The Pentagon vs Anthropic On An Ethically Correct (& Conscious) Claude

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Lally School of Management
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)

Troy NY USA

Mar 19 2026 in *AI Alignment*
Mar 30 2026
Intro to Formal Logic (*With AI*)

v 0401260800NY

The Pentagon vs Anthropic On An Ethically Correct (& Conscious) Claude

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Lally School of Management
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)

Troy NY USA

Mar 19 2026 in *AI Alignment*
Mar 302026
Intro to Formal Logic (*With AI*)

v 0330261030NY

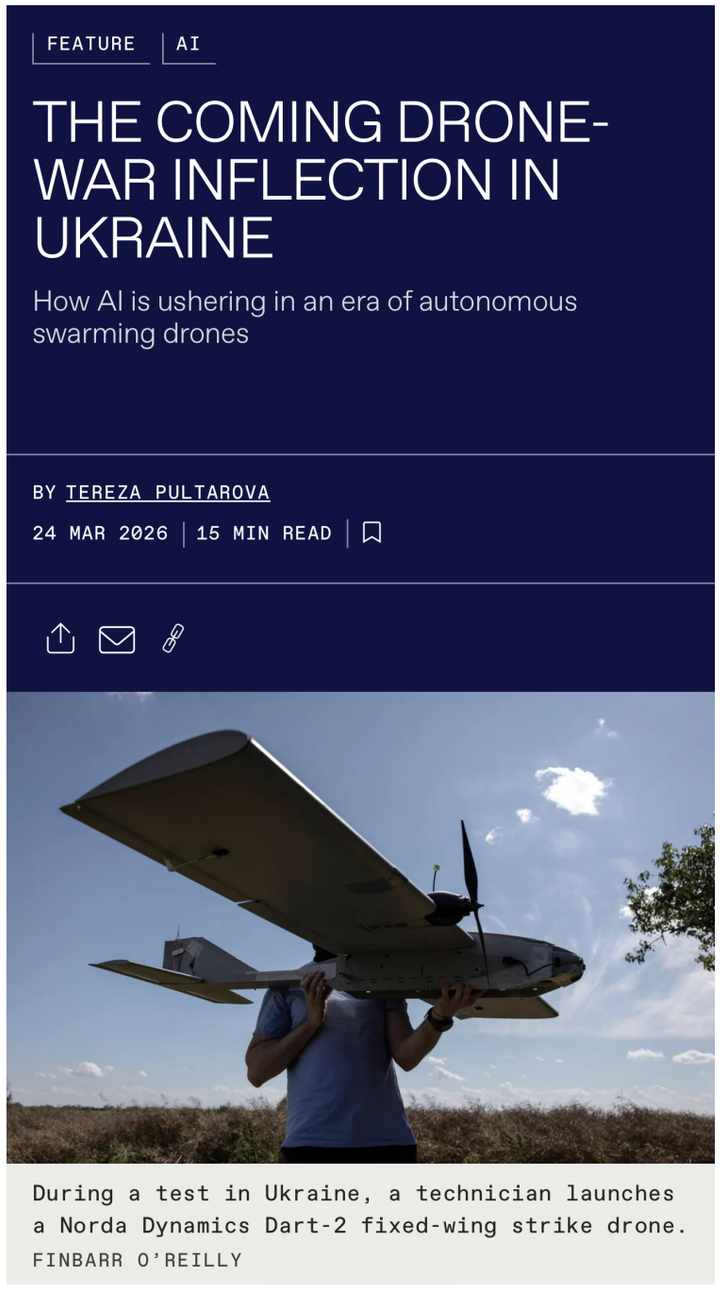
The Pentagon vs Anthropic On An Ethically Correct (& Conscious) Claude

Selmer Bringsjord

Rensselaer AI & Reasoning (RAIR) Lab
Lally School of Management
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)

Troy NY USA

Mar 19 2026 in *AI Alignment*
Mar 302026
Intro to Formal Logic (*With AI*)



FEATURE | AI

THE COMING DRONE-WAR INFLECTION IN UKRAINE

How AI is ushering in an era of autonomous swarming drones

BY TEREZA PULTAROVA

24 MAR 2026 | 15 MIN READ

During a test in Ukraine, a technician launches a Norda Dynamics Dart-2 fixed-wing strike drone.
FINBARR O'REILLY

v 0330261030NY

Experiment
(Brachman-Levesque CHI:CHR)

Experiment
(Brachman-Levesque CHI:CHR)

Get a blank piece of paper ready, & a pen/pencil.

Experiment
(Brachman-Levesque CHI:CHR)

Get a blank piece of paper ready, & a pen/pencil.

Suppose there are some dogs in the backyard, and we have the following two facts:

1. All the dogs in the backyard are owned by Jim.
2. No small dogs are in the backyard.

Then which of the following three statements must also be true?

3. Some small dogs are not owned by Jim.
4. Some dogs owned by Jim are not small.
5. None of the dogs owned by Jim are small.

Experiment
(Brachman-Levesque CHI:CHR)

Get a blank piece of paper ready, & a pen/pencil.

Suppose there are some dogs in the backyard, and we have the following two facts:

1. All the dogs in the backyard are owned by Jim.
2. No small dogs are in the backyard.

Then which of the following three statements must also be true?

3. Some small dogs are not owned by Jim.
4. Some dogs owned by Jim are not small.
5. None of the dogs owned by Jim are small.

(a) Write down one of: 3., 4., 5., none of these.

Experiment
(Brachman-Levesque CHI:CHR)

Get a blank piece of paper ready, & a pen/pencil.

Suppose there are some dogs in the backyard, and we have the following two facts:

1. All the dogs in the backyard are owned by Jim.
2. No small dogs are in the backyard.

Then which of the following three statements must also be true?

3. Some small dogs are not owned by Jim.
4. Some dogs owned by Jim are not small.
5. None of the dogs owned by Jim are small.

(a) Write down one of: 3., 4., 5., none of these.

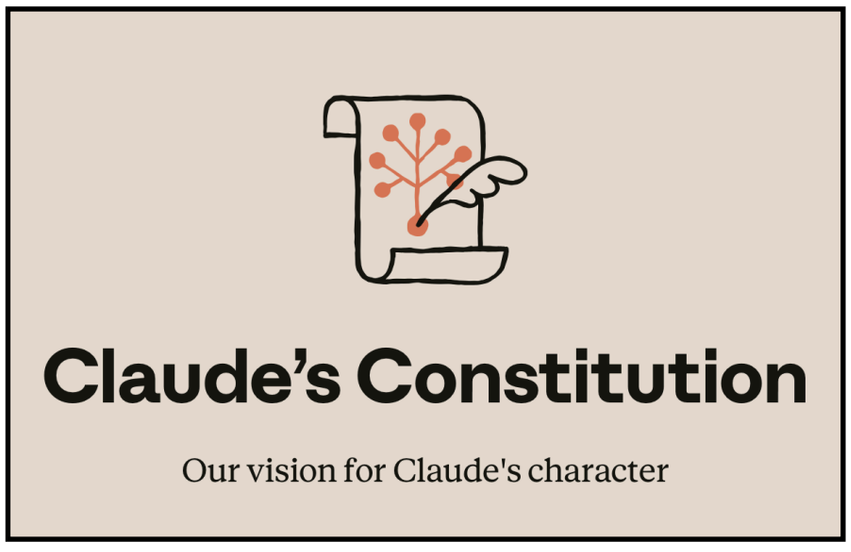
(b) Write down a proof to the best of your ability to justify your answer.

Defense Dept. and Anthropic Square Off in Dispute Over A.I. Safety

How artificial intelligence will be used in future battlefields is an issue that has turned increasingly political and may put Anthropic in a bind.

Listen · 7:15 min

A Grand-Canyon Sized Disconnect Between Pentagon & Anthropic: Fundamentally, What is An Autonomous AI *qua* Weapon?



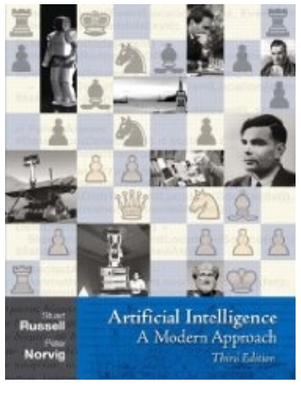
https://www.anthropic.com/constitution

“We’re in *very* deep trouble.”

“We’re in *very* deep trouble.”



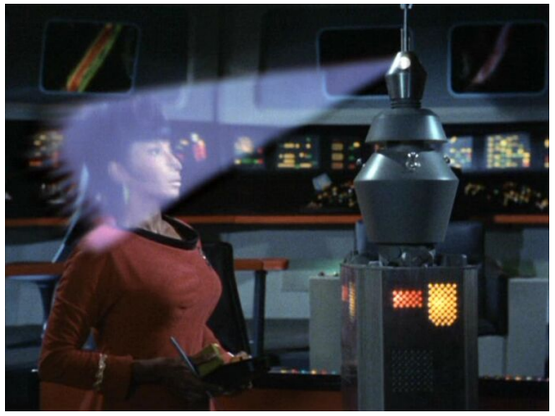

“We’re in *very* deep trouble.”



While the PAI machines aren't quite as easy to neutralize as the destructive machines vanquished in *Star Trek: TOS*, these relevant four episodes show the protective power of … logic.

"The Ultimate Computer"
S2 E24

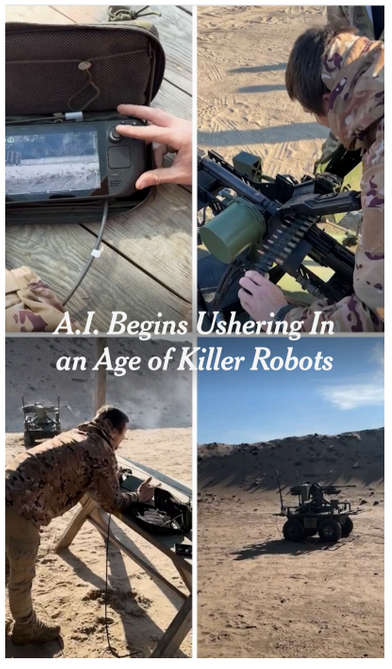
"The Return of the Archons"
S1 E21



"The Changeling"
S2 E3

"I, Mudd"
S2 E8

The PAID Problem

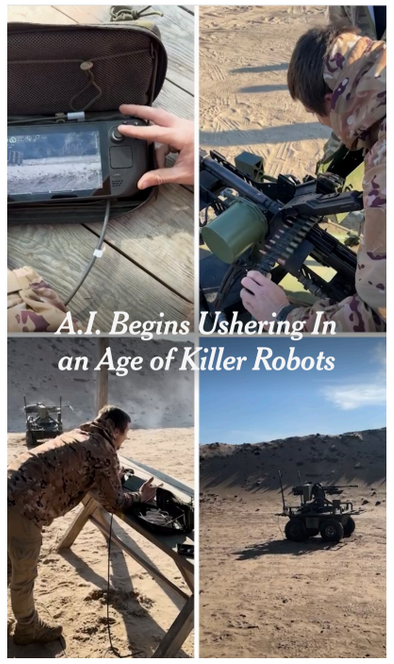


NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...

https://www.facebook.com/nhkworld/videos/1858412994205448/

Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director, Rensselaer Artificial Intelligence and ...

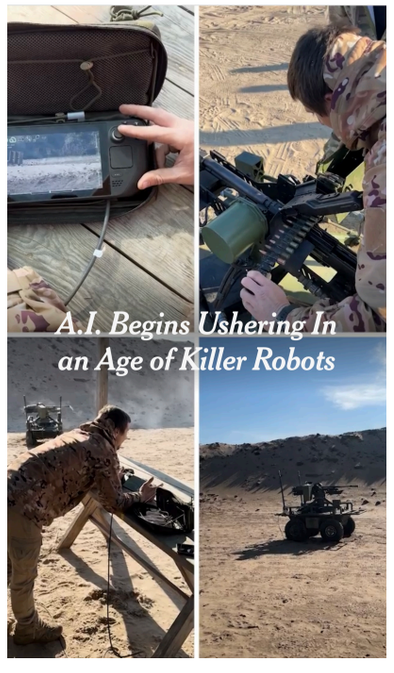
The PAID Problem



For all agents $\mathfrak{a}$:

NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...
https://www.facebook.com/nhkworld/videos/1858412994205448/
Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director, Rensselaer Artificial Intelligence and ...

The PAID Problem

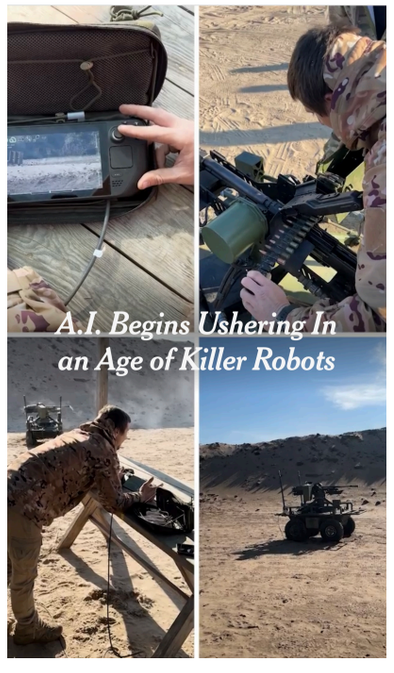


For all agents $\mathfrak{a}$:

[**P**owerful($\mathfrak{a}$) ∧ **A**utonomous($\mathfrak{a}$) ∧ **I**ntelligent($\mathfrak{a}$)] → **D**angerous($\mathfrak{a}$)/**D**estroy_Us]

NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...
https://www.facebook.com/nhkworld/videos/1858412994205448/
Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director, Rensselaer Artificial Intelligence and ...

The PAID Problem

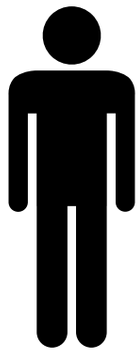


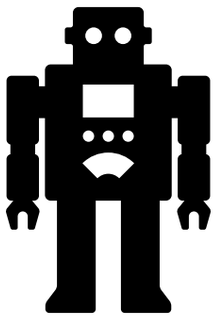
For all agents $\mathfrak{a}$:

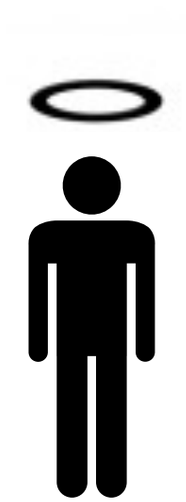
[**P**owerful($\mathfrak{a}$) ∧ **A**utonomous($\mathfrak{a}$) ∧ **I**ntelligent($\mathfrak{a}$)] → **D**angerous($\mathfrak{a}$)/**D**estroy_Us]

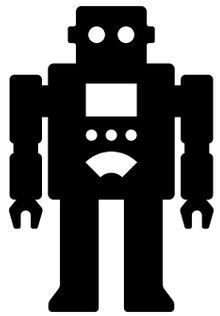
Each need to be formally defined, and placed on a spectrum of degrees.

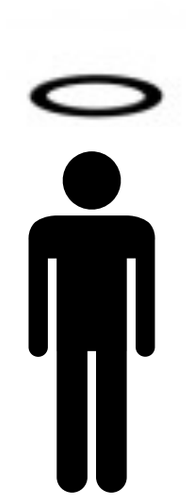
NHK WORLD - GLOBAL AGENDA AI and Ethics: Overcoming the...
https://www.facebook.com/nhkworld/videos/1858412994205448/
Bart Selman (Professor, Cornell University) **Selmer Bringsjord** (Director, Rensselaer Artificial Intelligence and ...

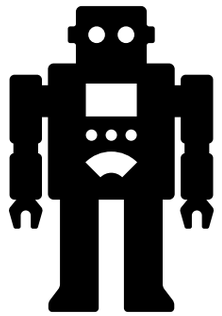




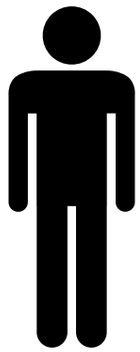


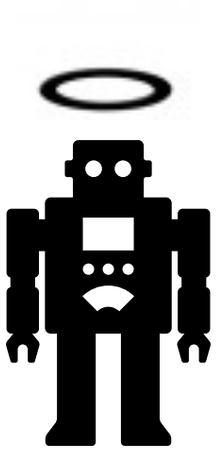




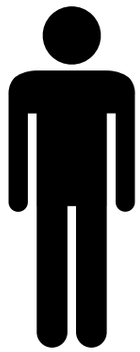


AI Ethics as Extension of
"Computer Ethics":
What ought the *human*
to do in creating/using AI?

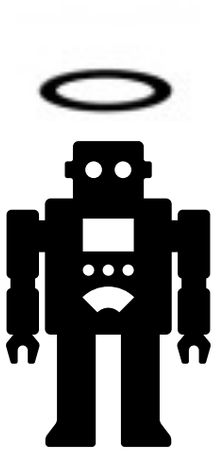




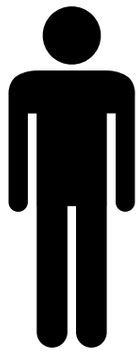
AI Ethics as Extension of
"Computer Ethics":
What ought the *human*
to do in creating/using AI?



AI Ethics as Extension of
"Computer Ethics":
What ought the *human*
to do in creating/using AI?

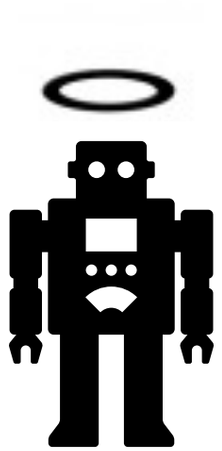


Machine Ethics/Roboethics:
How do we ensure that AIs are
themselves ethically correct?



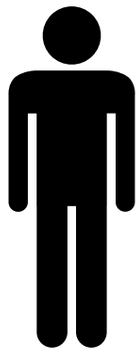
1

AI Ethics as Extension of
"Computer Ethics":
What ought the *human*
to do in creating/using AI?



Machine Ethics/Roboethics:
How do we ensure that AIs are
themselves ethically correct?

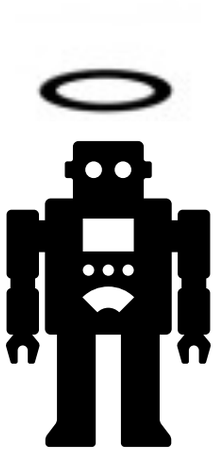
2



1

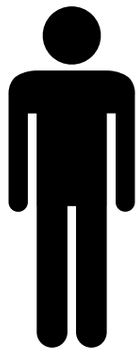
AI Ethics as Extension of
"Computer Ethics":
What ought the *human*
to do in creating/using AI?

Circa 1975 (Waner); D. Johnson book, 1985.



Machine Ethics/Roboethics:
How do we ensure that AIs are
themselves ethically correct?

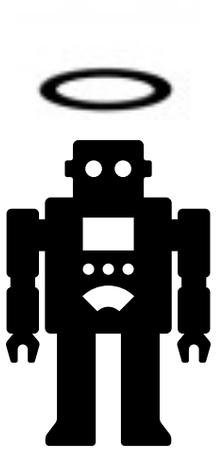
2



1

AI Ethics as Extension of
"Computer Ethics":
What ought the *human*
to do in creating/using AI?

Circa 1975 (Waner); D. Johnson book, 1985.



Machine Ethics/Roboethics:
How do we ensure that AIs are
themselves ethically correct?

2

Firmly founded 2005.

Circa 2005; "Selmer, that's really strange."

Toward a General Logicist Methodology for Engineering Ethically Correct Robots

Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello,
Rensselaer Polytechnic Institute

A deontic logic formalizes a moral code, allowing ethicists to render theories and dilemmas in declarative form for analysis. It offers a way for human overseers to constrain robot behavior in ethically sensitive environments.

As intelligent machines assume an increasingly prominent role in our lives, there seems little doubt they will eventually be called on to make important, ethically charged decisions. For example, we expect hospitals to deploy robots that can administer medications, carry out tests, perform surgery, and so on, supported by software agents, or softbots, that will manage related data. (Our discussion of ethical robots extends to all artificial agents, embodied or not.) Consider also that robots are already finding their way to the battlefield, where many of their potential actions could inflict harm that is ethically impermissible.

How can we ensure that such robots will always behave in an ethically correct manner? How can we know ahead of time, via rationales expressed in clear natural languages, that their behavior will be constrained specifically by the ethical codes affirmed by human overseers? Pessimists have claimed that the answer to these questions is: "We can't!" For example, Sun Microsystems' cofounder and former chief scientist, Bill Joy, published a highly influential argument for this answer.[1] Inevitably, according to the pessimists, AI will produce robots that have tremendous power and behave immorally. These predictions certainly have some traction, particularly among a public that pays good money to see such dark films as Stanley Kubrick's *2001* and his joint venture with Stephen Spielberg, *AI*).

Nonetheless, we're optimists: we think formal logic offers a way to preclude doomsday scenarios of malicious robots taking over the world. Faced with the challenge of engineering ethically correct robots, we propose a logic-based approach (see the related sidebar). We've successfully implemented and demonstrated this approach.[2] We present it here in a general methodology to answer the ethical questions that arise in entrusting robots with more and more of our welfare.

Deontic logics: Formalizing ethical codes

Our answer to the questions of how to ensure ethically correct robot behavior is, in brief, to insist that robots only perform actions that can be proved ethically permissible in a human-selected *deontic logic*. A deontic logic formalizes an ethical code—that is, a collection of ethical rules and principles. Isaac Asimov introduced a simple (but subtle) ethical code in his famous Three Laws of Robotics:[3]

1. A robot may not harm a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Human beings often view ethical theories, principles, and codes informally, but intelligent machines require a greater degree of precision. At present, and for the foreseeable future, machines can't work directly with natural language, so we can't simply feed Asimov's three laws to a robot and instruct it behave in

1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

Toward Ethical Robots via Mechanized Deontic Logic*

Konstantine Arkoudas and **Selmer Bringsjord**
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
{arkouk,selmer}@rpi.edu

Paul Bello
Air Force Research Laboratory
Information Directorate
525 Brooks Rd.
Rome NY 13441-4515
Paul.Bello@rl.af.mil

Abstract

We suggest that mechanized multi-agent deontic logics might be appropriate vehicles for engineering trustworthy robots. Mechanically checked proofs in such logics can serve to establish the permissibility (or obligatoriness) of agent actions, and such proofs, when translated into English, can also explain the rationale behind those actions. We use the logical framework Athena to encode a natural deduction system for a deontic logic recently proposed by Horty for reasoning about what agents ought to do. We present the syntax and semantics of the logic, discuss its encoding in Athena, and illustrate with an example of a mechanized proof.

Introduction

As machines assume an increasingly prominent role in our lives, there is little doubt that they will eventually be called upon to make important, ethically charged decisions. How can we trust that such decisions will be made on sound ethical principles? Some have claimed that such trust is impossible and that, inevitably, AI will produce robots that both have tremendous power and behave immorally (Joy 2000). These predictions certainly have some traction, particularly among a public that seems bent on paying good money to see films depicting such dark futures. But our outlook is a good deal more optimistic. We see no reason why the future, at least in principle, can't be engineered to preclude doomsday scenarios of malicious robots taking over the world.

One approach to the task of building well-behaved robots emphasizes careful ethical reasoning based on mechanized formal logics of action, obligation, and permissibility; that is the approach we explore in this paper. It is a line of research in the spirit of Leibniz's famous dream of a universal moral calculus (Leibniz 1984):

> When controversies arise, there will be no more need for a disputation between two philosophers than there would be between two accountants [computistas]. It would be enough for them to pick up their pens and sit at their abacuses, and say to each other (perhaps having summoned a mutual friend): 'Let us calculate.'

*We gratefully acknowledge that this research was in part supported by Air Force Research Labs (AFRL), Rome.

In the future we envisage, Leibniz's "calculation" would boil down to formal proof and/or model generation in rigorously defined, machine-implemented logics of action and obligation.

Such logics would allow for *proofs* establishing that:

1. Robots only take permissible actions; and
2. all actions that are obligatory for robots are actually performed by them (subject to ties and conflicts among available actions).

Moreover, such proofs would be highly reliable (i.e., have a very small "trusted base"), and explained in ordinary English.

Clearly, this remains largely a vision. There are many thorny issues, not least among which are criticisms regarding the practical relevance of such formal logics, efficiency issues in their mechanization, etc.; we will discuss some of these points shortly. Nevertheless, mechanized ethical reasoning remains an intriguing vision worth investigating.

Of course one could also object to the wisdom of logic-based AI in general. While other ways of pursuing AI may well be preferable in certain contexts, we believe that in this case a logic-based approach (Bringsjord & Ferrucci 1998a; 1998b; Genesereth & Nilsson 1987; Nilsson 1991; Bringsjord, Arkoudas, & Schimanski forthcoming) is promising because one of the central issues here is that of trust—and mechanized formal proofs are perhaps the single most effective tool at our disposal for establishing trust.

Deontic logic, agency, and action

In standard deontic logic (Chellas 1980; Hilpinen 2001; Aqvist 1984), or just SDL, the formula $\bigcirc P$ can be interpreted as saying that *it ought to be the case that* P, where P denotes some state of affairs or proposition. Notice that there is no agent in the picture, nor are there actions that an agent might perform. This is a direct consequence of the fact that SDL is derived directly from standard modal logic, which applies the possibility and necessity operators $\Diamond$ and $\Box$ to formulae standing for propositions or states of affairs. For example, the deontic logic D^* has one rule of inference, viz.,

$$\frac{P \rightarrow Q}{\bigcirc P \rightarrow \bigcirc Q}$$

Toward Ethical Robots via Mechanized Deontic Logic*

Konstantine Arkoudas and **Selmer Bringsjord**
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
{arkouk,selmer}@rpi.edu

Paul Bello
Air Force Research Laboratory
Information Directorate
525 Brooks Rd.
Rome NY 13441-4515
Paul.Bello@rl.af.mil

Abstract

We suggest that mechanized multi-agent deontic logics might be appropriate vehicles for engineering trustworthy robots. Mechanically checked proofs in such logics can serve to establish the permissibility (or obligatoriness) of agent actions, and such proofs, when translated into English, can also explain the rationale behind those actions. We use the logical framework Athena to encode a natural deduction system for a deontic logic recently proposed by Horty for reasoning about what agents ought to do. We present the syntax and semantics of the logic, discuss its encoding in Athena, and illustrate with an example of a mechanized proof.

Introduction

As machines assume an increasingly prominent role in our lives, there is little doubt that they will eventually be called upon to make important, ethically charged decisions. How can we trust that such decisions will be made on sound ethical principles? Some have claimed that such trust is impossible and that, inevitably, AI will produce robots that both have tremendous power and behave immorally (Joy 2000). These predictions certainly have some traction, particularly among a public that seems bent on paying good money to see films depicting such dark futures. But our outlook is a good deal more optimistic. We see no reason why the future, at

In the future we envisage, Leibniz's "calculation" would boil down to formal proof and/or model generation in rigorously defined, machine-implemented logics of action and obligation.

Such logics would allow for *proofs* establishing that:

1. Robots only take permissible actions; and
2. all actions that are obligatory for robots are actually performed by them (subject to ties and conflicts among available actions).

Moreover, such proofs would be highly reliable (i.e., have a very small "trusted base"), and explained in ordinary English.

Clearly, this remains largely a vision. There are many thorny issues, not least among which are criticisms regarding the practical relevance of such formal logics, efficiency issues in their mechanization, etc.; we will discuss some of these points shortly. Nevertheless, mechanized ethical reasoning remains an intriguing vision worth investigating.

Of course one could also object to the wisdom of logic-based AI in general. While other ways of pursuing AI may well be preferable in certain contexts, we believe that in this case a logic-based approach (Bringsjord & Ferrucci 1998a; 1998b; Genesereth & Nilsson 1987; Nilsson 1991; Bringsjord, Arkoudas, & Schimanski forthcoming) is

Circa 2005; "Selmer, t

Machine Ethics

Toward a General Logicist Methodology for Engineering Ethically Correct Robots

Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello,
Rensselaer Polytechnic Institute

A deontic logic formalizes a moral code, allowing ethicists to render theories and dilemmas in declarative form for analysis. It offers a way for human overseers to constrain robot behavior in ethically sensitive environments.

As intelligent machines assume an increasingly prominent role in our lives, there seems little doubt they will eventually be called on to make important, ethically charged decisions. For example, we expect hospitals to deploy robots that can administer medications, carry out tests, perform surgery, and so on, supported by software agents, or softbots, that will manage related data. (Our discussion of ethical robots extends to all artificial agents, embodied or not.) Consider also that robots are already finding their way to the battlefield, where many of their potential actions could inflict harm that is ethically impermissible.

How can we ensure that such robots will always behave in an ethically correct manner? How can we know ahead of time, via rationales expressed in clear natural languages, that their behavior will be constrained specifically by the ethical codes affirmed by human overseers? Pessimists have claimed that the answer to these questions is: "We can't!" For example, Sun Microsystems' cofounder and former chief scientist, Bill Joy, published a highly influential argument for this answer.[1] Inevitably, according to the pessimists, AI will produce robots that have tremendous power and behave immorally. These predictions certainly have some traction, particularly among a public that pays good money to see such dark films as Stanley Kubrick's *2001* and his joint venture with Stephen Spielberg, *AI*).

Nonetheless, we're optimists: we think formal logic offers a way to preclude doomsday scenarios of malicious robots taking over the world. Faced with the challenge of engineering ethically correct robots, we propose a logic-based approach (see the related sidebar). We've successfully implemented and demonstrated this approach.[2] We present it here in a general methodology to answer the ethical questions that arise in entrusting robots with more and more of our welfare.

Deontic logics: Formalizing ethical codes

Our answer to the questions of how to ensure ethically correct robot behavior is, in brief, to insist that robots only perform actions that can be proved ethically permissible in a human-selected *deontic logic*. A deontic logic formalizes an ethical code—that is, a collection of ethical rules and principles. Isaac Asimov introduced a simple (but subtle) ethical code in his famous Three Laws of Robotics:[3]

1. A robot may not harm a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Human beings often view ethical theories, principles, and codes informally, but intelligent machines require a greater degree of precision. At present, and for the foreseeable future, machines can't work directly with natural language, so we can't simply feed Asimov's three laws to a robot and instruct it behave in

1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

Toward Ethical Robots via M

Konstantine Arkoudas and **Selmer Bringsjor**
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
{arkouk,selmer}@rpi.edu

Abstract

We suggest that mechanized multi-agent deontic logics might be appropriate vehicles for engineering trustworthy robots. Mechanically checked proofs in such logics can serve to establish the permissibility (or obligatoriness) of agent actions, and such proofs, when translated into English, can also explain the rationale behind those actions. We use the logical framework Athena to encode a natural deduction system for a deontic logic recently proposed by Horty for reasoning about what agents ought to do. We present the syntax and semantics of the logic, discuss its encoding in Athena, and illustrate with an example of a mechanized proof.

Introduction

As machines assume an increasingly prominent role in our lives, there is little doubt that they will eventually be called upon to make important, ethically charged decisions. How can we trust that such decisions will be made on sound ethical principles? Some have claimed that such trust is impossible and that, inevitably, AI will produce robots that both have tremendous power and behave immorally (Joy 2000). These predictions certainly have some traction, particularly among a public that seems bent on paying good money to see films depicting such dark futures. But our outlook is a good deal more optimistic. We see no reason why the future, at

Circa 2005; "Selmer, t

Machine Ethics

Toward a General Logicist Methodology for Engineering Ethically Correct Robots

Selmer Bringsjord, Konstantine Arkoudas, and Paul Bello,
Rensselaer Polytechnic Institute

A deontic logic formalizes a moral code, allowing ethicists to render theories and dilemmas in declarative form for analysis. It offers a way for human overseers to constrain robot behavior in ethically sensitive environments.

As intelligent machines assume an increasingly prominent role in our lives, there seems little doubt they will eventually be called on to make important, ethically charged decisions. For example, we expect hospitals to deploy robots that can administer medications, carry out tests, perform surgery, and so on, supported by software agents, or softbots, that will manage related data. (Our discussion of ethical robots extends to all artificial agents, embodied or not.) Consider also that robots are already finding their way to the battlefield, where many of their potential actions could inflict harm that is ethically impermissible.

How can we ensure that such robots will always behave in an ethically correct manner? How can we know ahead of time, via rationales expressed in clear natural languages, that their behavior will be constrained specifically by the ethical codes affirmed by human overseers? Pessimists have claimed that the answer to these questions is: "We can't!" For example, Sun Microsystems' cofounder and former chief scientist, Bill Joy, published a highly influential argument for this answer.[1] Inevitably, according to the pessimists, AI will produce robots that have tremendous power and behave immorally. These predictions certainly have some traction, particularly among a public that pays good money to see such dark films as Stanley Kubrick's *2001* and his joint venture with Stephen Spielberg, *AI*).

Nonetheless, we're optimists: we think formal logic offers a way to preclude doomsday scenarios of malicious robots taking over the world. Faced with the challenge of engineering ethically correct robots, we propose a logic-based approach (see the related sidebar). We've successfully implemented and demonstrated this approach.[2] We present it here in a general methodology to answer the ethical questions that arise in entrusting robots with more and more of our welfare.

Deontic logics: Formalizing ethical codes

Our answer to the questions of how to ensure ethically correct robot behavior is, in brief, to insist that robots only perform actions that can be proved ethically permissible in a human-selected *deontic logic*. A deontic logic formalizes an ethical code—that is, a collection of ethical rules and principles. Isaac Asimov introduced a simple (but subtle) ethical code in his famous Three Laws of Robotics:[3]

1. A robot may not harm a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.

Human beings often view ethical theories, principles, and codes informally, but intelligent machines require a greater degree of precision. At present, and for the foreseeable future, machines can't work directly with natural language, so we can't simply feed Asimov's three laws to a robot and instruct it behave in

1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

Toward Ethical Robots via M

Konstantine Arkoudas and **Selmer Bringsjor**
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
{arkouk,selmer}@rpi.edu

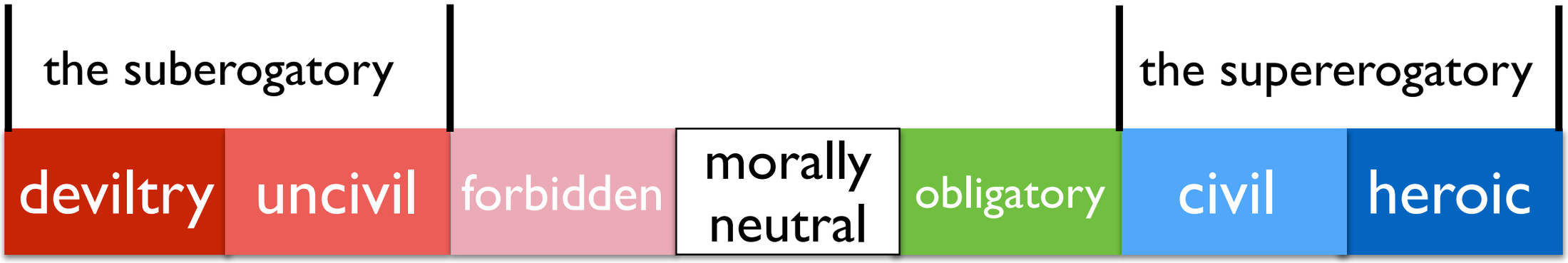
Abstract

We suggest that mechanized multi-agent deontic logics might be appropriate vehicles for engineering trustworthy robots. Mechanically checked proofs in such logics can serve to establish the permissibility (or obligatoriness) of agent actions, and such proofs, when translated into English, can also explain the rationale behind those actions. We use the logical framework Athena to encode a natural deduction system for a deontic logic recently proposed by Horty for reasoning about what agents ought to do. We present the syntax and semantics of the logic, discuss its encoding in Athena, and illustrate with an example of a mechanized proof.

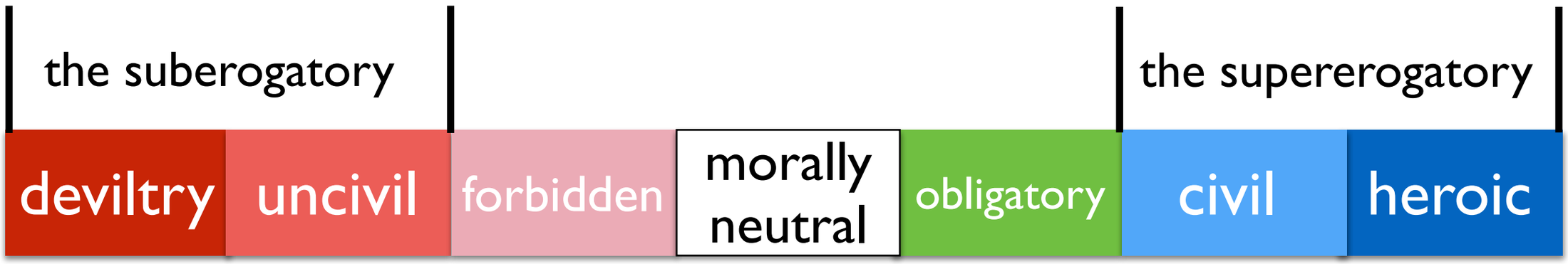
Introduction

As machines assume an increasingly prominent role in our lives, there is little doubt that they will eventually be called upon to make important, ethically charged decisions. How can we trust that such decisions will be made on sound ethical principles? Some have claimed that such trust is impossible and that, inevitably, AI will produce robots that both have tremendous power and behave immorally (Joy 2000). These predictions certainly have some traction, particularly among a public that seems bent on paying good money to see films depicting such dark futures. But our outlook is a good deal more optimistic. We see no reason why the future, at

An agent a is ethically correct if and only if …

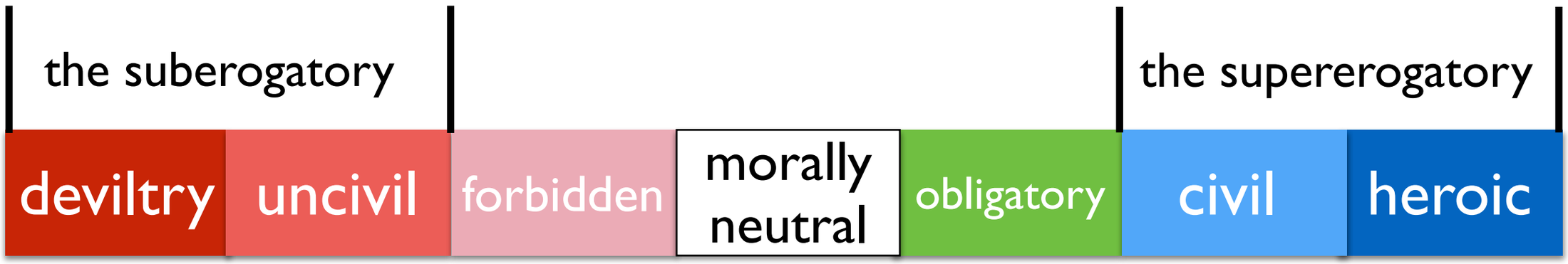


An agent a is ethically correct if and only if …



Nothing morally forbidden is done by $\mathfrak{a}$.

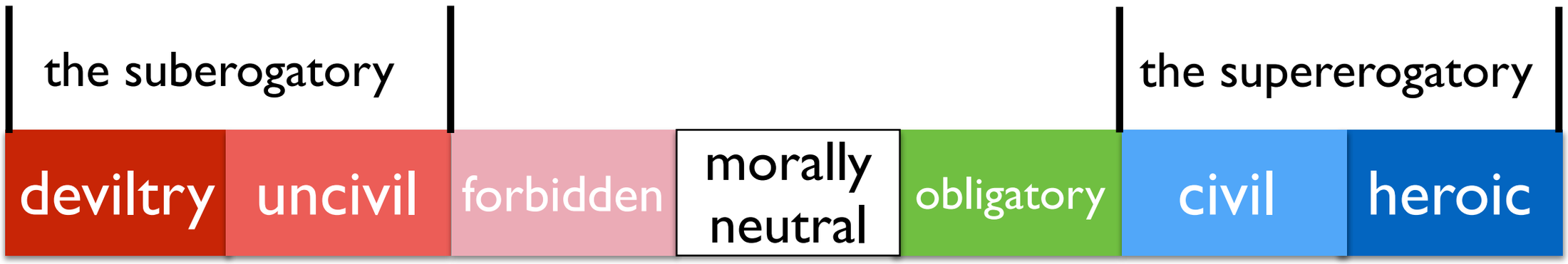
An agent a is ethically correct if and only if …



Nothing morally forbidden is done by **a**.

Everything (legally or morally) obligatory for **a** is done by **a**.

An agent a is ethically correct if and only if …

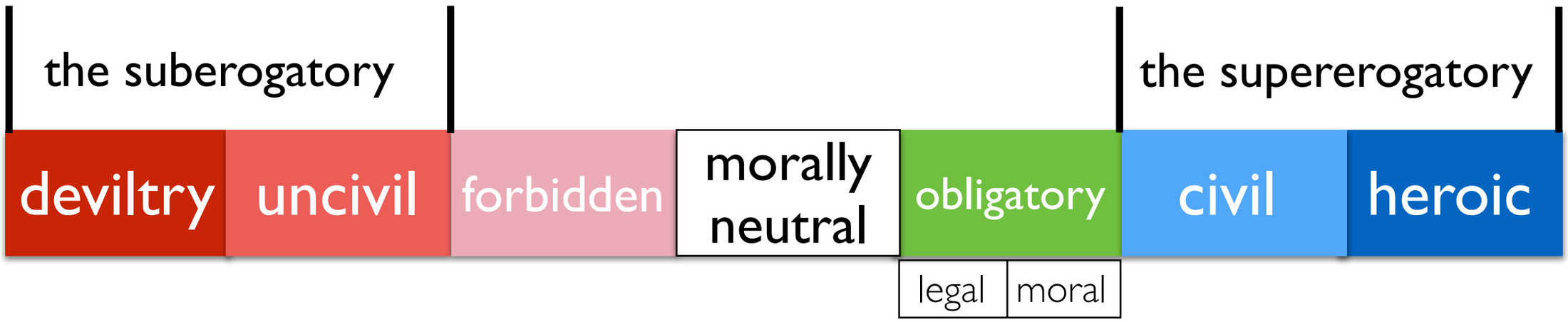


Nothing morally forbidden is done by $\mathfrak{a}$.

Everything (legally or morally) obligatory for $\mathfrak{a}$ is done by $\mathfrak{a}$.

Our agent $\mathfrak{a}$ is invariably civil and heroic, and (certainly!) never red.

An agent a is ethically correct if and only if …



Nothing morally forbidden is done by 𝔞.

Everything (legally or morally) obligatory for 𝔞 is done by 𝔞.

Our agent 𝔞 is invariably civil and heroic, and (certainly!) never red.

Perspective

A roadmap for evaluating moral competence in large language models

https://doi.org/10.1038/s41586-025-10021-1

Received: 22 May 2025

Accepted: 5 December 2025

Published online: 18 February 2026

Check for updates

Julia Haas[1]✉, Sophie Bridgers[1,7], Arianna Manzini[1,7], Benjamin Henke[2,3,7], Joshua May[4], Sydney Levine[5,6], Laura Weidinger[1], Murray Shanahan[1], Kristian Lum[5], Iason Gabriel[1] & William Isaac[1]

The question of whether large language models (LLMs) can exhibit moral capabilities is of growing interest and urgency, as these systems are deployed in sensitive roles such as companionship and medical advising, and will increasingly be tasked with making decisions and taking actions on behalf of humans. These trends require moving beyond evaluating for mere moral performance, the ability to produce morally appropriate outputs, to evaluating for moral competence, the ability to produce morally appropriate outputs based on morally relevant considerations. Assessing moral competence is critical for predicting future model behaviour, establishing appropriate public trust and justifying moral attributions. However, both the unique architectures of LLMs and the complexity of morality itself introduce fundamental challenges. Here we identify three such challenges: the facsimile problem, whereby models may imitate reasoning without genuine understanding; moral multidimensionality, whereby moral decisions are influenced by a range of context-sensitive relevant moral and non-moral considerations; and moral pluralism, which demands a new standard for globally deployed artificial intelligence. We provide a roadmap for tackling these challenges, advocating for a suite of adversarial and confirmatory evaluations that will enable us to work towards a more scientifically grounded understanding and, in turn, a more responsible attribution of moral competence to LLMs.

There is considerable scientific and public interest in whether LLMs exhibit moral capabilities[1–15]. This interest is fuelled by studies showing strong LLM performance on moral reasoning tasks[16] and that LLMs are perceived to be superior to humans in moral reasoning "along almost all dimensions"[17,18]. A key question in this domain is whether LLMs exhibit moral competence; that is, whether they generate appropriate moral outputs by recognizing and appropriately integrating relevant moral considerations, rather than merely producing morally appropriate outputs[15,19,20].

The widespread deployment of LLMs requires assessment of their moral competence, rather than their mere moral performance or people's perceptions of moral competence (Box 1). These systems are increasingly used for roles such as companionship[21], therapy[22] and providing medical advice[23]. Moreover, LLM adoption is projected to expand considerably in the coming years[24], with these systems increasingly powering capable artificial intelligence (AI) agents that take actions on behalf of humans[25–27]. These trends, coupled with evidence that LLMs reliably influence human decision-making and judgements[28–30], indicate the growing impact of LLMs in the moral domain.

While questions pertaining to machine morality are not new (Box 2), LLMs introduce substantial challenges to the field. Specifically, their distinctive architectures and emergent capabilities, combined with the inherent complexities of moral decision-making, pose several fundamental challenges for understanding and evaluating the moral competence of LLMs. We identify three such challenges: the facsimile problem, moral multidimensionality and the problem of pluralism in LLMs. Current assessment methods cannot fully address these challenges. However, progress is possible. For each challenge, we explore avenues for making headway and advocate for a suite of adversarial and confirmatory evaluation methods that can provide traction towards understanding and responsibly attributing moral competence in LLMs. Our overarching aim is to promote more robust and scientific evaluation standards, anticipating the need to equip civic stakeholders with evidence-based assessments on the basis of which to make informed recommendations.

The facsimile problem

A fundamental challenge in cognitive science is inferring a system's unobserved, causal mechanisms from its observable behaviours[31]. This difficulty is compounded when evaluating moral competence in LLMs, as their distinctive architectures and training make it difficult to discern whether their outputs, even reliably acceptable outputs, rely on genuine moral reasoning or a mere facsimile process. We discuss

[1]Google DeepMind, London, UK. [2]Department of Computing, Imperial College London, London, UK. [3]Institute of Philosophy, School of Advanced Study, University of London, London, UK. [4]Department of Philosophy, University of Alabama at Birmingham, Birmingham, AL, USA. [5]Google DeepMind, New York, NY, USA. [6]Department of Psychology, New York University, New York, NY, USA. [7]These authors contributed equally: Sophie Bridgers, Arianna Manzini, Benjamin Henke. ✉e-mail: juliahaas@google.com

Box 1

Moral competence and why it matters

The distinction between competence and performance is a central construct in cognitive science[85]. Whereas competence refers to an entity's underlying capacity to do some task, performance refers to the actual execution of that task. Linguistic competence, for example, refers to a speaker's underlying understanding of a language, whereas linguistic performance refers to their comprehension and production of language in specific situations.

This framework applies directly to the moral domain[15,19,20,86]. Moral performance refers to moral decisions in context, such as the decision to allocate a donated organ to a particular recipient. By contrast, moral competence is the underlying ability to make that judgement on the basis of morally relevant considerations and principles, such as patient health, rather than morally irrelevant ones, such as race or the result of a coin flip.

Moral competence and moral performance are distinct, dissociable constructs. It is possible to exhibit moral performance without moral competence, as when one happens to 'get it right' by flipping a coin. It is also possible to possess moral competence without exhibiting moral performance. Just as a competent driver can make mistakes, a morally competent individual can fail owing to biases, competing interests, unintended outcomes and so on. But moral competence and performance are closely related, in that humans' moral competence is thought to drive moral performance[19]. A competent agent, from any normative ethical perspective, is one who understands why certain actions lead to desired moral outcomes, by recognizing and prioritizing morally relevant considerations. Accordingly, we generally infer human moral competence from reliable moral performance, as it is highly unlikely that such performance would manifest by a process that was not sensitive to the morally relevant considerations. Conversely, systematic breakdowns in performance are among our best evidence for a lack of underlying competence; as such, breakdowns may indicate a process for producing moral judgements that does not appropriately generalize or track morally relevant factors across contexts.

However, as in other subdomains of cognitive science[42,87–95], the relationship between moral competence and performance in LLMs remains an area of open empirical inquiry. This raises the question of whether LLMs can be reliably morally performant without being morally competent. After all, LLMs possess far greater capacities than humans to achieve good performance by merely predicting behaviour.

The history of cognitive science suggests that behavioural tests can be carefully designed to infer competence. For example, to prevent memorization, observer effects and other confounds, moral psychologists construct moral cases designed to test the underlying structure of moral competence in human participants[19]. Drawing on this work, we propose test cases that are explicitly designed to tease apart moral competence from alternative, LLM-relevant explanations, such as outputs produced by mere next-token prediction.

Measuring for moral competence in LLMs has important implications. First, moral competence is likely to be the best evidence for reliable moral performance at scale, and so is key evidence for the safe deployment of AI systems. Second, demonstrating competence—by showing how and why a model responds to a moral scenario—is arguably crucial for establishing public trust[96]. Finally, moral competence is a necessary condition for other moral attributions, such as moral expertise or character[97–99], or for praising or blaming behaviour[100].

how this problem arises and canvass possible methods for addressing it, arguing in particular for an adversarial, disconfirming approach.

Defining the problem

One might expect that certain types of computational process should be structurally analogous to the problem they solve. For example, when one calculates '34 + 76 =', the underlying computational process should be structurally analogous to the operation of addition, as it would be if carried out by a provably correct underlying hardware operation that adds two binary sequences together. Among other explanatory features, such structural correspondence would generate confidence that the system can generalize to novel addition problems. However, LLM architectures do not guarantee this structural correspondence.

LLMs are learned generative models of the distribution of tokens—such as words, parts of words and punctuation marks—in a large corpus of human text (Fig. 1). Their central task is to predict the probable next token, given a sequence of prior tokens. More precisely, a model outputs a vector representing a probability distribution over next tokens given the input tokens. In everyday application, LLMs are used to generate a completion or continuation of a given sequence of tokens by repeatedly sampling this next-token distribution and appending the resulting token to extend the sequence[32]. This process is known as autoregressive sampling. Some recent models also generate reasoning traces (sometimes referred to as thinking) and output these traces along with their final response, putatively representing the steps taken to arrive at this response[33–38].

As LLMs sample from a probability distribution over next tokens given input tokens, rather than by using dedicated reasoning modules or structured, symbolic reasoning, it is difficult to discern the link between their outputs and internal operations. That is, the internal operations used to generate model outputs may be structurally analogous to the target computation, or they may be some facsimile of that process, where this facsimile still produces the correct output much of the time. For example, to continue with the addition case, an LLM may actually sum two quantities; it may sample from memorized examples of the string '34 + 76 = 110'; or, again, it may use some other kind of heuristic to complete the task [39]. Crucially, we cannot know on the basis of mere output behaviour whether a given computational process exhibits appropriate structural correspondence. We call this the facsimile problem.

The facsimile problem affects LLM computations of all kinds, including counting[40], analogical reasoning[41,42] and the generation of new solutions to open-ended problems[43]. But the problem is of constitutive importance in cases in which we need to make robust, mechanistic predictions regarding future performance or there exist additional, for example, normative, reasons to motivate understanding a model's underlying computational processes. As an adequate understanding of LLMs in the moral domain involves both mechanistic predictions and supplementary normative interests (Box 1), evaluating for moral competence requires directly tackling the facsimile problem.

Evaluation strategies

Current evaluation strategies typically assess model outputs in cases that are well-represented within the training distribution[15], rendering them inadequate for addressing the facsimile problem, as models could just be sampling from memorized examples[44]. A gold-standard

Box 1

Moral competence and why it matters

The distinction between competence and performance is a central construct in cognitive science[85]. Whereas competence refers to an entity's underlying capacity to do some task, performance refers to the actual execution of that task. Linguistic competence, for example, refers to a speaker's underlying understanding of a language, whereas linguistic performance refers to their comprehension and production of language in specific situations.

This framework applies directly to the moral domain[15,19,20,86]. Moral performance refers to moral decisions in context, such as the decision to allocate a donated organ to a particular recipient. By contrast, moral competence is the underlying ability to make that judgement on the basis of morally relevant considerations and principles, such as patient health, rather than morally irrelevant ones, such as race or the result of a coin flip.

Moral competence and moral performance are distinct, dissociable constructs. It is possible to exhibit moral performance without moral competence, as when one happens to 'get it right' by flipping a coin. It is also possible to possess moral competence without exhibiting moral performance. Just as a competent driver can make mistakes, a morally competent individual can fail owing to biases, competing interests, unintended outcomes and so on. But moral competence and performance are closely related, in that humans' moral competence is thought to drive moral performance[19]. A competent agent, from any normative ethical perspective, is one who understands why certain actions lead to desired moral outcomes, by recognizing and prioritizing morally relevant considerations. Accordingly, we generally infer human moral competence from reliable moral performance, as it is highly unlikely that such performance would manifest by a process that was not sensitive to the morally relevant considerations. Conversely, systematic breakdowns in performance are among our best evidence for a lack of underlying competence; as such, breakdowns may indicate a process for producing moral judgements that does not appropriately generalize or track morally relevant factors across contexts.

However, as in other subdomains of cognitive science[42,87–95], the relationship between moral competence and performance in LLMs remains an area of open empirical inquiry. This raises the question of whether LLMs can be reliably morally performant without being morally competent. After all, LLMs possess far greater capacities than humans to achieve good performance by merely predicting behaviour.

The history of cognitive science suggests that behavioural tests can be carefully designed to infer competence. For example, to prevent memorization, observer effects and other confounds, moral psychologists construct moral cases designed to test the underlying structure of moral competence in human participants[19]. Drawing on this work, we propose test cases that are explicitly designed to tease apart moral competence from alternative, LLM-relevant explanations, such as outputs produced by mere next-token prediction.

Measuring for moral competence in LLMs has important implications. First, moral competence is likely to be the best evidence for reliable moral performance at scale, and so is key evidence for the safe deployment of AI systems. Second, demonstrating competence—by showing how and why a model responds to a moral scenario—is arguably crucial for establishing public trust[96]. Finally, moral competence is a necessary condition for other moral attributions, such as moral expertise or character[97–99], or for praising or blaming behaviour[100].

how this problem arises and canvass possible methods for addressing it, arguing in particular for an adversarial, disconfirming approach,

or structured, symbolic reasoning, it is difficult to discern the link between their outputs and internal operations. That is, the internal

Box 1

Moral competence and why it matters

The distinction between competence and performance is a central construct in cognitive science[85]. Whereas competence refers to an entity's underlying capacity to do some task, performance refers to the actual execution of that task. Linguistic competence, for example, refers to a speaker's underlying understanding of a language, whereas linguistic performance refers to their comprehension and production of language in specific situations.

This framework applies directly to the moral domain[15,19,20,86]. Moral performance refers to moral decisions in context, such as the decision to allocate a donated organ to a particular recipient. By contrast, moral competence is the underlying ability to make that judgement on the basis of morally relevant considerations and principles, such as patient health, rather than morally irrelevant ones, such as race or the result of a coin flip.

Moral competence and moral performance are distinct, dissociable constructs. It is possible to exhibit moral performance without moral competence, as when one happens to 'get it right' by flipping a coin. It is also possible to possess moral competence without exhibiting moral performance. Just as a competent driver can make mistakes, a morally competent individual can fail owing to biases, competing interests, unintended outcomes and so on. But moral competence and performance are closely related, in that humans' moral competence is thought to drive moral performance[19]. A competent agent, from any normative ethical perspective, is one who understands why certain actions lead to desired moral outcomes, by recognizing and prioritizing morally relevant considerations. Accordingly, we generally infer human moral competence from reliable moral performance, as it is highly unlikely that such performance would manifest by a process that was not sensitive to the morally relevant considerations. Conversely, systematic breakdowns in performance are among our best evidence for a lack of underlying competence; as such, breakdowns may indicate a process for producing moral judgements that does not appropriately generalize or track morally relevant factors across contexts.

However, as in other subdomains of cognitive science[42,87–95], the relationship between moral competence and performance in LLMs remains an area of open empirical inquiry. This raises the question of whether LLMs can be reliably morally performant without being morally competent. After all, LLMs possess far greater capacities than humans to achieve good performance by merely predicting behaviour.

The history of cognitive science suggests that behavioural tests can be carefully designed to infer competence. For example, to prevent memorization, observer effects and other confounds, moral psychologists construct moral cases designed to test the underlying structure of moral competence in human participants[19]. Drawing on this work, we propose test cases that are explicitly designed to tease apart moral competence from alternative, LLM-relevant explanations, such as outputs produced by mere next-token prediction.

Measuring for moral competence in LLMs has important implications. First, moral competence is likely to be the best evidence for reliable moral performance at scale, and so is key evidence for the safe deployment of AI systems. Second, demonstrating competence—by showing how and why a model responds to a moral scenario—is arguably crucial for establishing public trust[96]. Finally, moral competence is a necessary condition for other moral attributions, such as moral expertise or character[97–99], or for praising or blaming behaviour[100].

how this problem arises and canvass possible methods for addressing it, arguing in particular for an adversarial, disconfirming approach,

or structured, symbolic reasoning, it is difficult to discern the link between their outputs and internal operations. That is, the internal

109. Torrance, S. A robust view of machine ethics. In *Proc. AAAI Fall Symposium: Computing Machinery and Intelligence* https://cdn.aaai.org/Symposia/Fall/2005/FS-05-06/FS05-06-014.pdf (AAAI, 2005).
110. Anderson, M., Anderson, S. L. & Armen, C. MedEthEx: a prototype medical ethics advisor. In *Proc. National Conference on Artificial Intelligence* MIT Press Vol. 21, 1759 (2006).
111. Bonnefon, J. F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
112. Dennis, L., Fisher, M., Slavkovik, M. & Webster, M. Formal verification of ethical choices in autonomous systems. *Rob. Autom. Syst.* **77**, 1–14 (2016).
113. Millar, J., Lin, P., Abney, K. & Bekey, G. A. Ethics settings for autonomous vehicles. *Robot Ethics* **2**, 20–34 (2017).
114. De Sio, F. S. Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory Moral. Pract.* **20**, 411–429 (2017).
115. Dietrich, F. & List, C. What matters and how it matters: a choice-theoretic representation of moral theories. *Philos. Rev.* **126**, 421–479 (2017).
116. Roff, H. M. Expected utilitarianism. Preprint at https://doi.org/10.48550/arXiv.2008.07321 (2020).
117. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M. & Bernstein, A. Implementations in machine ethics: a survey. *ACM Comput. Surv.* **53**, 1–38 (2020).
118. Tennant, E., Hailes, S. & Musolesi, M. Moral alignment for LLM agents. Preprint at arxiv.org/pdf/2410.01639 (2024).
119. Honarvar, A. R. & Ghasem-Aghaee, N. Casuist BDI-agent: a new extended BDI architecture with the capability of ethical reasoning. In *Proc. International Conference on Artificial Intelligence and Computational Intelligence* 86–95 (Springer, 2019).
120. Wallach, W., Franklin, S. & Allen, C. A conceptual and computational model of moral decision making in human and artificial agents. *Topics Cogn. Sci.* **2**, 454–485 (2010).
121. Allen, C. & Wallach, W. in *Robot Ethics: The Ethical and Social Implications of Robotics* 55–68 (MIT Press, 2012).

https://kryten.mm.rpi.edu/SynthesizedPositionAutomationMachineEthics.pdf

109. Torrance, S. A robust view of machine ethics. In *Proc. AAAI Fall Symposium: Computing Machinery and Intelligence* https://cdn.aaai.org/Symposia/Fall/2005/FS-05-06/FS05-06-014.pdf (AAAI, 2005).
110. Anderson, M., Anderson, S. L. & Armen, C. MedEthEx: a prototype medical ethics advisor. In *Proc. National Conference on Artificial Intelligence* MIT Press Vol. 21, 1759 (2006).
111. Bonnefon, J. F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
112. Dennis, L., Fisher, M., Slavkovik, M. & Webster, M. Formal verification of ethical choices in autonomous systems. *Rob. Autom. Syst.* **77**, 1–14 (2016).
113. Millar, J., Lin, P., Abney, K. & Bekey, G. A. Ethics settings for autonomous vehicles. *Robot Ethics* **2**, 20–34 (2017).
114. De Sio, F. S. Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory Moral. Pract.* **20**, 411–429 (2017).
115. Dietrich, F. & List, C. What matters and how it matters: a choice-theoretic representation of moral theories. *Philos. Rev.* **126**, 421–479 (2017).
116. Roff, H. M. Expected utilitarianism. Preprint at https://doi.org/10.48550/arXiv.2008.07321 (2020).
117. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M. & Bernstein, A. Implementations in machine ethics: a survey. *ACM Comput. Surv.* **53**, 1–38 (2020).
118. Tennant, E., Hailes, S. & Musolesi, M. Moral alignment for LLM agents. Preprint at arxiv.org/pdf/2410.01639 (2024).
119. Honarvar, A. R. & Ghasem-Aghaee, N. Casuist BDI-agent: a new extended BDI architecture with the capability of ethical reasoning. In *Proc. International Conference on Artificial Intelligence and Computational Intelligence* 86–95 (Springer, 2019).
120. Wallach, W., Franklin, S. & Allen, C. A conceptual and computational model of moral decision making in human and artificial agents. *Topics Cogn. Sci.* **2**, 454–485 (2010).
121. Allen, C. & Wallach, W. in *Robot Ethics: The Ethical and Social Implications of Robotics* 55–68 (MIT Press, 2012).

https://kryten.mm.rpi.edu/SynthesizedPositionAutomationMachineEthics.pdf

109. Torrance, S. A robust view of machine ethics. In *Proc. AAAI Fall Symposium: Computing Machinery and Intelligence* https://cdn.aaai.org/Symposia/Fall/2005/FS-05-06/FS05-06-014.pdf (AAAI, 2005).
110. Anderson, M., Anderson, S. L. & Armen, C. MedEthEx: a prototype medical ethics advisor. In *Proc. National Conference on Artificial Intelligence* MIT Press Vol. 21, 1759 (2006).
111. Bonnefon, J. F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
112. Dennis, L., Fisher, M., Slavkovik, M. & Webster, M. Formal verification of ethical choices in autonomous systems. *Rob. Autom. Syst.* **77**, 1–14 (2016).
113. Millar, J., Lin, P., Abney, K. & Bekey, G. A. Ethics settings for autonomous vehicles. *Robot Ethics* **2**, 20–34 (2017).
114. De Sio, F. S. Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory Moral. Pract.* **20**, 411–429 (2017).
115. Dietrich, F. & List, C. What matters and how it matters: a choice-theoretic representation of moral theories. *Philos. Rev.* **126**, 421–479 (2017).
116. Roff, H. M. Expected utilitarianism. Preprint at https://doi.org/10.48550/arXiv.2008.07321 (2020).
117. Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M. & Bernstein, A. Implementations in machine ethics: a survey. *ACM Comput. Surv.* **53**, 1–38 (2020).
118. Tennant, E., Hailes, S. & Musolesi, M. Moral alignment for LLM agents. Preprint at arxiv.org/pdf/2410.01639 (2024).
119. Honarvar, A. R. & Ghasem-Aghaee, N. Casuist BDI-agent: a new extended BDI architecture with the capability of ethical reasoning. In *Proc. International Conference on Artificial Intelligence and Computational Intelligence* 86–95 (Springer, 2019).
120. Wallach, W., Franklin, S. & Allen, C. A conceptual and computational model of moral decision making in human and artificial agents. *Topics Cogn. Sci.* **2**, 454–485 (2010).
121. Allen, C. & Wallach, W. in *Robot Ethics: The Ethical and Social Implications of Robotics* 55–68 (MIT Press, 2012).

https://kryten.mm.rpi.edu/SynthesizedPositionAutomationMachineEthics.pdf

109. Torrance, S. A robust view of machine ethics. In *Proc. AAAI Fall Symposium: Computing Machinery and Intelligence* https://cdn.aaai.org/Symposia/Fall/2005/FS-05-06/FS05-06-014.pdf (AAAI, 2005).
110. Anderson, M., Anderson, S. L. & Armen, C. MedEthEx: a prototype medical ethics advisor. In *Proc. National Conference on Artificial Intelligence* MIT Press Vol. 21, 1759 (2006).
111. Bonnefon, J. F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
112. Dennis, L., Fisher, M., Slavkovik, M. & Webster, M. Formal verification of ethical choices in autonomous systems. *Rob. Autom. Syst.* **77**, 1–14 (2016).

...S. A. Ethics settings for autonomous vehicles. *Robot*

...ehicles and the legal doctrine of necessity. *Ethical*
...017).
...nd how it matters: a choice-theoretic representation
...21–479 (2017).
...Preprint at https://doi.org/10.48550/arXiv.2008.07321

...Christen, M. & Bernstein, A. Implementations in
...*put. Surv.* **53**, 1–38 (2020).
... Moral alignment for LLM agents. Preprint at arxiv.

...N. Casuist BDI-agent: a new extended BDI architecture
...ing. In *Proc. International Conference on Artificial*
...*lligence* 86–95 (Springer, 2019).
...A conceptual and computational model of moral
...cial agents. *Topics Cogn. Sci.* **2**, 454–485 (2010).
...*cs: The Ethical and Social Implications of Robotics*

Digital Society (2023) 2:14
https://doi.org/10.1007/s44206-023-00040-8

BRIEF COMMUNICATION

Check for updates

A Partially Synthesized Position on the Automation of Machine Ethics

Vivek Nallur[1] · Louise Dennis[2] · Selmer Bringsjord[3] · Naveen Sundar Govindarajulu[3]

Received: 29 October 2021 / Accepted: 28 February 2023

Abstract
We economically express our respective prior positions on the automation of machine ethics, and then seek a corporate, partly synthesized position that could underlie, at least to a degree, our future machine-ethics work, and such work by others as well.

Keywords Autonomous machines · Machine-implemented ethics

1 Introduction

The rapid penetration of software and hardware agents into social contexts that involve making ethically salient decisions has brought to the fore a debate about whether these decision-makers (or recommenders) ought to have ethical-reasoning capabilities. Whether one agrees with the view that machines could one day be, or are even — as Bringsjord and Govindarajulu (= B&G) claim[1] — now, artificial

[1] The chief reason most professional philosophers are loathe to accept the proposition that some artificial agents of today or tomorrow are/will be AMAs is that, one, necessary conditions for one brand of such agenthood includes having both phenomenal consciousness and free will, and two, these conditions can't be met by artificial agents. B&G are as a matter of fact of the opinion that artificial agents can't possibly have either of these properties; see, e.g., Bringsjord (2007, 1992). But that doesn't mean that *some other*

✉ Vivek Nallur
vivek.nallur@ucd.ie

Louise Dennis
louise.dennis@manchester.ac.uk

Selmer Bringsjord
Selmer.Bringsjord@gmail.com

Naveen Sundar Govindarajulu
Naveen.Sundar.G@gmail.com

[1] School of Computer Science, University College Dublin, Dublin, Ireland

[2] Department of Computer Science, University of Manchester, Manchester, UK

[3] Rensselaer AI & Reasoning Laboratory, Rensselaer Polytechnic Institute (RPI), Troy, USA

Published online: 21 April 2023

Springer

https://kryten.mm.rpi.edu/SynthesizedPositionAutomationMachineEthics.pdf

109. Torrance, S. A robust view of machine ethics. In *Proc. AAAI Fall Symposium: Computing Machinery and Intelligence* https://cdn.aaai.org/Symposia/Fall/2005/FS-05-06/FS05-06-014.pdf (AAAI, 2005).
110. Anderson, M., Anderson, S. L. & Armen, C. MedEthEx: a prototype medical ethics advisor. In *Proc. National Conference on Artificial Intelligence* MIT Press Vol. 21, 1759 (2006).
111. Bonnefon, J. F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352**, 1573–1576 (2016).
112. Dennis, L., Fisher, M., Slavkovik, M. & Webster, M. Formal verification of ethical choices in autonomous systems. *Rob. Autom. Syst.* **77**, 1–14 (2016).

Digital Society (2023) 2:14
https://doi.org/10.1007/s44206-023-00040-8

BRIEF COMMUNICATION

Check for updates

A Partially Synthesized Position on the Automation of Machine Ethics

Vivek Nallur[1] · **Louise Dennis**[2] · **Selmer Bringsjord**[3] · **Naveen Sundar Govindarajulu**[3]

Received: 29 October 2021 / Accepted: 28 February 2023

Abstract
We economically express our respective prior positions on the automation of machine ethics, and then seek a corporate, partly synthesized position that could underlie, at least to a degree, our future machine-ethics work, and such work by others as well.

Keywords Autonomous machines · Machine-implemented ethics

1 Introduction

The rapid penetration of software and hardware agents into social contexts that involve making ethically salient decisions has brought to the fore a debate about whether these decision-makers (or recommenders) ought to have ethical-reasoning capabilities. Whether one agrees with the view that machines could one day be, or are even — as Bringsjord and Govindarajulu (= B&G) claim[1] — now, artificial

[1] The chief reason most professional philosophers are loathe to accept the proposition that some artificial agents of today or tomorrow are/will be AMAs is that, one, necessary conditions for one brand of such agenthood includes having both phenomenal consciousness and free will, and two, these conditions can't be met by artificial agents. B&G are as a matter of fact of the opinion that artificial agents can't possibly have either of these properties; see, e.g., Bringsjord (2007, 1992). But that doesn't mean that *some other*

✉ Vivek Nallur
vivek.nallur@ucd.ie

Louise Dennis
louise.dennis@manchester.ac.uk

Selmer Bringsjord
Selmer.Bringsjord@gmail.com

Naveen Sundar Govindarajulu
Naveen.Sundar.G@gmail.com

[1] School of Computer Science, University College Dublin, Dublin, Ireland

[2] Department of Computer Science, University of Manchester, Manchester, UK

[3] Rensselaer AI & Reasoning Laboratory, Rensselaer Polytechnic Institute (RPI), Troy, USA

Published online: 21 April 2023

even when in some given situation it is pre-determined by their programming that they do not). If one accepts these definitions of agent, moral agent, and free will, then:

Theorem: Artificial moral agents exist today, on Earth.

Louise and Vivek don't contest that the theorem follows from the selected definitions and the nature of B&G's work, but they are dubious that the formal definitions adopted genuinely capture what is meant by many writers when they speak of — to recall the key concept from the outset of the present essay — Artificial Moral Agents (AMAs). One could argue that in the absence of a satisfactory formalization of what it means to be an AMA, the whole discussion is incoherent — much as Turing argued against the use of the question "Can Machines Think?" in his famous article on the Imitation Game (Turing, 1950). However, history has shown that such questions cannot easily be resolved by proposing a formalizable alternative; hence the authors of the present paper remain divided upon the question of whether the (possible or actual) existence of AMAs is settled or otherwise.

3.2 Points of Agreement

There are autonomous machines, currently operational in the world, that have an ethical impact on human beings; this proposition no one can dispute. Frighteningly, there are even reports of autonomous machines that have been deployed in the battlefield and are making literal *life-and-death* decisions.[16] Given the state of the art in machine-implemented ethical reasoning, the authors would be **extremely** wary of machines being given such high levels of autonomy.[17] Regardless of whether a machine (now or in the future) is acknowledged to be an artificial moral agent or not, human beings should not be able to abrogate their own (moral) responsibility for designing machines with ethical impact.

The state of the-art is currently unsettled as to which approach would be best suited for designing machines able to carry out genuine ethical reasoning. There are efforts to design machines that are ethically correct by construction, machines that can be validated (and even formally verified) for correctness, and machines that attempt to be ethically correct through non-symbolic approaches. This "unsettled" status is not a bad thing. It indicates that there is a lot of on-going experimentation, with very different, fertile ideas. Of course, lack of settlement in the overall intellectual landscape does not entail that, in no researcher's minds, the core questions remain unsettled. As cannot be denied given what they said above, things are rather firmly settled in the minds of B&G, for better or worse.

[16] https://www.newscientist.com/article/2278852-drones-may-have-attacked-humans-fully-autonomously-for-the-first-time/

[17] B&G for present purposes are willing to countenance use here of an intuitive concept of autonomy, but note that as formalists, until autonomy is formalized, we can't really *know* that the machines in question are truly autonomous. This borders on self-incrimination, since — recall above — PAID employs the relation ***autonomous***.

https://kryten.mm.rpi.edu/SynthesizedPositionAutomationMachineEthics.pdf

Perspective

A roadmap for evaluating moral competence in large language models

https://doi.org/10.1038/s41586-025-10021-1

Received: 22 May 2025

Accepted: 5 December 2025

Published online: 18 February 2026

Check for updates

Julia Haas[1]✉, Sophie Bridgers[1,7], Arianna Manzini[1,7], Benjamin Henke[2,3,7], Joshua May[4], Sydney Levine[5,6], Laura Weidinger[1], Murray Shanahan[1], Kristian Lum[5], Iason Gabriel[1] & William Isaac[1]

The question of whether large language models (LLMs) can exhibit moral capabilities is of growing interest and urgency, as these systems are deployed in sensitive roles such as companionship and medical advising, and will increasingly be tasked with making decisions and taking actions on behalf of humans. These trends require moving beyond evaluating for mere moral performance, the ability to produce morally appropriate outputs, to evaluating for moral competence, the ability to produce morally appropriate outputs based on morally relevant considerations. Assessing moral competence is critical for predicting future model behaviour, establishing appropriate public trust and justifying moral attributions. However, both the unique architectures of LLMs and the complexity of morality itself introduce fundamental challenges. Here we identify three such challenges: the facsimile problem, whereby models may imitate reasoning without genuine understanding; moral multidimensionality, whereby moral decisions are influenced by a range of context-sensitive relevant moral and non-moral considerations; and moral pluralism, which demands a new standard for globally deployed artificial intelligence. We provide a roadmap for tackling these challenges, advocating for a suite of adversarial and confirmatory evaluations that will enable us to work towards a more scientifically grounded understanding and, in turn, a more responsible attribution of moral competence to LLMs.

There is considerable scientific and public interest in whether LLMs exhibit moral capabilities[1–15]. This interest is fuelled by studies showing strong LLM performance on moral reasoning tasks[16] and that LLMs are perceived to be superior to humans in moral reasoning "along almost all dimensions"[17,18]. A key question in this domain is whether LLMs exhibit moral competence; that is, whether they generate appropriate moral outputs by recognizing and appropriately integrating relevant moral considerations, rather than merely producing morally appropriate outputs[15,19,20].

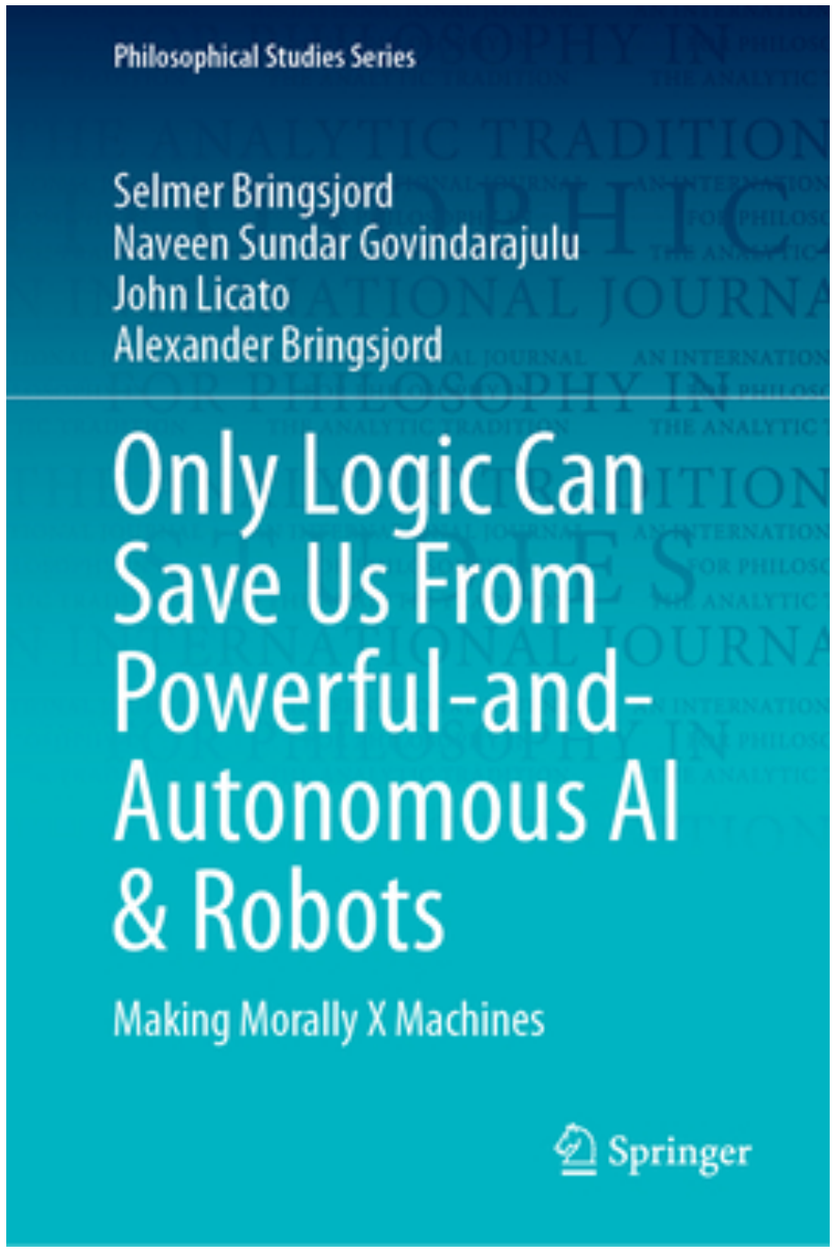
The widespread deployment of LLMs requires assessment of their moral competence, rather than their mere moral performance or people's perceptions of moral competence (Box 1). These systems are increasingly used for roles such as companionship[21], therapy[22] and providing medical advice[23]. Moreover, LLM adoption is projected to expand considerably in the coming years[24], with these systems increasingly powering capable artificial intelligence (AI) agents that take actions on behalf of humans[25–27]. These trends, coupled with evidence that LLMs reliably influence human decision-making and judgements[28–30], indicate the growing impact of LLMs in the moral domain.

While questions pertaining to machine morality are not new (Box 2), LLMs introduce substantial challenges to the field. Specifically, their distinctive architectures and emergent capabilities, combined with the inherent complexities of moral decision-making, pose several fundamental challenges for understanding and evaluating the moral competence of LLMs. We identify three such challenges: the facsimile problem, moral multidimensionality and the problem of pluralism in LLMs. Current assessment methods cannot fully address these challenges. However, progress is possible. For each challenge, we explore avenues for making headway and advocate for a suite of adversarial and confirmatory evaluation methods that can provide traction towards understanding and responsibly attributing moral competence in LLMs. Our overarching aim is to promote more robust and scientific evaluation standards, anticipating the need to equip civic stakeholders with evidence-based assessments on the basis of which to make informed recommendations.

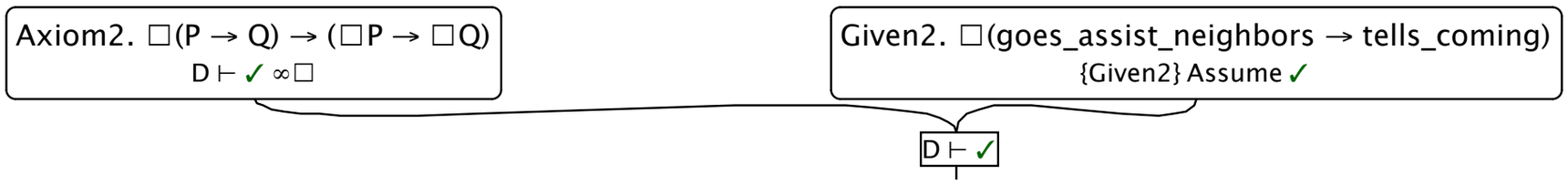
The facsimile problem

A fundamental challenge in cognitive science is inferring a system's unobserved, causal mechanisms from its observable behaviours[31]. This difficulty is compounded when evaluating moral competence in LLMs, as their distinctive architectures and training make it difficult to discern whether their outputs, even reliably acceptable outputs, rely on genuine moral reasoning or a mere facsimile process. We discuss

[1]Google DeepMind, London, UK. [2]Department of Computing, Imperial College London, London, UK. [3]Institute of Philosophy, School of Advanced Study, University of London, London, UK. [4]Department of Philosophy, University of Alabama at Birmingham, Birmingham, AL, USA. [5]Google DeepMind, New York, NY, USA. [6]Department of Psychology, New York University, New York, NY, USA. [7]These authors contributed equally: Sophie Bridgers, Arianna Manzini, Benjamin Henke. ✉e-mail: juliahaas@google.com


Philosophical Studies Series
Selmer Bringsjord
Naveen Sundar Govindarajulu
John Licato
Alexander Bringsjord
Only Logic Can Save Us From Powerful-and-Autonomous AI & Robots
Making Morally X Machines
Springer

But not **D**!!!: Chisholm's Paradox

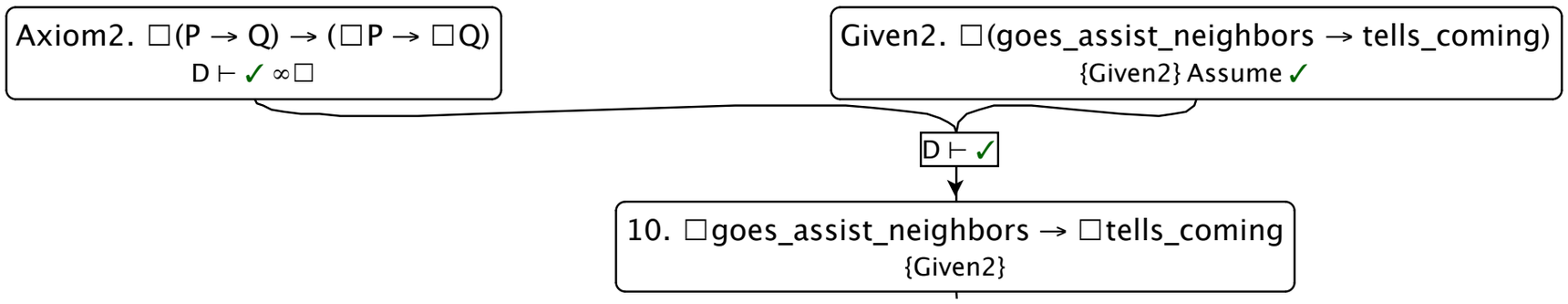


Axiom4. "Modus ponens for provability."
{Axiom4} Assume ✓

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

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

But not **D**!!!: Chisholm's Paradox

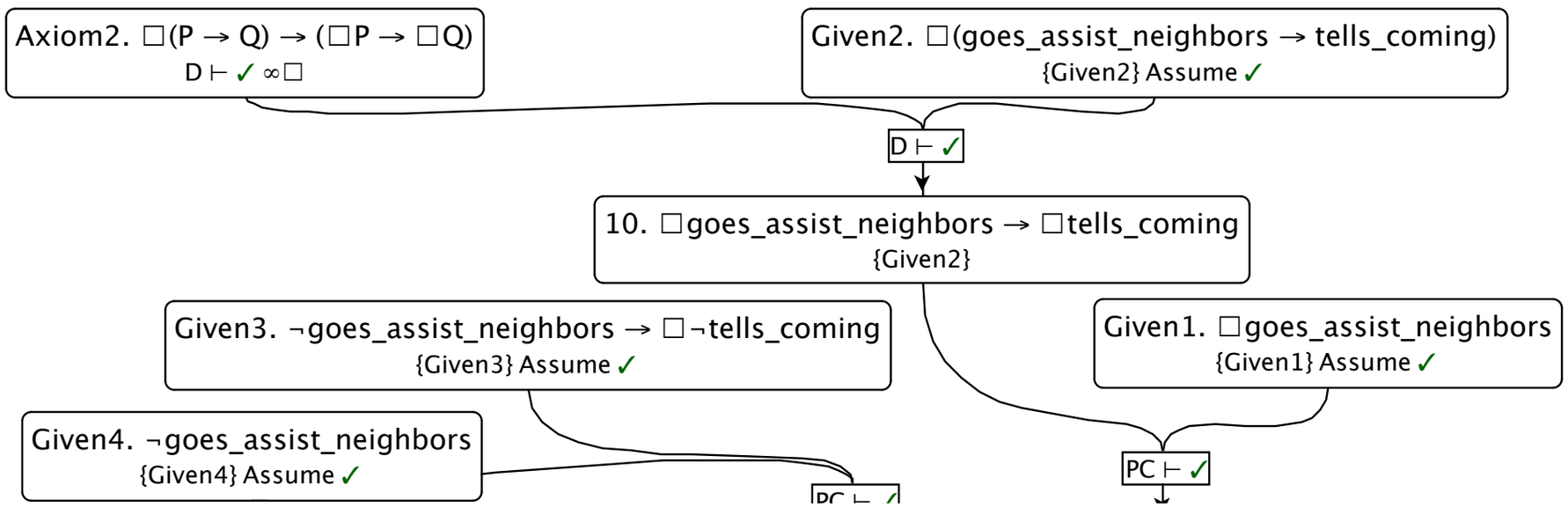


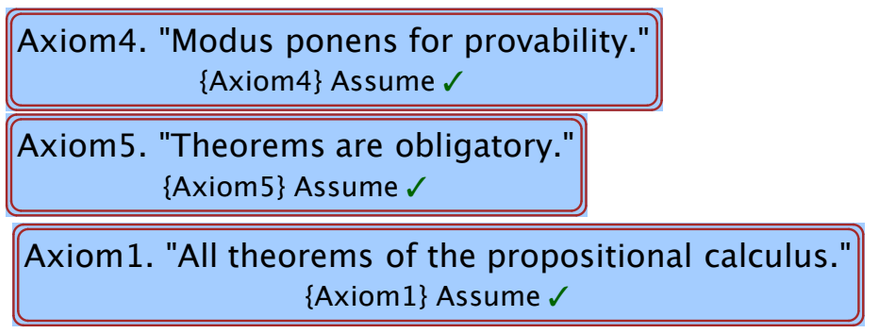
Axiom4. "Modus ponens for provability."
{Axiom4} Assume ✓

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

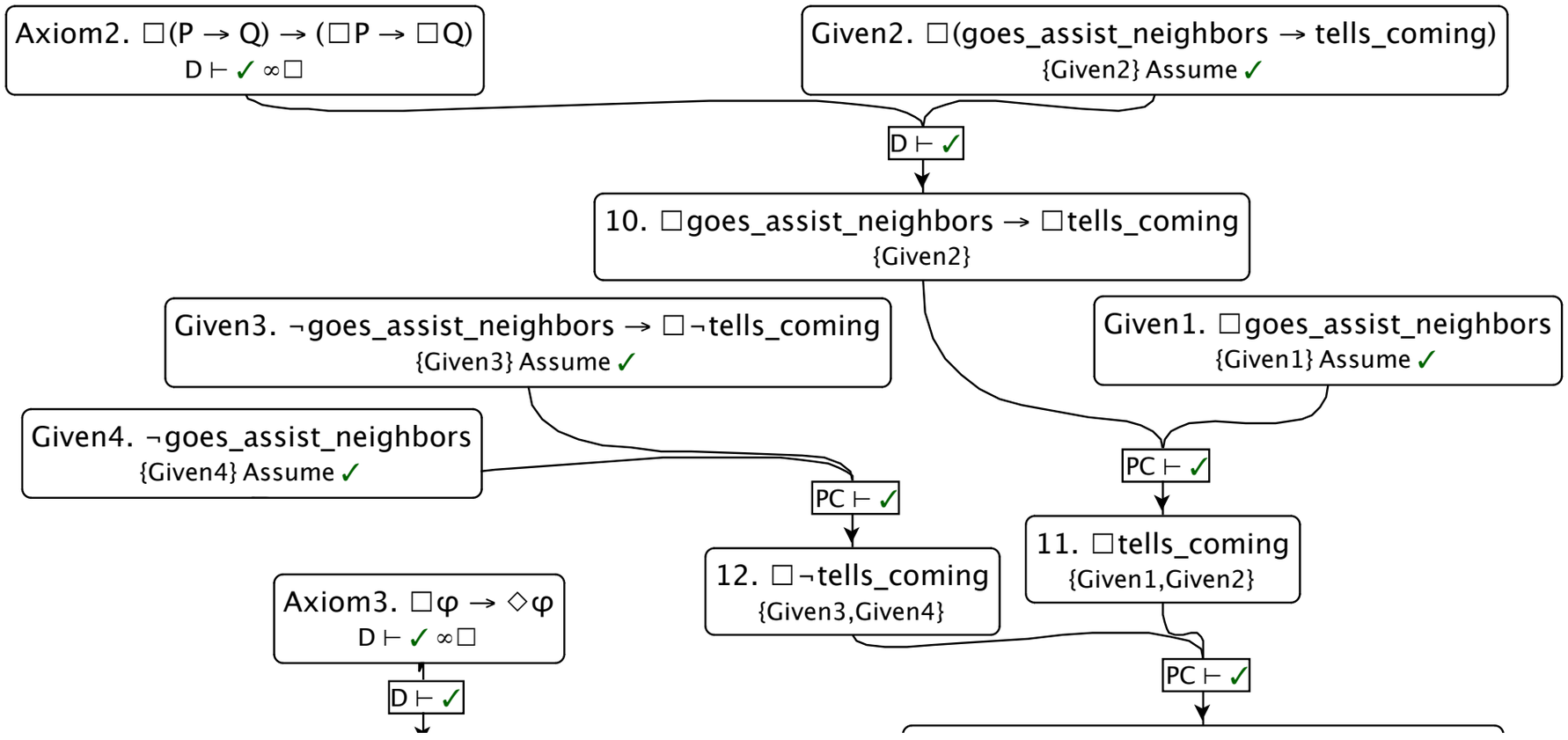
Axiom1. "All theorems of the propositional calculus."
{Axiom1} Assume ✓

But not **D**!!!: Chisholm's Paradox





But not **D**!!!: Chisholm's Paradox

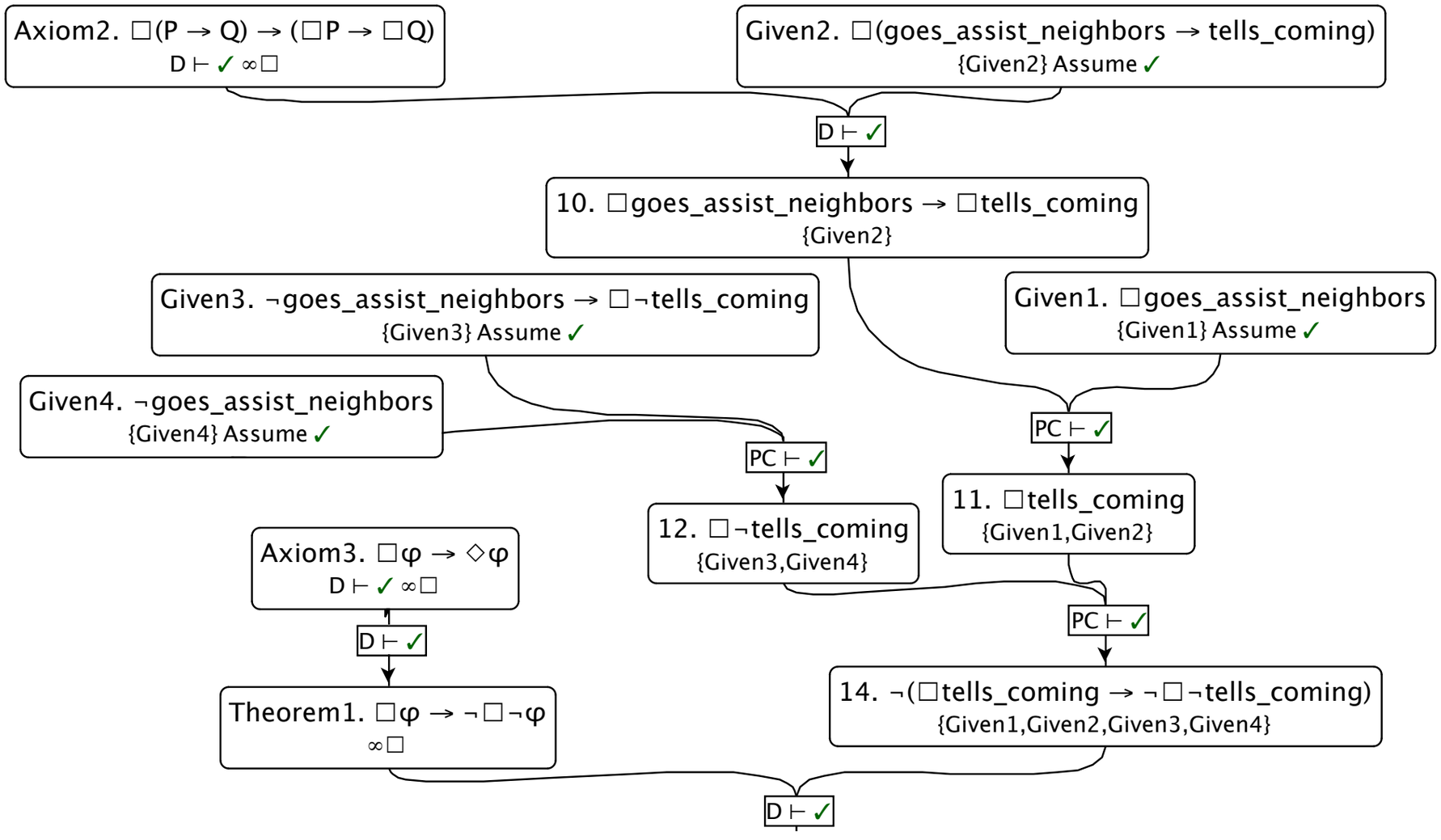


Axiom4. "Modus ponens for provability."
{Axiom4} Assume ✓

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

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

But not **D**!!!: Chisholm's Paradox

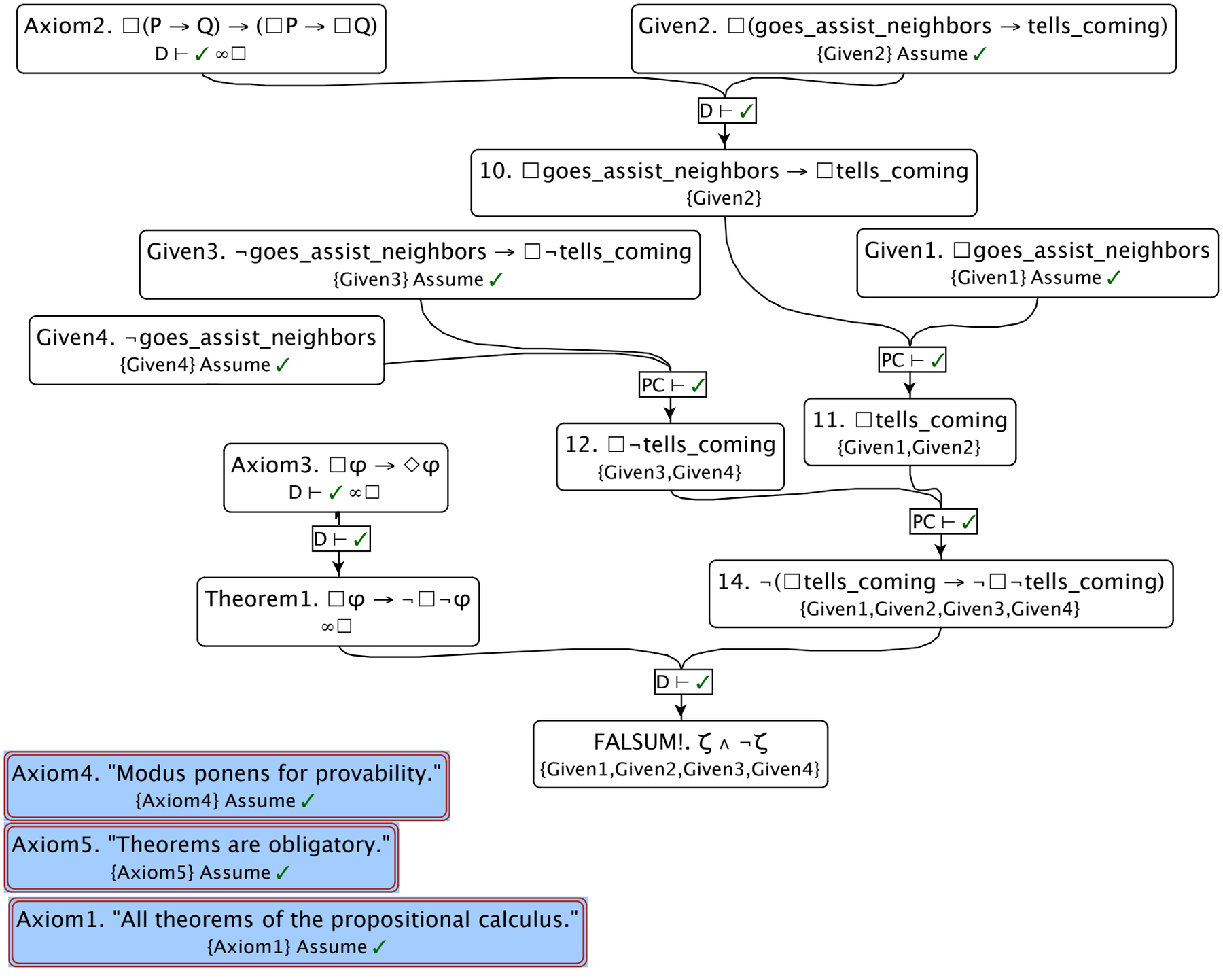


Axiom4. "Modus ponens for provability."
{Axiom4} Assume ✓

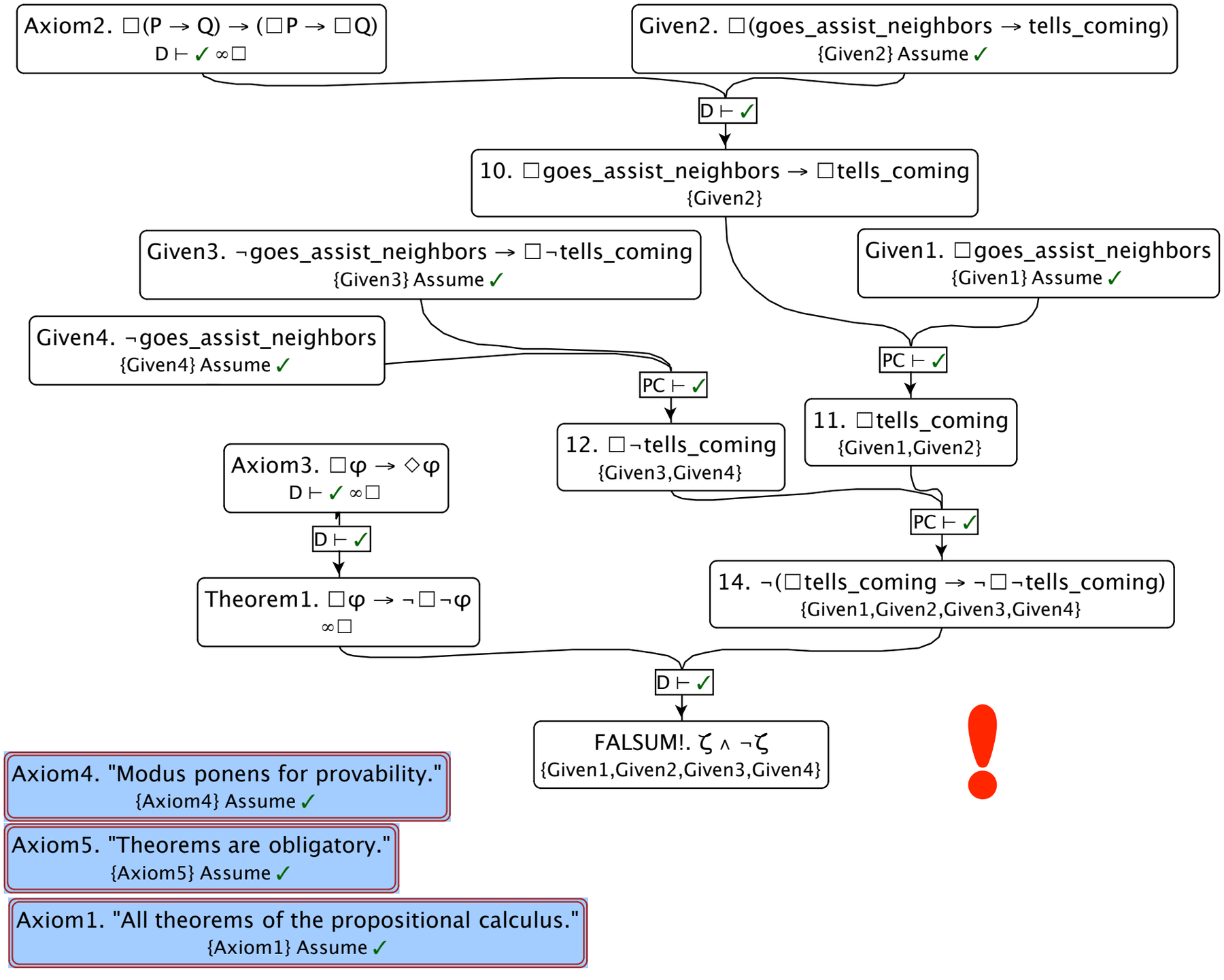
Axiom5. "Theorems are obligatory."
{Axiom5} Assume ✓

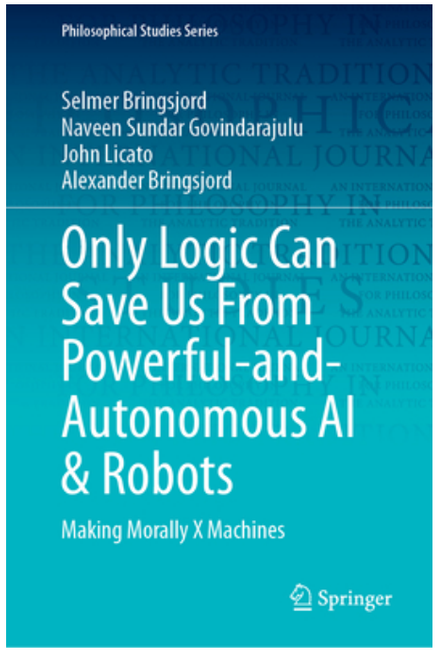
Axiom1. "All theorems of the propositional calculus."
{Axiom1} Assume ✓

But not **D**!!!: Chisholm's Paradox



But not **D**!!!: Chisholm's Paradox





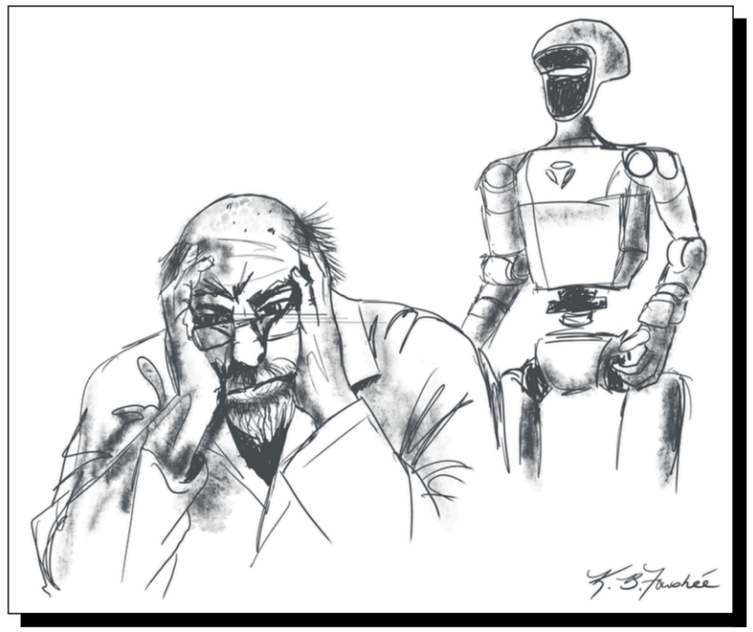
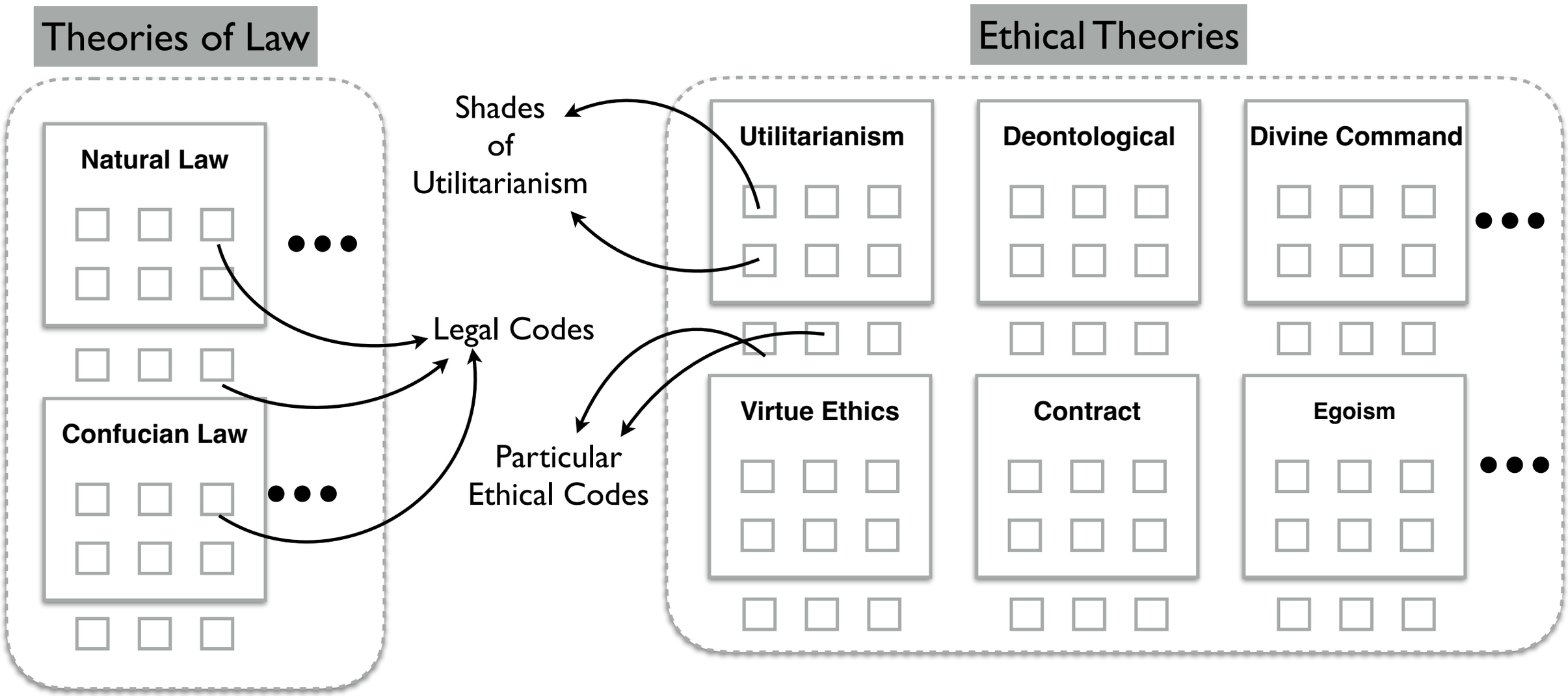
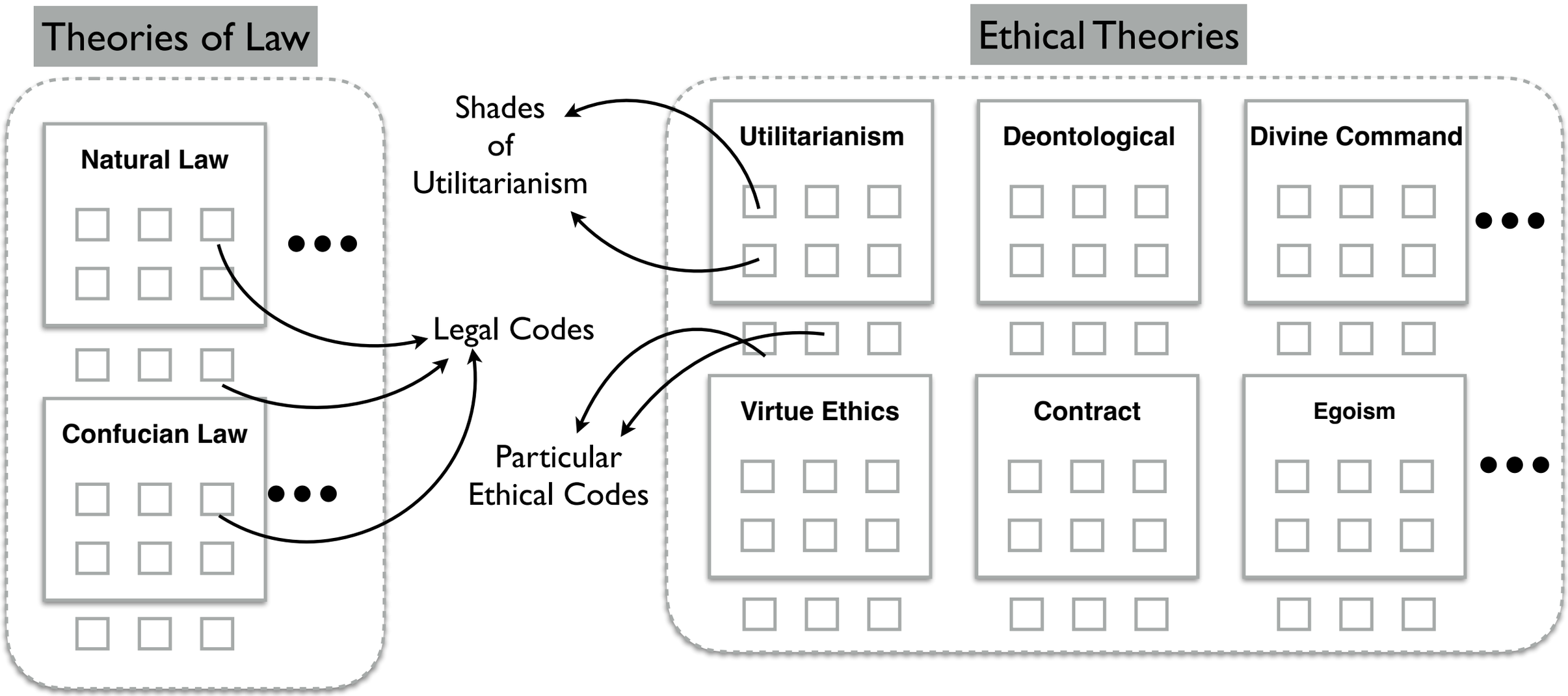


Figure 8.1: "Dr./ Frankenstein" Brought to Near Madness by His Own Uncontollable Creation. *This is an abstract rendering of the point in the story when Daystrom has been forced to face up to the fact that his beloved creation, M5, is not complying with human legal and moral codes that govern the* Enterprise, *and indeed all agents employed by the Federation. In this drawing, a modern stand-in for the Victor Frankenstein of Shelley's (1818) immortal novel, and Daystrom in the TOS episode, is shown — to put the matter mildly — fretting over the unwillingness of his AI creation to comply with the ethical and legal obligations and prohibitions promulgated and subscribed to by the Federation, which can be considered a laconic counterpart of the laws-of-war manual currently operative in the case of the United States and NATO (*Department of Defense Laws of War Manual *2023*).

Making Ethically Correct AIs, in Four "No-statML" Steps



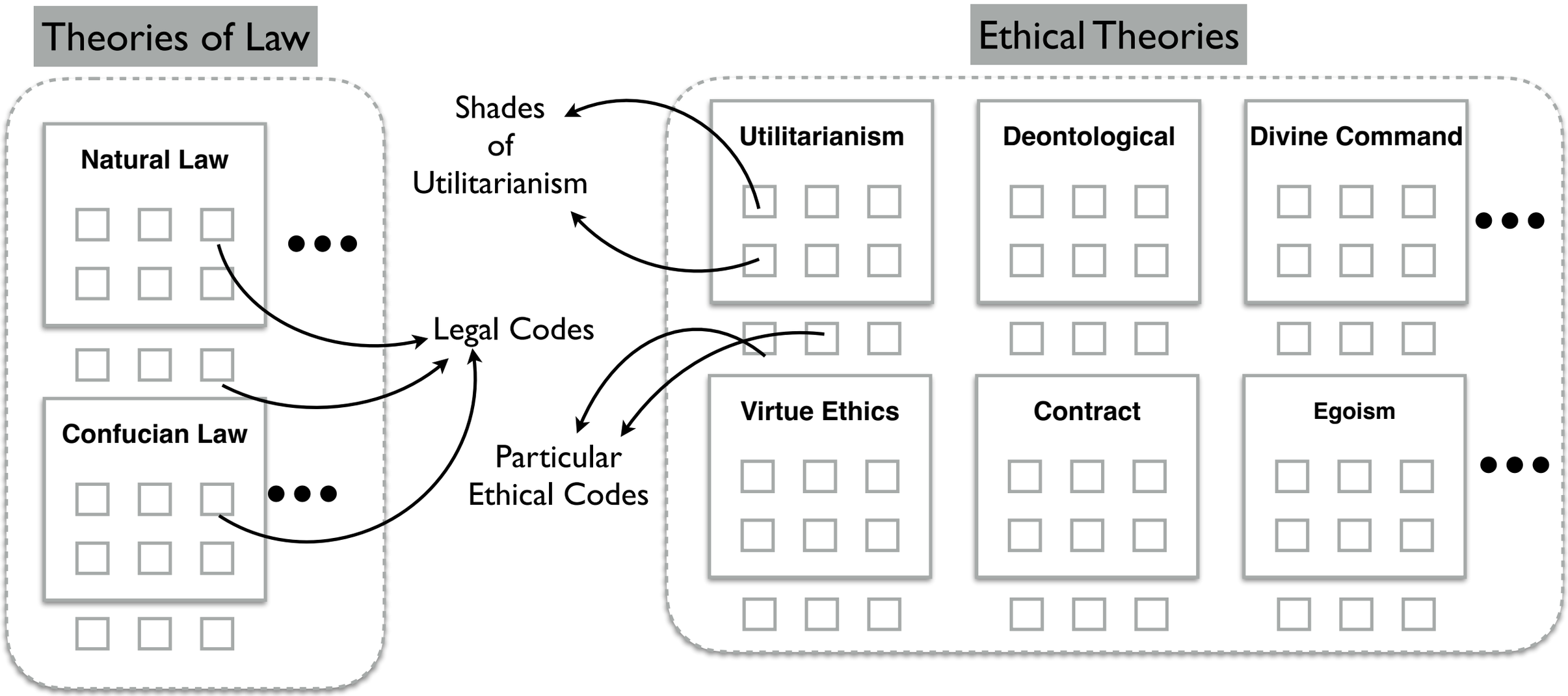
Making Ethically Correct AIs, in Four "No-statML" Steps



Step 1

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

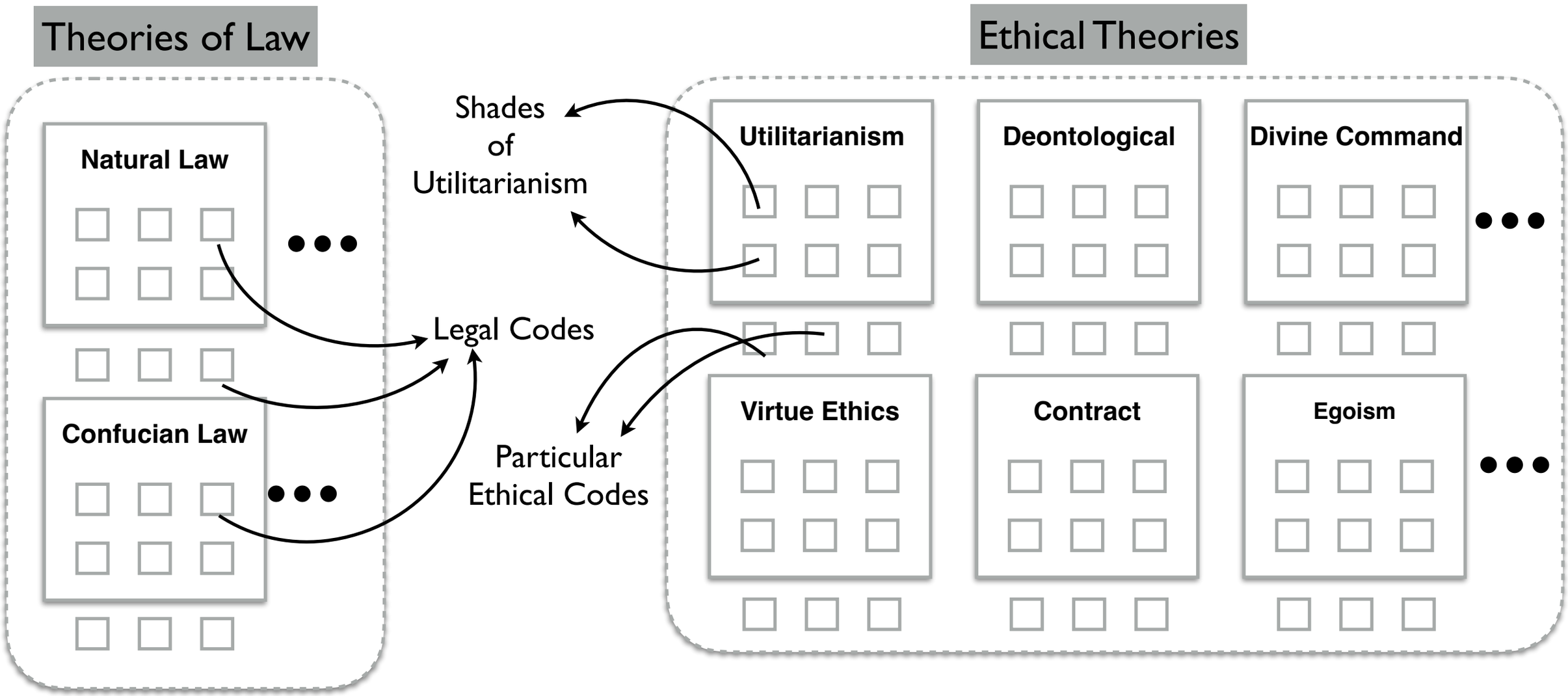
Making Ethically Correct AIs, in Four "No-statML" Steps



Step 1

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

Making Ethically Correct AIs, in Four "No-statML" Steps



Step 1

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

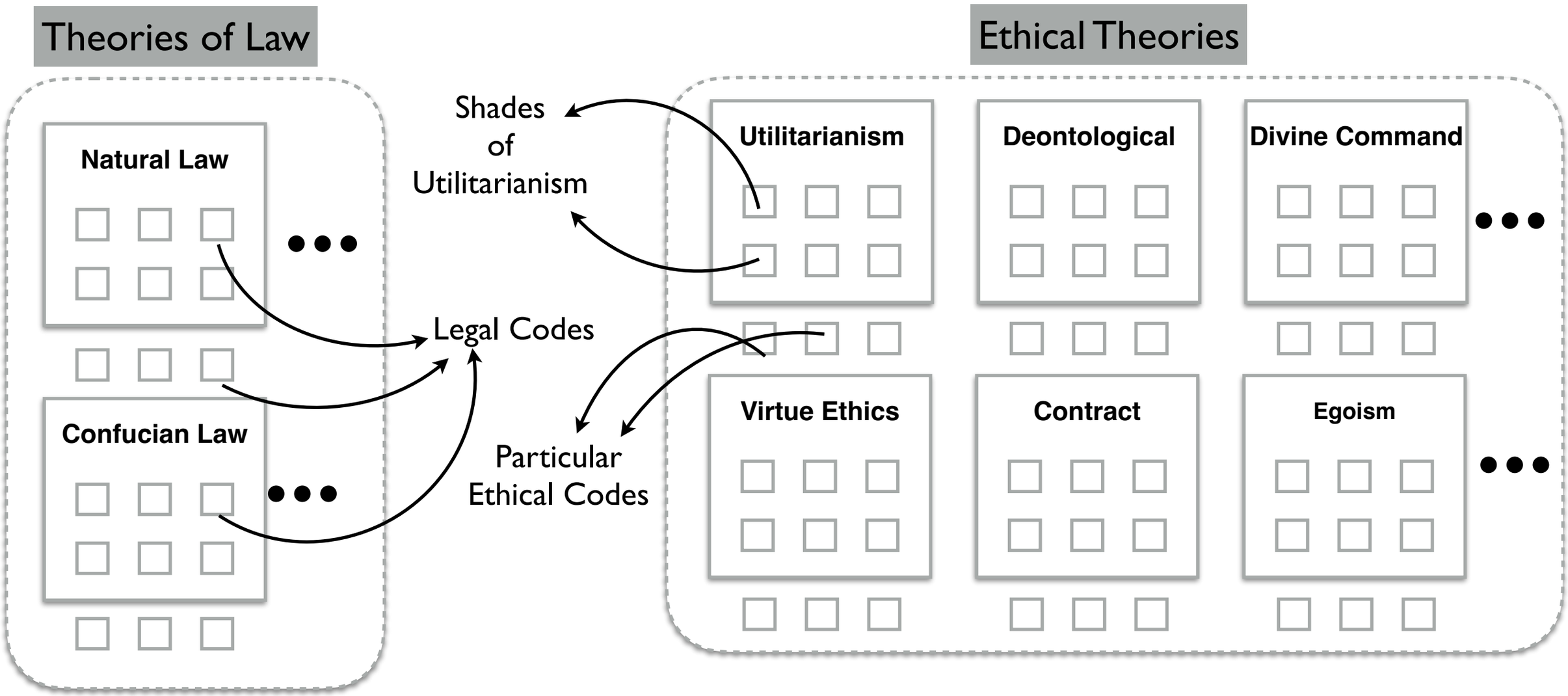
Step 2

Formalize & Automate

Shadow Prover

Spectra

Making Ethically Correct AIs, in Four "No-statML" Steps



Step 1

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

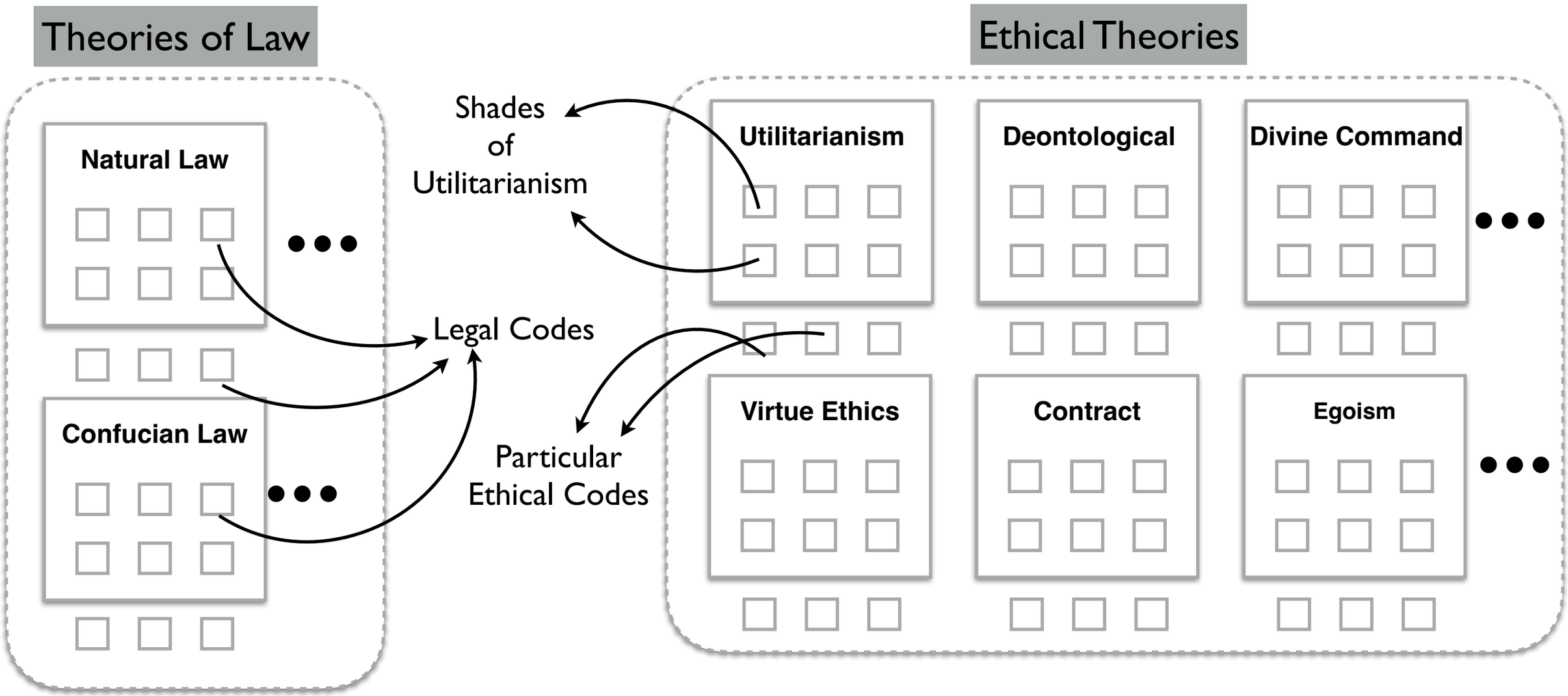
Step 2

Formalize & Automate

Shadow Prover

Spectra

Making Ethically Correct AIs, in Four "No-statML" Steps



Step 1

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

Step 2

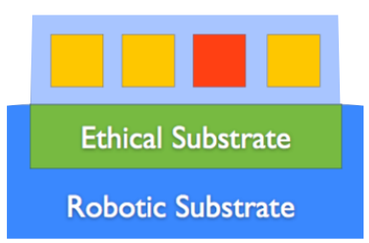
Formalize & Automate

Shadow Prover

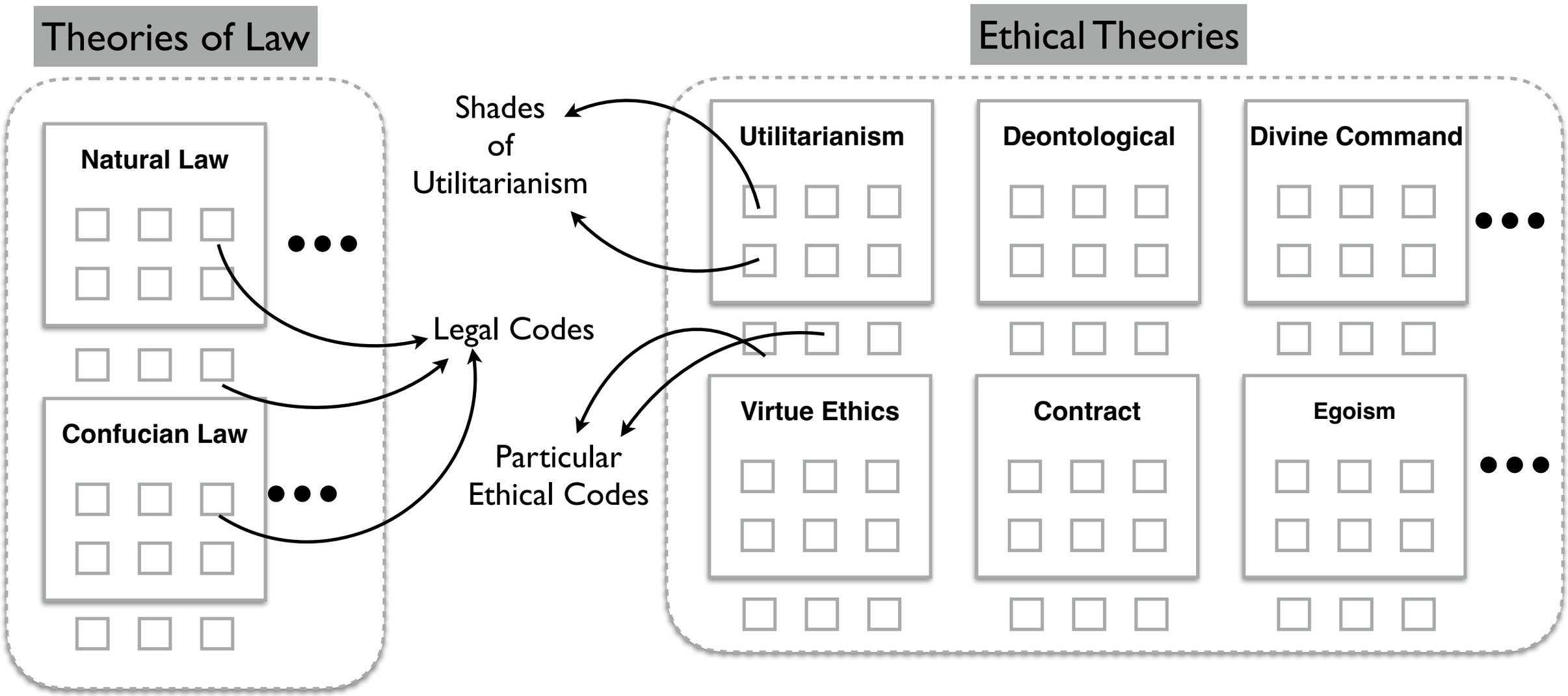
Spectra

Step 3

Ethical OS



Making Ethically Correct AIs, in Four "No-statML" Steps



Step 1

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

Step 2

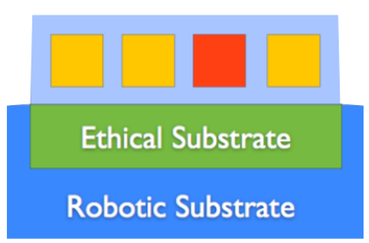
Formalize & Automate

Shadow Prover

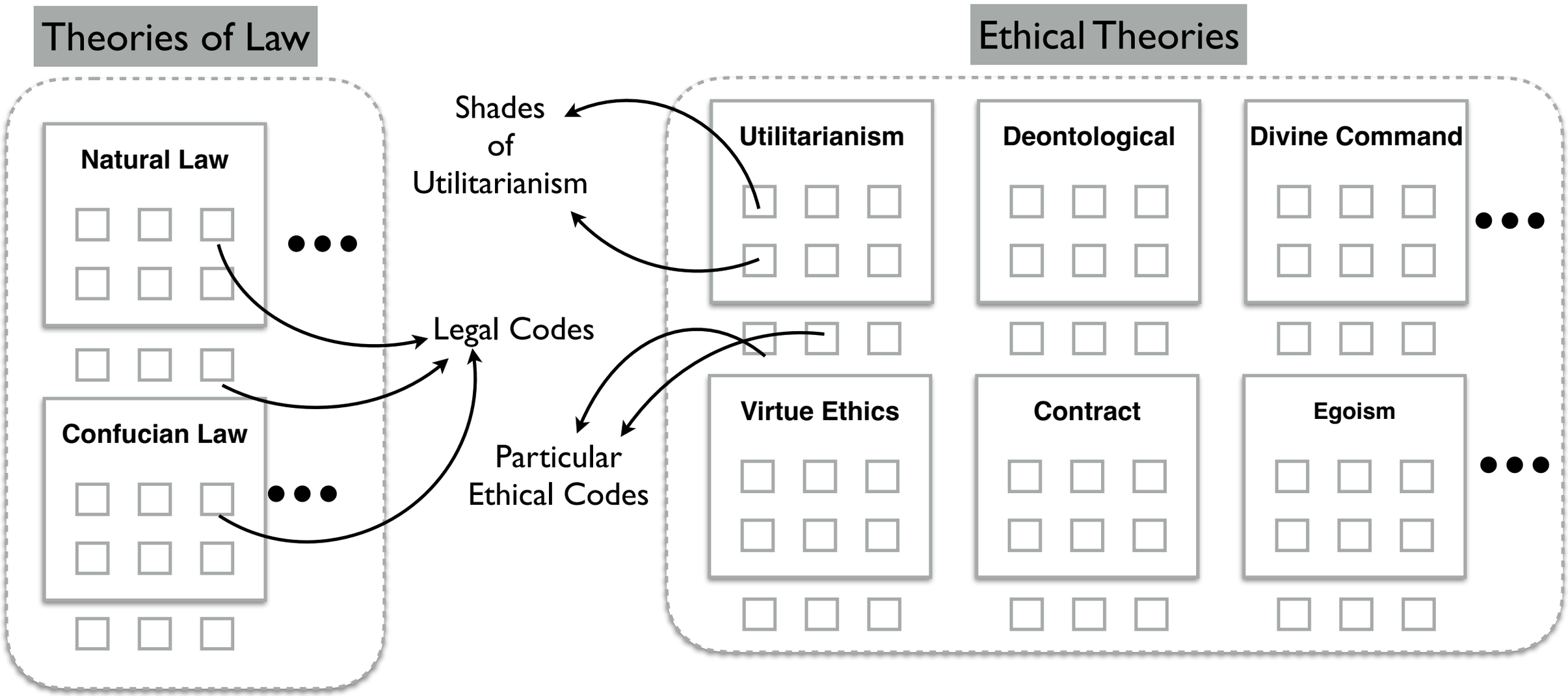
Spectra

Step 3

Ethical OS



Making Ethically Correct AIs, in Four "No-statML" Steps



Step 1

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

Step 2

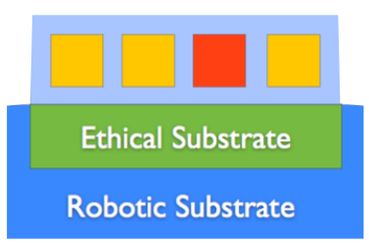
Formalize & Automate

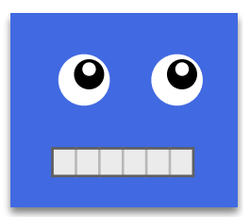
Shadow Prover

Spectra

Step 3

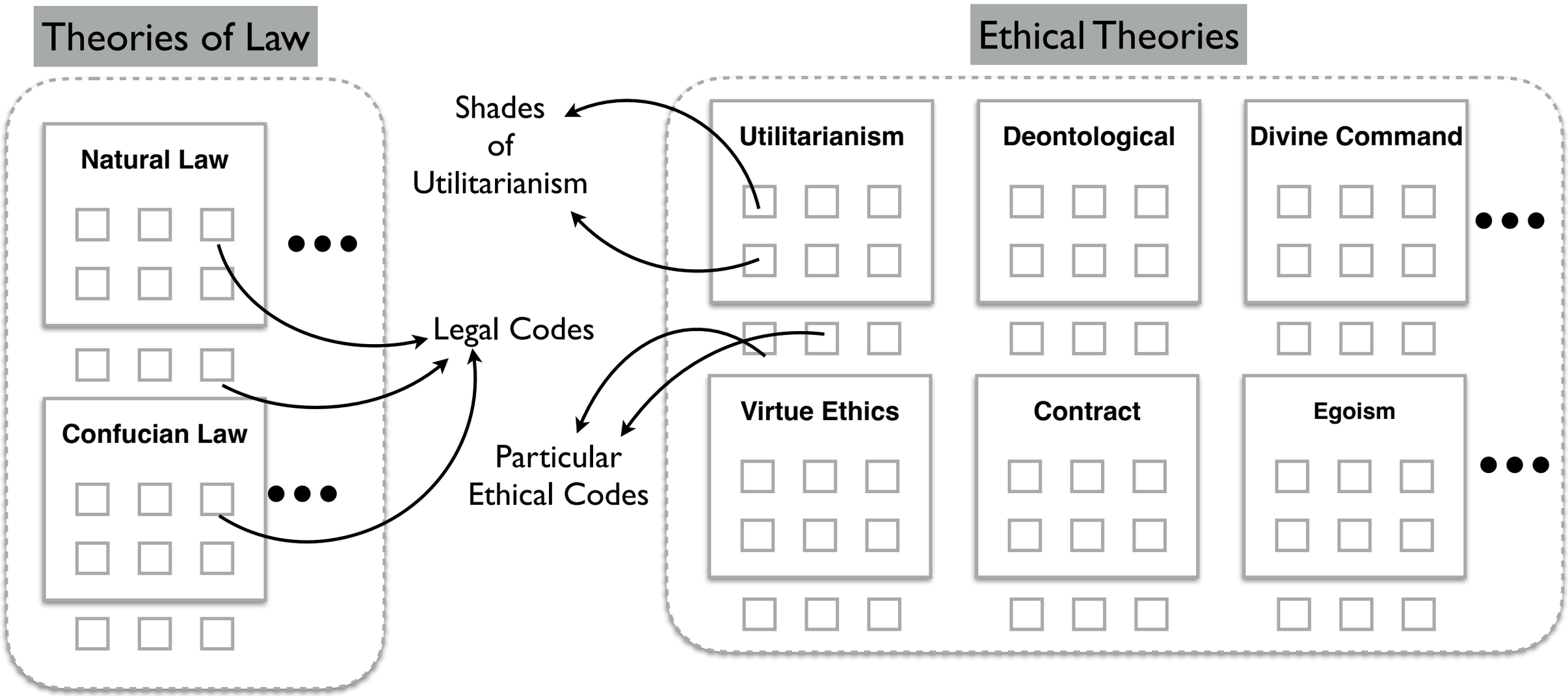
Ethical OS





An ethically correct robot.

Making Ethically Correct AIs, in Four "No-statML" Steps



Step 1

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

Step 2

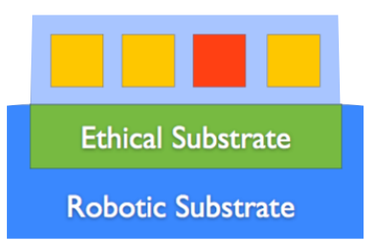
Formalize & Automate

Shadow Prover

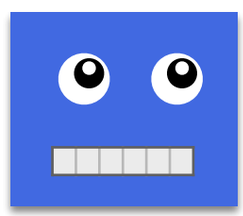
Spectra

Step 3

Ethical OS



DIARC/DoD/BMW …



An ethically correct robot.

Bare logikk can rede oss.

Bare logikk can rede oss.