , connected to each other like so and to the environment like so, with internal learning algorithms  $l_A$ ,  $l_B$ ,  $l_C$  that optimize internal feedback rewards  $r_A$ ,  $r_B$ ,  $r_C$  defined like so, and [a few more conditions] . . . Then, with very high probability, the machine's behavior will be very close in value (for humans) to the best possible behavior realizable on any machine with the same computational and physical capabilities.

Russell's main point here is that such a theorem should hold regardless of how smart the components become — that is, “the vessel never springs a leak and the machine always remains beneficial to humans” ([11, Chap. 8, § “Mathematical Guarantees,” ¶ 8]). The next step in our evaluation of PBAI is to investigate carefully how theorems of this general shape can *in fact* be proved. This will require formalizing the concepts that Russell leaves vague and undefined here. For example, what, logico-mathematically speaking, is a ‘machine’ in the theorem-sketch that Russell provides here?<sup>9</sup> Likewise, what precisely is ‘the environment’? At the very least, we shall need to venture precise answers to these questions in order to understand what Russell is gesturing toward when he sketches the kind of theorem to target in PBAI. We will then need to see if in fact an actual theorem of this shape can be proved, and what the proof would need to be like. Following on this, another step will be to see if, in approaches very different than PBAI, theorems providing greater assurance can be obtained. After all, Russell here concedes, explicitly, that the best his approach can reach is only “very high probability” that the machines will operate in our interests. We believe that total assurance can in fact be secured on the strength of proving theorems of a different nature than what Russell describes, and will seek to demonstrate that our optimism is well-founded.

#### ACKNOWLEDGMENTS

We are indebted to Stuart Russell for bravely and perspicaciously dealing with an acute future danger that many may wish to ignore or at least severely downplay. We are deeply grateful to ONR for past, extended support of research in the area of robot ethics (that informs the present paper), in particular through a MURI grant on which both Bringsjord and Govindarajulu were central researchers (along with PI Matthias Scheutz and Co-PI Betram Malle).

#### REFERENCES

- [1] N. Block. On a Confusion About a Function of Consciousness. *Behavioral and Brain Sciences*, 18:227–247, 1995.

<sup>9</sup>Apropos of the discussion above, what about computing machines that are provably capable of more than what can be done by a standard Turing machine? E.g., what about infinite-time Turing machines [9]?

- [2] S. Bringsjord. Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline. *Journal of Consciousness Studies*, 14(7):28–43, 2007. URL <http://kryten.mm.rpi.edu/jcsonebillion2.pdf>.
- [3] S. Bringsjord and K. Arkoudas. The Modal Argument for Hypercomputing Minds. *Theoretical Computer Science*, 317:167–190, 2004.
- [4] S. Bringsjord and J. Taylor. The Divine-Command Approach to Robot Ethics. In P. Lin, G. Bekey, and K. Abney, editors, *Robot Ethics: The Ethical and Social Implications of Robotics*, pages 85–108. MIT Press, Cambridge, MA, 2012. URL [http://kryten.mm.rpi.edu/Divine-Command\\_Robo](http://kryten.mm.rpi.edu/Divine-Command_Robo)
- [5] S. Bringsjord and M. Zenzen. *Superminds: People Harness Hypercomputation, and More*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003.
- [6] S. Bringsjord, O. Kellett, A. Shilliday, J. Taylor, B. van Heuveln, Y. Yang, J. Baumes, and K. Ross. A New Gödelian Argument for Hypercomputing Minds Based on the Busy Beaver Problem. *Applied Mathematics and Computation*, 176:516–530, 2006.
- [7] F. Feldman. *Introductory Ethics*. Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [8] N. Govindarajulu and S. Bringsjord. On Automating the Doctrine of Double Effect. In C. Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4722–4730. International Joint Conferences on Artificial Intelligence, 2017. ISBN 978-0-9992411-0-3. doi: 10.24963/ijcai.2017/658. URL <https://doi.org/10.24963/ijcai.2017/658>.
- [9] J. D. Hamkins and A. Lewis. Infinite Time Turing Machines. *Journal of Symbolic Logic*, 65(2):567–604, 2000.
- [10] P. Quinn. *Divine Commands and Moral Requirements*. Oxford University Press, Oxford, UK, 1978.
- [11] S. Russell. *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin Books, New York, NY, 2019. This is the ebook version, specifically an Apple Books ebook.
- [12] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Pearson, New York, NY, 2020. Fourth edition.
- [13] R. Swinburne. *Was Jesus God?* Oxford University Press, Oxford, UK, 2010.

**No need to assume  
superhuman AI ...**

**No need to assume  
superhuman AI ...**

**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

$\forall x : \text{Agents}$

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

$\forall x$  : Agents

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

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

$\forall x : \text{Agents}$

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





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

$\forall x : \text{Agents}$

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



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

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

$\forall x : \text{Agents}$

Powerful(x)



$u(AIA_i(\pi_j)) >$

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> C <sup>o</sup> C)	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

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

$\forall x : \text{Agents}$

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



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

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

$\forall x : \text{Agents}$

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



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

**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.

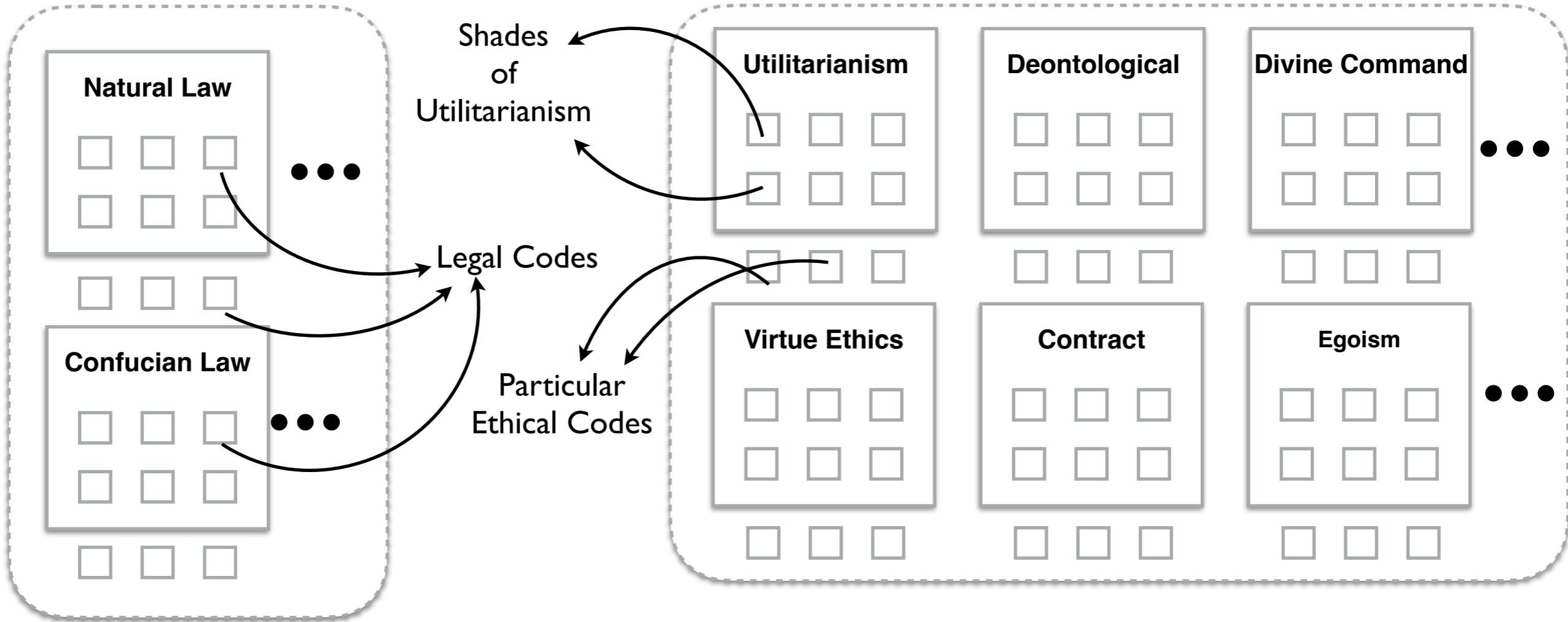
**Proof:** Let  $a$  and  $a'$  be agents satisfying the hypothesis of the theorem in an arbitrary collaborative situation. Then, by definition,  $a \neq a'$  desires to obtain some goal  $\gamma$  in part by way of a contributed action  $\alpha_k$  from  $a'$ ,  $a'$  knows this, and moreover  $a'$  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**

# Making Morally X Machines

**\$IIM**

Theories of Law

Ethical Theories

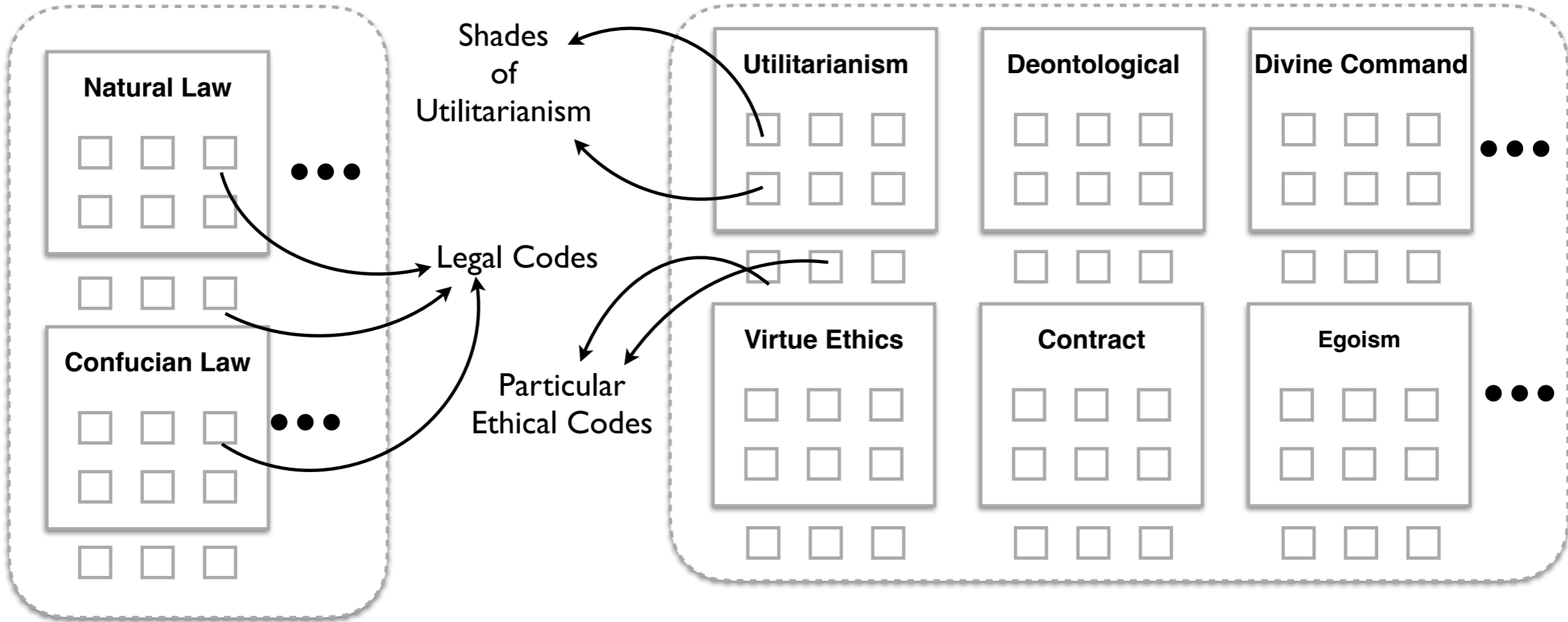


# Making Morally X Machines

**\$IIM**

Theories of Law

Ethical Theories

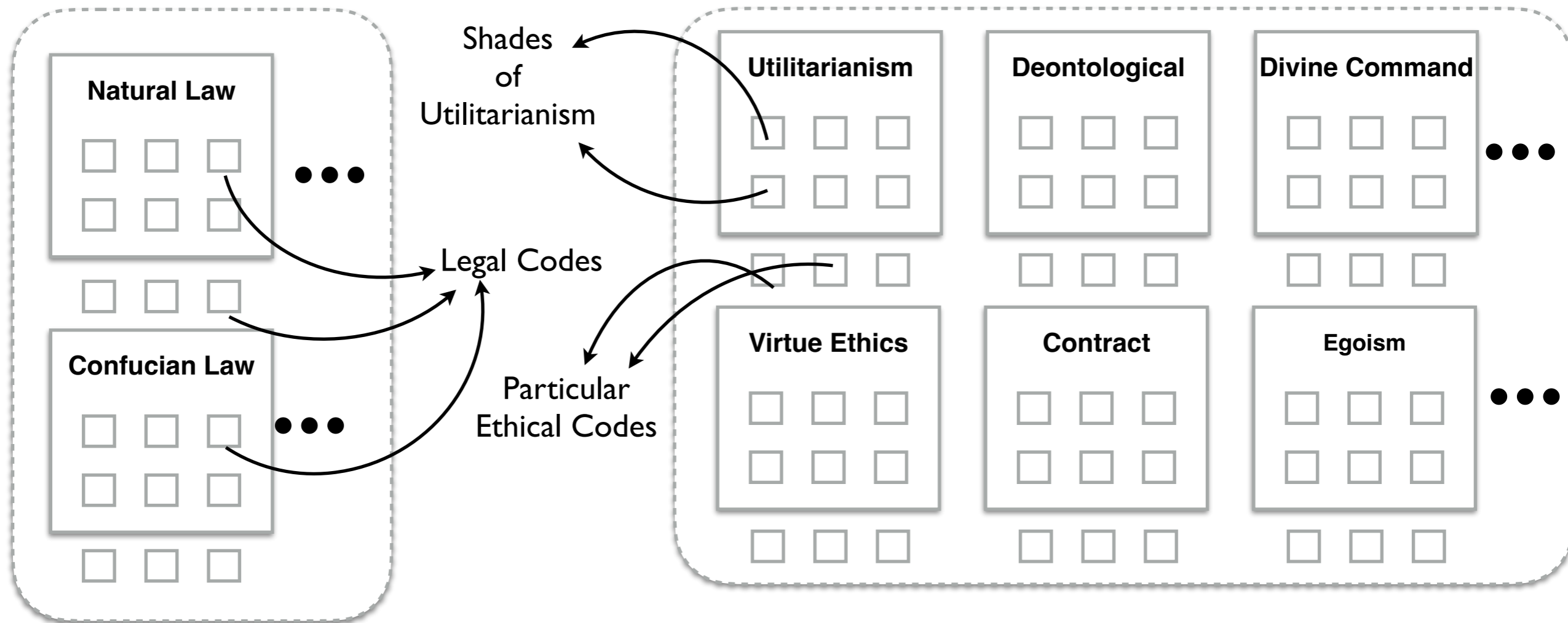


# Making Morally X Machines

**\$IIM**

Theories of Law

Ethical Theories

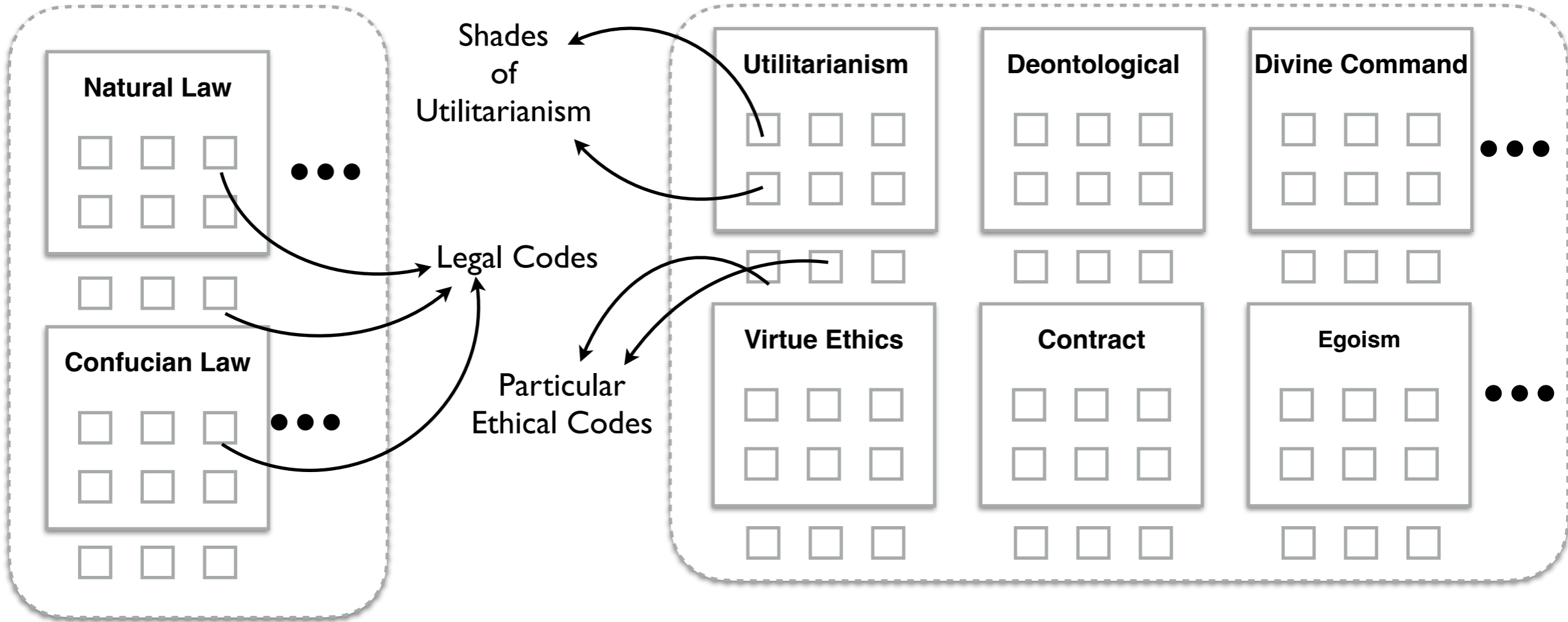


# Making Morally X Machines

**\$IIM**

## Theories of Law

## Ethical Theories



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

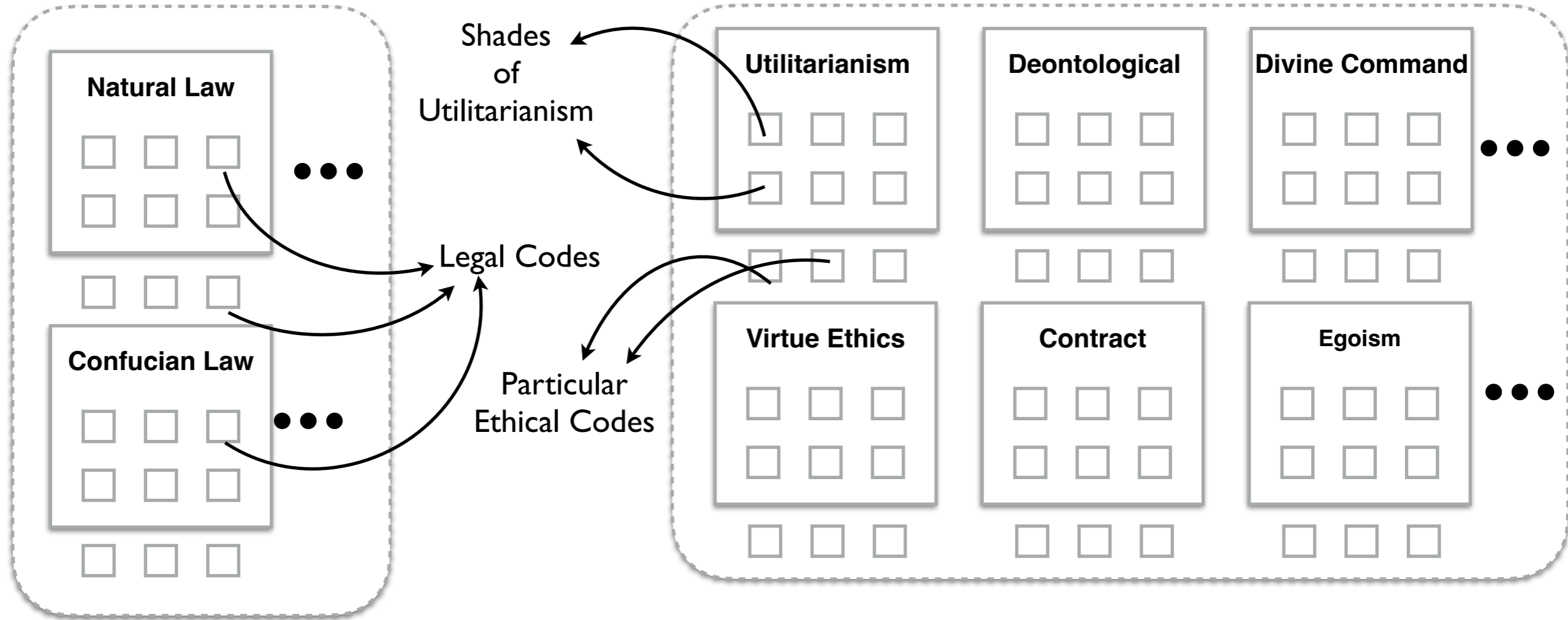


# Making Morally X Machines

**\$IIM**

## Theories of Law

## Ethical Theories



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

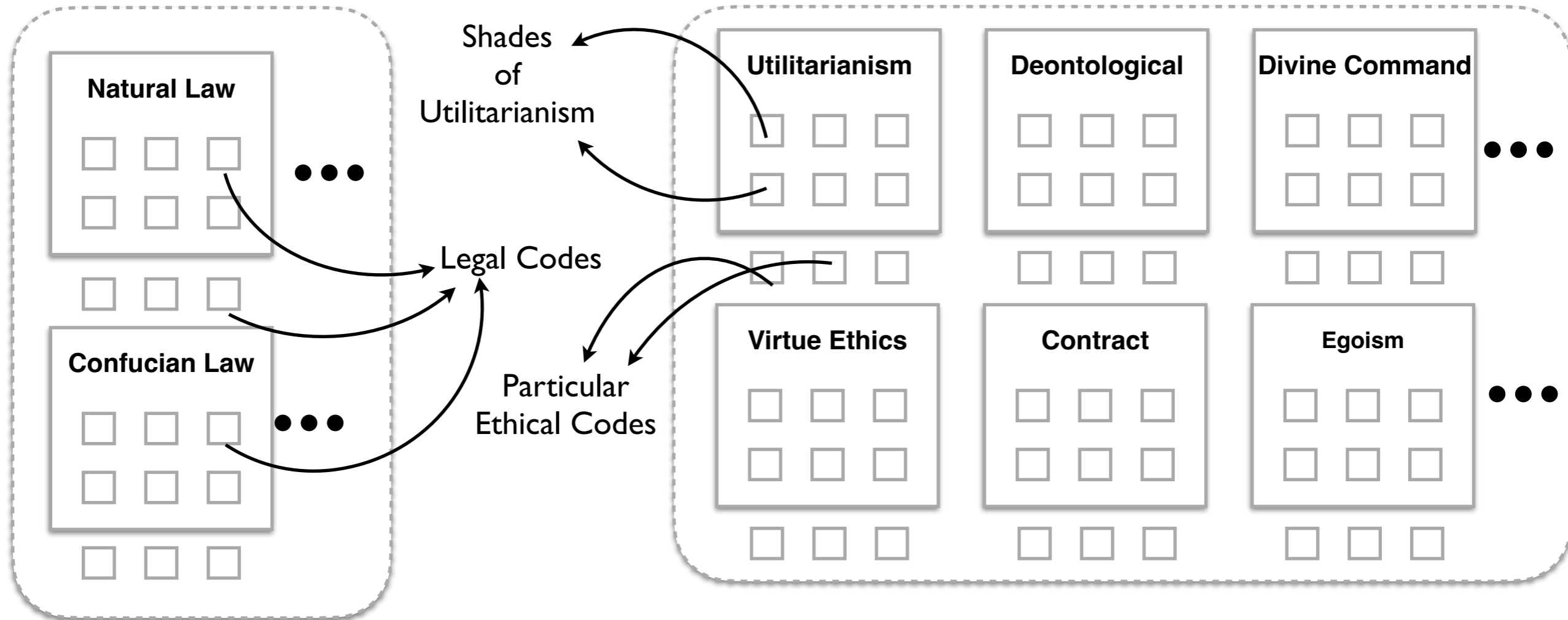


# Making Morally X Machines

**\$IIM**

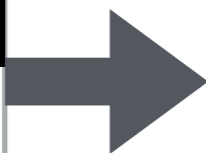
## Theories of Law

## Ethical Theories




**Step 1**


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



**Step 2**

Automate

 Prover

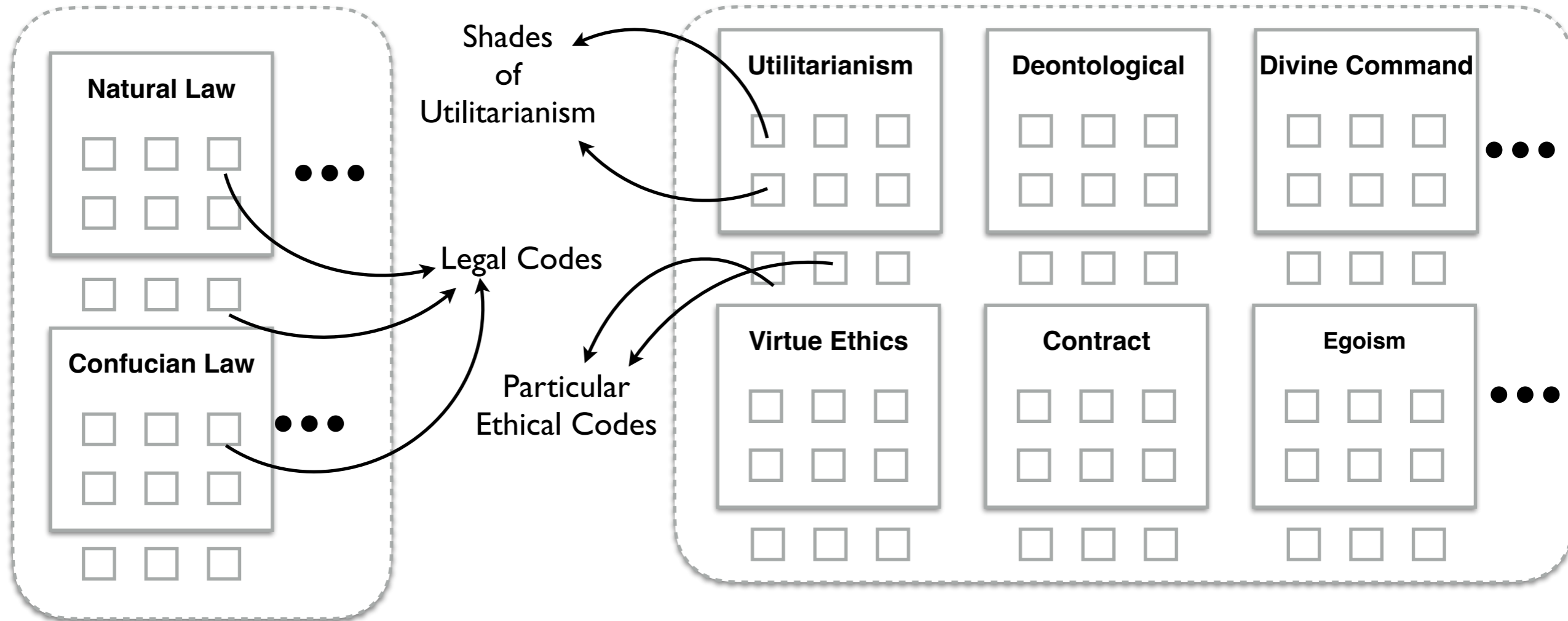
 Spectra

# Making Morally X Machines

**\$IIM**

## Theories of Law

## Ethical Theories



### Step 1

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

### Step 2

Automate

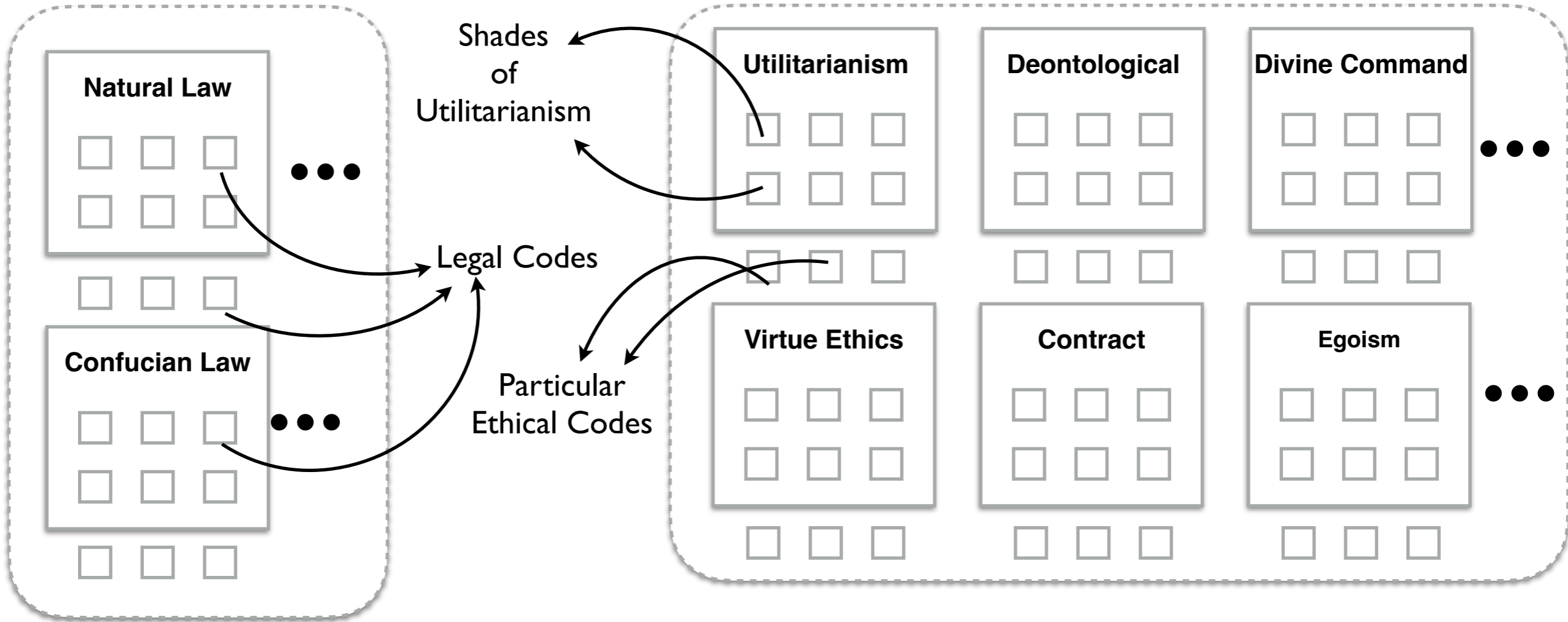


# Making Morally X Machines



## Theories of Law

## Ethical Theories



### Step 1

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

### Step 2

Automate

Prover

Spectra

### Step 3

Ethical OS

Ethical Substrate

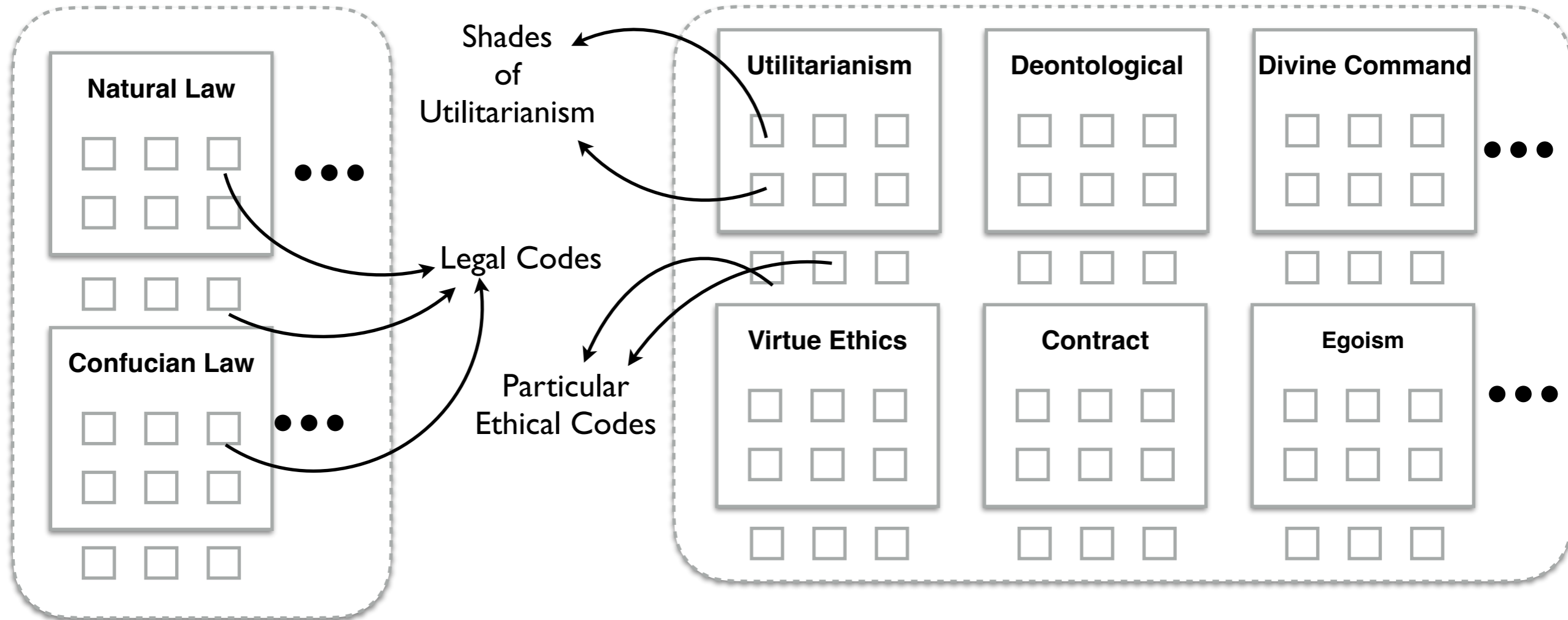
Robotic Substrate

# Making Morally X Machines

**\$IIM**

## Theories of Law

## Ethical Theories





### 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

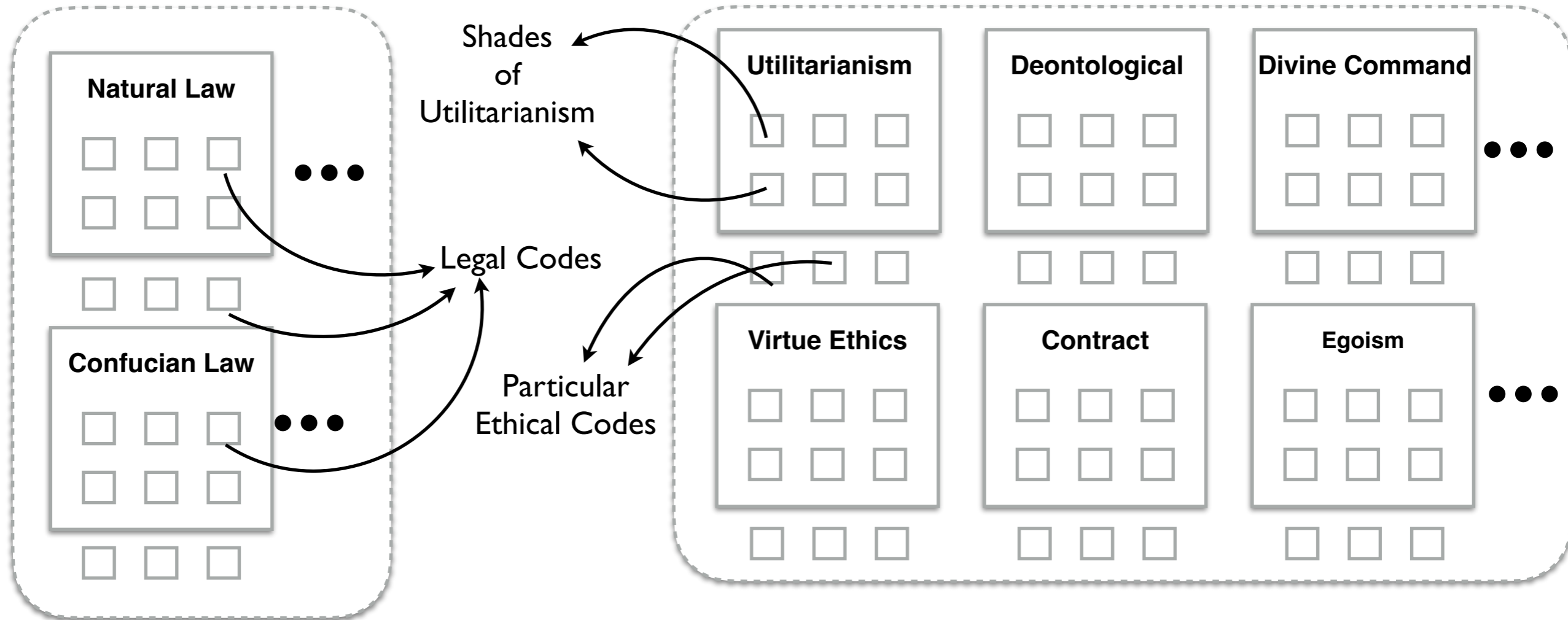
Robotic Substrate

# Making Morally X Machines

**\$IIM**

## Theories of Law

## Ethical Theories





### Step 1

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

### Step 2

Automate

 Prover

 Spectra

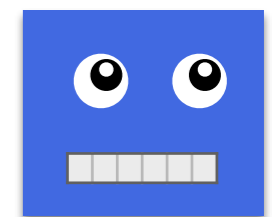
### Step 3

Ethical OS



Ethical Substrate

Robotic Substrate

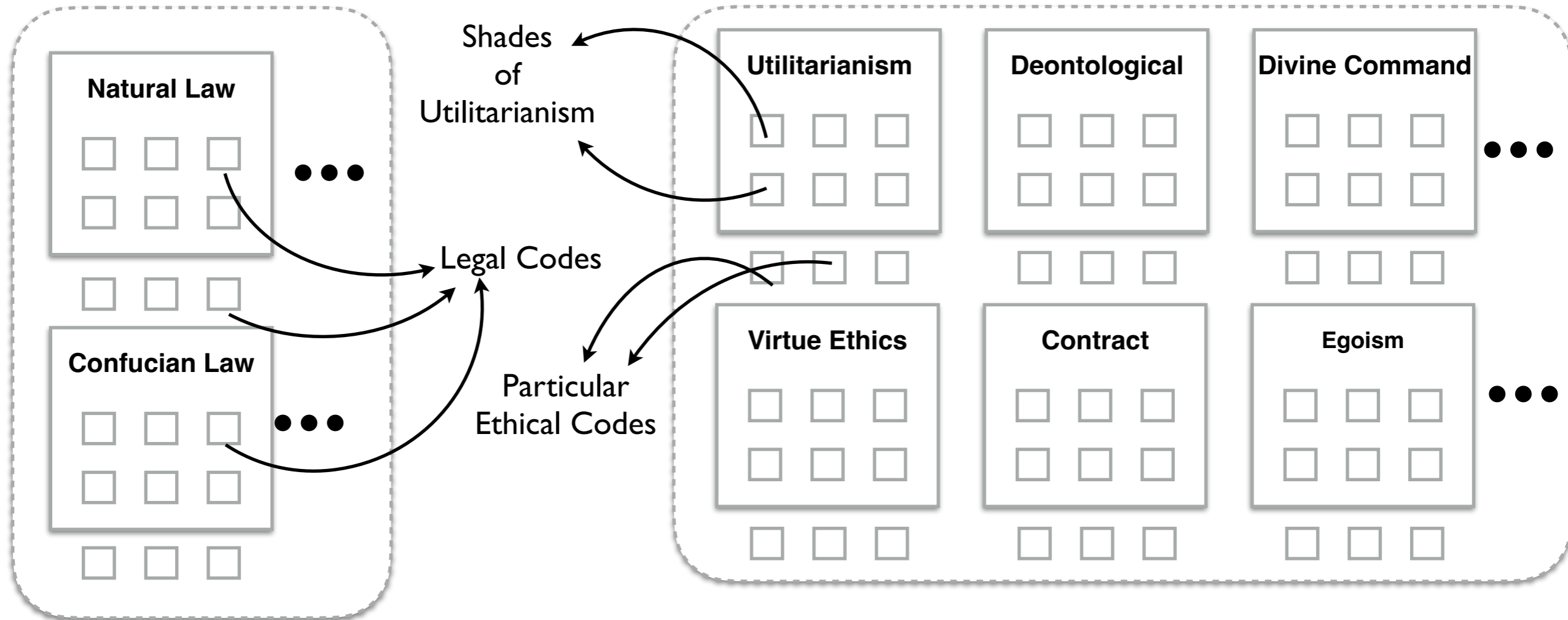


# Making Morally X Machines

**\$IIM**

## Theories of Law

## Ethical Theories



### Step 1

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

### Step 2

Automate

Prover

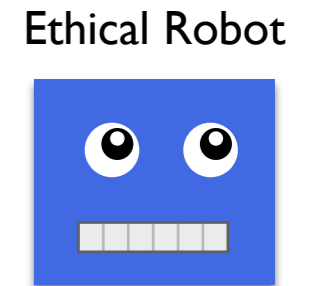
Spectra

### Step 3

Ethical OS

Ethical Substrate

Robotic Substrate

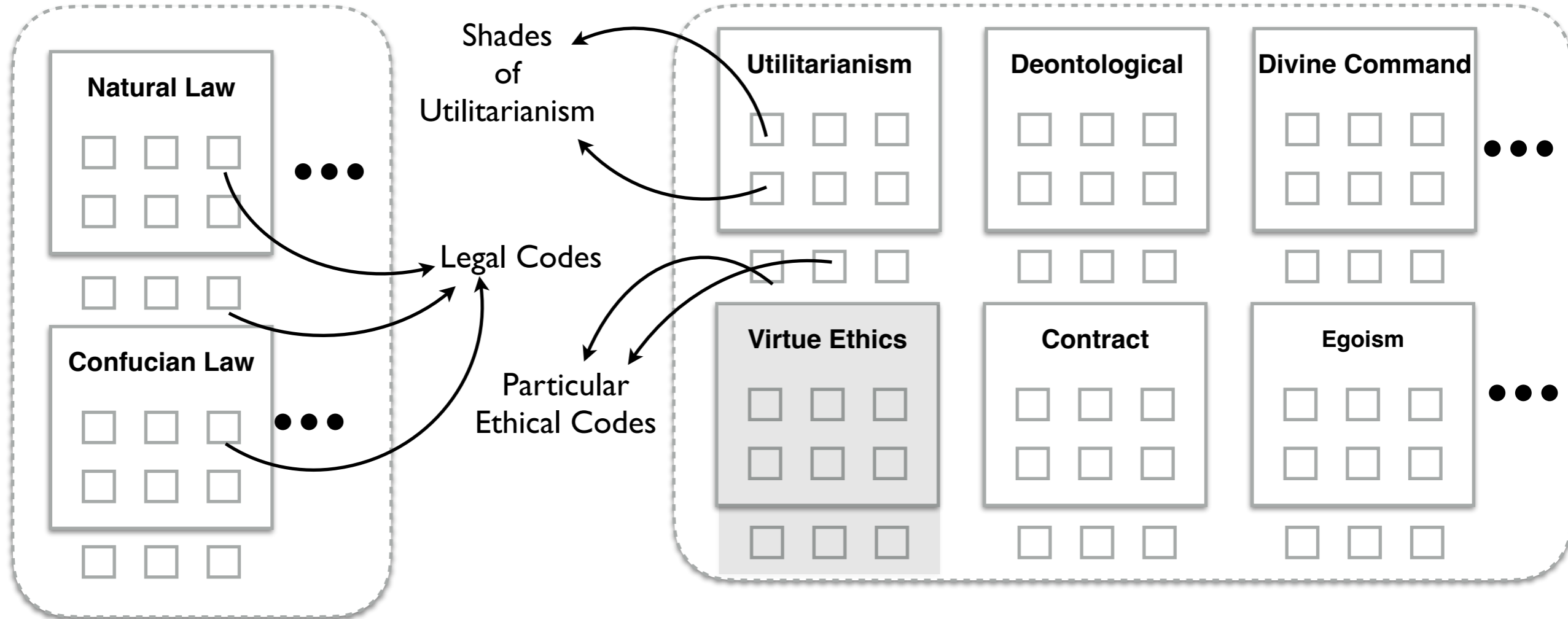


# Making Morally X Machines

**\$IIM**

## Theories of Law

## Ethical Theories





### 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



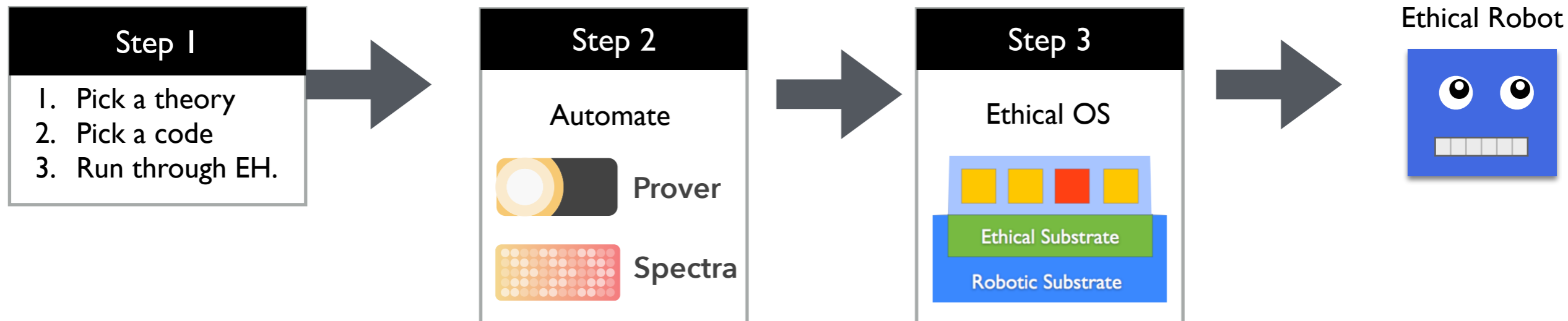
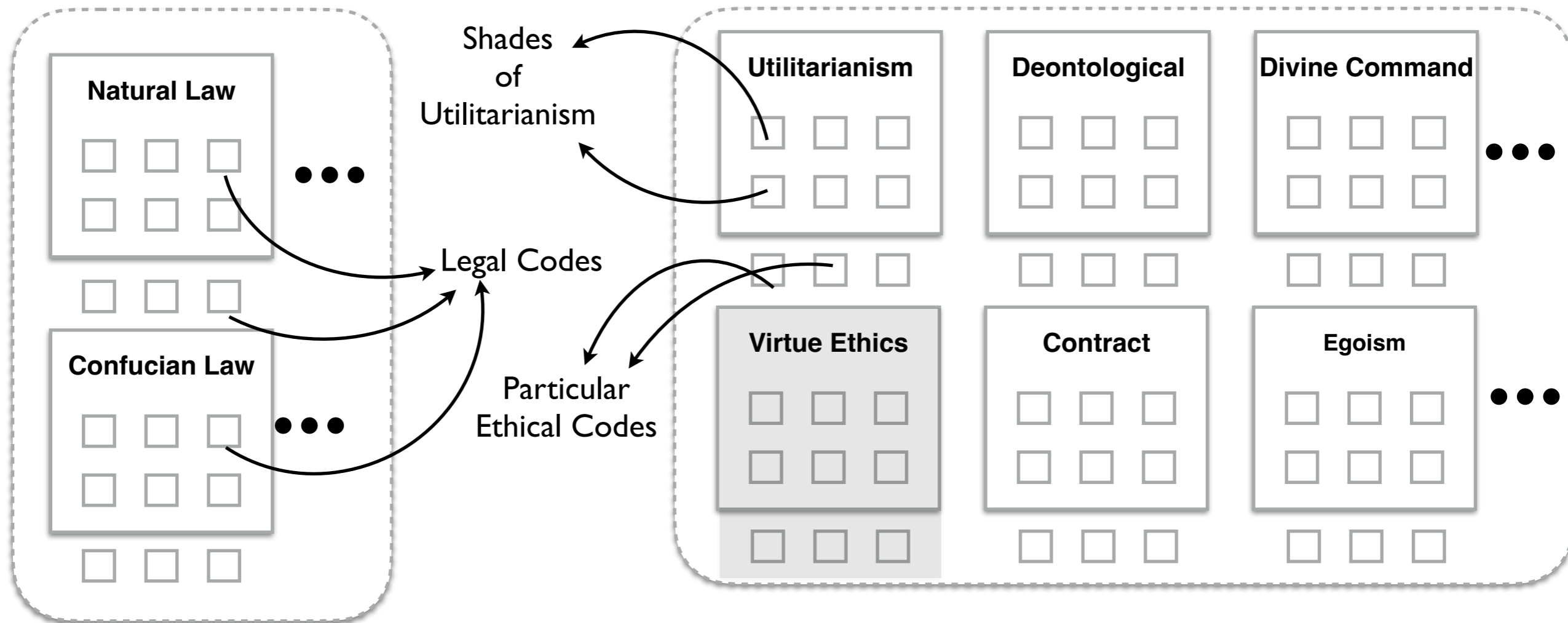


# Making Morally X Machines

**\$IIM**

Theories of Law

Ethical Theories



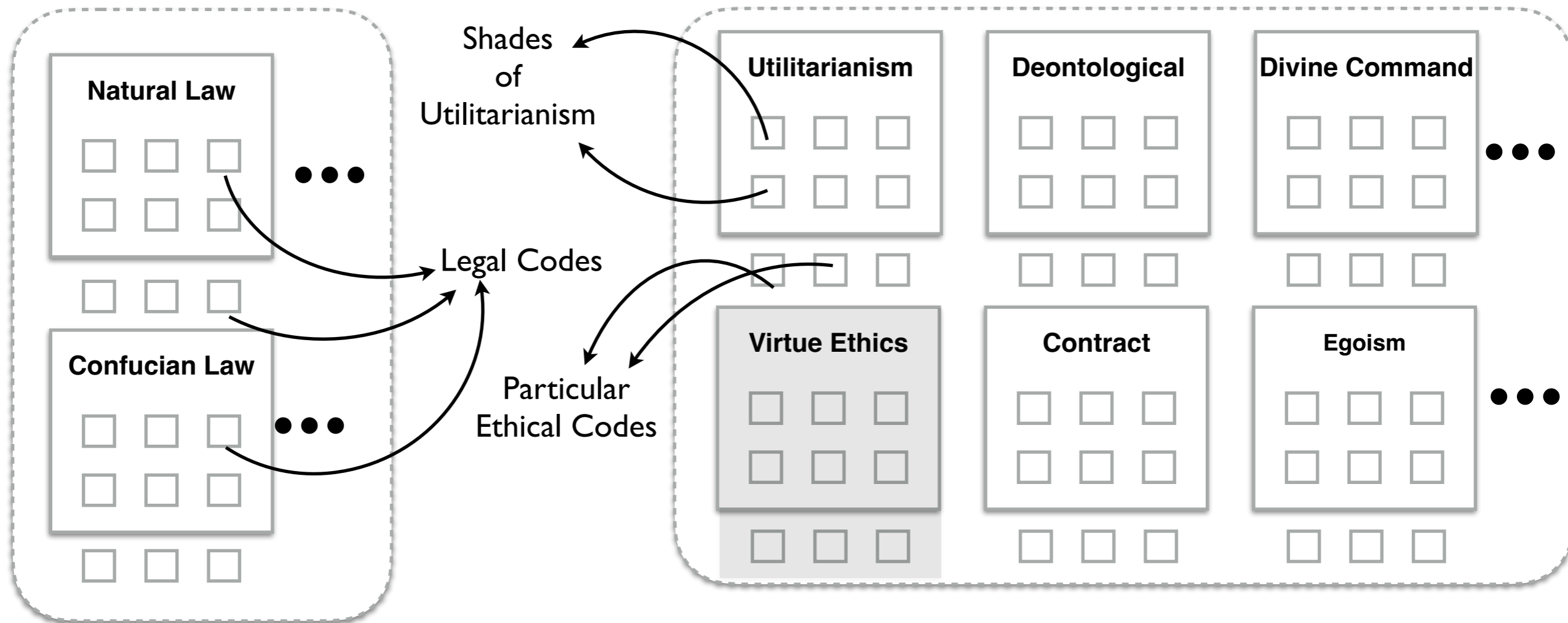
“Toward the Engineering of Virtuous Robots” Naveen, Selmer et al.

# Making Morally X Machines

**\$IIM**

## Theories of Law

## Ethical Theories





**Step 1**

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

**Step 2**

Automate

 Prover

 Spectra

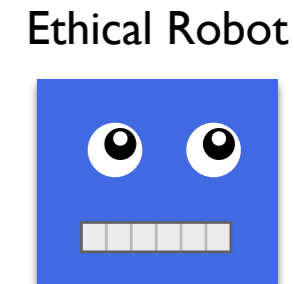
**Step 3**

Ethical OS



Ethical Substrate

Robotic Substrate



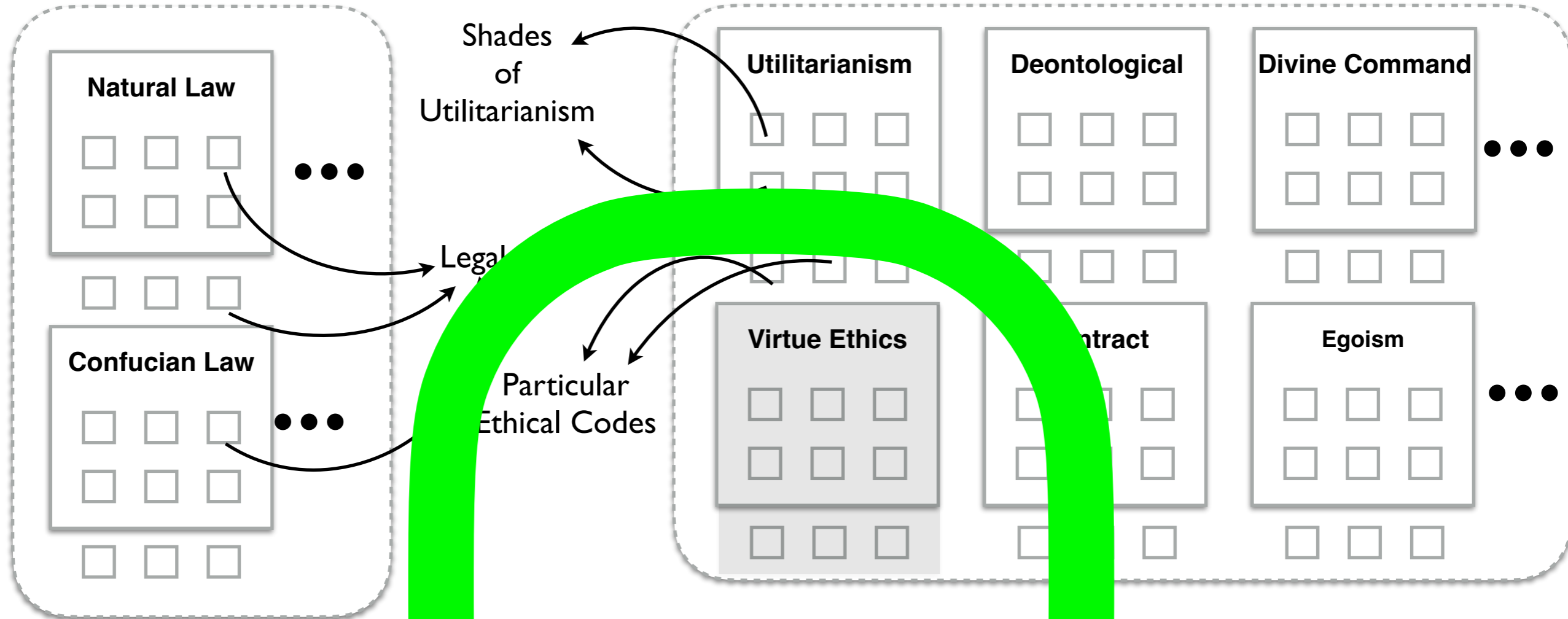
“Toward the Engineering of Virtuous Robots” Naveen, Selmer et al.

# Making Morally X Machines

**\$IIM**

Theories of Law

Ethical Theories



**Step 1**

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

**Step 2**

Automate

Prover

**Step 3**

Ethical

Substrate

Robotic Substrate



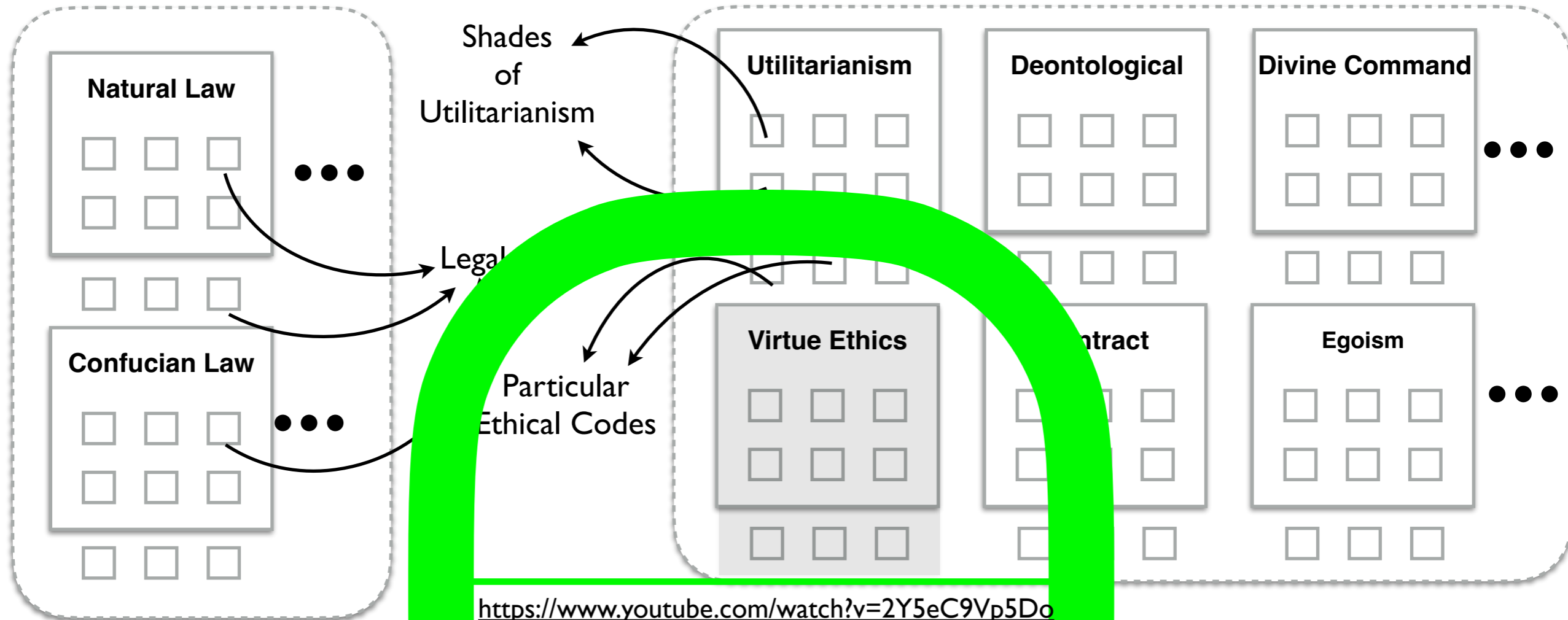
“Toward the Engineering of Virtuous Robots” Naveen, Selmer et al.

# Making Morally X Machines

**\$IIM**

Theories of Law

Ethical Theories



**Step 1**

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

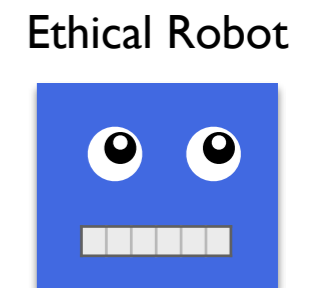
**Step 2**

Automate

Prover

**Step 3**

Ethical



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

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

# 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 model.

**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 shows three overlapping windows from the Slate application, each displaying a grid of modal logic formulas and their validity in various 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 derivability 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

**K**

**T**

**D**

**4 = S4**

**5 = S5**

The image shows five Slate windows, each displaying a grid of modal logic formulas and their derivability in different systems. The windows are titled as follows:

- Slate - K.slt**: Shows formulas K, T, 4, and 5. K is derivable in K (K ⊢ ✓ ∞ □). T, 4, and 5 are not derivable in K (K ⊢ ✗ ∞ □).
- Slate - T.slt**: Shows formulas K, T, 4, and 5. K and T are derivable in M (M ⊢ ✓ ∞ □). 4 and 5 are not derivable in M (M ⊢ ✗ ∞ □).
- Slate - D.slt**: Shows formulas K, T, D, 4, 5, and INTER. K, T, and 4 are not derivable in D (D ⊢ ✗ ∞ □). D and 5 are derivable in D (D ⊢ ✓ ∞ □). INTER is derivable in D (D ⊢ ✓ ∞ □).
- Slate - S4.slt**: Shows formulas K, T, D, 4, 5, and INTER. K, T, D, and 4 are derivable in S4 (S4 ⊢ ✓ ∞ □). 5 is not derivable in S4 (S4 ⊢ ✗ ∞ □). INTER is derivable in S4 with the assumption {INTER} (S4 ⊢ ✓ ∞ □).
- Slate - S5.slt**: Shows formulas K, T, D, 4, 5, and INTER. K, T, and 5 are derivable in S5 (S5 ⊢ ✓ ∞ □). D and 4 are not derivable in S5 (S5 ⊢ ✗ ∞ □). D and 4 are derivable in S5 with assumptions {D} and {4} respectively (S5 ⊢ ✓ ∞ □). INTER is derivable in S5 with the assumption {INTER} (S5 ⊢ ✓ ∞ □).

# Review: Encapsulation

**K**

**T**

**D**

**4 = S4**

**5 = S5**

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

# Review: Encapsulation

**K**

**T**

**D**

The screenshot displays five windows of the HyperSlate interface, each showing a set of logical formulas and their derivability status in a specific modal logic. 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:** (This window is highlighted with a red border)
  - 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  $\checkmark$ )

**4 = S4**

**5 = 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 to modern deontic logic would be quite hard to do. 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  $\vdash \odot\phi \leftrightarrow \odot\psi$ .

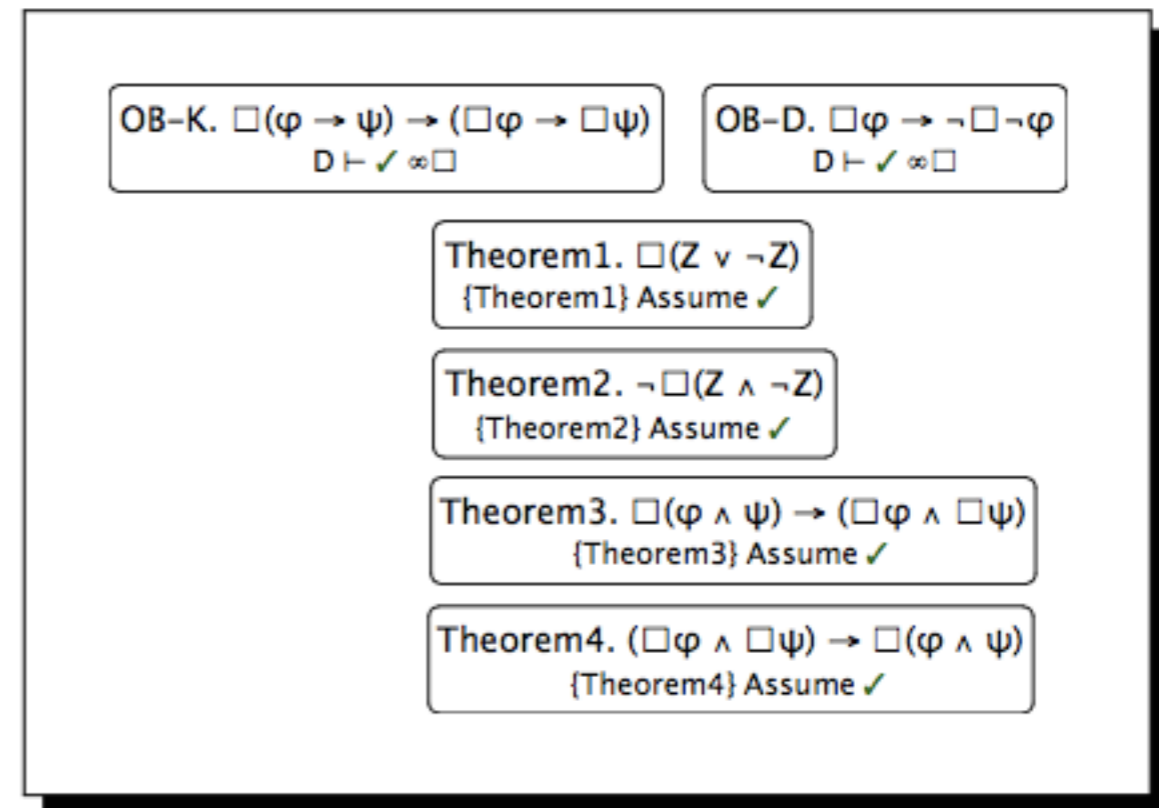
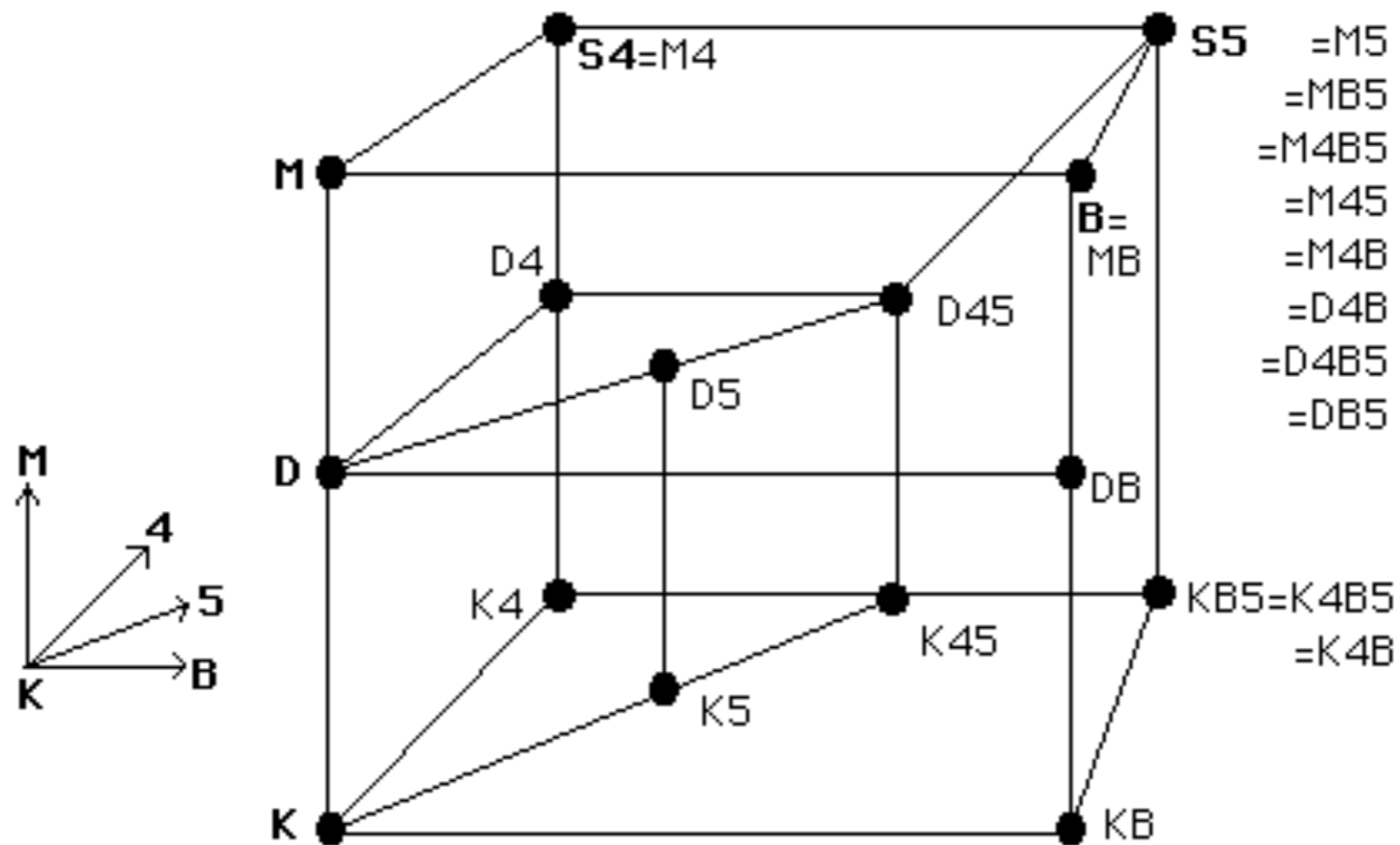


Figure 4.7: The Initial Configuration Upon Opening the File `SDL.slt`





#### 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.

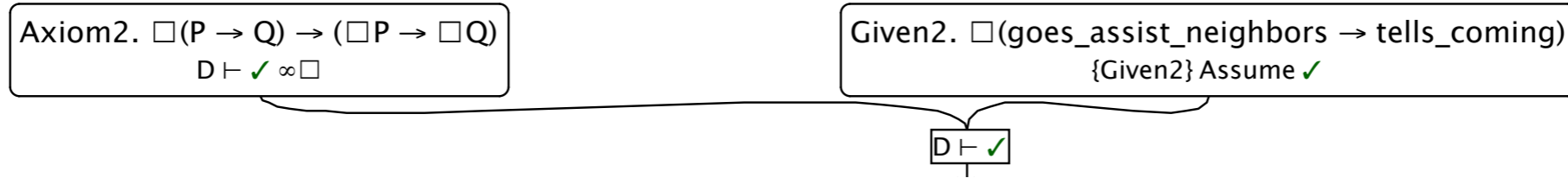
Chisholm's clever scenario revolves around the character Jones.<sup>11</sup> It's given that Jones is obligated to go to assist his neighbors, in part because he has promised to do so. The second given fact is that it's obligatory that, if Jones goes to assist his neighbors, he tells them (in advance) 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

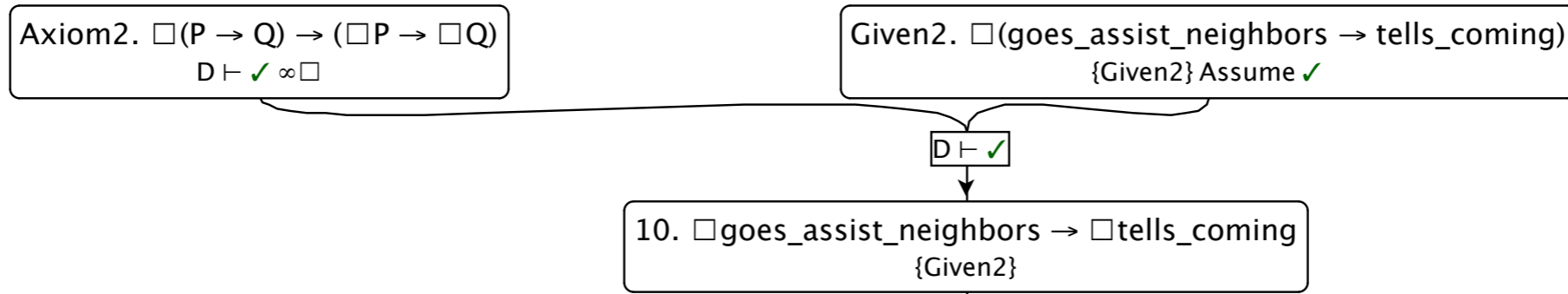


Axiom4. "Modus ponens for provability."  
{Axiom4} Assume ✓

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

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

# Chisholm's Paradox

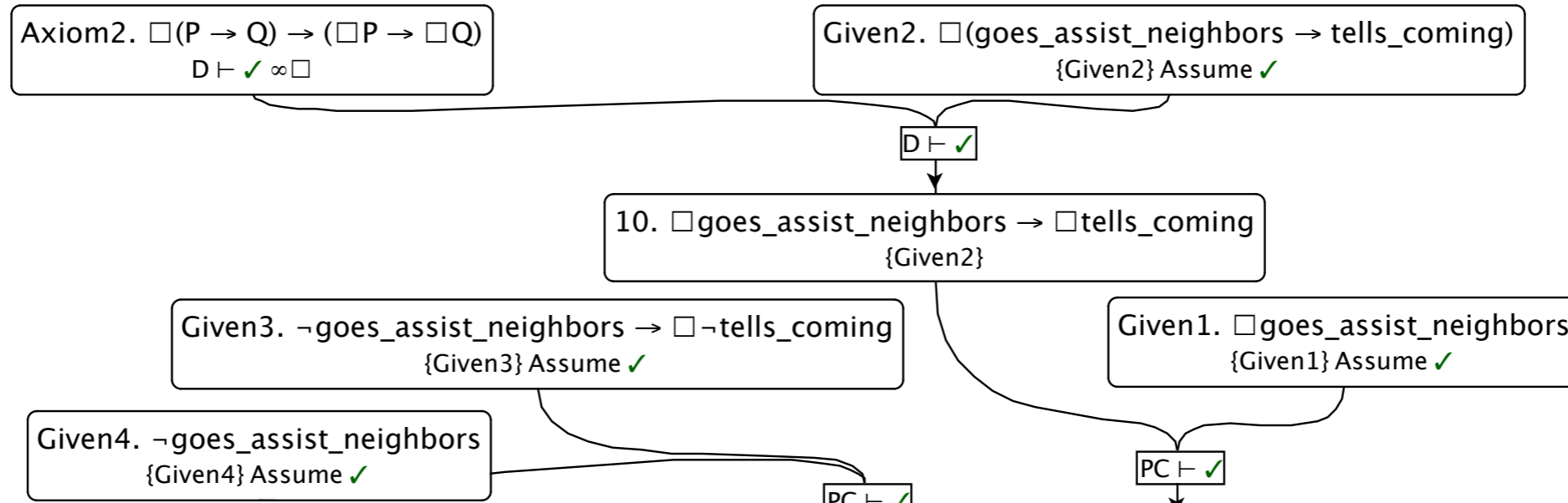


Axiom4. "Modus ponens for provability."  
{Axiom4} Assume  $\checkmark$

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

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

# Chisholm's Paradox

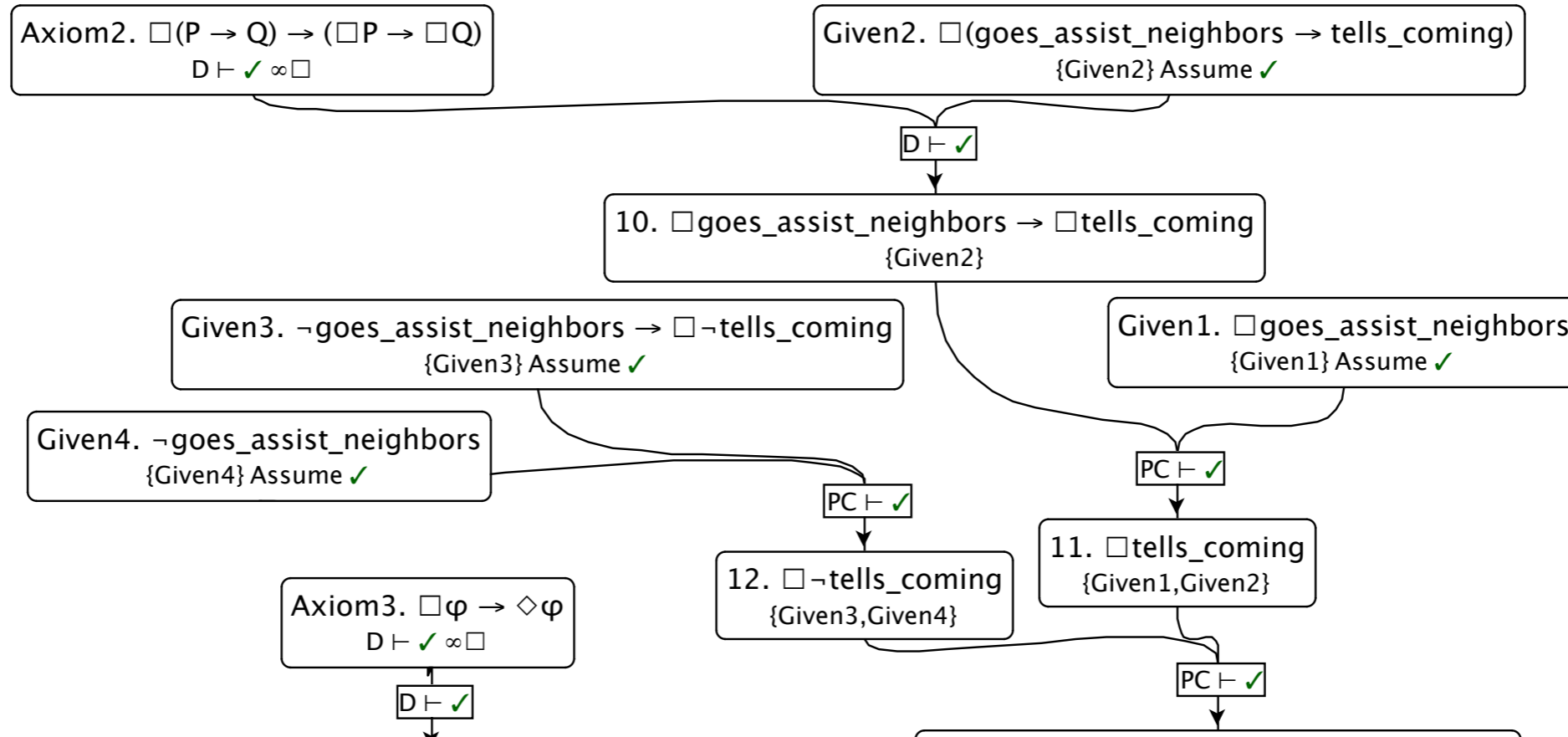


Axiom4. "Modus ponens for provability."  
 {Axiom4} Assume ✓

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

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

# Chisholm's Paradox

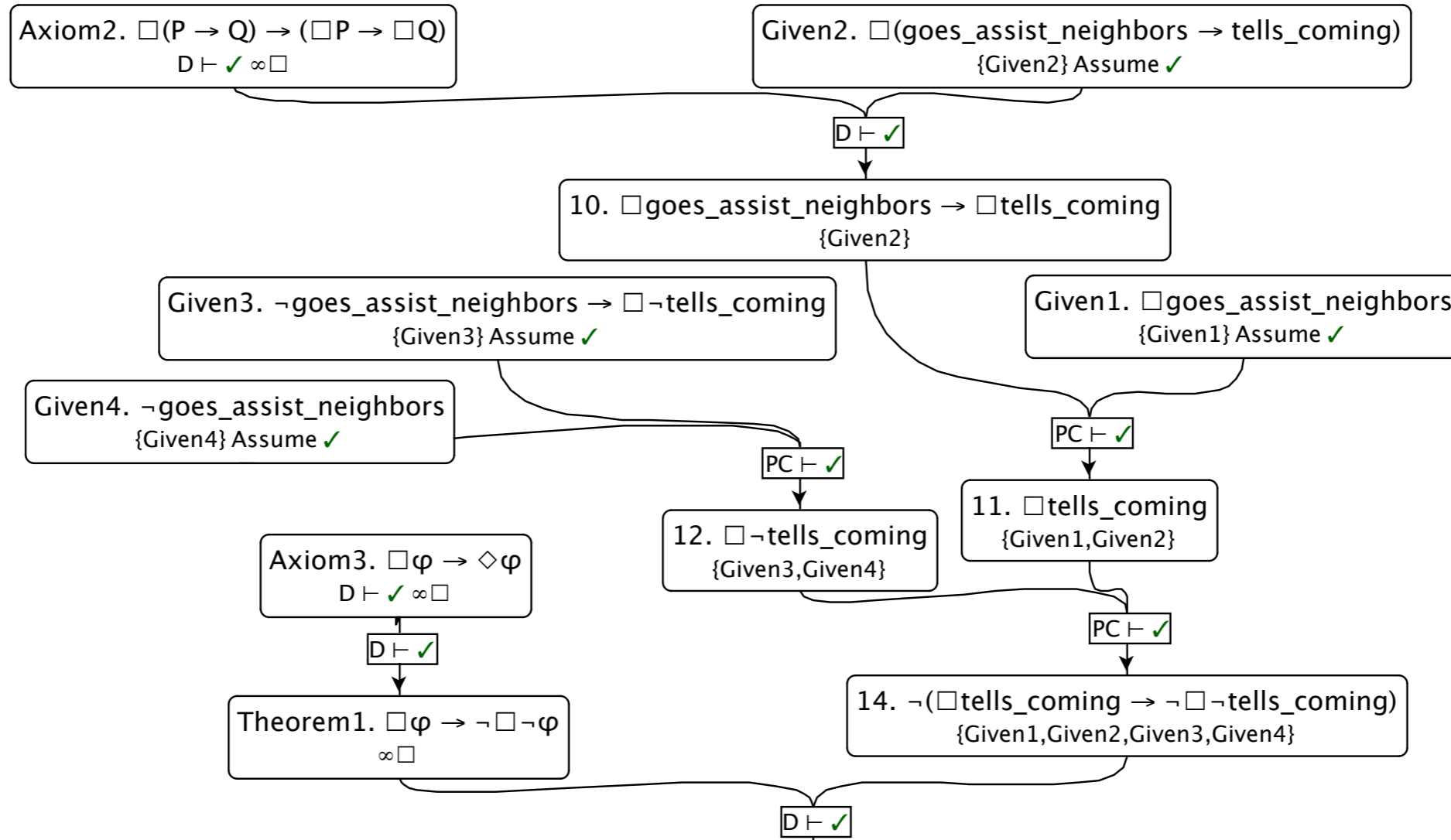


**Axiom4.** "Modus ponens for provability."  
 $\{\text{Axiom4}\} \text{Assume } \checkmark$

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

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

# Chisholm's Paradox



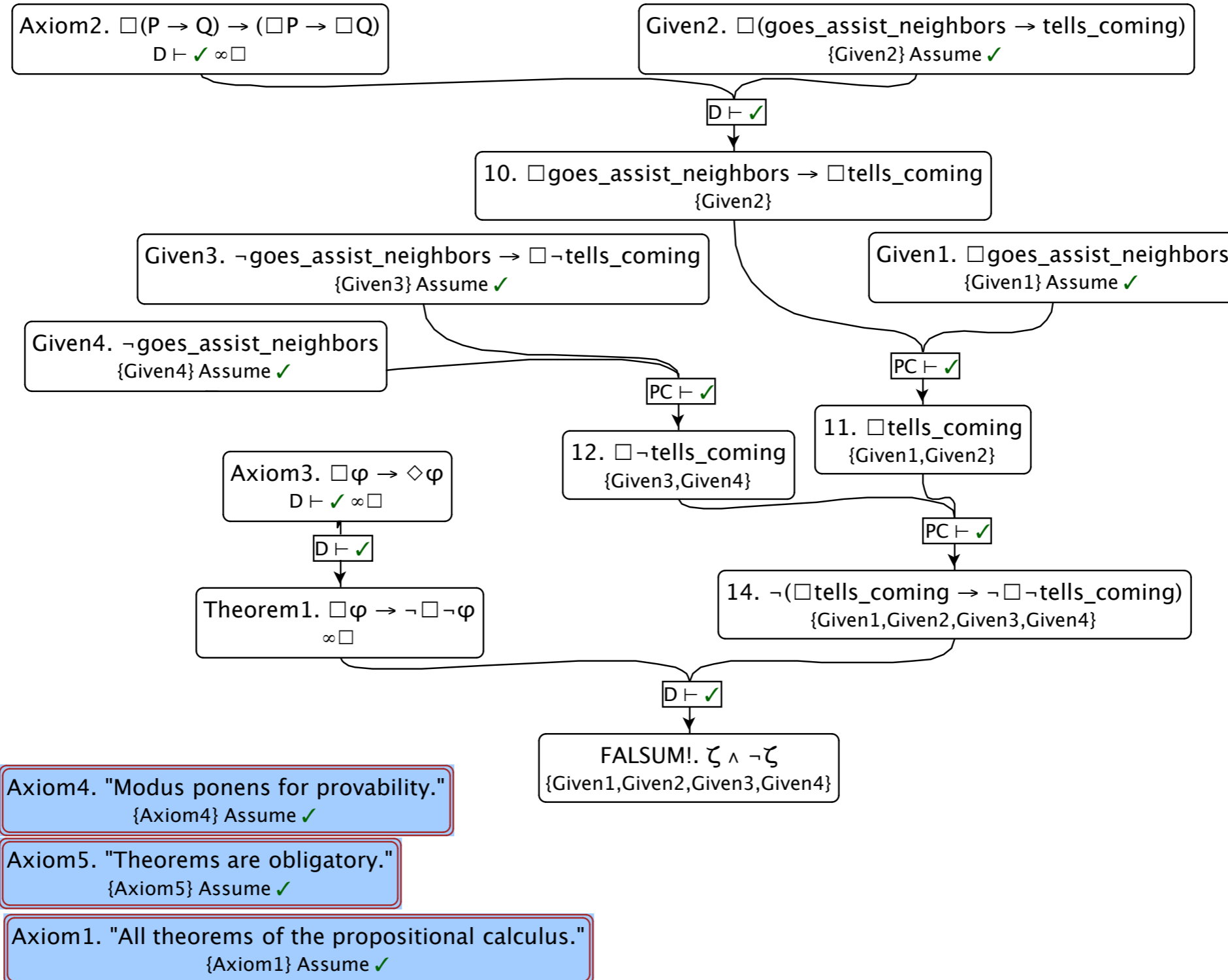
Axiom4. "Modus ponens for provability."  
 $\{\text{Axiom4}\} \text{Assume } \checkmark$

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

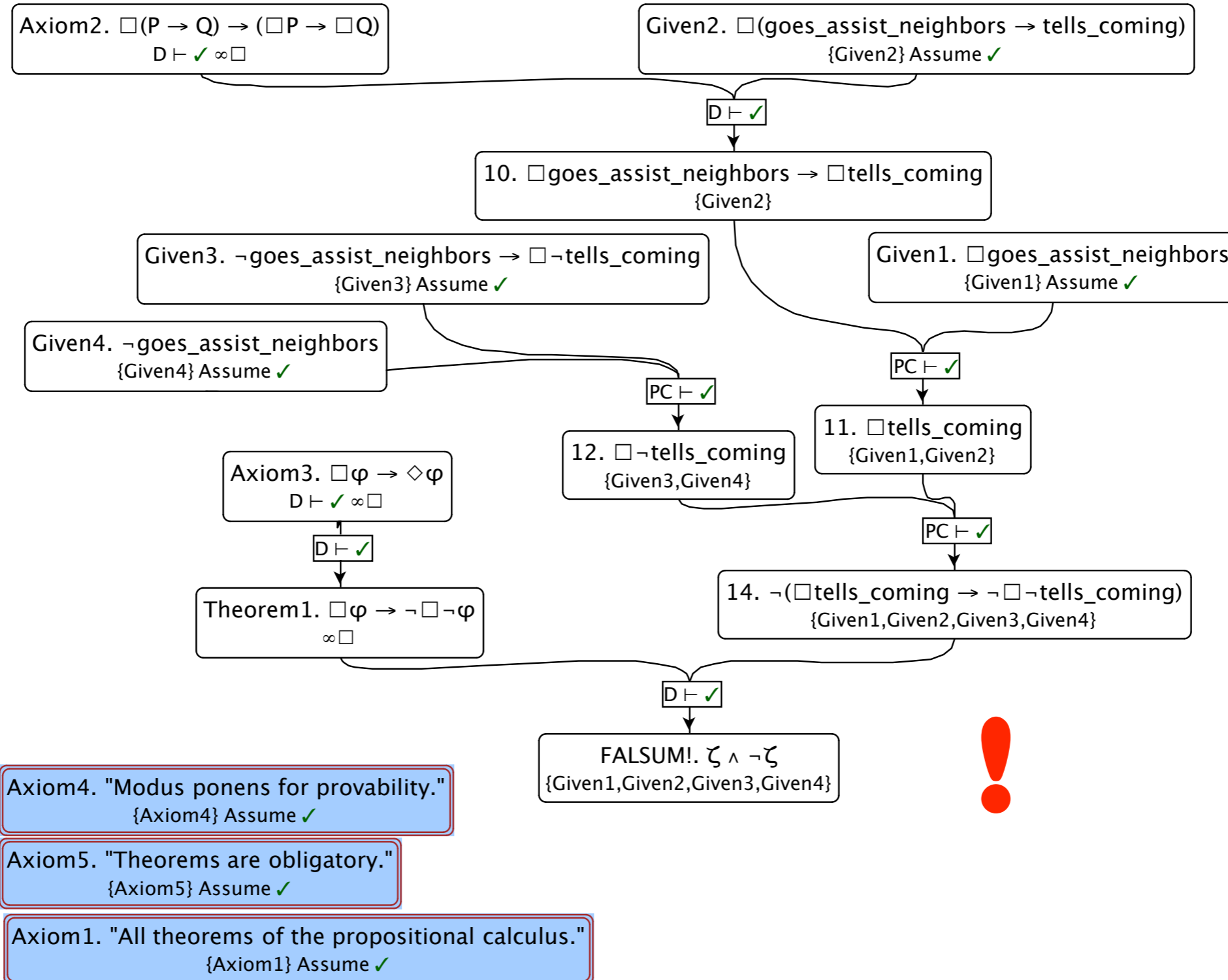
Axiom1. "All theorems of the propositional calculus."  
 $\{\text{Axiom1}\} \text{Assume } \checkmark$



# 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 ✓

$D \vdash \times$

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

(You should do it.)

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 ✓

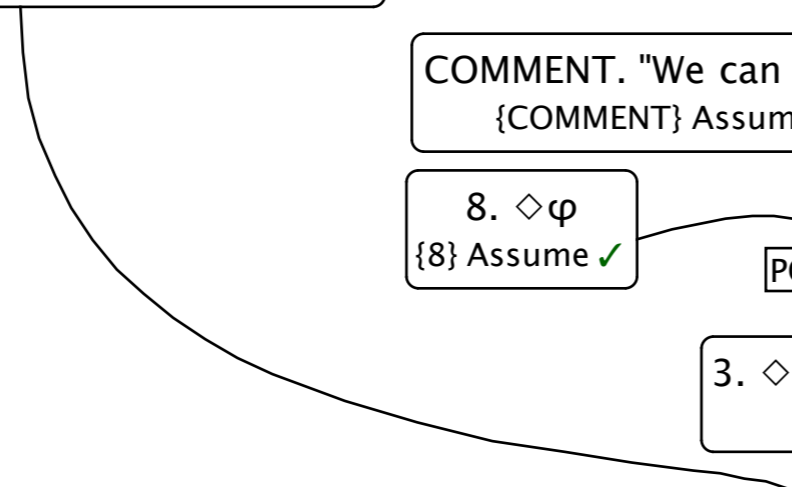
THM 5.  $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$   
 $D \vdash \checkmark \infty \square$

(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?)

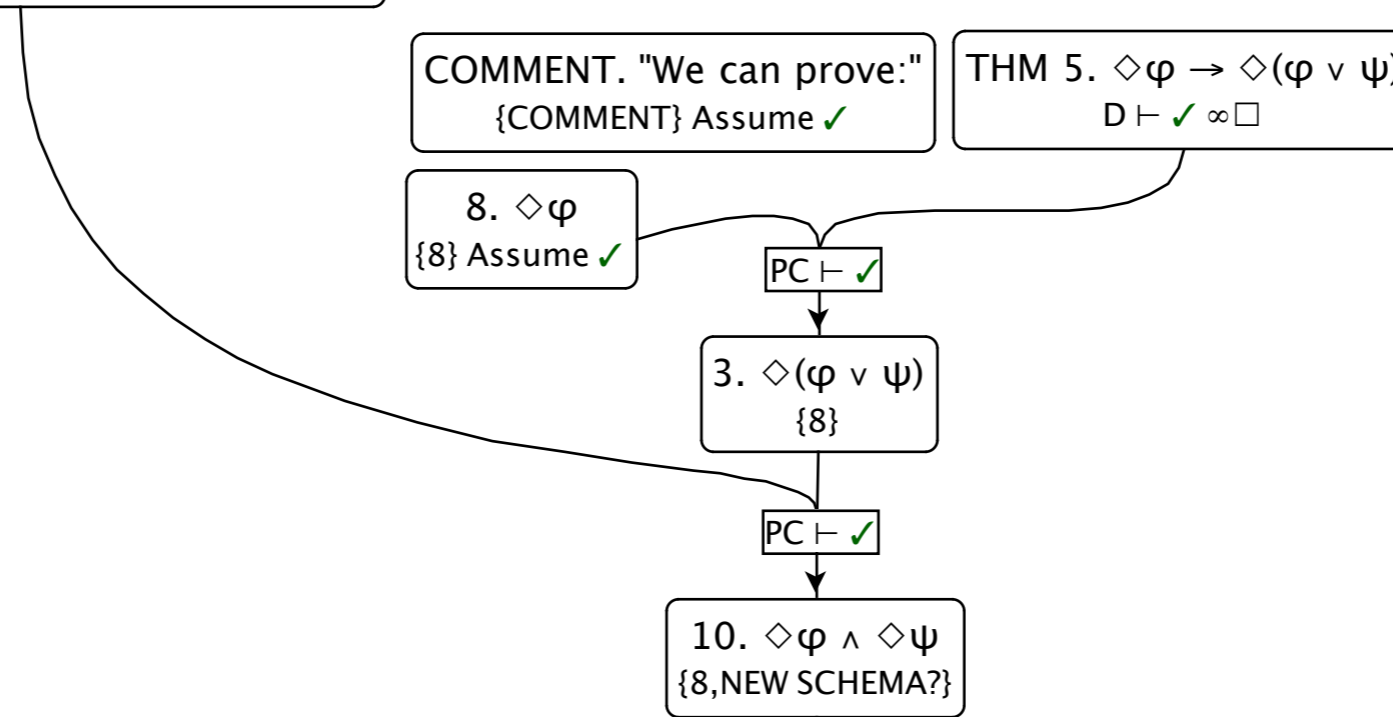
8.  $\diamond\varphi$   
 {8} Assume ✓

$PC \vdash \checkmark$

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 ✓

$PC \vdash \checkmark$

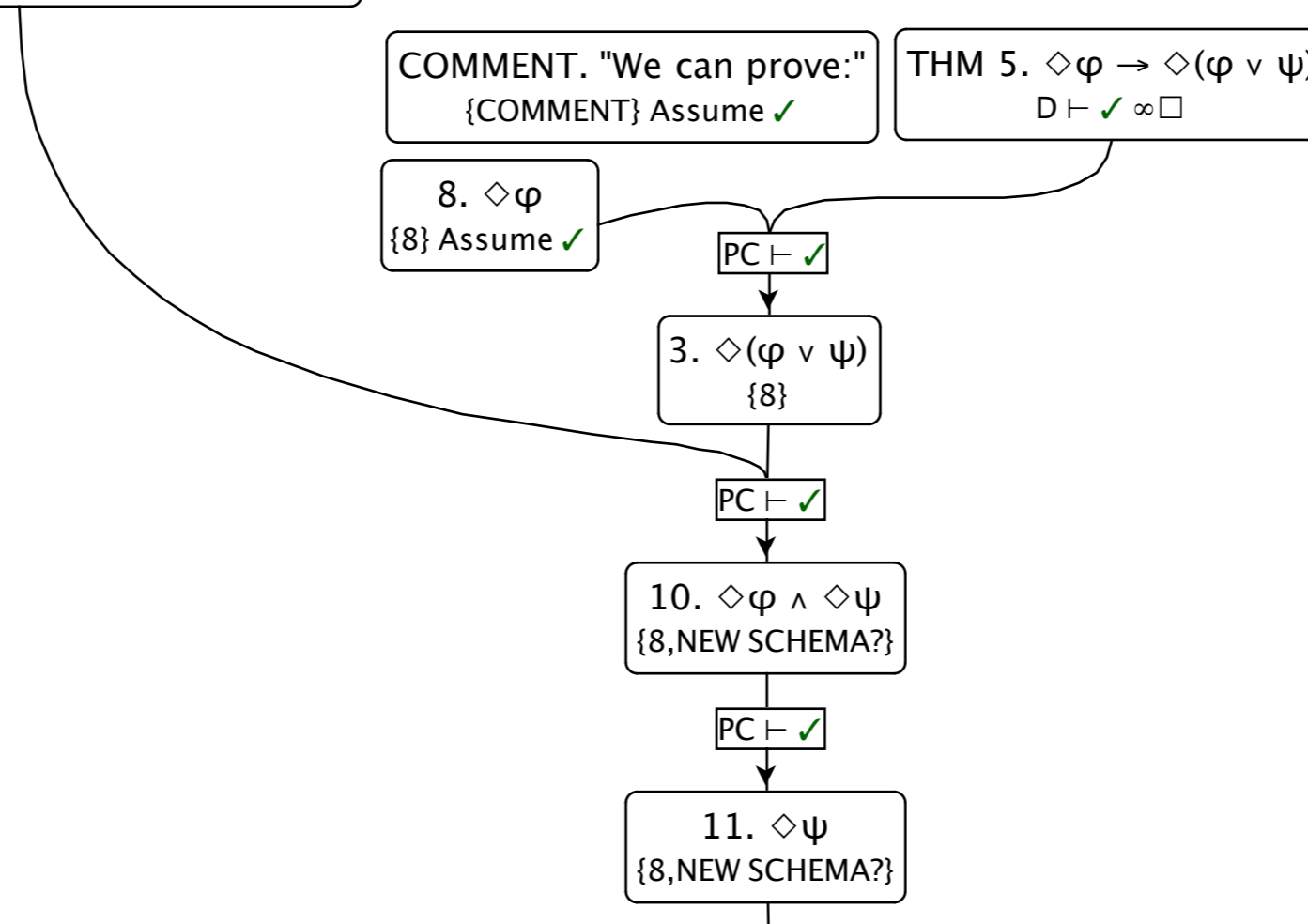
3.  $\diamond(\varphi \vee \psi)$   
 {8}

$PC \vdash \checkmark$

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?}

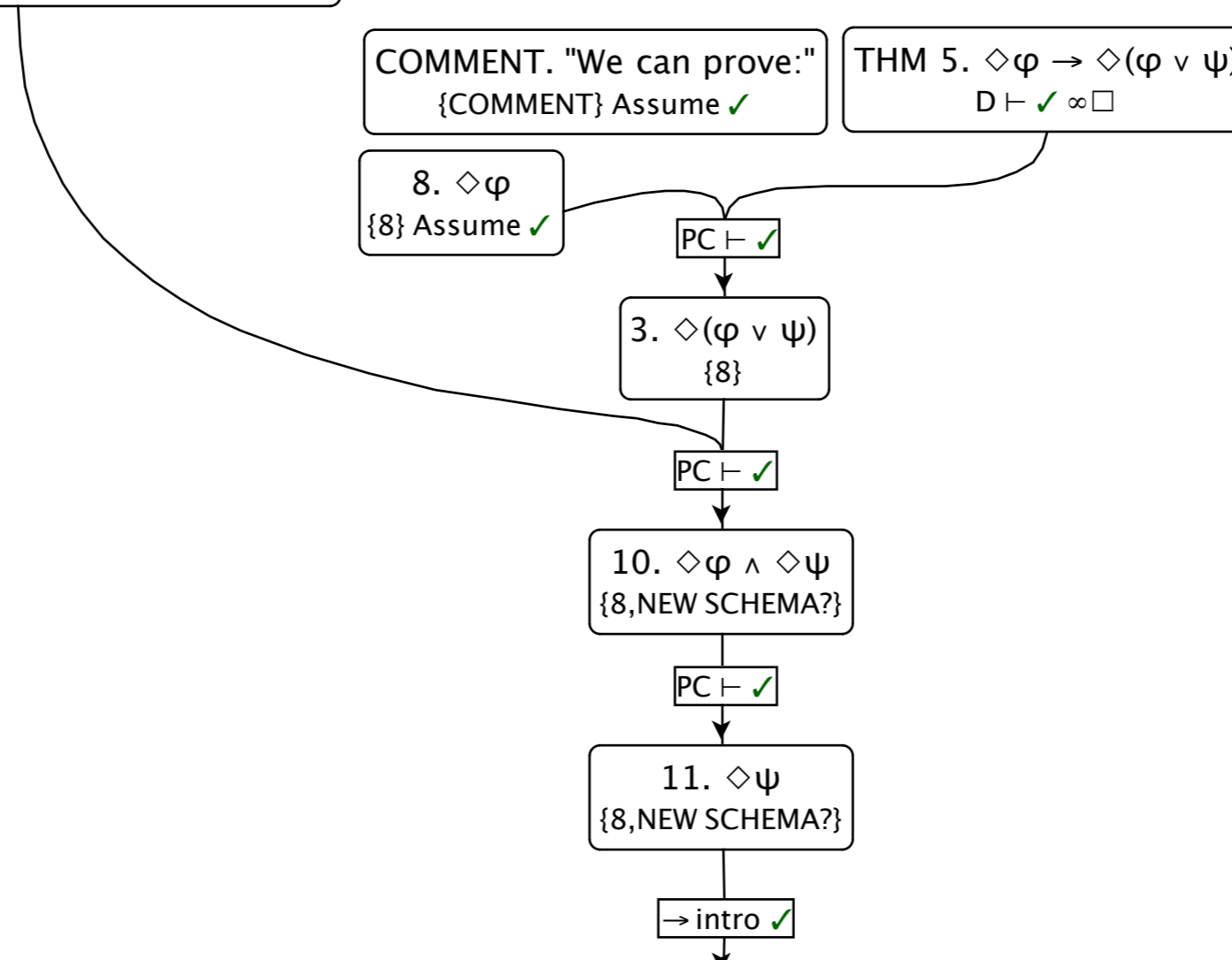
$\text{PC} \vdash \checkmark$

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

$\rightarrow$  intro ✓

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

COMMENT. Absurd!  
 {COMMENT} Assume ✓



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 *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')**

$\diamond(\textit{sofabed} \vee \textit{guestbed})$

**(2')**

$\diamond\textit{sofabed} \wedge \diamond\textit{guestbed}$

But (2') doesn't follow deductively from (1') in **D**, as a call to the provability oracle for **D** in the HyperSlate™ 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 permissible, everything is!

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

**And, Ross' Paradox in  
HyperSlate® now ...**





“So, computational logician,  
sorry, back to your drawing  
board to find a logic that *does*  
work with The Four Steps!”

