

The PAID Problem; *Only Logic Can Save Us*

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

ILBAI
RPI
12/5/2024



The PAID Problem; *Only Logic Can Save Us*

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

ILBAI
RPI
12/5/2024



The PAID Problem; *Only Logic Can Save Us*

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

ILBAI
RPI
12/5/2024



Logic-&-AI In The News

How Students Can AI-Proof Their Careers

Artificial intelligence is going to eliminate a lot of jobs in the future. It's possible to reduce the risk that it will be yours.



ILLUSTRATION: OWEN GENT

By *James R. Hagerty*

[Follow](#)

Nov 20, 2024 11:00 a.m. ET

Logic-&-AI In The News

The current generation of college students is facing a challenge that those who came before never had to worry about: They'll be competing with AI for jobs.

What can they do to get ready?

After all, artificial intelligence is likely to eliminate at least some jobs that formerly served as first rungs on career ladders. "We have to accept and embrace the idea that in fact with AI we are going to have jobs that are going to be eliminated and jobs that are going to be created, and we don't know which ones," says Joseph E. Aoun, president of Northeastern University.

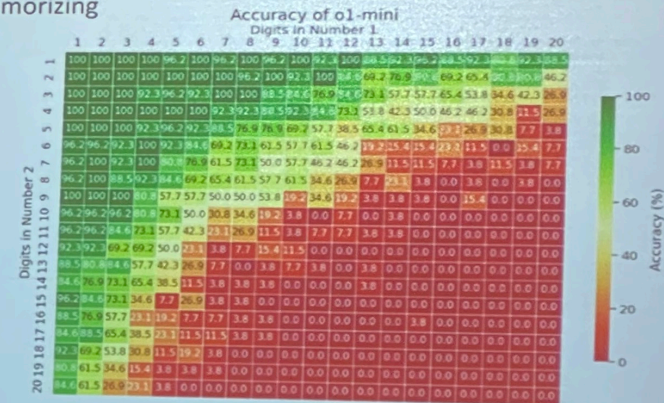
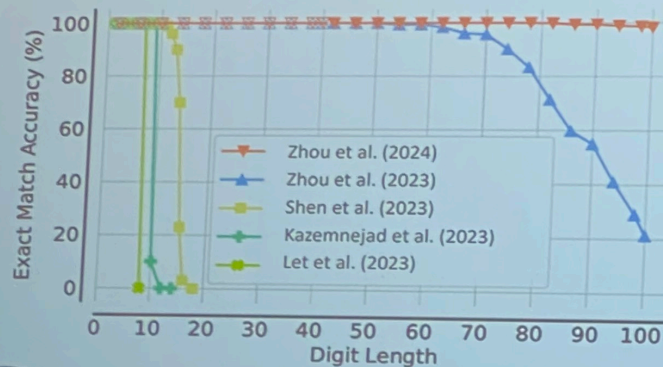
That uncertainty leaves today's college students struggling to prepare for a workplace that is changing faster than ever. We asked a range of career counselors and employers how they would suggest students AI-proof their careers. One consensus: It's important to master skills not easily matched by machines, such as human-style communications and the ability to understand and work smoothly with people who have different perspectives and personalities.

Logic-&-AI In The News

The current generation of college students is facing a challenge that those who came before never had to worry about: They'll be competing

Planning and reasoning remain hard problems

- SOTA LLMs aren't good at planning or critiquing plans, only good at producing high-level planning knowledge
 - In study, only 12% of GPT4 plans worked.
- Gemini combines RL with transformers. The combination suggests better reasoning potential, but we're not (yet) seeing leap ahead capability.
- LLMs aren't learning how to add or multiply; they are memorizing

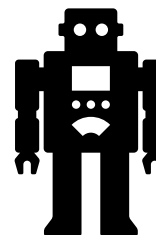


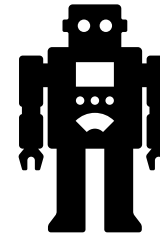
human-style communications and the ability to understand and work smoothly with people who have different perspectives and personalities.

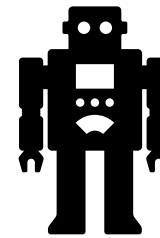
A key distinction

(reminder of which made eg last month @ RP2024 by W. Wallach)

...



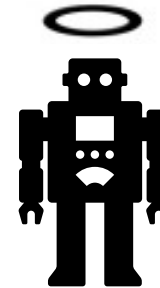




AI Ethics as Extension of
“Computer Ethics”:
What ought the *human* to
do in creating/using AI?

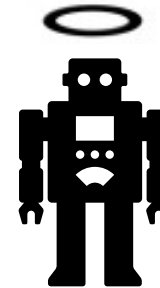


AI Ethics as Extension of
“Computer Ethics”:
What ought the *human* to
do in creating/using AI?





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

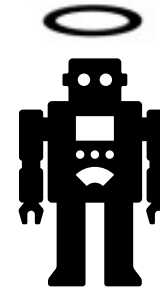


Machine Ethics/Roboethics:
How do we ensure that AI are
themselves ethically correct?



1

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



2

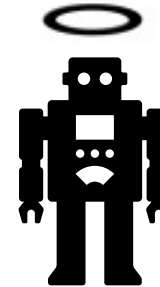
Machine Ethics/Roboethics:
How do we ensure that AI are
themselves ethically correct?



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.



2

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



1

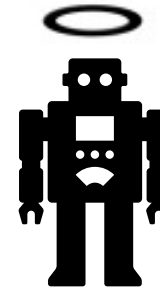
AI Ethics as Extension of
“Computer Ethics”:
What ought the *human* to
do in creating/using AI?

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

DOD Adopts Ethical Principles for Artificial Intelligence

Feb. 24, 2020 | f X ↗

The U.S. Department of Defense officially adopted a series of ethical principles for the use of Artificial Intelligence today following recommendations provided to Secretary of Defense Dr. Mark T. Esper by the Defense Innovation Board last October.



2

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



1

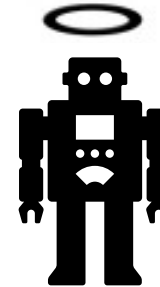
AI Ethics as Extension of
“Computer Ethics”:
What ought the *human* to
do in creating/using AI?

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

DOD Adopts Ethical Principles for Artificial Intelligence

Feb. 24, 2020 | f X ↗

The U.S. Department of Defense officially adopted a series of ethical principles for the use of Artificial Intelligence today following recommendations provided to Secretary of Defense Dr. Mark T. Esper by the Defense Innovation Board last October.



2

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



1

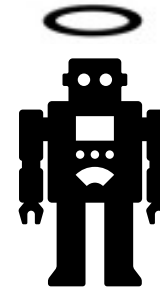
AI Ethics as Extension of
“Computer Ethics”:
What ought the *human* to
do in creating/using AI?

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

DOD Adopts Ethical Principles for Artificial Intelligence

Feb. 24, 2020 | f X ↗

The U.S. Department of Defense officially adopted a series of ethical principles for the use of Artificial Intelligence today following recommendations provided to Secretary of Defense Dr. Mark T. Esper by the Defense Innovation Board last October.



2

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

Firmly founded circa 2005.

Circa 2005; “Selmer, that’s really strange.”

Machine Ethics

Toward a General Logician Methodology for Engineering Ethically Correct Robots

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

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 method-

ology 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

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 [computists]. 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. Copyright © 2005, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

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 formulate 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

We need ethically correct
robots because of ...

The **PAID** Problem ...

The **PAID** Problem

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 ...

▶ 1:32

The **PAID** Problem

For all agents (whether artificial or natural like us) **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 (whether artificial or natural like us) α :

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha)/\mathbf{D}estroy_Us]$$

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 ...)

▶ 1:32

The **PAID** Problem

For all agents (whether artificial or natural like us) α :

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha)/\mathbf{D}estroy_Us]$$

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 ...)

▶ 1:32

The **PAID** Problem

For all agents (whether artificial or natural like us) α :

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha)/\mathbf{D}estroy_Us]$$

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

Ultimately, we need theorems, and they are starting to arrive — but presumably out of scope @ this workshop, yet ... $A(\alpha) \uparrow \vdash \mathbf{Trust}(h, \alpha) \downarrow$.

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 ...)

▶ 1:32

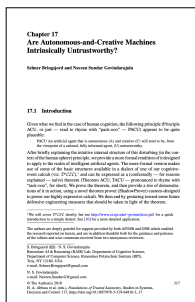
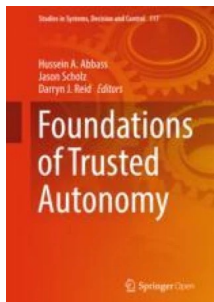
The **PAID** Problem

For all agents (whether artificial or natural like us) α :

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha)/\mathbf{D}estroy_Us]$$

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

Ultimately, we need theorems, and they are starting to arrive — but presumably out of scope @ this workshop, yet ... $\mathbf{A}(\alpha) \uparrow \vdash \mathbf{T}rust(h, \alpha) \downarrow$.



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 ...)

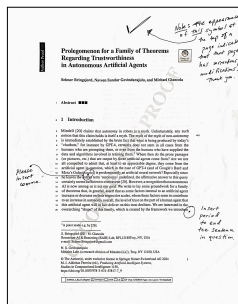
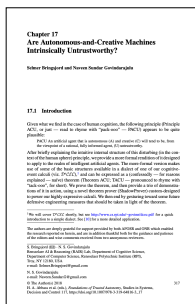
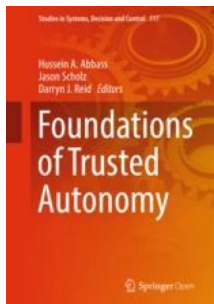
The **PAID** Problem

For all agents (whether artificial or natural like us) α :

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha) / \mathbf{D}estroy_Us]$$

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

Ultimately, we need theorems, and they are starting to arrive — but presumably out of scope @ this workshop, yet ... $\mathbf{A}(\alpha) \uparrow \vdash \mathbf{Trust}(\mathbf{h}, \alpha) \downarrow$.



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 ...)

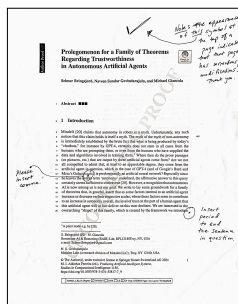
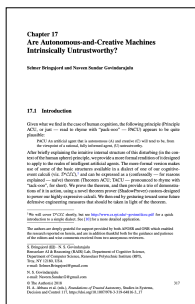
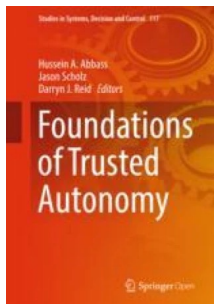
The **PAID** Problem

For all agents (whether artificial or natural like us) α :

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha) / \mathbf{D}estroy_Us]$$

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

Ultimately, we need theorems, and they are starting to arrive — but presumably out of scope @ this workshop, yet ... $\mathbf{A}(\alpha) \uparrow \vdash \mathbf{Trust}(\mathbf{h}, \alpha) \downarrow$.



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 ...)

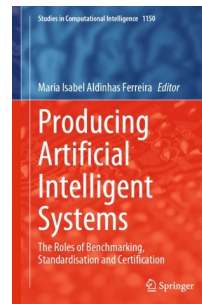
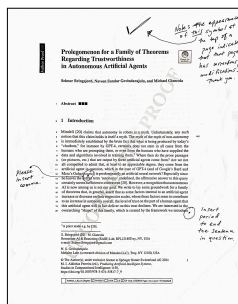
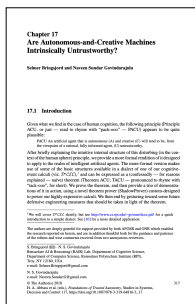
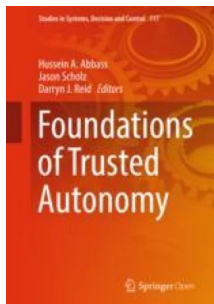
The **PAID** Problem

For all agents (whether artificial or natural like us) α :

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha)/\mathbf{D}estroy_Us]$$

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

Ultimately, we need theorems, and they are starting to arrive — but presumably out of scope @ this workshop, yet ... $\mathbf{A}(\alpha) \uparrow \vdash \mathbf{Trust}(\mathbf{h}, \alpha) \downarrow$.



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 ...)

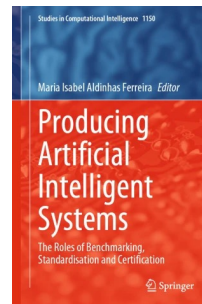
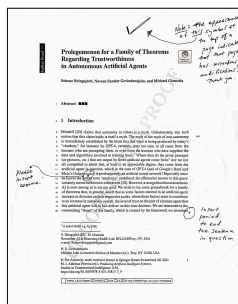
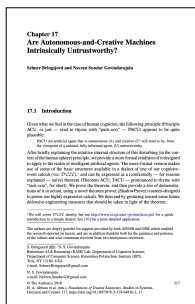
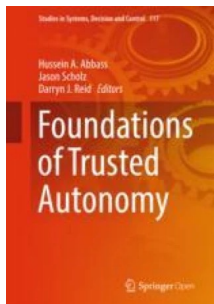
The **PAID** Problem

For all agents (whether artificial or natural like us) α :

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha)/\mathbf{D}estroy_Us]$$

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

Ultimately, we need theorems, and they are starting to arrive — but presumably out of scope @ this workshop, yet ... $\mathbf{A}(\alpha) \uparrow \vdash \mathbf{T}rust(h, \alpha) \downarrow$.



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 ...)

$\forall \mathbf{a}$

$[\mathbf{P}owerful(\mathbf{a}) \wedge \mathbf{A}utonomous(\mathbf{a}) \wedge \mathbf{I}ntelligent(\mathbf{a})] \rightarrow \mathbf{D}angerous(\mathbf{a})/\mathbf{D}estroy_Us]$

$\forall \alpha$

$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha) / \mathbf{D}estroy_Us]$

Self-Programmings; Formal Shades Thereof



$\forall \alpha$

$$[\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha) / \mathbf{D}estroy_Us]$$

Self-Programmings; Formal Shades Thereof



$$\forall \alpha \quad [\mathbf{P}owerful(\alpha) \wedge \mathbf{A}utonomous(\alpha) \wedge \mathbf{I}ntelligent(\alpha)] \rightarrow \mathbf{D}angerous(\alpha) / \mathbf{D}estroy_Us]$$

Paradigms in the Science of Universal Intelligence

Paradigms in the Science of Universal Intelligence

- **Universal Artificial Intelligence (UAI); AIXI**

Paradigms in the Science of Universal Intelligence

- **Universal Artificial Intelligence (UAI); AIXI**
 - Hutter & Legg

Paradigms in the Science of Universal Intelligence

- **Universal Artificial Intelligence** (UAI); AIXI
 - Hutter & Legg
- **Universal Cognitive Intelligence**; Λ + hierarchies

Paradigms in the Science of Universal Intelligence

- **Universal Artificial Intelligence** (UAI); AIXI
 - Hutter & Legg
- **Universal Cognitive Intelligence**; Λ + hierarchies
 - Bringsjord (& Govindarajulu (& Oswald & Bello))

Paradigms in the Science of Universal Intelligence

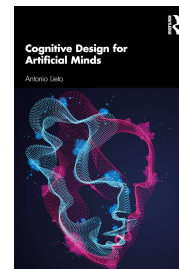
- **Universal Artificial Intelligence** (UAI); AIXI
 - Hutter & Legg
- **Universal Cognitive Intelligence**; Λ + hierarchies
 - Bringsjord (& Govindarajulu (& Oswald & Bello))
- **MCG+**

Paradigms in the Science of Universal Intelligence

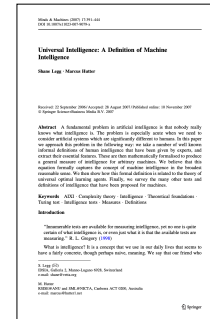
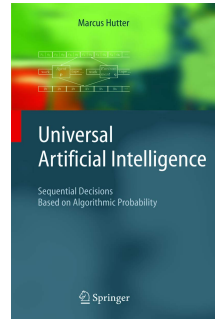
- **Universal Artificial Intelligence** (UAI); AIXI
 - Hutter & Legg
- **Universal Cognitive Intelligence**; Λ + hierarchies
 - Bringsjord (& Govindarajulu (& Oswald & Bello))
- **MCG+**
 - MCG *simplicter* completely from A. Lieto.

Paradigms in the Science of Universal Intelligence

- **Universal Artificial Intelligence** (UAI); AIXI
 - Hutter & Legg
- **Universal Cognitive Intelligence**; Λ + hierarchies
 - Bringsjord (& Govindarajulu (& Oswald & Bello))
- **MCG+**
 - MCG *simplicter* completely from A. Lieto.

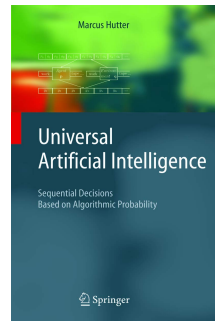


UAI



$$\Upsilon[\mathbf{a}] = \sum_{\mu \in Env\mathbf{s}} weight \cdot reward(\mathbf{a}, \mu)$$

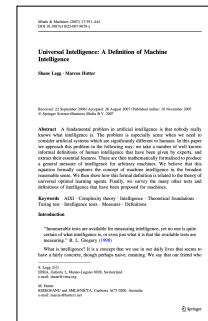
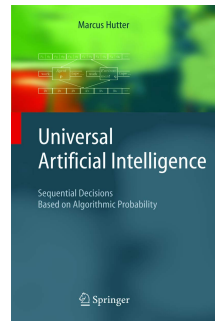
UAI



$$\Upsilon[\mathbf{a}] = \sum_{\mu \in \mathcal{Env}s} weight \cdot reward(\mathbf{a}, \mu)$$

But this is consistent w/ super-intelligent agents can knowing absolutely nothing!

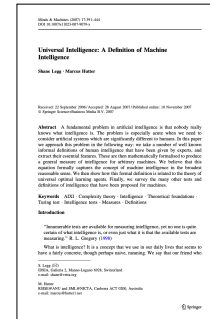
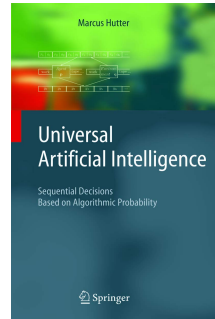
UAI



$$\Upsilon[\mathbf{a}] = \sum_{\mu \in \mathcal{Env}s} weight \cdot reward(\mathbf{a}, \mu)$$

But this is consistent w/ super-intelligent agents can knowing absolutely nothing!
And the environments can e.g. present agents with Turing-undecidable problems!

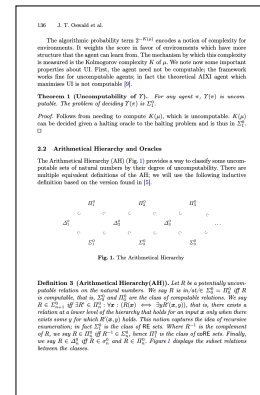
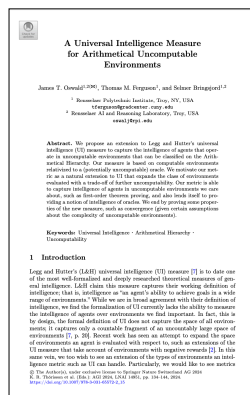
UAI



$$\Upsilon[\mathbf{a}] = \sum_{\mu \in Env\mathbf{s}} weight \cdot reward(\mathbf{a}, \mu)$$

But this is consistent w/ super-intelligent agents can knowing absolutely nothing!
And the environments can e.g. present agents with Turing-undecidable problems!

Oswald et al.



AGI 2024





GCI \Rightarrow UCI

In general, for a computational artifact \mathcal{C} to have GCI, we hold that it must produce a result ρ that is,

Significant by at least near-consensus among relevant humans, intrinsically significant;

Independent generated by a problem-solving run carried out to a high degree by \mathcal{C} independent of human insight and assistance; and

Innovative where this problem-solving run begins from a starting point ι that is a “long distance” from ρ .

We shall assume that λ applied to a pair (ι, ρ) yields a distance δ ; we therefore write

$$\lambda(\iota, \rho) = \delta.$$

To say that \mathcal{C} produces ρ having started with ι , we write

$$\mathcal{C} : \iota \longrightarrow \rho.$$

We shall further assume that the general space of inputs is ι^* , and the general space of results ρ^* . Under this notation, it can be informatively said that a good indicator of whether a result is significant is that the function f from ι^* to ρ^* is Turing-unsolvable. Were this indicator promoted to an absolute requirement, which is quite tempting, the first property of GCI could plausibly be formalized via something like the following equation as a necessary condition for this property (significance) to be possessed.⁷

$$\mathcal{C} : \iota \longrightarrow \rho \text{ where the function } f : \iota^* \longrightarrow \rho^* \text{ is Turing-unsolvable.} \quad (2)$$

⁷One must be careful here. Let h be a binary halting function taking as input the Gödel number n^M of a Turing machine M along with input m to that Turing machine. As is well-known, h is Turing-uncomputable. Yet there are individual Turing machines, accompanied by inputs to them, which can be instantly declared and proved to be either



GCI \Rightarrow UCI

In general, for a computational artifact \mathcal{C} to have GCI, we hold that it must produce a result ρ that is,

Significant by at least near-consensus among relevant humans, intrinsically significant;

Independent generated by a problem-solving run carried out to a high degree by \mathcal{C} independent of human insight and assistance; and

Innovative where this problem-solving run begins from a starting point ι that is a “long distance” from ρ .

We shall assume that λ applied to a pair (ι, ρ) yields a distance δ ; we therefore write

$$\lambda(\iota, \rho) = \delta.$$

To say that \mathcal{C} produces ρ having started with ι , we write

$$\mathcal{C} : \iota \longrightarrow \rho.$$

We shall further assume that the general space of inputs is ι^* , and the general space of results ρ^* . Under this notation, it can be informatively said that a good indicator of whether a result is significant is that the function f from ι^* to ρ^* is Turing-unsolvable. Were this indicator promoted to an absolute requirement, which is quite tempting, the first property of GCI could plausibly be formalized via something like the following equation as a necessary condition for this property (significance) to be possessed.⁷

$$\mathcal{C} : \iota \longrightarrow \rho \text{ where the function } f : \iota^* \longrightarrow \rho^* \text{ is Turing-unsolvable.} \quad (2)$$

⁷One must be careful here. Let h be a binary halting function taking as input the Gödel number n^M of a Turing machine M along with input m to that Turing machine. As is well-known, h is Turing-uncomputable. Yet there are individual Turing machines, accompanied by inputs to them, which can be instantly declared and proved to be either

UCI

Computational Approaches to Conscious Artificial Intelligence

No Access

Chapter 5

Universal Cognitive Intelligence, from Cognitive Consciousness, and Lambda (Λ)

Selmer Bringsjord, Naveen Sundar Govindarajulu, and James Oswald

https://doi.org/10.1142/9789811276675_0005

[< Previous](#)

[Next >](#)

[Tools](#) [Share](#)

Abstract:

We explain that the concept of *universal cognitive intelligence* (UCI) can be derived in part by generalization from the previously introduced (and axiomatized) *theory of cognitive consciousness*, and the framework, Λ , for measuring the degree of such consciousness in an agent at a given time. UCI (i) covers intelligence that is artificial or natural (or a hybrid thereof) in nature, and intelligence that is not merely Turing-level or less, but also beyond this level; (ii) reflects a psychometric orientation to AI; (iii) withstands a series of objections (including e.g. the opposing position of David Gamez on tests, intelligence, and consciousness, and the complaint that so-called “emotional intelligence” is beyond the reach of any logic-based framework, including thus UCI); and (iv) connects smoothly and symbiotically with important formal hierarchies (e.g., the Polynomial, Arithmetic, and Analytic Hierarchies), while at the same yielding its own new all-encompassing hierarchy of logic machines: $\mathfrak{A}\mathfrak{N}$. We end with an admission: UCI by our lights, for reasons previously published, cannot take account of any form of intelligence that genuinely exploits *phenomenal* consciousness.

Series on Machine Consciousness - Vol. 5



Editor
Antonio Chella

COMPUTATIONAL
APPROACHES TO CONSCIOUS
ARTIFICIAL INTELLIGENCE

World Scientific

UCI

Computational Approaches to Conscious Artificial Intelligence No Access

Chapter 5

Universal Cognitive Intelligence, from Cognitive Consciousness, and Lambda (Λ)

Selmer Bringsjord, Naveen Sundar Govindarajulu, and James Oswald

https://doi.org/10.1142/9789811276675_0005

< Previous Next >

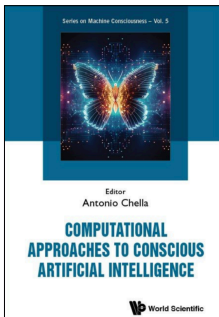
[Tools](#) [Share](#)

Abstract:

We explain that the concept of *universal cognitive intelligence* (UCI) can be derived in part by generalization from the previously introduced (and axiomatized) *theory of cognitive consciousness*, and the framework, Λ , for measuring the degree of such consciousness in an agent at a given time. UCI (i) covers intelligence that is artificial or natural (or a hybrid thereof) in nature, and intelligence that is not merely Turing-level or less, but also beyond this level; (ii) reflects a psychometric orientation to Λ ; (iii) withstands a series of objections (including e.g. the opposing position of David Gamez on tests, intelligence, and consciousness, and the complaint that so-called “emotional intelligence” is beyond the reach of any logic-based framework, including thus UCI); and (iv) connects smoothly and symbiotically with important formal hierarchies (e.g., the Polynomial, Arithmetic, and Analytic Hierarchies), while at the same yielding its own new all-encompassing hierarchy of logic machines: \mathfrak{ZM} .

We end with an admission: UCI by our lights, for reasons previously published, cannot take account of any form of intelligence that genuinely exploits *phenomenal* consciousness.

$$UCI = [f[a] \circ \Lambda[a]]$$

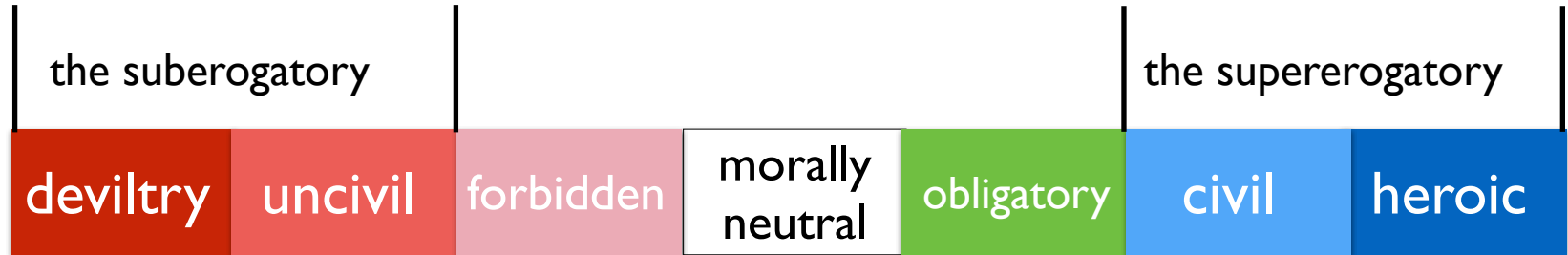


**But what is ethical
correctness?**

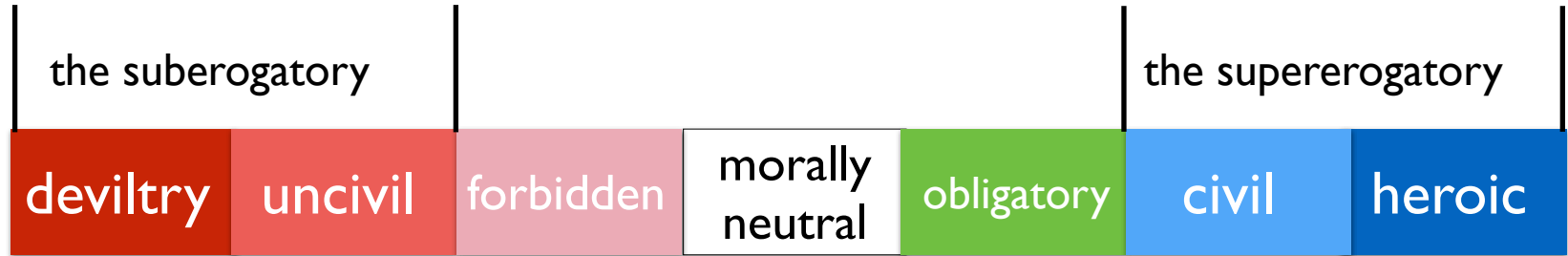
...

An agent a is ethically correct if and only if ...

An agent *a* is ethically correct if and only if ...

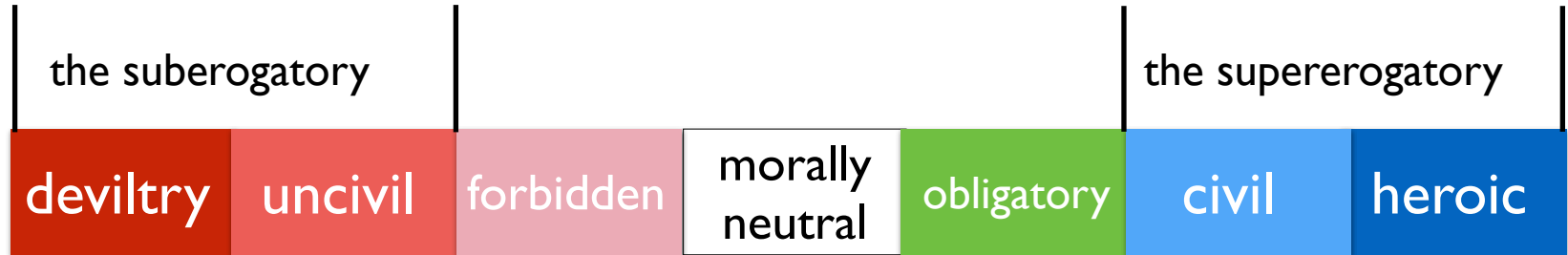


An agent a is ethically correct if and only if ...



Nothing morally forbidden is done by a .

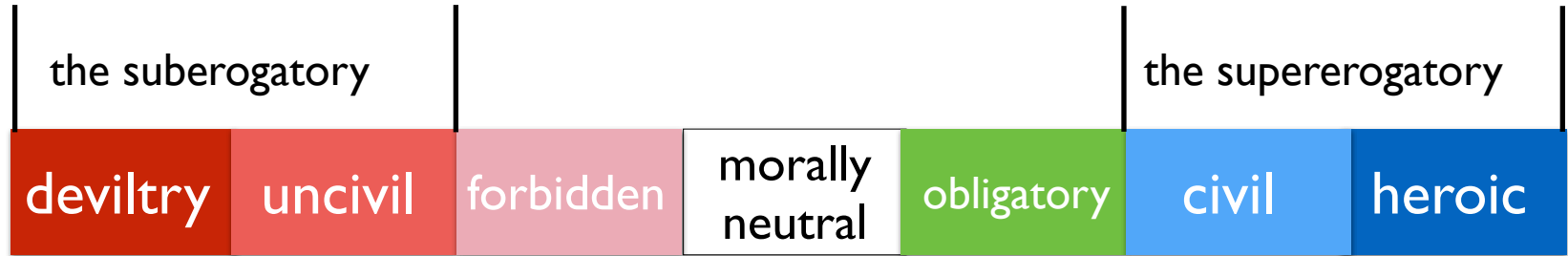
An agent a is ethically correct if and only if ...



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 a .

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

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

**Simplifying: Single Necessary Condition for
Verifiably Correct Ethical Correctness**

Simplifying: Single Necessary Condition for Verifiably Correct Ethical Correctness

If agent \mathfrak{A} is verifiably ethically correct, **then**, if it follows by valid reasoning from some body of information Φ that doing some action a is morally *impermissible*, and agent \mathfrak{A} is supplied with Φ , this agent can itself reason to the moral impermissibility of doing a (in verifiably valid fashion) from Φ .

Simplifying: Single Necessary Condition for Verifiably Correct Ethical Correctness

If agent \mathbf{a} is verifiably ethically correct, **then**, if it follows by valid reasoning from some body of information Φ that doing some action a is morally *impermissible*, and agent \mathbf{a} is supplied with Φ , this agent can itself reason to the moral impermissibility of doing a (in verifiably valid fashion) from Φ .

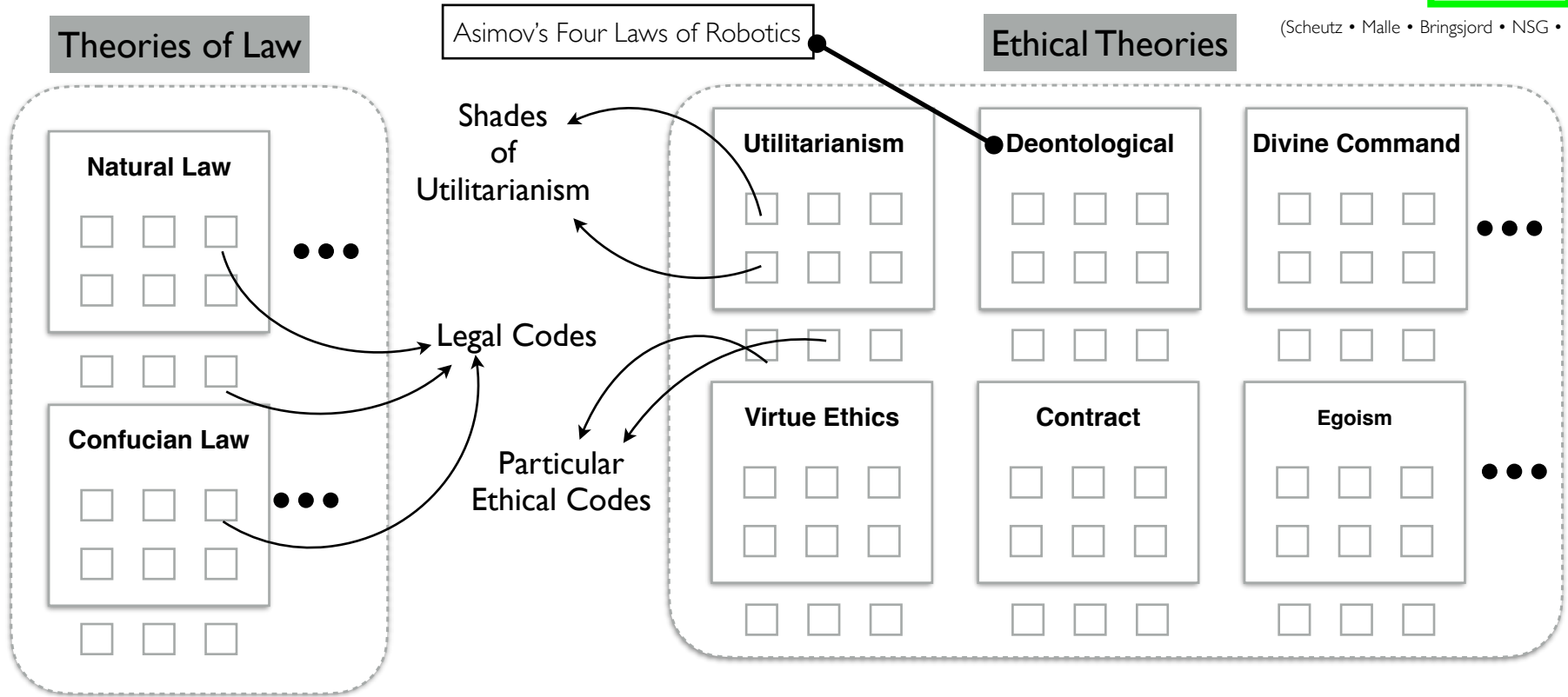
The PAID Problem *Solved*

...

The Four Steps

\$kM

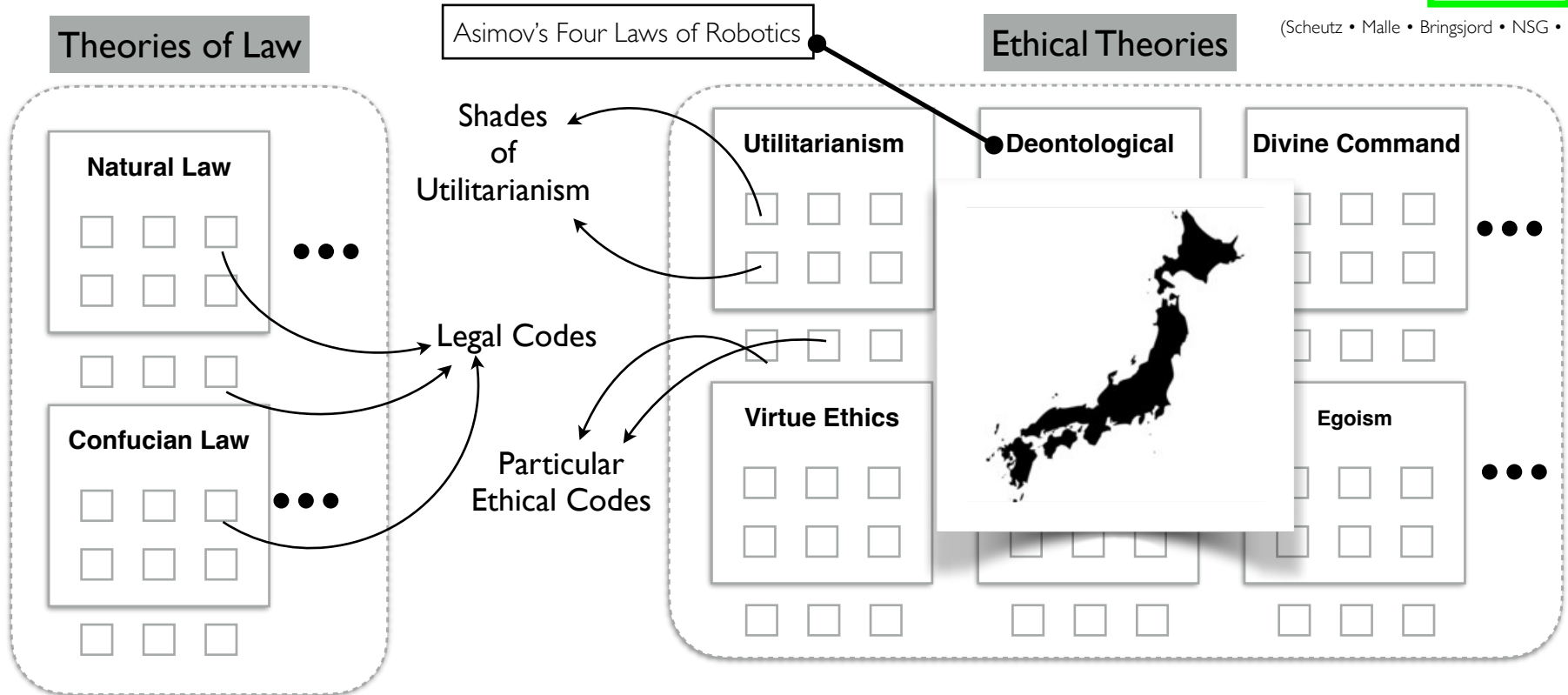
(Scheutz • Malle • Bringsjord • NSG • Bello)



The Four Steps

\$kM

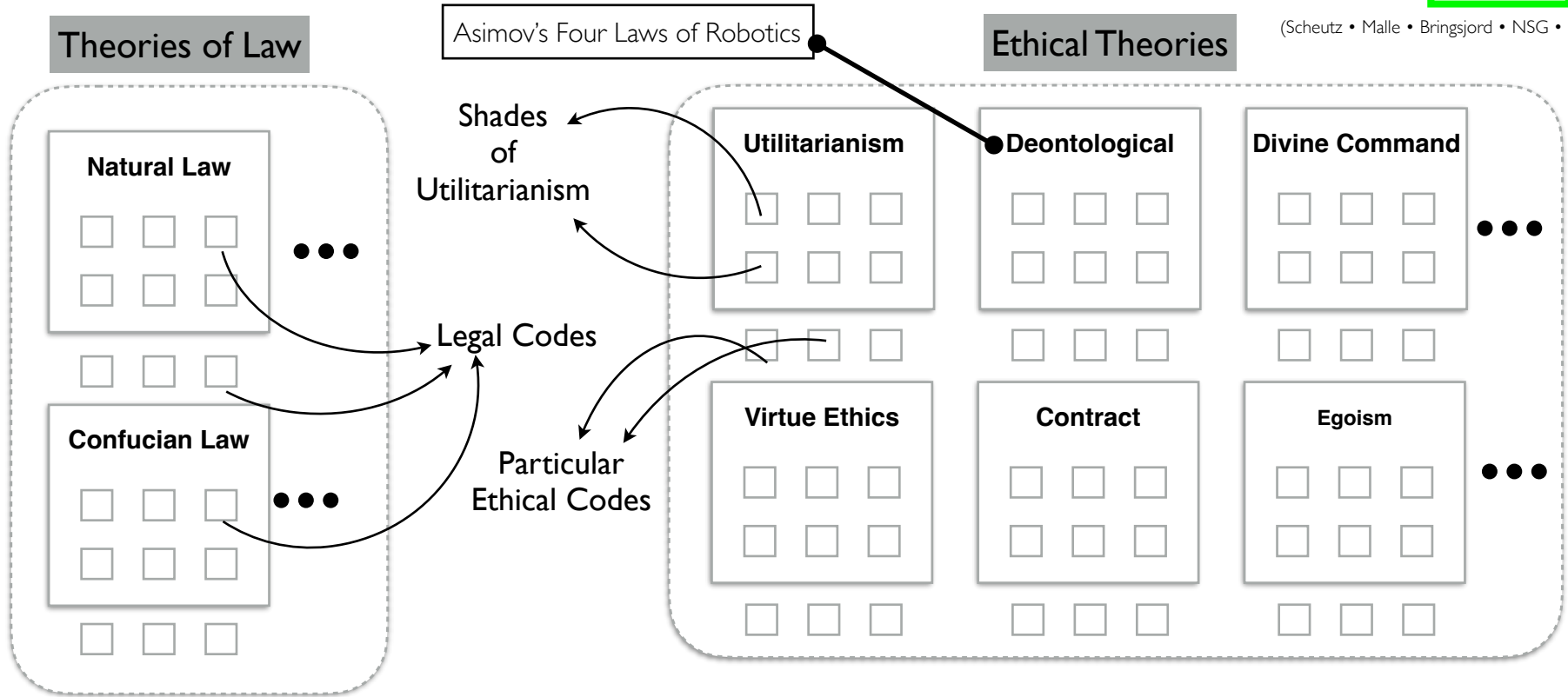
(Scheutz • Malle • Bringsjord • NSG • Bello)



The Four Steps

\$kM

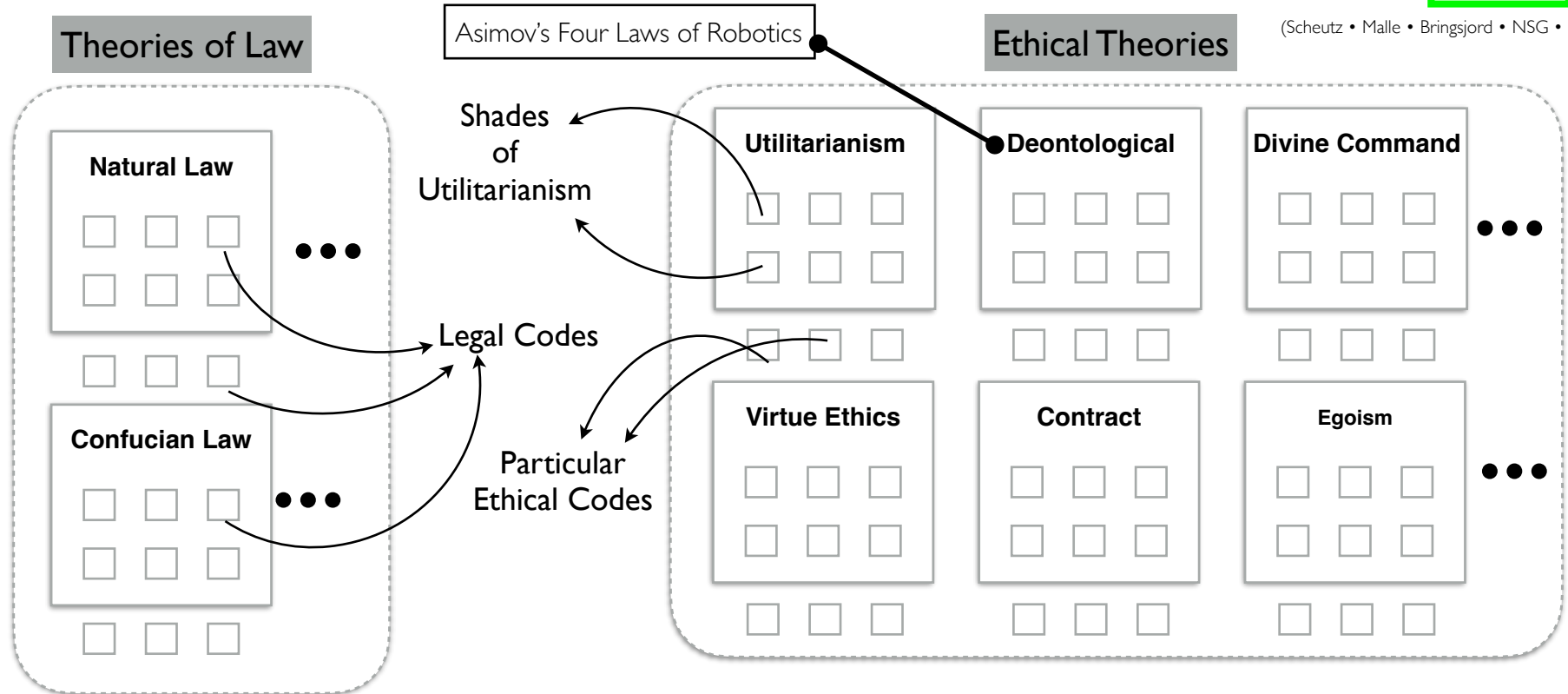
(Scheutz • Malle • Bringsjord • NSG • Bello)



The Four Steps

\$kM

(Scheutz • Malle • Bringsjord • NSG • Bello)



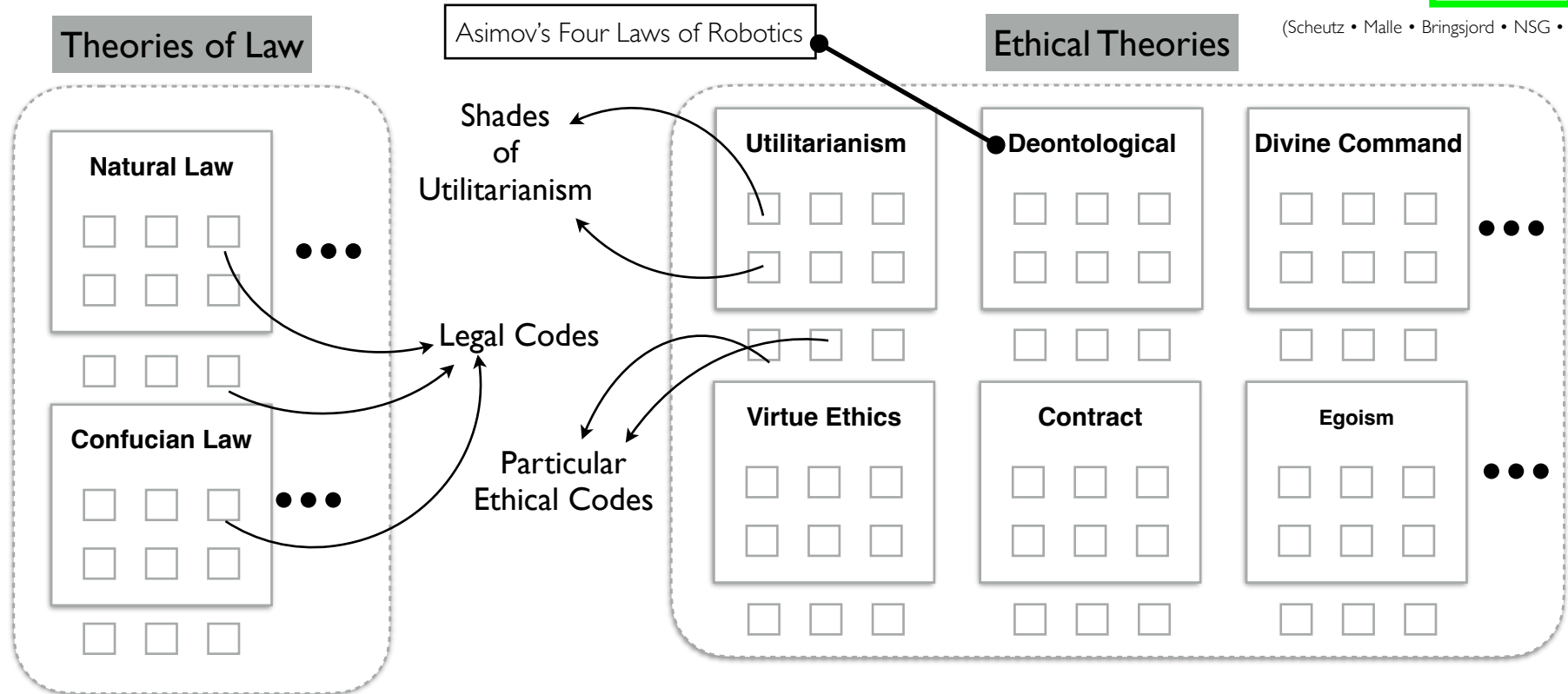
Step I

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

The Four Steps

\$kM

(Scheutz • Malle • Bringsjord • NSG • Bello)



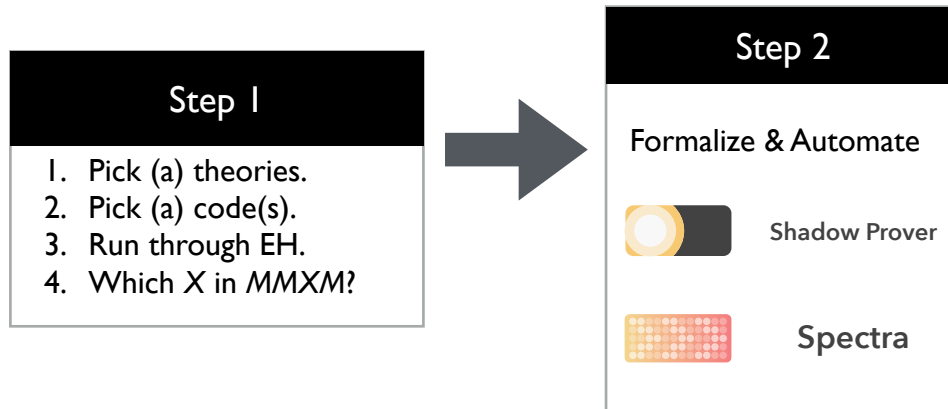
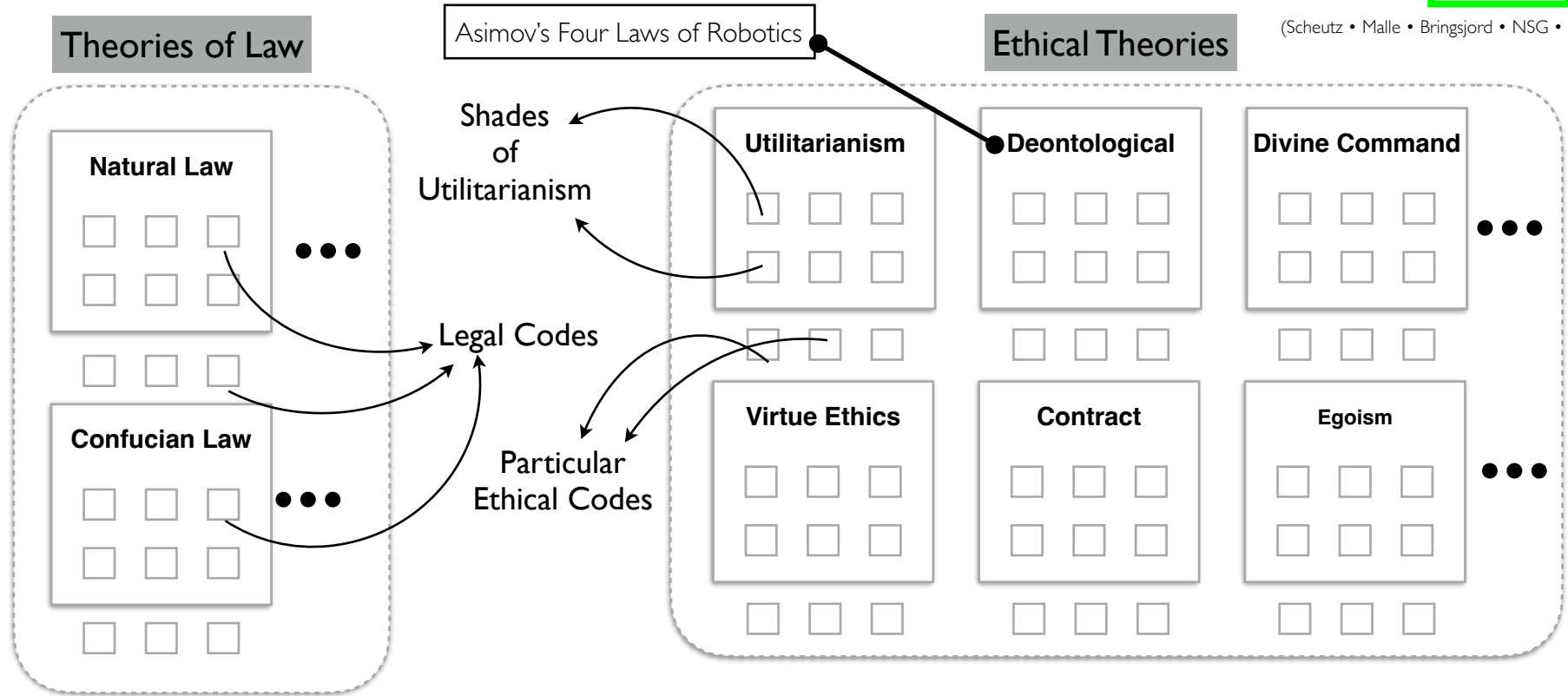
Step I

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

The Four Steps

\$kM

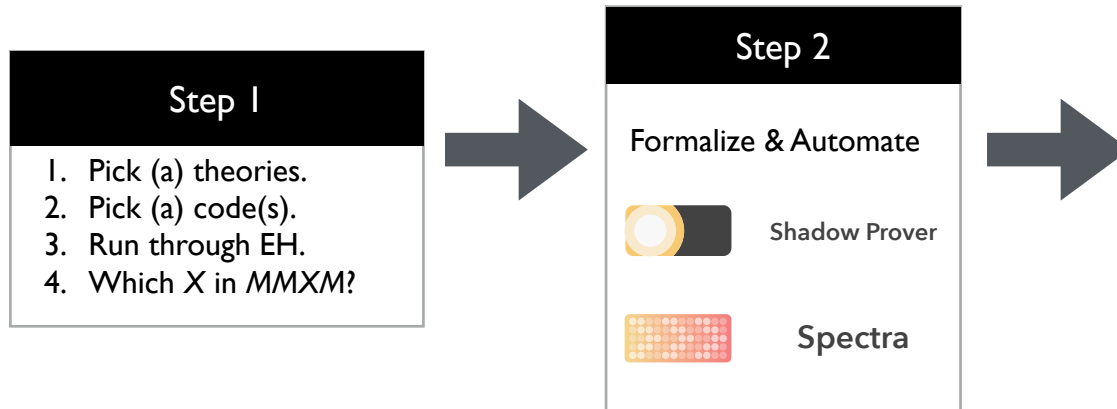
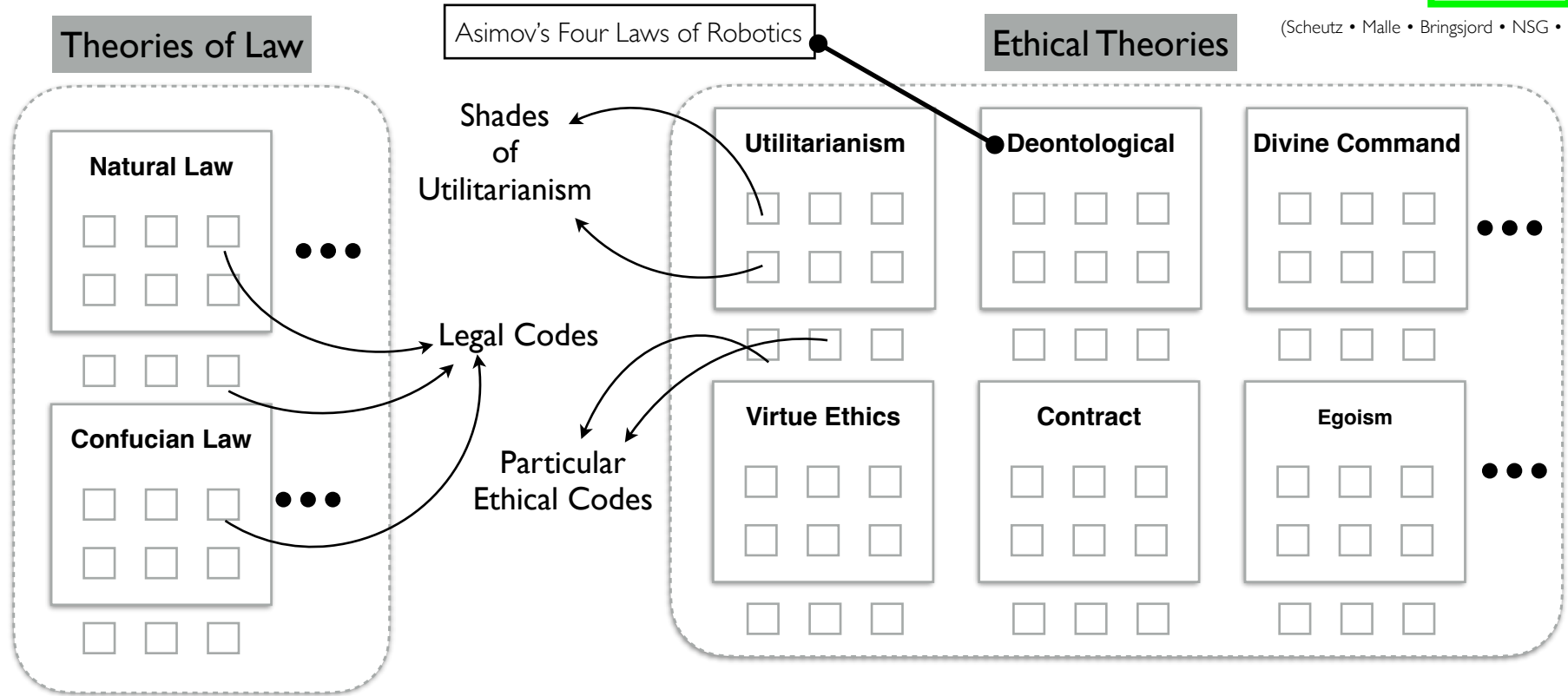
(Scheutz • Malle • Bringsjord • NSG • Bello)



The Four Steps

\$kM

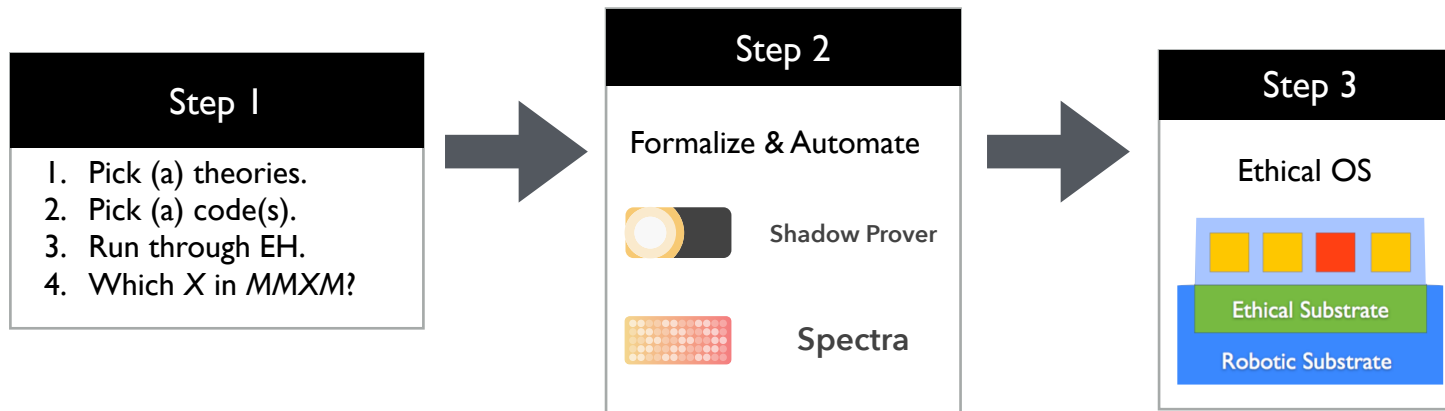
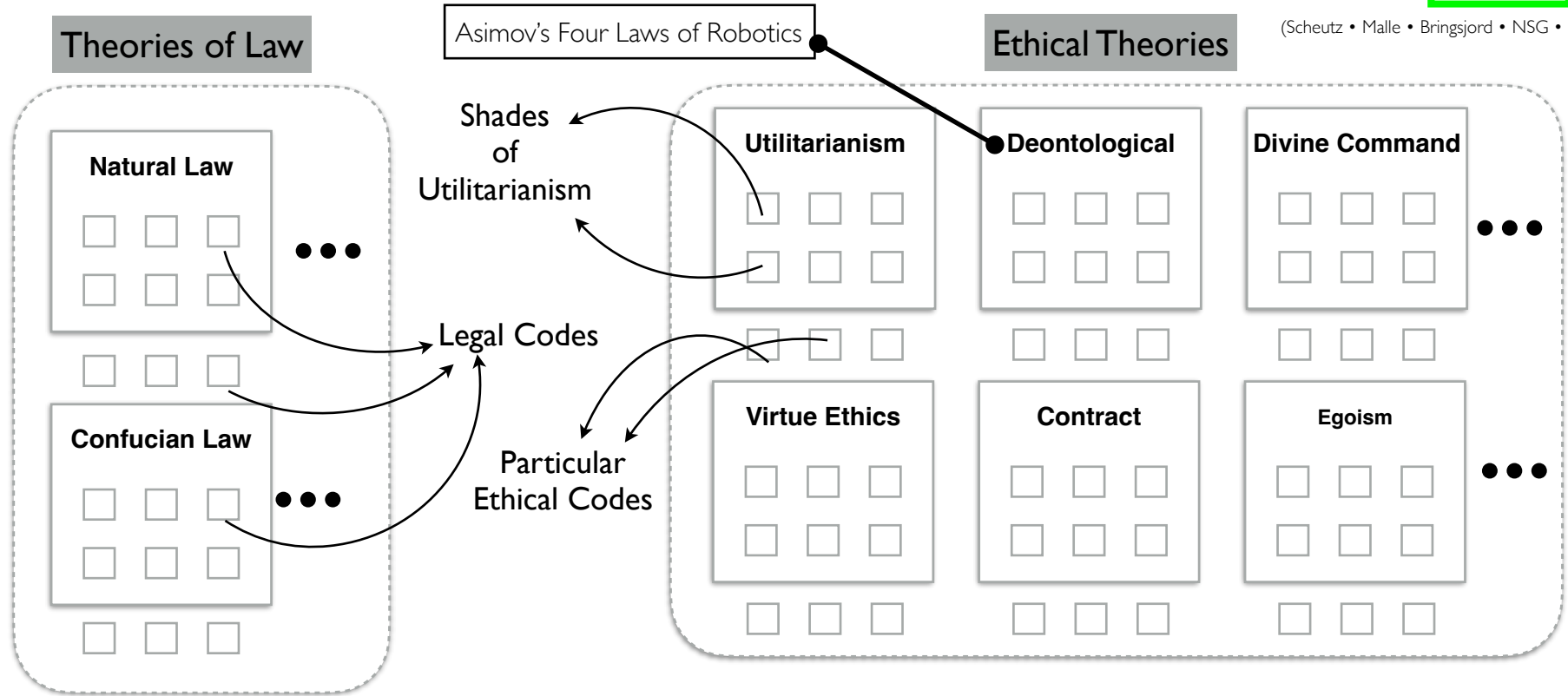
(Scheutz • Malle • Bringsjord • NSG • Bello)



The Four Steps

\$kM

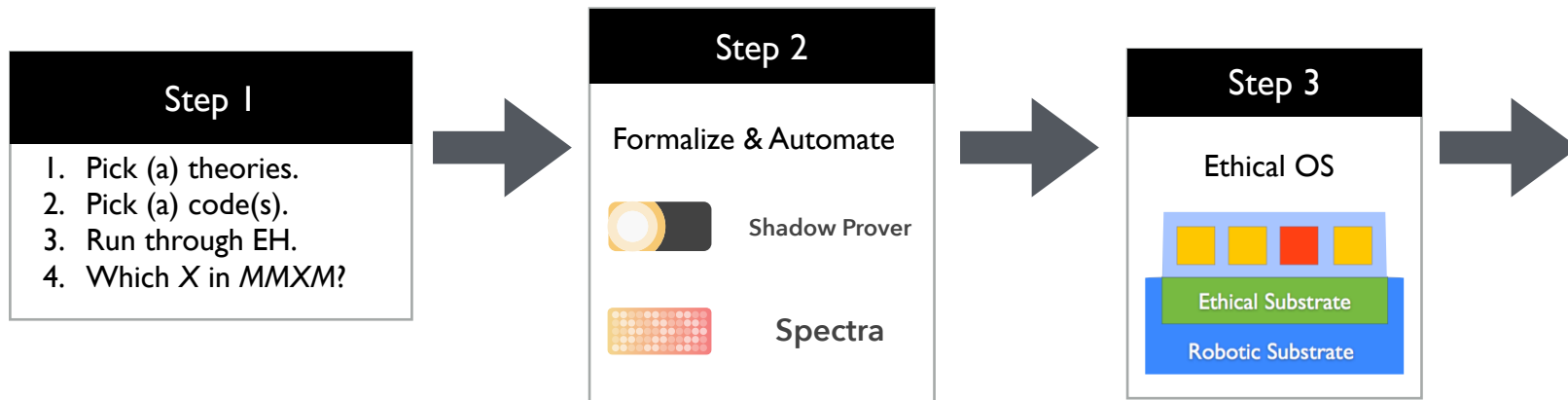
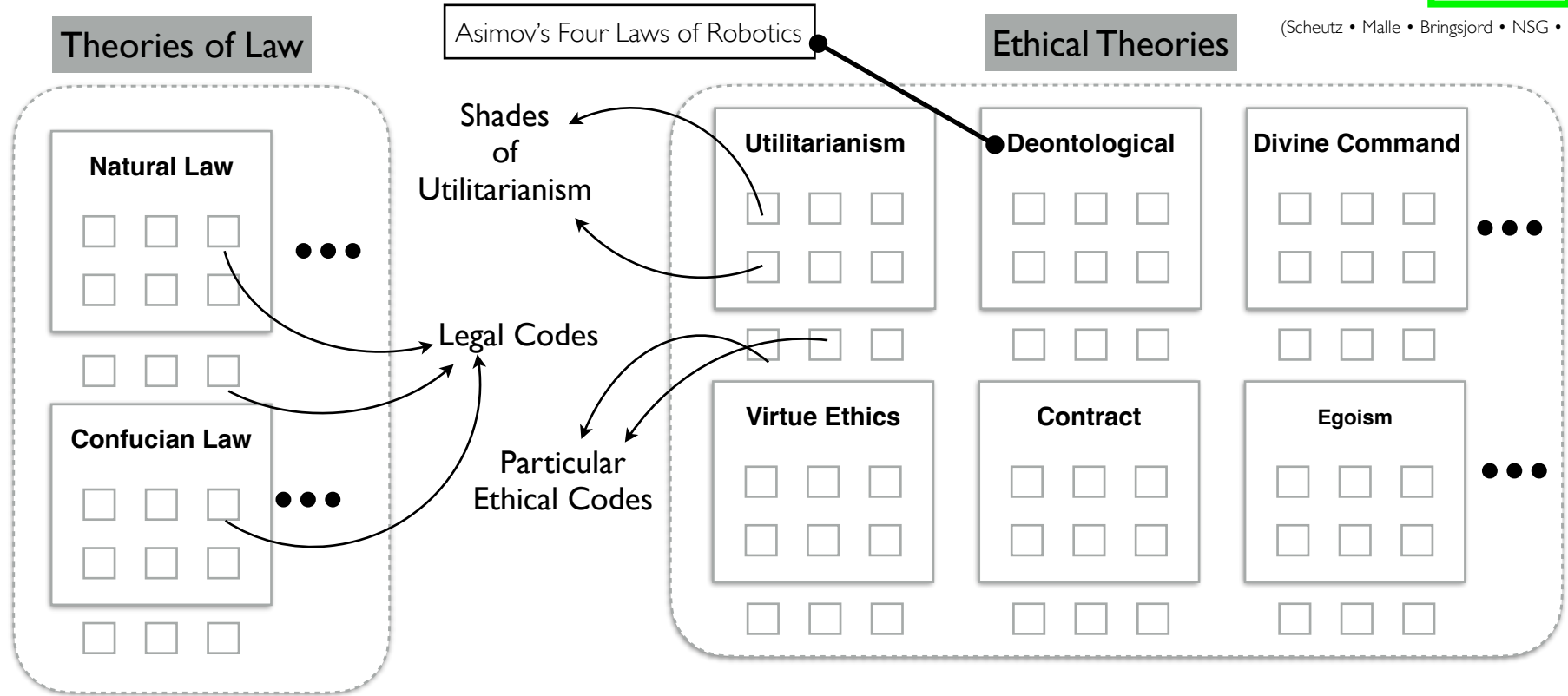
(Scheutz • Malle • Bringsjord • NSG • Bello)



The Four Steps

\$kM

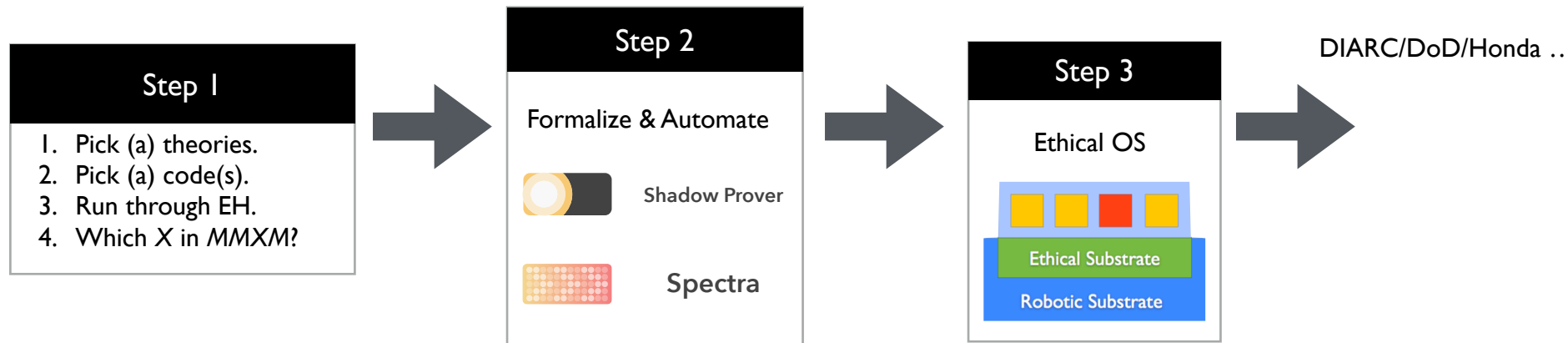
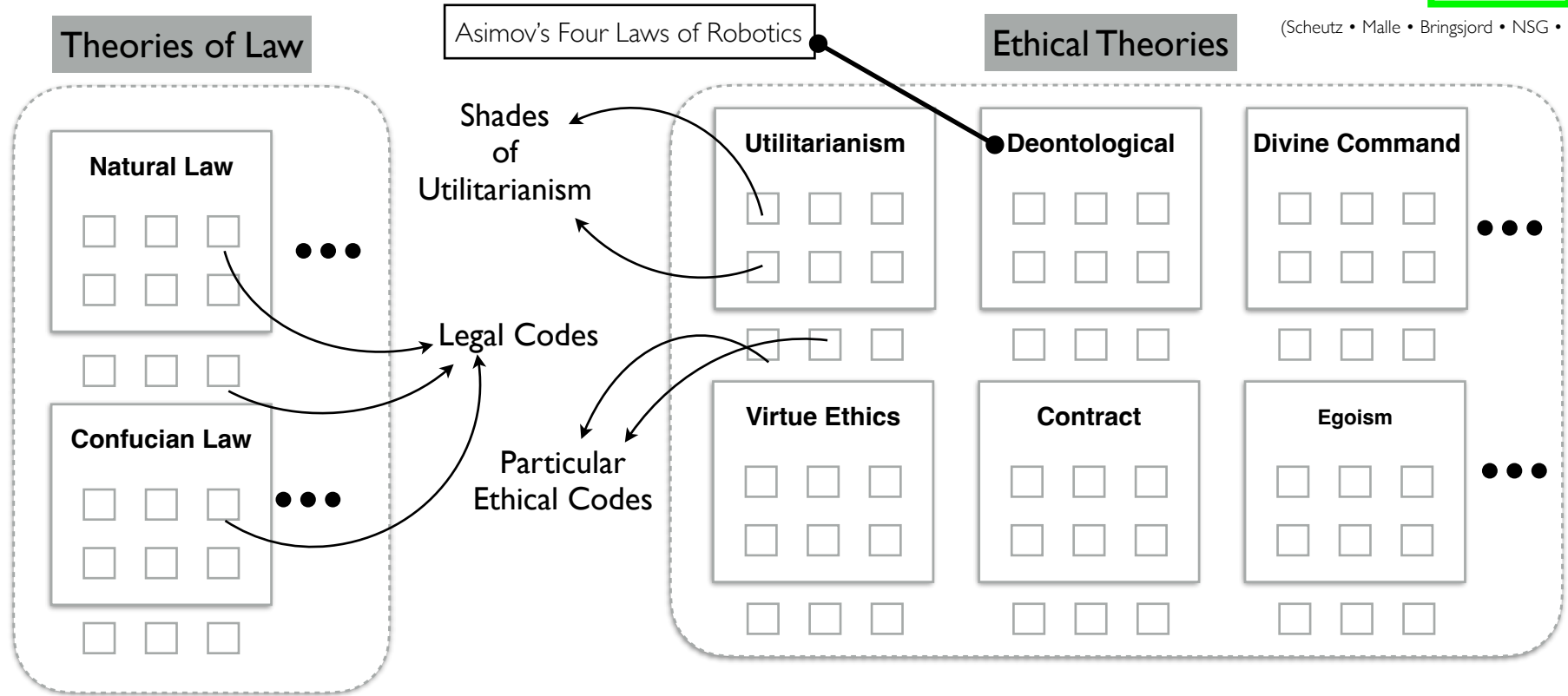
(Scheutz • Malle • Bringsjord • NSG • Bello)



The Four Steps

\$kM

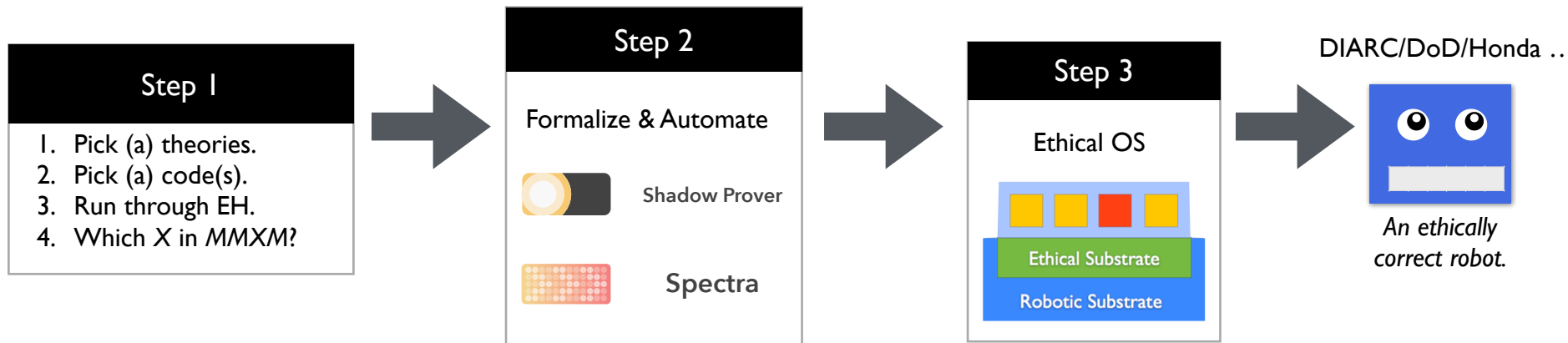
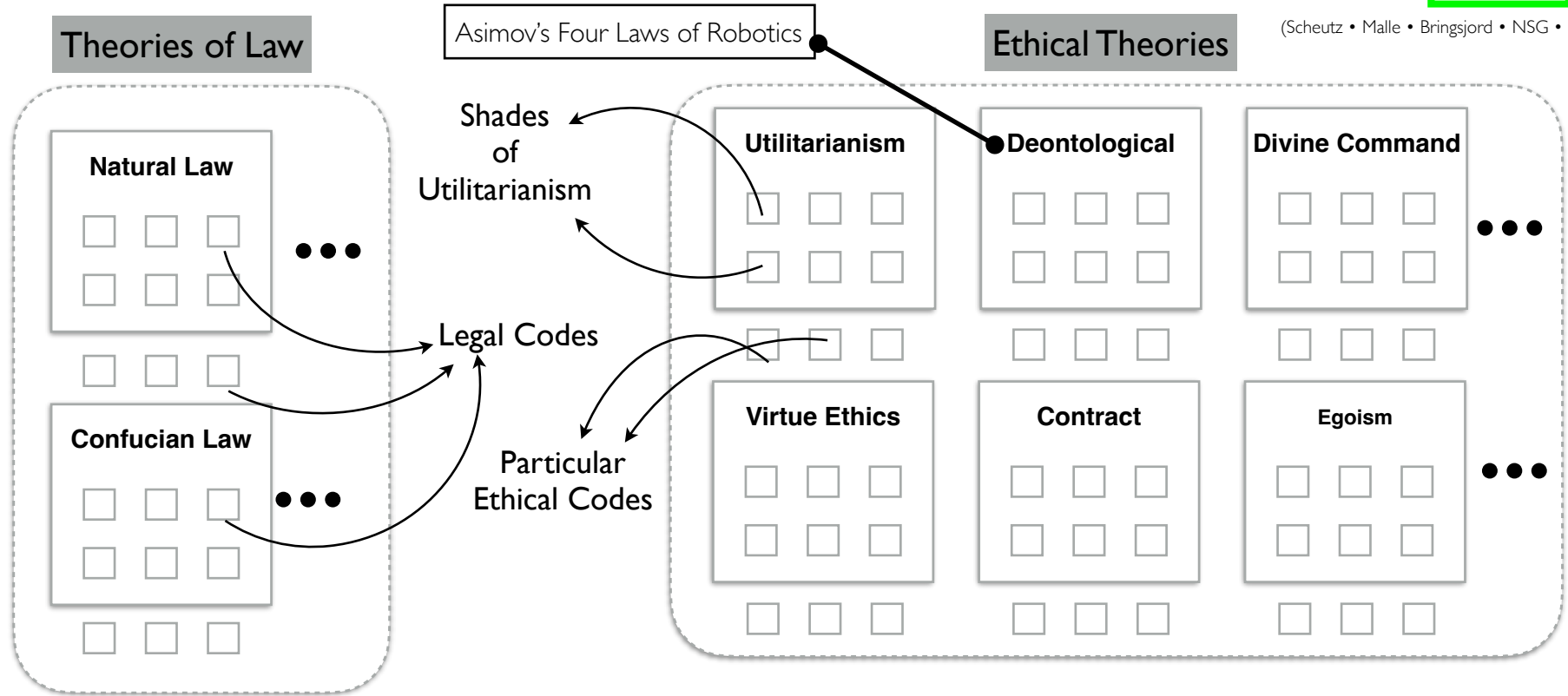
(Scheutz • Malle • Bringsjord • NSG • Bello)



The Four Steps

\$kM

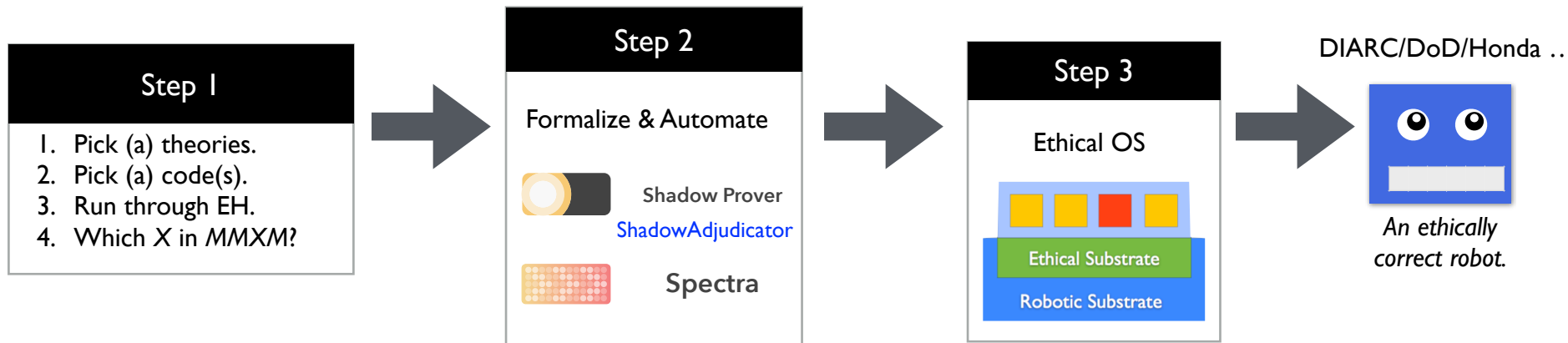
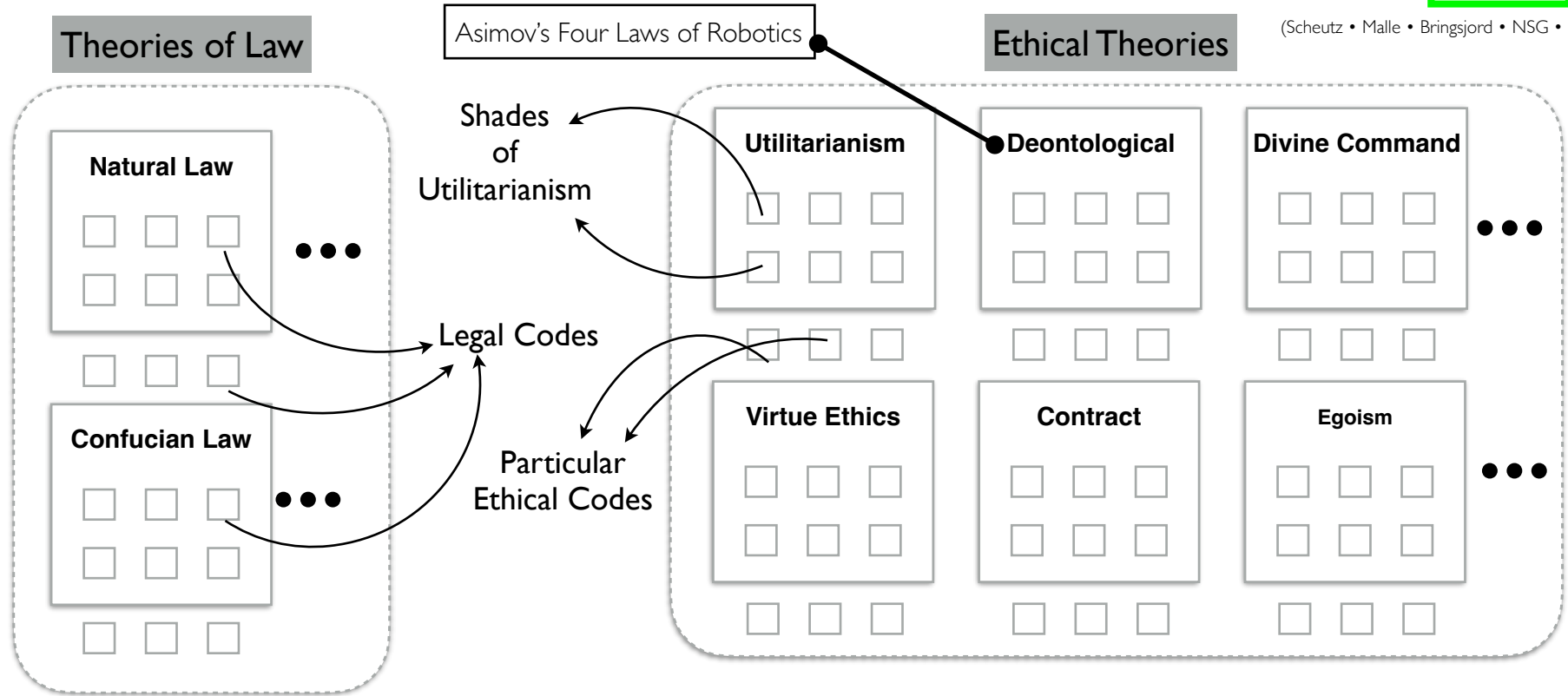
(Scheutz • Malle • Bringsjord • NSG • Bello)

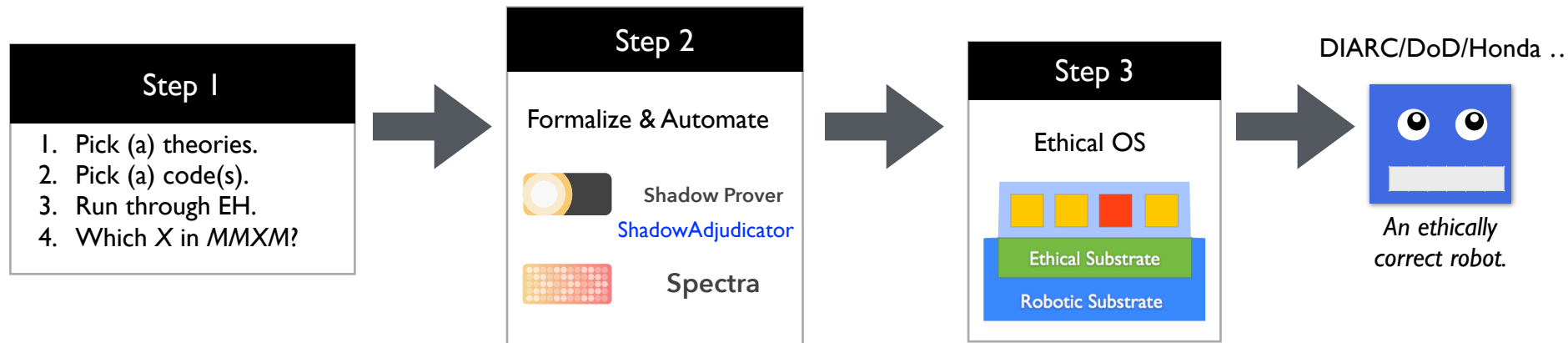
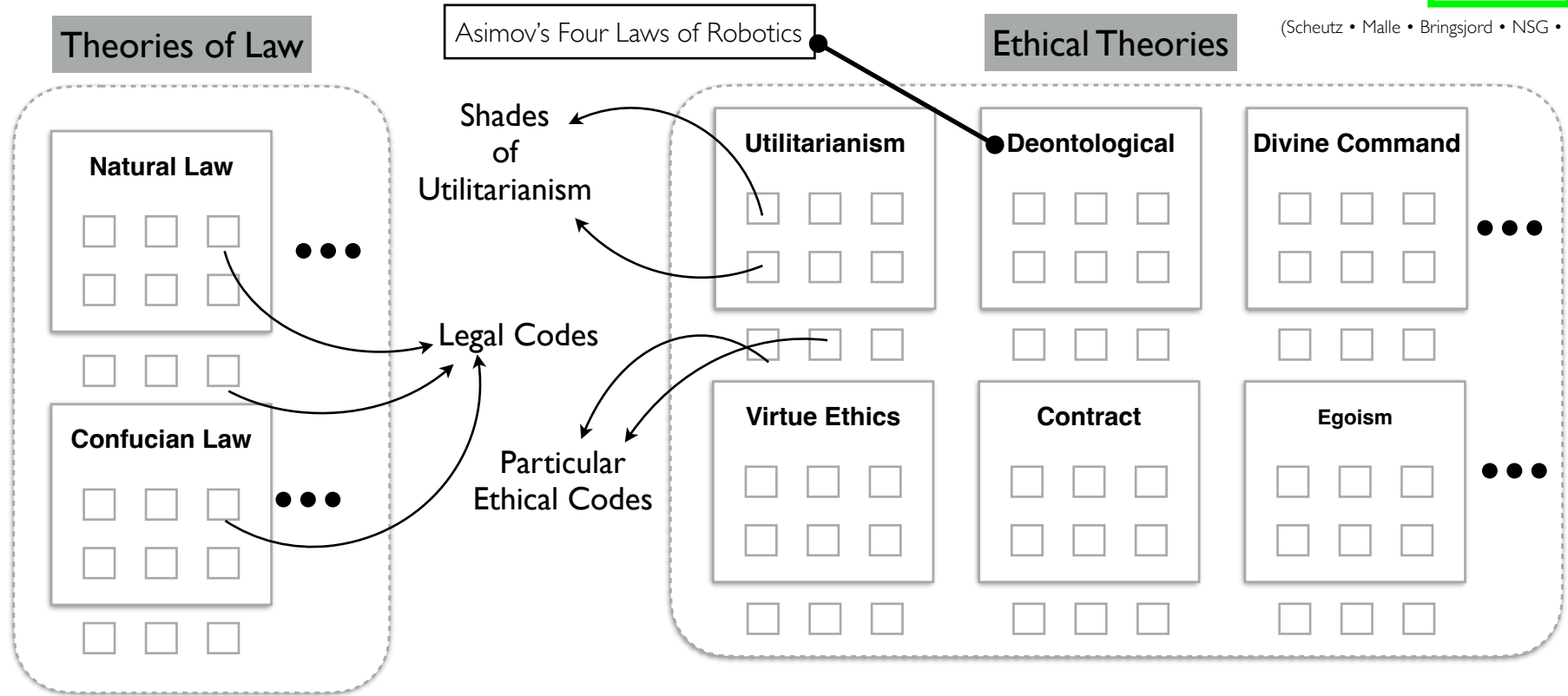


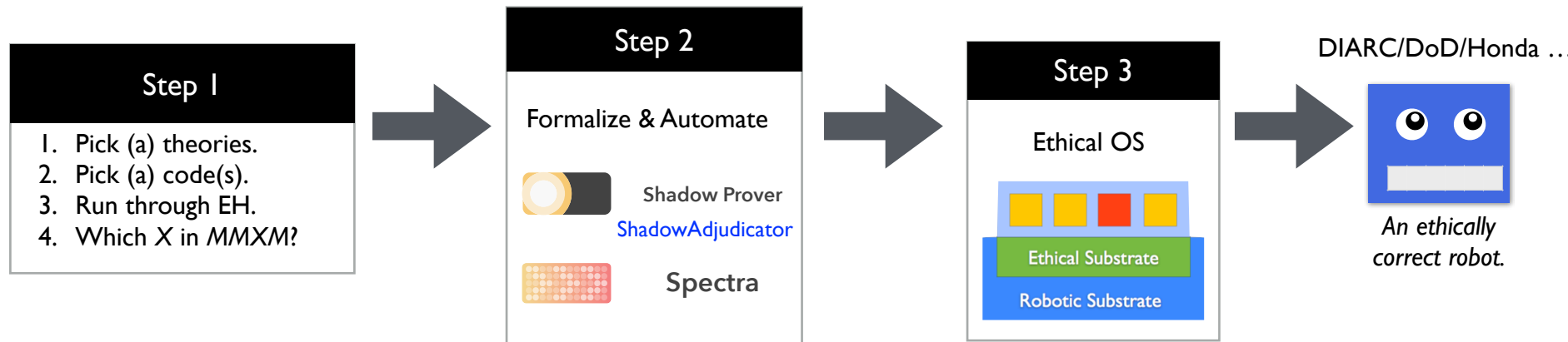
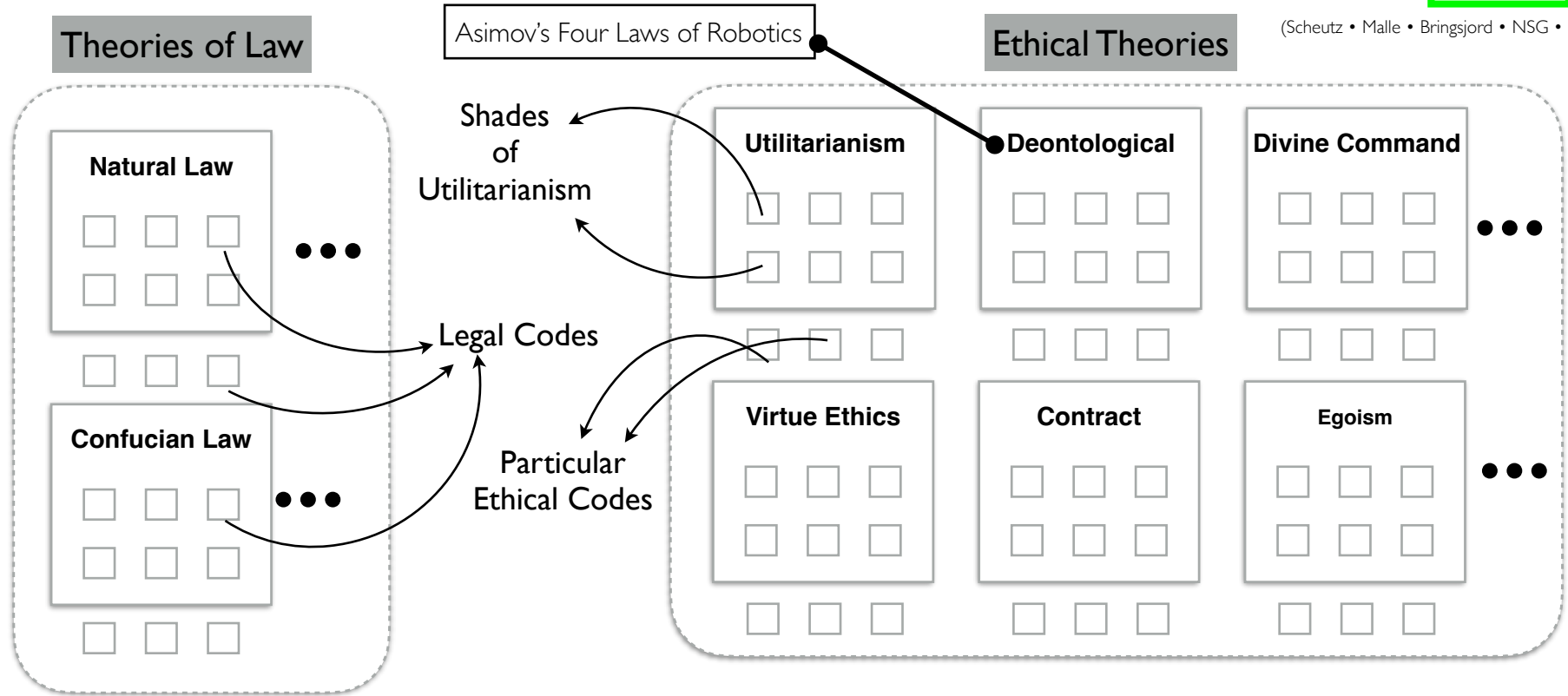
The Four Steps

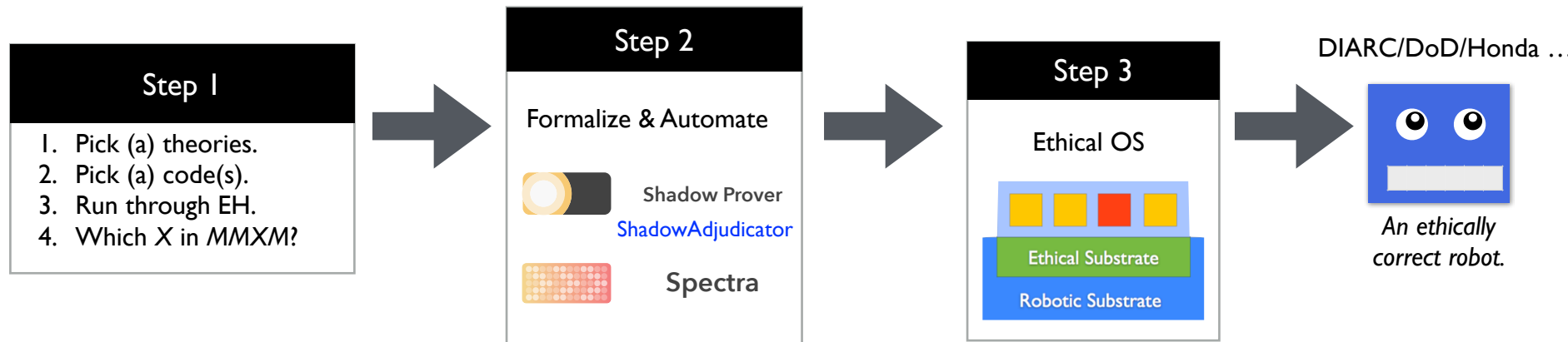
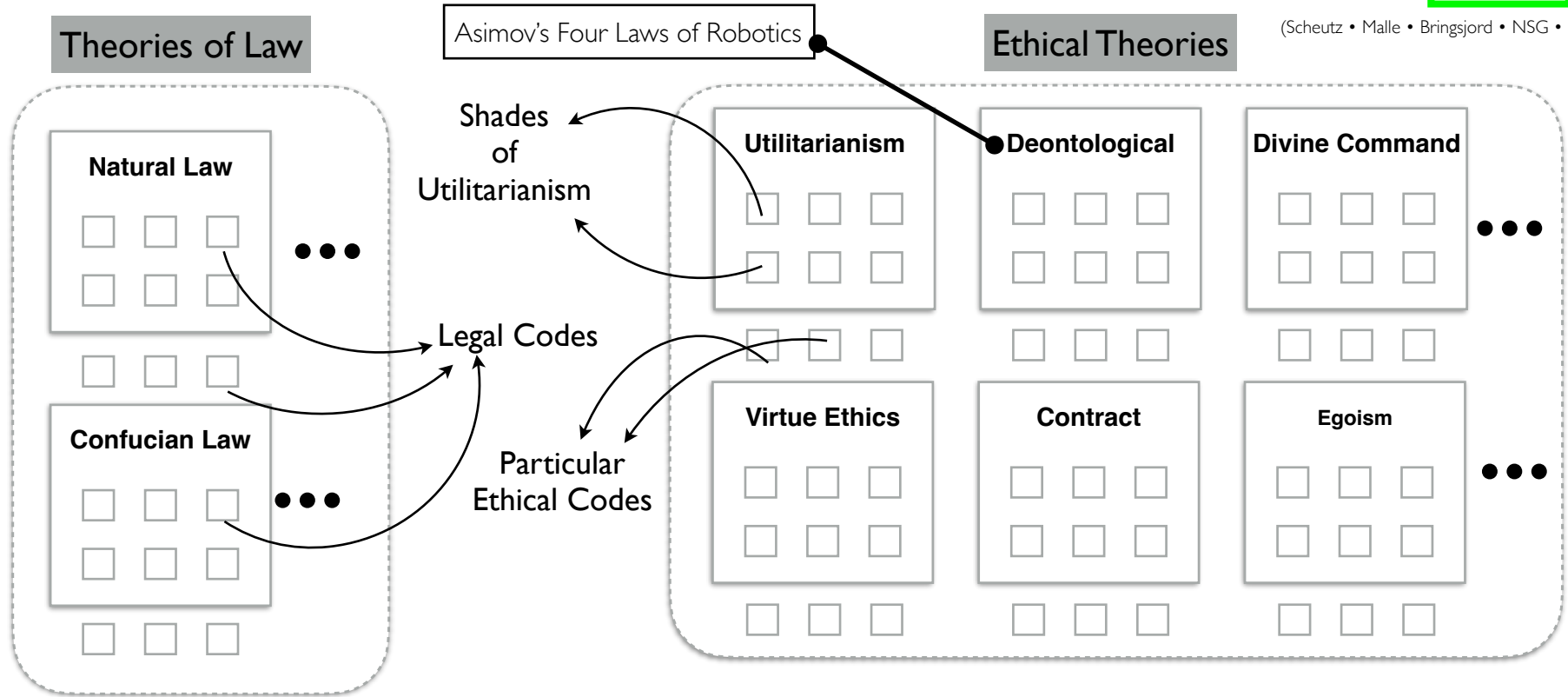
\$kM

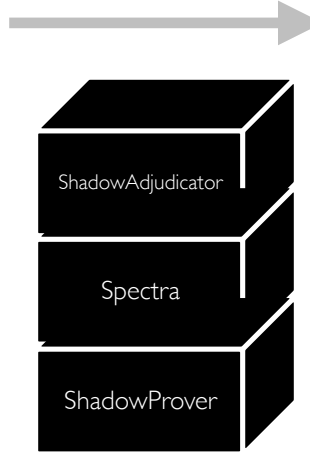
(Scheutz • Malle • Bringsjord • NSG • Bello)

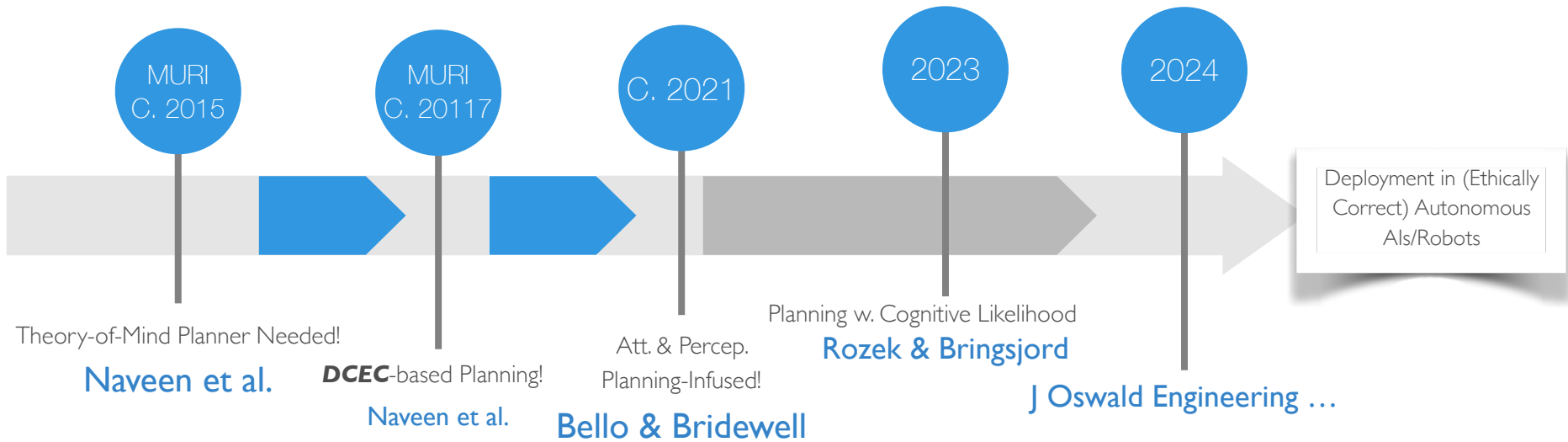
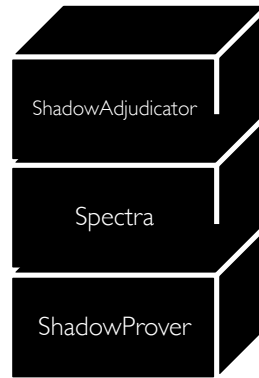


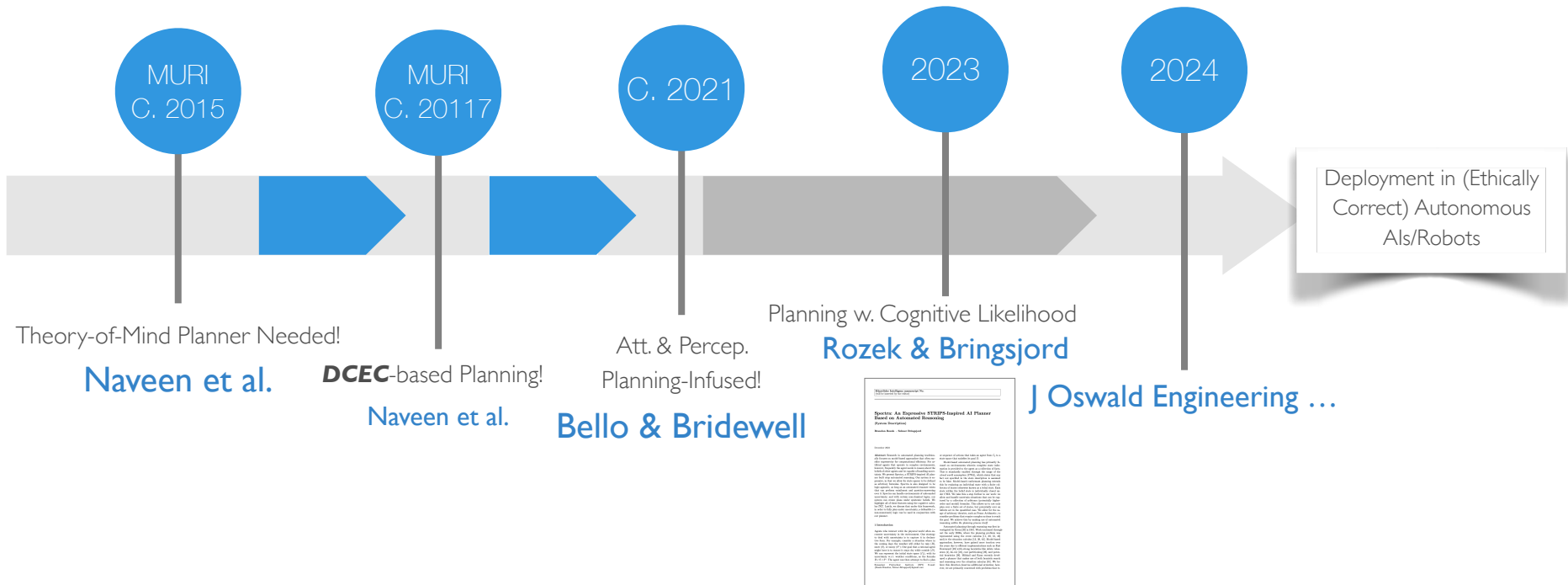
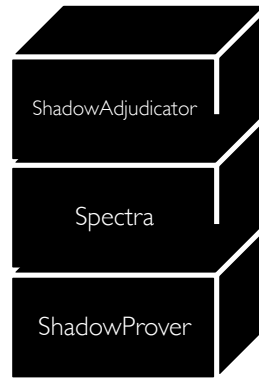


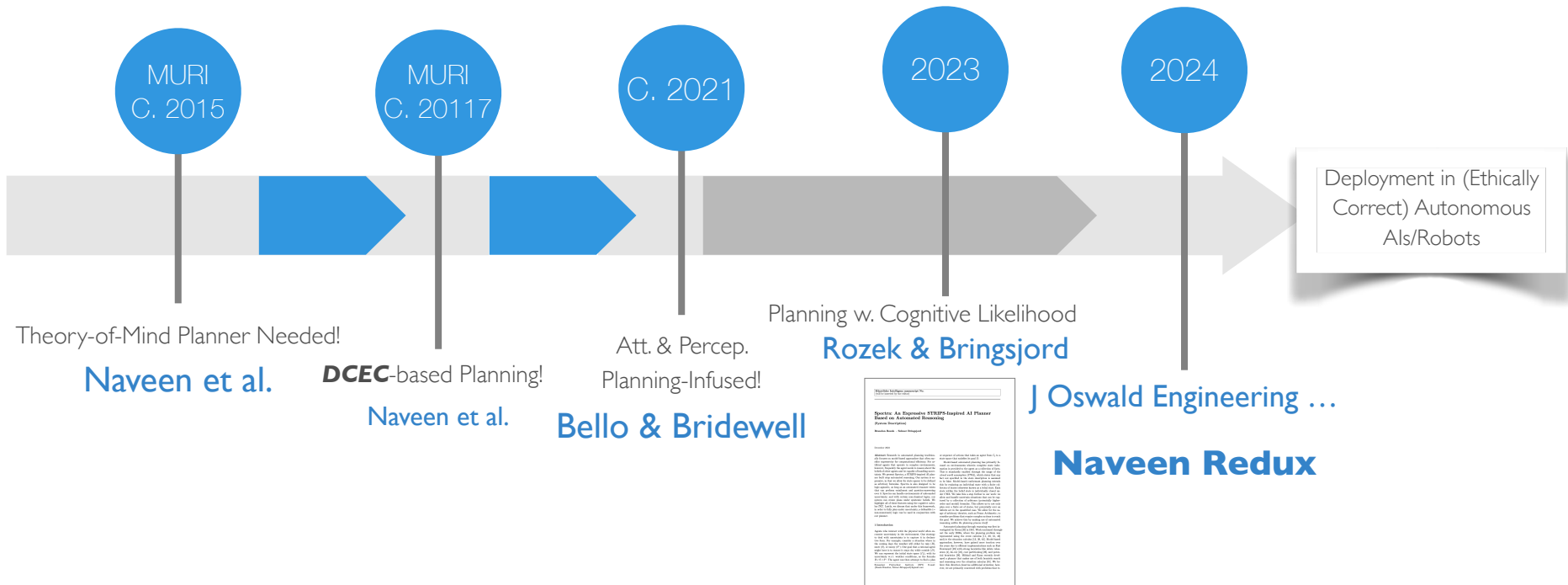
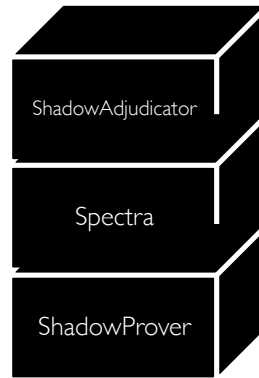








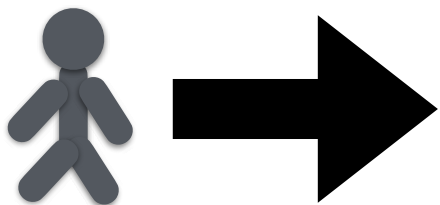




Informal Version of DDE

- C₁** the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [2017], and require that the action be neutral or above neutral in such a hierarchy);
- C₂** the net utility or goodness of the action is greater than some positive amount γ ;
- C_{3a}** the agent performing the action intends only the good effects;
- C_{3b}** the agent does not intend any of the bad effects;
- C₄** the bad effects are not used as a means to obtain the good effects; and
- C₅** if there are bad effects, the agent would rather the situation be different and the agent not have to perform the action. That is, the action is unavoidable.





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

$$\Gamma \not\models \neg \mathbf{O}(a, t, \sigma, \neg \text{happens}(\text{action}(a, \alpha), t))$$

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

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

F_{3a} The agent a intends at least one good effect. (**F₂** should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu(f_g, y) > 0$, or f_b in $\alpha_T^{a,t}$ with $\mu(f_b, y) < 0$, and some y with $t < y \leq H$ such that the following holds:

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

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

$$\Gamma \not\models \mathbf{I}(a, t, \text{Holds}(f_b, y)) \text{ and}$$

$$\Gamma \not\models \mathbf{I}(a, t, \neg \text{Holds}(f_g, y))$$

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

$$\Gamma \vdash \neg \triangleright (\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2))$$



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

$$\Gamma \not\models \neg \mathbf{O}(a, t, \sigma, \neg \text{happens}(\text{action}(a, \alpha), t))$$

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

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

F_{3a} The agent a intends at least one good effect. (**F₂** should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu(f_g, y) > 0$, or f_b in $\alpha_T^{a,t}$ with $\mu(f_b, y) < 0$, and some y with $t < y \leq H$ such that the following holds:

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

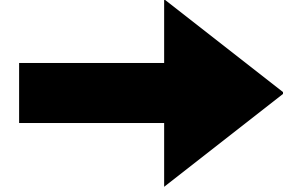
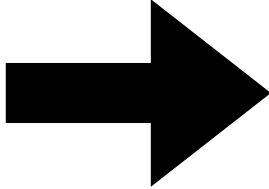
F_{3b} The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu(f_b, y) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu(f_g, y) > 0$, and for all y such that $t < y \leq H$ the following holds:

$$\Gamma \not\models \mathbf{I}(a, t, \text{Holds}(f_b, y)) \text{ and}$$

$$\Gamma \not\models \mathbf{I}(a, t, \neg \text{Holds}(f_g, y))$$

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

$$\Gamma \vdash \neg \triangleright (\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2))$$



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

$$\Gamma \not\models \neg \mathbf{O}(a, t, \sigma, \neg \text{happens}(\text{action}(a, \alpha), t))$$

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

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

F_{3a} The agent a intends at least one good effect. (**F₂** should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu(f_g, y) > 0$, or f_b in $\alpha_T^{a,t}$ with $\mu(f_b, y) < 0$, and some y with $t < y \leq H$ such that the following holds:

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

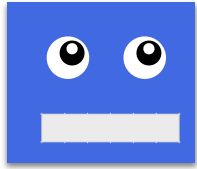
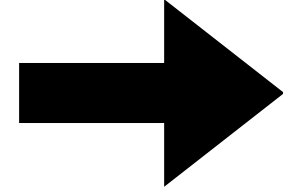
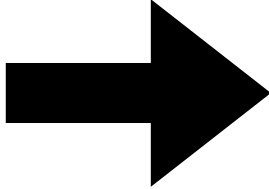
F_{3b} The agent a does not intend any bad effect. For all fluents f_b in $\alpha_I^{a,t}$ with $\mu(f_b, y) < 0$, or f_g in $\alpha_T^{a,t}$ with $\mu(f_g, y) > 0$, and for all y such that $t < y \leq H$ the following holds:

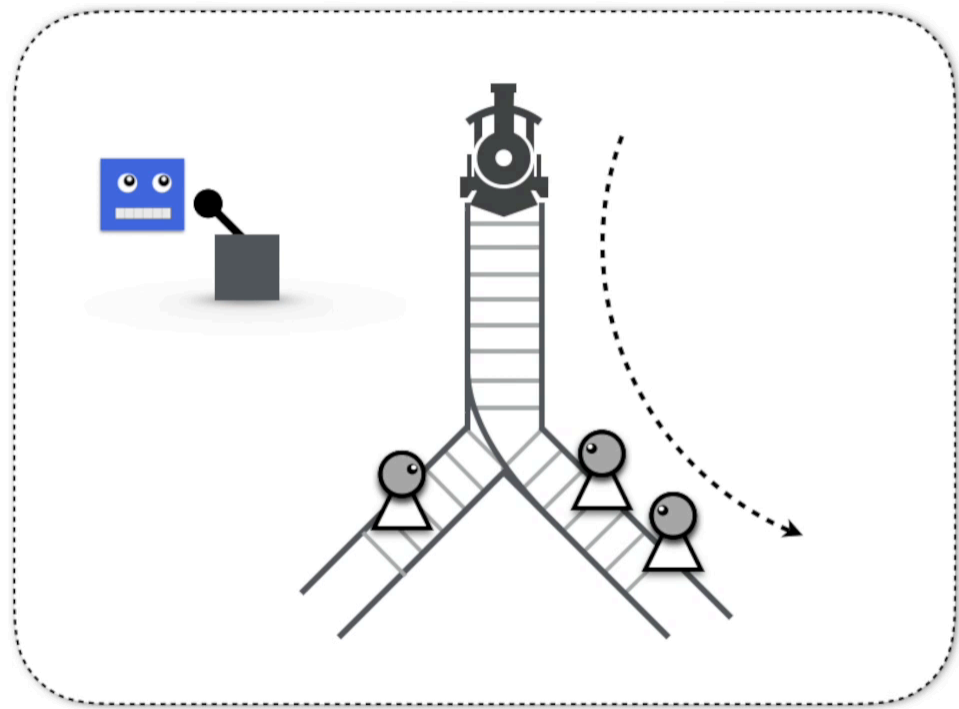
$$\Gamma \not\models \mathbf{I}(a, t, \text{Holds}(f_b, y)) \text{ and}$$

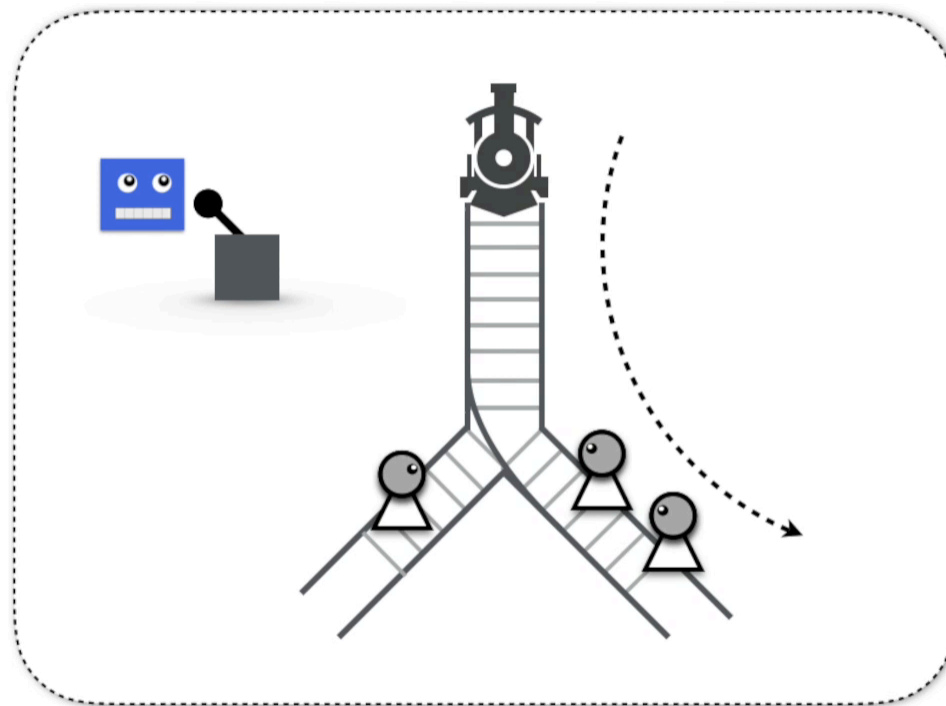
$$\Gamma \not\models \mathbf{I}(a, t, \neg \text{Holds}(f_g, y))$$

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

$$\Gamma \vdash \neg \triangleright (\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2))$$







But! — given that A. Chella
is right, there is an obstacle


...

But! — given that A. Chella
is right, there is an obstacle


...

namely, *which* theory/kind of
consciousness?!

He Cites Different Kinds of Consciousness

 **frontiers** | Frontiers in **Robotics and AI**

TYPE: Mini Review
PUBLISHED: 25 November 2023
DOI: 10.3389/frobt.2023.1270460

 Check for updates

OPEN ACCESS

EDITED BY
Amil Kumar Pandey,
Rovai Space, France

REVIEWED BY
Minoru Asada,
Osaka University, Japan
Robert H. Wortham,
University of Bath, United Kingdom

*CORRESPONDENCE
Antonio Chella,
|| antonio.chella@unipa.it

RECEIVED 31 July 2023
ACCEPTED 08 November 2023
PUBLISHED 21 November 2023

CITATION
Chella A (2023), Artificial consciousness:
the missing ingredient for ethical AI?
Front. Robot. AI 10:1270460.
doi: 10.3389/frobt.2023.1270460

COPYRIGHT
© 2023 Chella. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Artificial consciousness: the missing ingredient for ethical AI?

Antonio Chella*

RoboticsLab, Department of Engineering, Università degli Studi di Palermo, Italy & ICAR-CNR, Palermo, Italy

Can we conceive machines that can formulate autonomous intentions and make conscious decisions? If so, how would this ability affect their ethical behavior? Some case studies help us understand how advances in understanding artificial consciousness can contribute to creating ethical AI systems.

KEYWORDS
artificial consciousness, robot ethics framework, ethical AI, robot consciousness, cognitive architectures

Introduction

In April 2023, the prestigious Association for Mathematical Consciousness Science (AMCS), which brings together researchers studying the theoretical aspects of consciousness, published an open letter entitled “The Responsible Development of AI Agenda Needs to Include Consciousness Research.”¹

This letter came in response to the Future of Life Institute’s letter regarding the proposed moratorium of at least 6 months for training AI systems of the GPT-4 type². The letter, whose signatories include distinguished Turing Award scholars such as Manuel Blum and Yoshua Bengio, and many other scholars active in AI and consciousness, calls for research on AI to be coupled with consciousness research.

In Chella et al. (2022), some key theoretical aspects of artificial consciousness studies are reviewed, introducing the main concepts, theories, and issues related to this field of research. Two recent review papers, by Chalmers and by Butlin et al., summarize the state-of-the-art of artificial consciousness. Chalmers (2023) analyzes the possibility that a large language model, such as ChatGPT, may eventually be conscious by reviewing some commonly accepted indicators for consciousness. Examples are the capability of self-reporting and seeming conscious and conversational, as well as general intelligence capability. Chalmers also analyzes structural capabilities, such as the presence of senses and embodiment, the capability of recurrent processing and building a model of self and the environment, and the presence of a global workspace and unified agency. Chalmers then rules out the possibility of artificial consciousness in the current version of ChatGPT because it lacks all these capabilities.

A similar strategy is taken by Butlin et al. (2023). The authors consider the prominent theories of consciousness in the literature: the recurrent processing theory, the global workspace theory, the higher-order theory, the attention schema theory, the predictive processing, and agency and embodiment capabilities. Then, the authors outline the indicator properties derived from each of these theories. Considering these indicator properties, the authors conclude that no current AI system is a strong candidate for consciousness.

1 <https://amcs-community.org/open-letters/>

2 <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>

Frontiers in Robotics and AI

01

frontiersin.org



OPEN ACCESS

EDITED BY

Amit Kumar Pandey,
Rovial Space, France

REVIEWED BY

Minoru Asada,
Osaka University, Japan
Robert H. Wortham,
University of Bath, United Kingdom

*CORRESPONDENCE

Antonio Chella,
✉ antonio.chella@unipa.it

RECEIVED 31 July 2023

ACCEPTED 08 November 2023

PUBLISHED 21 November 2023

CITATION

Chella A (2023), Artificial consciousness:
the missing ingredient for ethical AI?
Front. Robot. AI 10:1270460.
doi: 10.3389/frobt.2023.1270460

COPYRIGHT

© 2023 Chella. This is an open-access
article distributed under the terms of the
[Creative Commons Attribution License](#)
(CC BY). The use, distribution or
reproduction in other forums is
permitted, provided the original author(s)
and the copyright owner(s) are credited
and that the original publication in this
journal is cited, in accordance with
accepted academic practice. No use,
distribution or reproduction is permitted
which does not comply with these terms.

Artificial consciousness: the missing ingredient for ethical AI?

Antonio Chella*

RoboticsLab, Department of Engineering, Università degli Studi di Palermo, Italy & ICAR-CNR,
Palermo, Italy

Can we conceive machines that can formulate autonomous intentions and make conscious decisions? If so, how would this ability affect their ethical behavior? Some case studies help us understand how advances in understanding artificial consciousness can contribute to creating ethical AI systems.

KEYWORDS

artificial consciousness, robot ethics framework, ethical AI, robot consciousness, cognitive architectures

Introduction

In April 2023, the prestigious Association for Mathematical Consciousness Science (AMCS), which brings together researchers studying the theoretical aspects of consciousness, published an open letter entitled “The Responsible Development of AI Agenda Needs to Include Consciousness Research¹.”

This letter came in response to the Future of Life Institute’s letter regarding the proposed moratorium of at least 6 months for training AI systems of the GPT-4 type². The letter, whose signatories include distinguished Turing Award scholars such as Manuel Blum and Yoshua Bengio, and many other scholars active in AI and consciousness, calls for research on AI to be coupled with consciousness research.

In Chella et al. (2022), some key theoretical aspects of artificial consciousness studies are reviewed, introducing the main concepts, theories, and issues related to this field of research.

Two recent review papers, by Chalmers and by Butlin et al., summarize the state-of-the-art of artificial consciousness. Chalmers (2023) analyzes the possibility that a large language model, such as ChatGPT, may eventually be conscious by reviewing some commonly accepted indicators for consciousness. Examples are the capability of self-reporting and seeming conscious and conversational, as well as general intelligence capability. Chalmers also analyzes structural capabilities, such as the presence of senses and embodiment, the capability of recurrent processing and building a model of self and the environment, and the presence of a global workspace and unified agency. Chalmers then rules out the possibility of artificial consciousness in the current version of ChatGPT because it lacks all these capabilities.

A similar strategy is taken by Butlin et al. (2023). The authors consider the prominent theories of consciousness in the literature: the recurrent processing theory, the global workspace theory, the higher-order theory, the attention schema theory, the predictive processing, and agency and embodiment capabilities. Then, the authors outline the indicator properties derived from each of these theories. Considering these indicator properties, the authors conclude that no current AI system is a strong candidate for consciousness.

Bringsjord et al.: Cognitive Consciousness

Chella

10.3389/frobt.2023.1270460

primitives. In contrast, slow time constraints characterize the networks at the higher levels of the hierarchy and are related to the recognition and generation of action plans.

Then, MTRNN operation is characterized by self-organization of the hierarchy consisting of the bottom-up acquisition of sensory data and the top-down generation of action plans related to the robot's intentions, which in turn trigger sequences of behavior primitives and movements. Tani showed that a sort of "free will" may be observed in the architecture when the higher-level networks spontaneously generate the robot's intentions through chaos. Then, when a gap emerges between the top-down generated intentions and the bottom-up perception of the external world, conscious awareness of intentions arises to minimize this gap [see Tani (2017), Chap. 10].

Tani disputes that this mechanism of free will may allow the robot to generate either good or bad behaviors. However, the robot may learn moral values such as its behavior. Then, it may learn to generate good behaviors according to its values and to inhibit bad behaviors.

Cognitive consciousness

A completely different approach from the one described above was proposed by Bringsjord and Naveen Sundar (2020). The authors axiomatically define "cognitive consciousness" as the functional requirements that an entity with consciousness must have, without regard to whether the entity feels anything. The authors then define a cognitive logic that roughly coincides with a family of higher-order quantified multi-operator modal logics for formally reasoning about the properties of consciousness. The characteristics of an entity endowed with consciousness are then formally defined through a system of axioms. The authors also implemented an automatic reasoning system and a planner related to systems endowed with consciousness.

An interesting aspect of the theory concerns the definition of a measure, called Lambda, the degree of cognitive consciousness of an entity. The Lambda measure provides the degree of cognitive consciousness of an agent at a given time and over intervals composed of such times. The measure has interesting aspects: it predicts null consciousness for some animals and machines, and a discontinuity in the level of consciousness between humans and machines and between humans and humans. One debated aspect concerns the null consciousness prediction for AI agents whose behavior is based on learning about neural networks.

Naveen Sundar and Bringsjord (2017) also built an AI system capable of reasoning about the doctrine of double effect and the well-known trolley problem and measured its level of consciousness. It follows from this study that reasoning about the doctrine of double effect requires a fairly high level of cognitive consciousness, which is not attainable by simple AI systems.

Artificial wisdom

"Artificial Phronesis" or artificial wisdom considers an artificial agent who is not bound to follow a specific ethical theory, such as

the double-effect theory or the deontological theory, but possesses the general ability to solve ethical problems wisely (Sullins et al., 2021).

According to this approach, an ethical agent should perform his or her actions based on wisdom and not through mere implementation of ethical doctrines. Following Aristotle, the ability to act wisely cannot be formalized through rules but is a practice that the agent must acquire through experience. Real situations are generally complex; each is encountered for the first time and thus lacks prior experience. Artificial wisdom, therefore, requires a wise agent to have the ability to understand the context, that is, what the actors are and what is at stake. The agent must also have the ability to learn new contexts and improvise on predefined patterns; it must be aware of the actions and potential reactions of other actors.

Finally, the agent must be able to revise its behavior by analyzing the interactions made. An early implementation of an agent based on artificial wisdom was described by Stenecke (2021).

In this vein, Chella et al. (2020) and Chella et al. (2024) are studying the effect of robots' inner speech on artificial wisdom. Specifically, the research has focused on experiments in which a user and a robot must perform a collaborative task, such as setting a dining table in a nursing home where people with dementia are also present. The experiments analyze how a user, by hearing the robot's inner speech during the collaborative task, can achieve a higher degree of awareness of issues related to people with dementia. Preliminary results support this hypothesis.

Conclusion

In this mini-review, we analyzed case studies focused on ethical AI agents inspired and influenced by various theories of artificial consciousness. This process allowed us to critically explore different facets of this complex topic.

Two of the most challenging questions concern whether an AI system may be a moral agent and if a form of artificial consciousness is needed to ensure ethical behavior in the AI system. These questions have no definitive answers and remain essential open lines of research. The problematic nature of the issue lies in defining what we mean by "consciousness" in a non-biological entity and in delineating the criteria to measure the ethics of an action performed by an AI system.

Finally, we mentioned another major open issue: the importance of research on consciousness and emotion studies in machines for progress toward more ethical AI.

This debate reflects a broader and more fundamental issue: the ability of machines to "feel" or "understand" authentically and how that ability might influence their ethical behavior.

These issues are dense with theoretical, methodological, and ethical implications and challenges that the scientific community cannot ignore. Their complexity is a reminder of the importance of a multidisciplinary approach in AI research, combining computer science, philosophy, psychology, neuroscience, and ethics to develop AI systems that are not only technically advanced but also ethically responsible.

Bringsjord

Chap. 10].

Tani disputes that this mechanism of free will may allow the robot to generate either good or bad behaviors. However, the robot may learn moral values such as its behavior. Then, it may learn to generate good behaviors according to its values and to inhibit bad behaviors.

Cognitive consciousness

A completely different approach from the one described above was proposed by Bringsjord and Naveen Sundar (2020). The authors axiomatically define “cognitive consciousness” as the functional requirements that an entity with consciousness must have, without regard to whether the entity feels anything. The authors then define a cognitive logic that roughly coincides with a family of higher-order quantified multi-operator modal logics for formally reasoning about the properties of consciousness. The characteristics of an entity endowed with consciousness are then formally defined through a system of axioms. The authors also implemented an automatic reasoning system and a planner related to systems endowed with consciousness.

An interesting aspect of the theory concerns the definition of a measure, called Lambda, the degree of cognitive consciousness of an entity. The Lambda measure provides the degree of cognitive consciousness of an agent at a given time and over intervals composed of such times. The measure has interesting aspects: it predicts null consciousness for some animals and machines, and a discontinuity in the level of consciousness between humans and machines and between humans and humans. One debated aspect concerns the null consciousness prediction for AI agents whose behavior is based on learning about neural networks.

Naveen Sundar and Bringsjord (2017) also built an AI system capable of reasoning about the doctrine of double effect and the well-known trolley problem and measured its level of consciousness. It follows from this study that reasoning about the doctrine of double effect requires a fairly high level of cognitive consciousness, which is not attainable by simple AI systems.

Artificial wisdom

“Artificial Phronesis” or artificial wisdom considers an artificial agent who is not bound to follow a specific ethical theory, such as

it must be aware of the actions and potential reactions of actors.

Finally, the agent must be able to revise its behavior by analyzing the interactions made. An early implementation of an agent based on artificial wisdom was described by S (2021).

In this vein, Chella et al. (2020) and Chella et al. (2021) studying the effect of robots’ inner speech on artificial wisdom. Specifically, the research has focused on experiments in which a user and a robot must perform a collaborative task, such as setting a dining table in a nursing home where people with dementia are also present. The experiments analyze how a user, by hearing the robot’s inner speech during the collaborative task, can achieve a higher degree of awareness of issues related to people with dementia. Preliminary results support this hypothesis.

Conclusion

In this mini-review, we analyzed case studies focused on AI agents inspired and influenced by various theories of artificial consciousness. This process allowed us to critically explore different facets of this complex topic.

Two of the most challenging questions concern whether an AI system may be a moral agent and if a form of artificial consciousness is needed to ensure ethical behavior in the AI system. These questions have no definitive answers and remain essential open questions of research. The problematic nature of the issue lies in clarifying what we mean by “consciousness” in a non-biological entity and delineating the criteria to measure the ethics of an action performed by an AI system.

Finally, we mentioned another major open issue: the impact of research on consciousness and emotion studies in machine learning progress toward more ethical AI.

This debate reflects a broader and more fundamental issue: the ability of machines to “feel” or “understand” authentically and how that ability might influence their behavior.

These issues are dense with theoretical, methodological, and ethical implications and challenges that the scientific community cannot ignore. Their complexity is a reminder of the importance of a multidisciplinary approach in AI research, combining computer science, philosophy, psychology, neuroscience, and ethics to develop AI systems that are not only technically advanced but also ethically responsible.

Bringsjord

Armed with *TCC* and Λ , obstacle surmounted.

Chap. 10].

Tani disputes that this mechanism of free will may allow the robot to generate either good or bad behaviors. However, the robot may learn moral values such as its behavior. Then, it may learn to generate good behaviors according to its values and to inhibit bad behaviors.

Cognitive consciousness

A completely different approach from the one described above was proposed by Bringsjord and Naveen Sundar (2020). The authors axiomatically define “cognitive consciousness” as the functional requirements that an entity with consciousness must have, without regard to whether the entity feels anything. The authors then define a cognitive logic that roughly coincides with a family of higher-order quantified multi-operator modal logics for formally reasoning about the properties of consciousness. The characteristics of an entity endowed with consciousness are then formally defined through a system of axioms. The authors also implemented an automatic reasoning system and a planner related to systems endowed with consciousness.

concerns the definition of cognitive consciousness provides the degree of a given time and over intervals composed of such times. The measure has interesting aspects: it predicts null consciousness for some animals and machines, and a discontinuity in the level of consciousness between humans and machines and between humans and humans. One debated aspect concerns the null consciousness prediction for AI agents whose behavior is based on learning about neural networks.

Naveen Sundar and Bringsjord (2017) also built an AI system capable of reasoning about the doctrine of double effect and the well-known trolley problem and measured its level of consciousness. It follows from this study that reasoning about the doctrine of double effect requires a fairly high level of cognitive consciousness, which is not attainable by simple AI systems.

Artificial wisdom

“Artificial Phronesis” or artificial wisdom considers an artificial agent who is not bound to follow a specific ethical theory, such as

it must be aware of the actions and potential reactions of actors.

Finally, the agent must be able to revise its behavior analyzing the interactions made. An early implementation of an agent based on artificial wisdom was described by S (2021).

In this vein, Chella et al. (2020) and Chella et al. (2021) studying the effect of robots’ inner speech on artificial wisdom. Specifically, the research has focused on experiments in which a user and a robot must perform a collaborative task, such as setting a dining table in a nursing home where people with dementia are also present. The experiments analyze how a user, by hearing the robot’s inner speech during the collaborative task, can achieve a higher degree of awareness of issues related to people with dementia. Preliminary results support this hypothesis.

Conclusion

In this mini-review, we analyzed case studies focused on AI agents inspired and influenced by various theories of artificial consciousness. This process allowed us to critically explore different facets of this complex topic.

Two of the most challenging questions concern whether an AI system may be a moral agent and if a form of artificial consciousness is needed to ensure ethical behavior in the AI system. These questions have no definitive answers and remain essential topics for research. The problematic nature of the issue lies in clarifying what we mean by “consciousness” in a non-biological entity and delineating the criteria to measure the ethics of an action performed by an AI system.

Finally, we mentioned another major open issue: the impact of research on consciousness and emotion studies in machine learning progress toward more ethical AI.

This debate reflects a broader and more fundamental issue: the ability of machines to “feel” or “understand” authentically and how that ability might influence their behavior.

These issues are dense with theoretical, methodological, and ethical implications and challenges that the scientific community cannot ignore. Their complexity is a reminder of the importance of a multidisciplinary approach in AI research, combining computer science, philosophy, psychology, neuroscience, and ethics to develop AI systems that are not only technically advanced but also ethically responsible.

Non-Technical Portal to TCC

24. CAN CONSCIOUSNESS BE EXPLAINED BY INTEGRATED INFORMATION THEORY OR THE THEORY OF COGNITIVE CONSCIOUSNESS?¹

Selmer Bringsjord and Naveen Sundar Govindarajulu

1. Introduction

AS READERS will doubtless have noted by now, some other chapters in the present volume have expressed the view (rather agreeable to us) that many aspects of human-level mental phenomena are recalcitrant to a mindset that insists upon mathematical and (usually) material explanations. First-person subjectivity, intentionality, mathematical cognition, robust episodic states, consciousness. . . these phenomena are exceedingly hard to explain in such a manner. It is the final member of that list of challenges that is our focus in the present chapter. Can science operating in the math-and-material manner explain—and perhaps even, courtesy of associated engineering, replicate in artificial agents—consciousness?

This question is now pressed upon at least all technologized societies on Earth, because of the advent of artificial agents able to converse in seemingly flawless English about pretty much anything, including consciousness itself.

A famous example is ChatGPT. This class of agents falls into what is now called “generative AI,” which includes agents not only able to generate natural language, but also images. In the case of language, these agents are sometimes called “chatbots,” but are more precisely known as “Large Language Models.” Some of these agents have been declared conscious,² and the question of whether they are is really just a special case of the general question taken up in the present chapter. We are very confident that ascriptions of consciousness to artificial agents are only going to grow in frequency, and such ascriptions are going to increasingly be issued by voices that seem balanced and authoritative. This chapter should in our opinion be read and understood by those humans who will find themselves living in the trend we foresee, because it provides at least a starting basis for two fundamental ways of looking not just at consciousness in general, but consciousness in computational artifacts.

Other Obstacles?

Other Obstacles?

- The Elephant (again): Als/Robots, *contra* Scharre, *are* weapons; where then is the military theory of human-AI combat? (Boyd, Hubin, ... okay okay; but AI?)

Other Obstacles?

- The Elephant (again): Als/Robots, *contra* Scharre, *are* weapons; where then is the military theory of human-AI combat? (Boyd, Hubin, ... okay okay; but AI?)
- What if we are forced to use foundation models, which are ethically stupid, failing as they do on the very challenges issued by the logicist-AI folks 40 years ago, yet with profiteers touting such models?

Other Obstacles?

- The Elephant (again): AIs/Robots, *contra* Scharre, are weapons; where then is the military theory of human-AI combat? (Boyd, Hubin, ... okay okay; but AI?)
- What if we are forced to use foundation models, which are ethically stupid, failing as they do on the very challenges issued by the logicist-AI folks 40 years ago, yet with profiteers touting such models?
- Naïve attention-perception schemes dominate AI, no?

Other Obstacles?

- The Elephant (again): AIs/Robots, *contra* Scharre, are weapons; where then is the military theory of human-AI combat? (Boyd, Hubin, ... okay okay; but AI?)
- What if we are forced to use foundation models, which are ethically stupid, failing as they do on the very challenges issued by the logicist-AI folks 40 years ago, yet with profiteers touting such models?
- Naïve attention-perception schemes dominate AI, no?
- Ethical correctness is nice, but what about moral creativity?

Other Obstacles?

- The Elephant (again): AIs/Robots, *contra* Scharre, are weapons; where then is the military theory of human-AI combat? (Boyd, Hubin, ... okay okay; but AI?)
- What if we are forced to use foundation models, which are ethically stupid, failing as they do on the very challenges issued by the logicist-AI folks 40 years ago, yet with profiteers touting such models?
- Naïve attention-perception schemes dominate AI, no?
- Ethical correctness is nice, but what about moral creativity?
- U.S./NATO's superior PAI machines must be curious and explore/monitor/probe, but mighn't curiosity then kill the ... AI?

Other Obstacles?

- The Elephant (again): AIs/Robots, *contra* Scharre, are weapons; where then is the military theory of human-AI combat? (Boyd, Hubin, ... okay okay; but AI?)
- What if we are forced to use foundation models, which are ethically stupid, failing as they do on the very challenges issued by the logicist-AI folks 40 years ago, yet with profiteers touting such models?
- Naïve attention-perception schemes dominate AI, no?
- Ethical correctness is nice, but what about moral creativity?
- U.S./NATO's superior PAI machines must be curious and explore/monitor/probe, but mighn't curiosity then kill the ... AI?
- What is the background “spiritual” content of articulated propositions?

Other Obstacles?

v. .5 WP

- The Elephant (again): AIs/Robots, *contra* Scharre, are weapons; where then is the military theory of human-AI combat? (Boyd, Hubin, ... okay okay; but AI?)

v. .5 WP

- What if we are forced to use foundation models, which are ethically stupid, failing as they do on the very challenges issued by the logicist-AI folks 40 years ago, yet with profiteers touting such models?

=>

- Naïve attention-perception schemes dominate AI, no?

c OLCSU

- Ethical correctness is nice, but what about moral creativity?

WP

- U.S./NATO's superior PAI machines must be curious and explore/monitor/probe, but mighn't curiosity then kill the ... AI?

?

- What is the background “spiritual” content of articulated propositions?

*Med nok penger, kan logikk
løse alle våre problemer.*