On International Paternalistic Taxation to Address The Mess That "Machine Learning" is Making

Selmer Bringsjord
(with Alexander Bringsjord of PwC & Naveen Sundar G., RAIR Lab)

Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

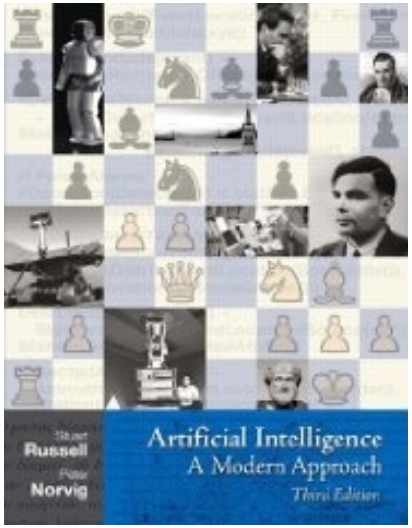
ICRES; 2020
ver. 0927200830NY

RAIR
Rensselaer AI and Reasoning Lab

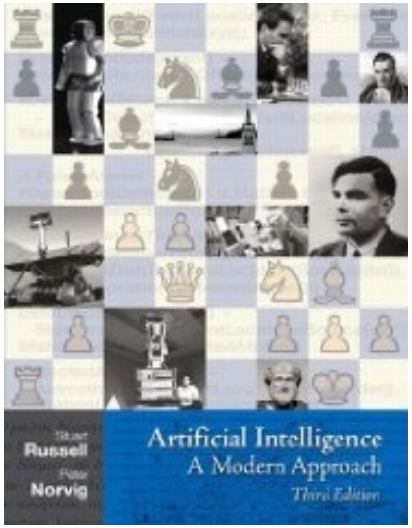
AI, ML, and MLn …

AI:

AI:


Russell
Norvig
Artificial Intelligence
A Modern Approach
Third Edition

AI:



Stanford Encyclopedia of Philosophy

Browse About Support SEP

Search SEP

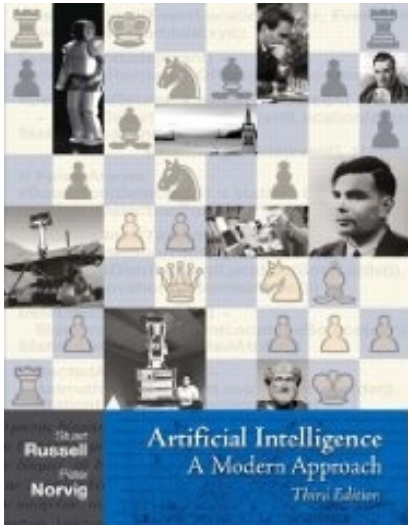
Entry Contents
Bibliography
Academic Tools
Friends PDF Preview
Author and Citation Info
Back to Top

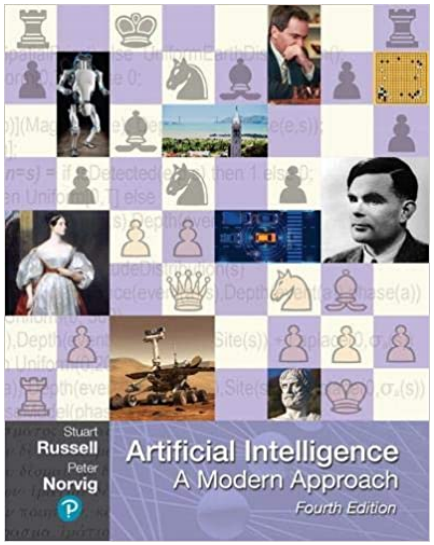
Artificial Intelligence

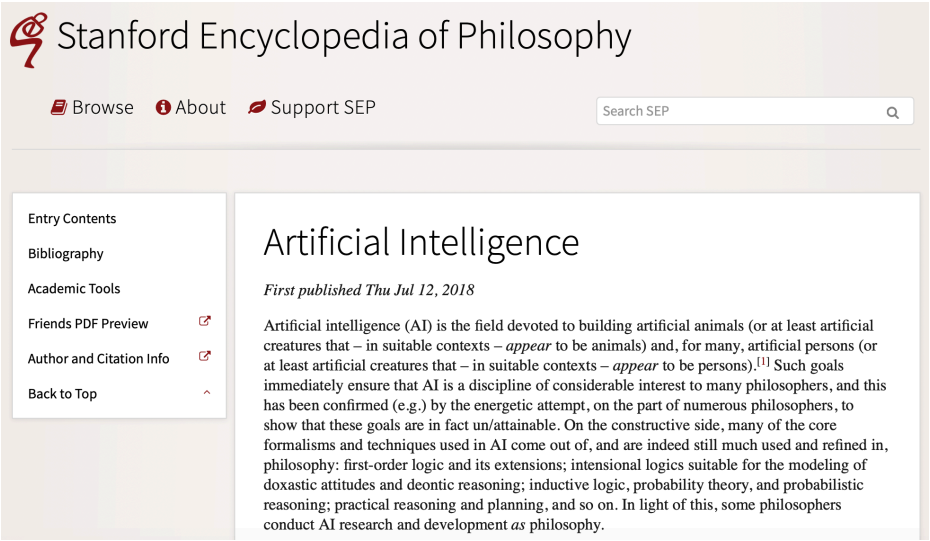
First published Thu Jul 12, 2018

Artificial intelligence (AI) is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons).[1] Such goals immediately ensure that AI is a discipline of considerable interest to many philosophers, and this has been confirmed (e.g.) by the energetic attempt, on the part of numerous philosophers, to show that these goals are in fact un/attainable. On the constructive side, many of the core formalisms and techniques used in AI come out of, and are indeed still much used and refined in, philosophy: first-order logic and its extensions; intensional logics suitable for the modeling of doxastic attitudes and deontic reasoning; inductive logic, probability theory, and probabilistic reasoning; practical reasoning and planning, and so on. In light of this, some philosophers conduct AI research and development *as* philosophy.

AI:







Stanford Encyclopedia of Philosophy

Browse About Support SEP

Search SEP

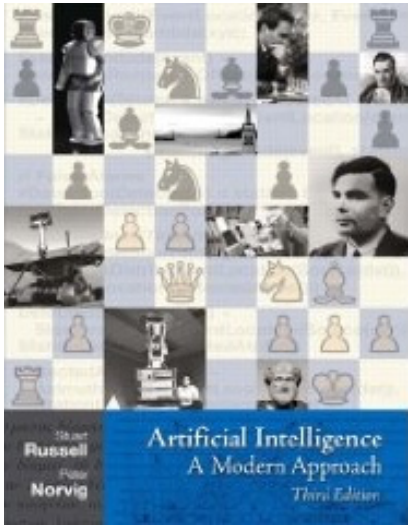
Entry Contents
Bibliography
Academic Tools
Friends PDF Preview
Author and Citation Info
Back to Top

Artificial Intelligence

First published Thu Jul 12, 2018

Artificial intelligence (AI) is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons).[1] Such goals immediately ensure that AI is a discipline of considerable interest to many philosophers, and this has been confirmed (e.g.) by the energetic attempt, on the part of numerous philosophers, to show that these goals are in fact un/attainable. On the constructive side, many of the core formalisms and techniques used in AI come out of, and are indeed still much used and refined in, philosophy: first-order logic and its extensions; intensional logics suitable for the modeling of doxastic attitudes and deontic reasoning; inductive logic, probability theory, and probabilistic reasoning; practical reasoning and planning, and so on. In light of this, some philosophers conduct AI research and development *as* philosophy.

AI:



Stanford Encyclopedia of Philosophy

Browse About Support SEP

Search SEP

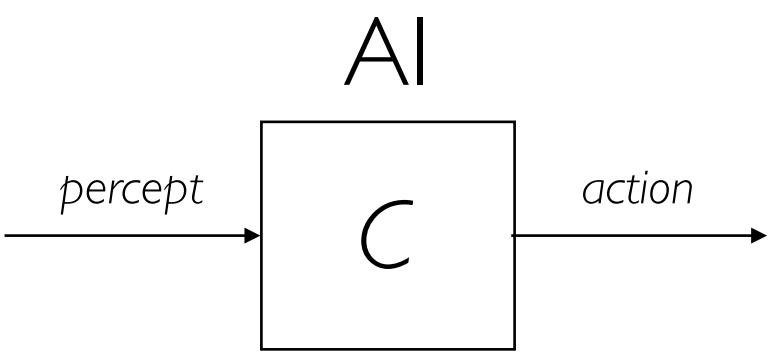
Entry Contents
Bibliography
Academic Tools
Friends PDF Preview
Author and Citation Info
Back to Top

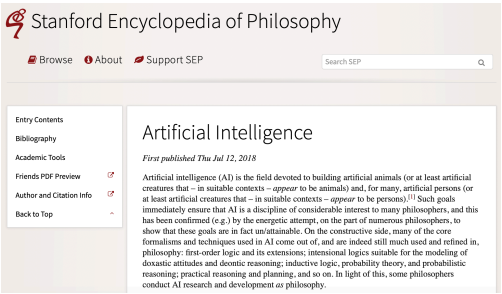
Artificial Intelligence

First published Thu Jul 12, 2018

Artificial intelligence (AI) is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons).[1] Such goals immediately ensure that AI is a discipline of considerable interest to many philosophers, and this has been confirmed (e.g.) by the energetic attempt, on the part of numerous philosophers, to show that these goals are in fact un/attainable. On the constructive side, many of the core formalisms and techniques used in AI come out of, and are indeed still much used and refined in, philosophy: first-order logic and its extensions; intensional logics suitable for the modeling of doxastic attitudes and deontic reasoning; inductive logic, probability theory, and probabilistic reasoning; practical reasoning and planning, and so on. In light of this, some philosophers conduct AI research and development *as* philosophy.

AI:





Stanford Encyclopedia of Philosophy

Browse About Support SEP

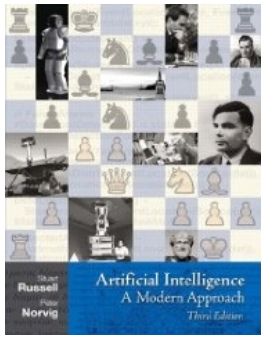
Search SEP

Entry Contents
Bibliography
Academic Tools
Friends PDF Preview
Author and Citation Info
Back to Top

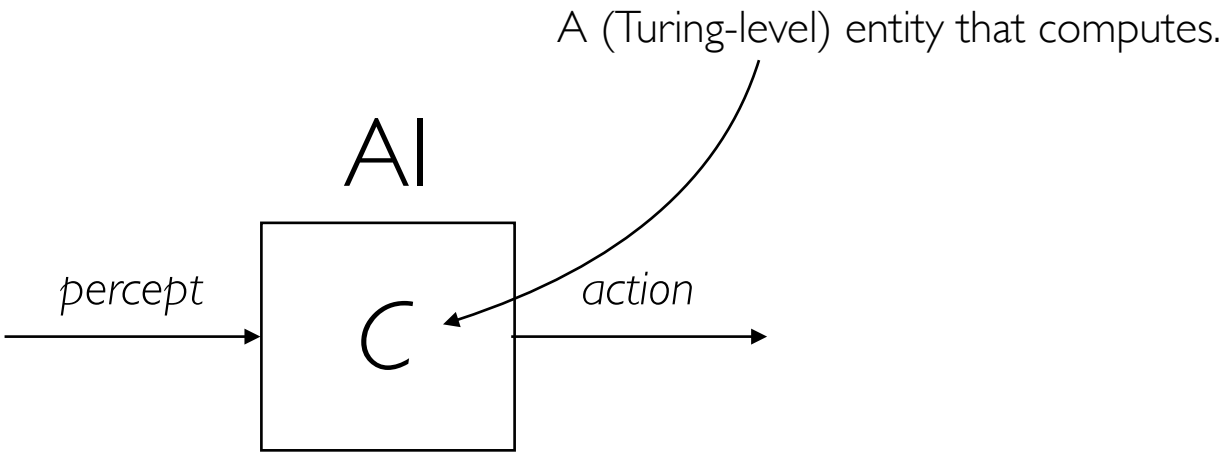
Artificial Intelligence

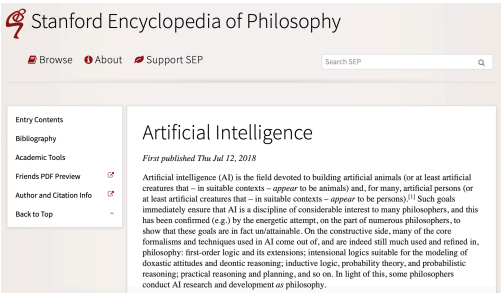
First published Thu Jul 12, 2018

Artificial intelligence (AI) is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons).[1] Such goals immediately ensure that AI is a discipline of considerable interest to many philosophers, and this has been confirmed (e.g.) by the energetic attempt, on the part of numerous philosophers, to show that these goals are in fact un/attainable. On the constructive side, many of the core formalisms and techniques used in AI come out of, and are indeed still much used and refined in, philosophy: first-order logic and its extensions; intensional logics suitable for the modeling of doxastic attitudes and deontic reasoning; inductive logic, probability theory, and probabilistic reasoning; practical reasoning and planning, and so on. In light of this, some philosophers conduct AI research and development *as* philosophy.



AI:





Stanford Encyclopedia of Philosophy

Browse About Support SEP

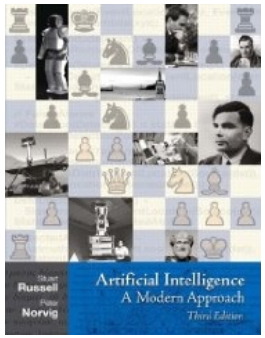
Search SEP

Entry Contents
Bibliography
Academic Tools
Friends PDF Preview
Author and Citation Info
Back to Top

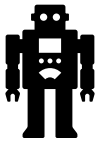
Artificial Intelligence

First published Thu Jul 12, 2018

Artificial intelligence (AI) is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons).[1] Such goals immediately ensure that AI is a discipline of considerable interest to many philosophers, and this has been confirmed (e.g.) by the energetic attempt, on the part of numerous philosophers, to show that these goals are in fact un/attainable. On the constructive side, many of the core formalisms and techniques used in AI come out of, and are indeed still much used and refined in, philosophy: first-order logic and its extensions; intensional logics suitable for the modeling of doxastic attitudes and deontic reasoning; inductive logic, probability theory, and probabilistic reasoning; practical reasoning and planning, and so on. In light of this, some philosophers conduct AI research and development *as* philosophy.

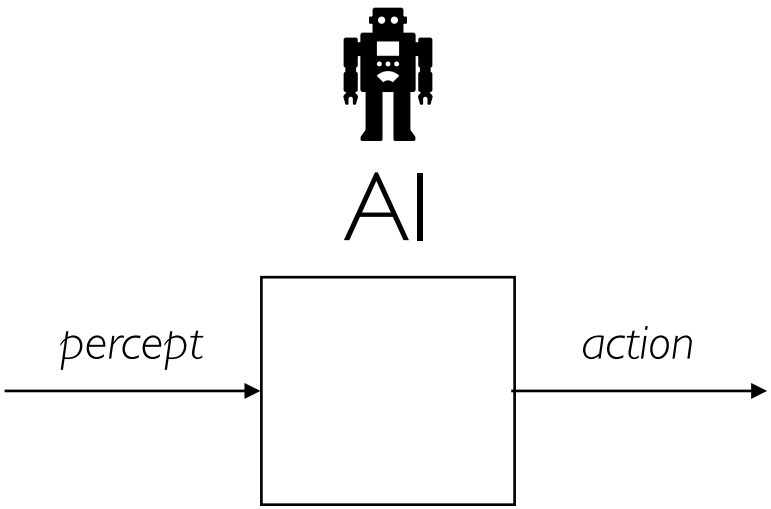


AI:MLn



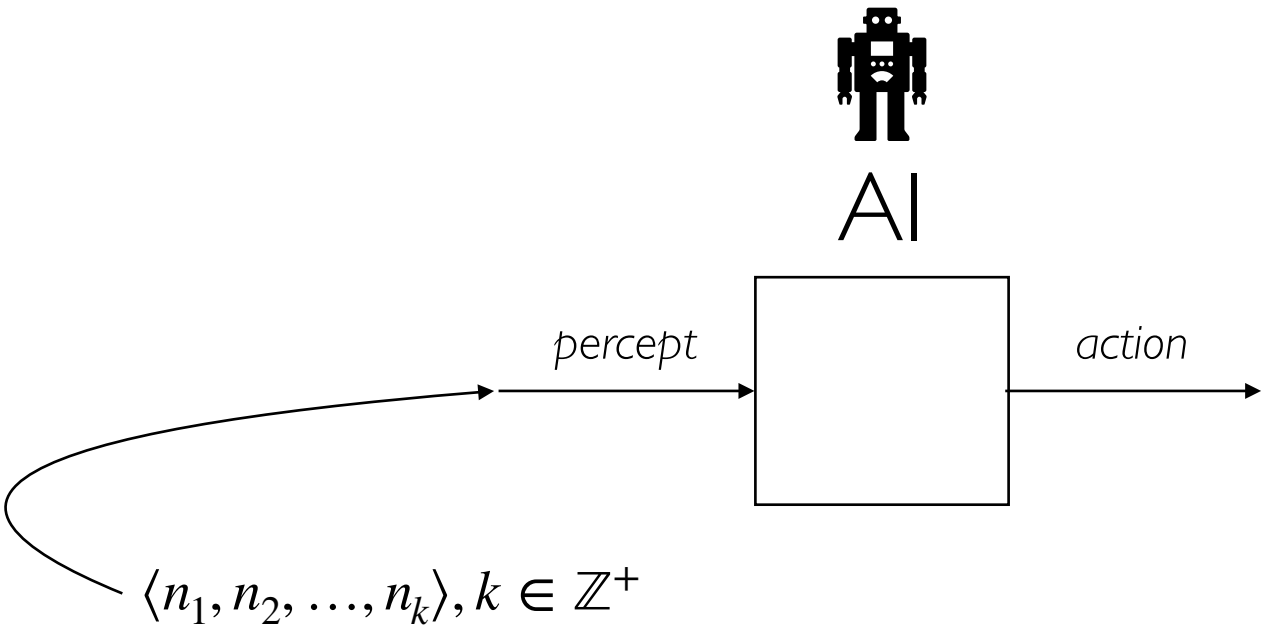
AI

AI:MLn

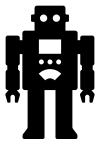


$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$

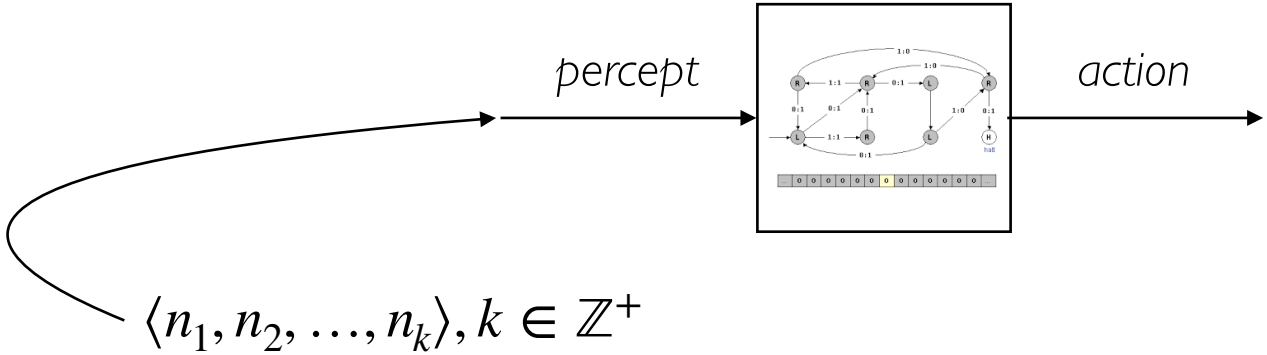
AI:MLn



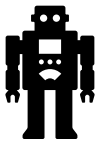
AI:MLn



AI



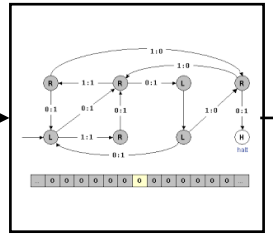
AI:MLn



AI

A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.

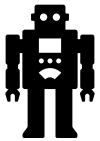
percept



action

$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$

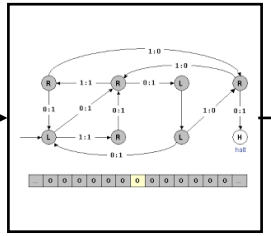
AI:MLn



AI

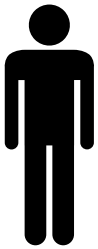
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.

percept

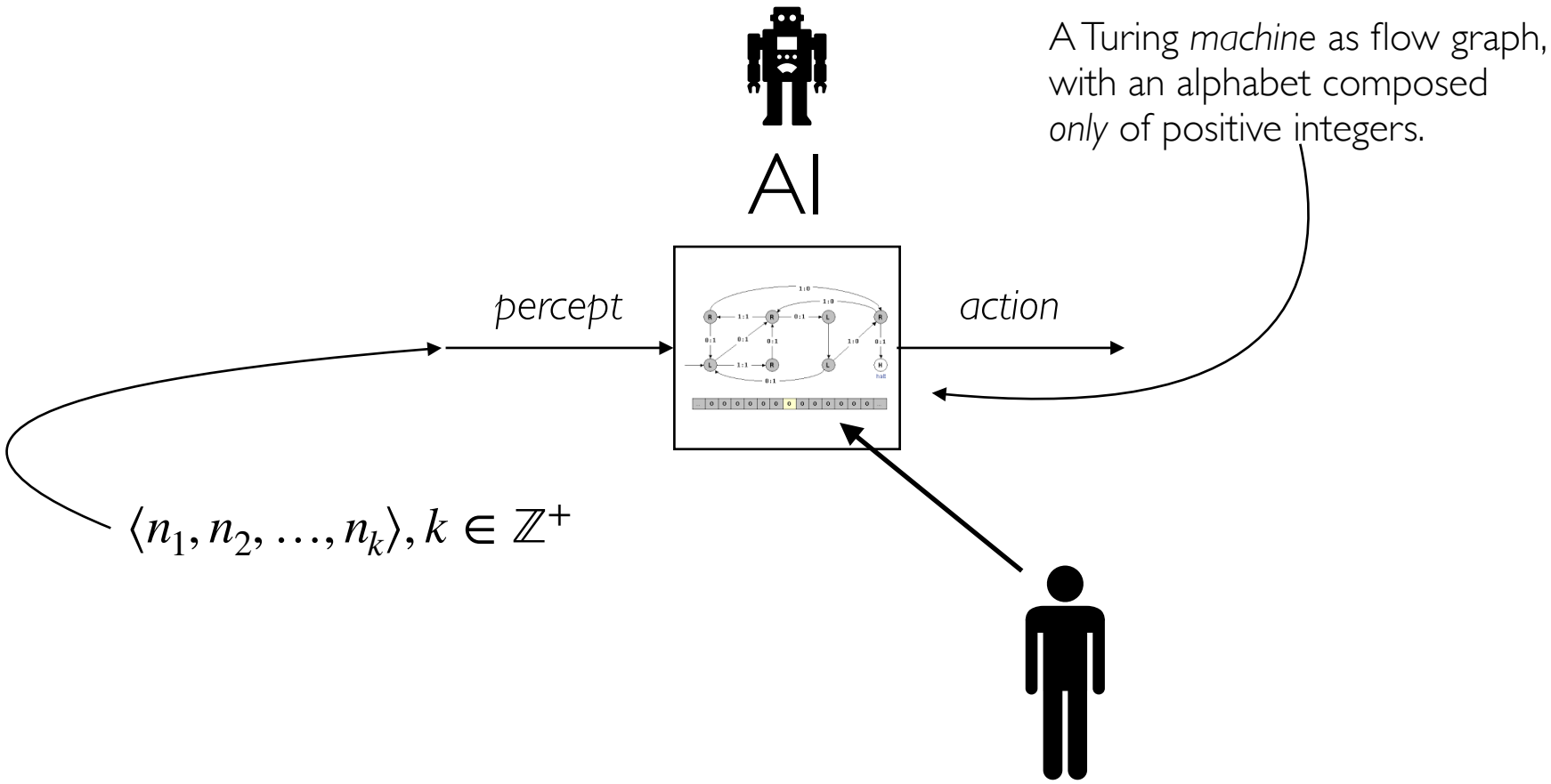


action

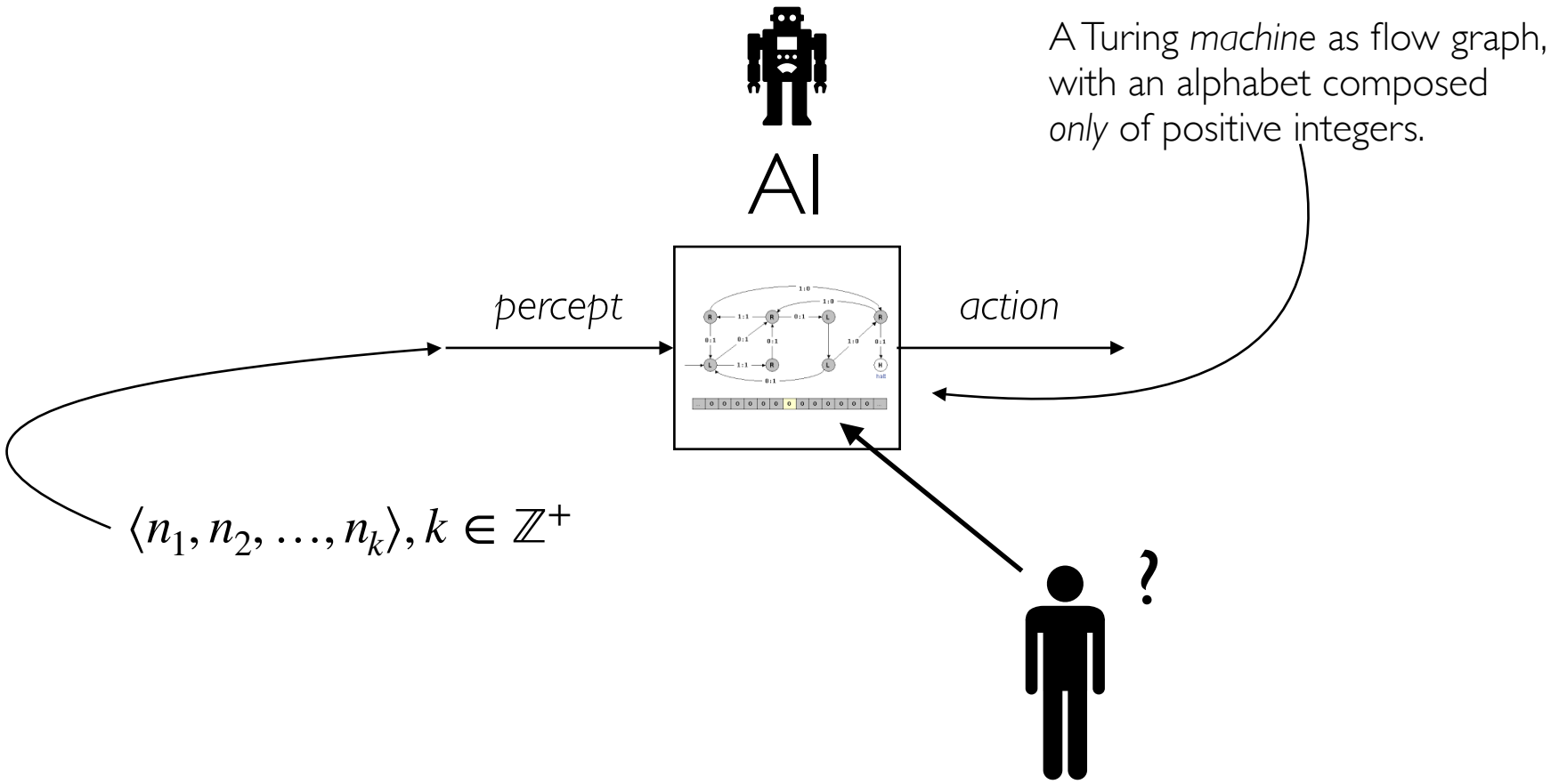
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$



AI:MLn


AI
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.
percept
action
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$

AI:MLn


AI
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.
percept
action
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$
?

J. Expt. Theor. Artif. Intell. 1(1991)319–349.

Research Paper

Is the connectionist–logicist clash one of AI's wonderful red herrings?[1]

SELMER BRINGSJORD

Abstract. A careful adjudication of the connectionist–logicist clash in AI and cognitive science seems to disclose that it is a mirage.

Keywords: logicism, connectionism, symbol systems, neural networks

Received 20 March 1991; revised 10 July 1991

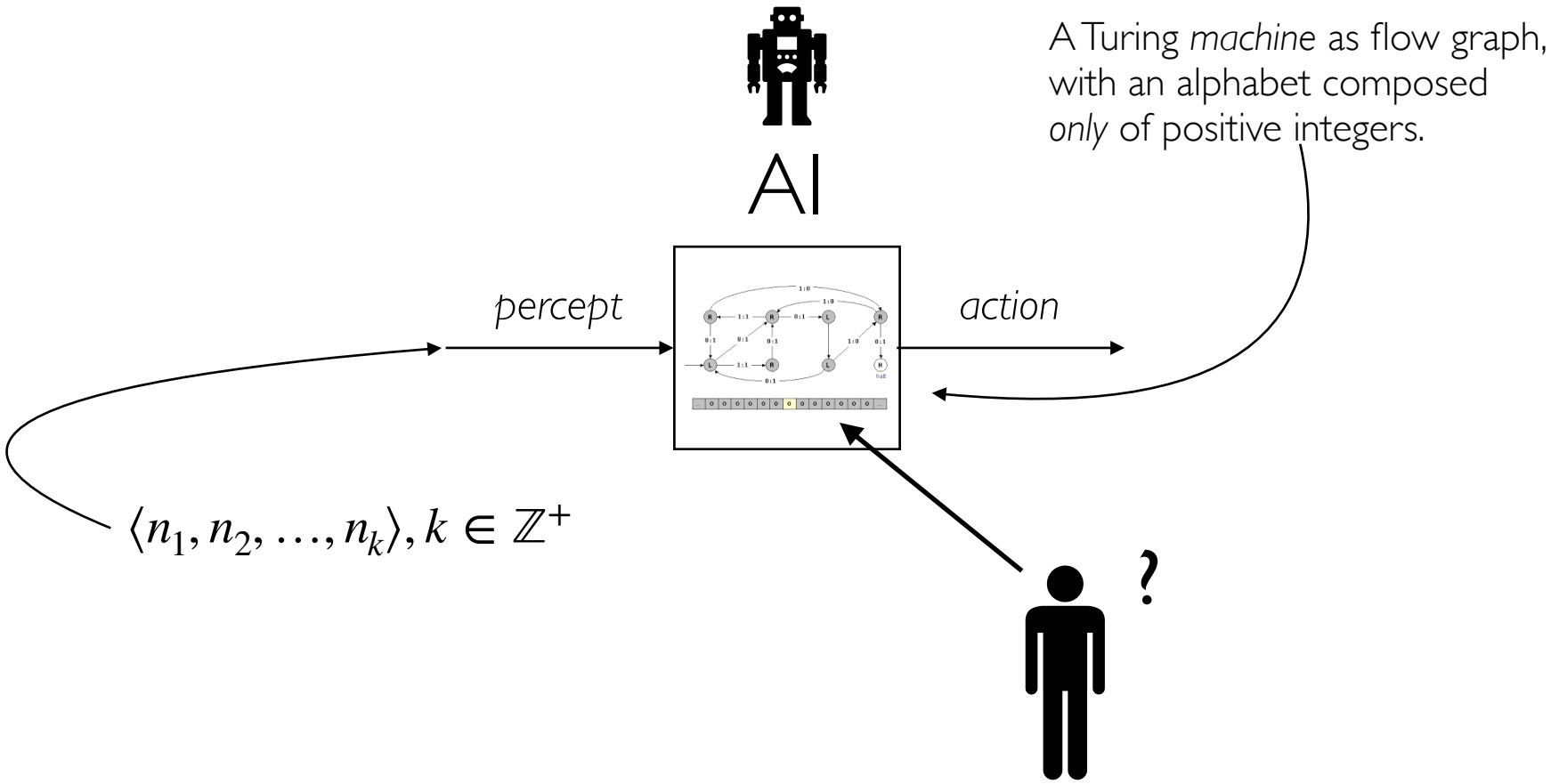
1. Introduction

An often unfriendly debate continues to rage in AI and cognitive science between 'logicists' and 'connectionists.' Many connectionists (e.g. Smolensky 1988a, Churchland & Churchland 1990, Waltz 1988, Schwartz 1988, Kaplan *et al.* 1990, Horgan and Tienson 1989) hold that their doctrines ought to supplant or at least supplement the logic-based ones of traditional logicist or symbolicist (or 'strong,' Searle 1980a,b, 1982; 'good old-fashioned,' Haugeland 1986; 'old hand,' Doyle 1988; 'person building,' Charniak & McDermott 1985, Nilsson & Genesereth 1988, Pollock 1989) AI. On the other hand, many logicists (e.g. Fodor & Pylyshyn 1988 hold that any successful AI model of human cognition, and *a fortiori* any sentient artificial intelligence itself, must use classical, logic-driven architecture.

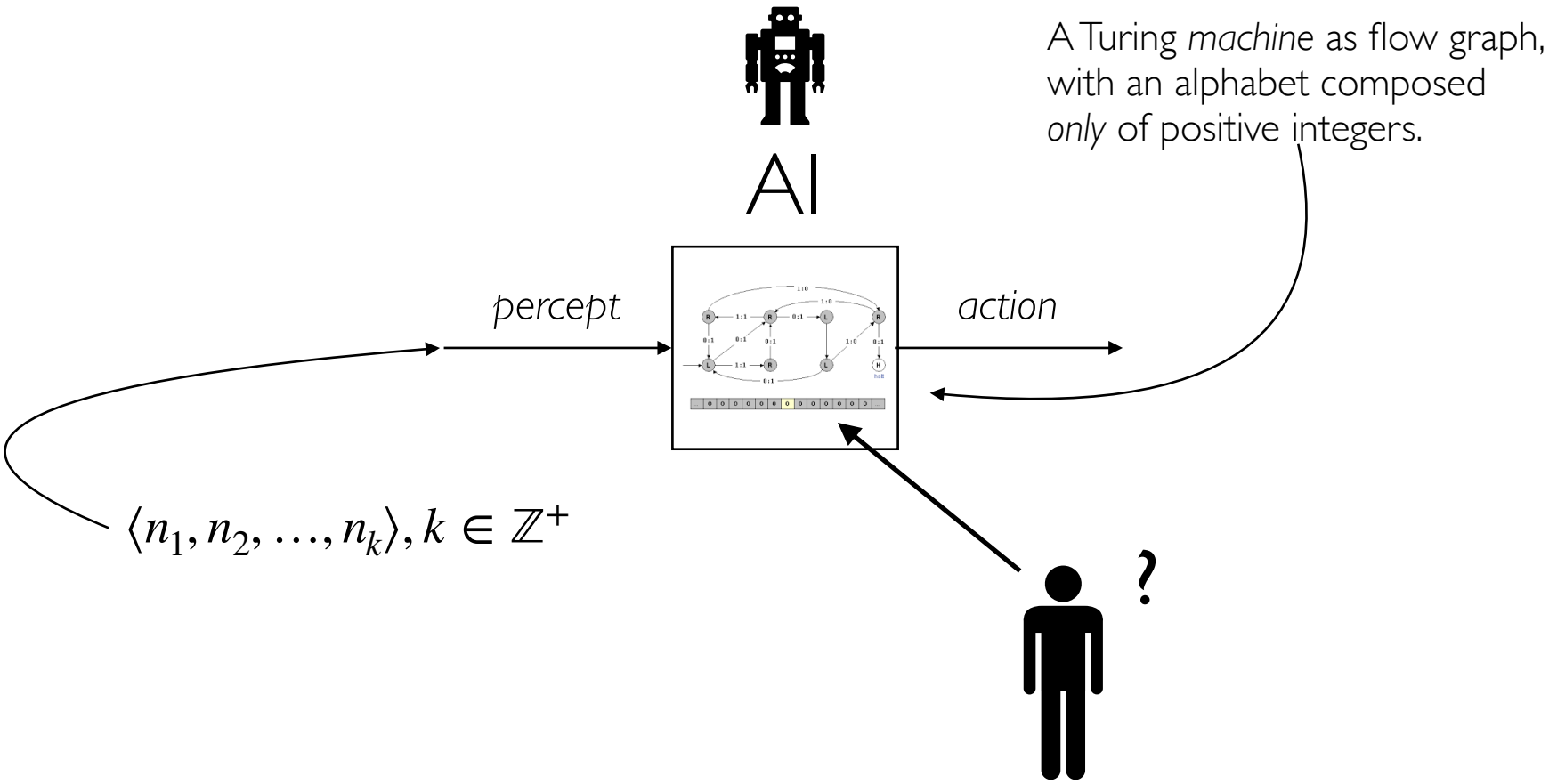
Herein I will attempt to adjudicate this clash—by, in a word, showing it to be 'one of AI's wonderful red herrings.'[2] In order to carry out this adjudication, I will need to adopt an approach that relies heavily on the formalization of declarative English sentences within first-order logic. There is considerable irony in the fact that the only rigorous method open to one seeking adjudication of the clash in question is one with which the logicist is likely to be comfortable, but one that is, if not anathema, then at least a bit foreign, to a connectionist more at home with differential equations than syllogisms. While human cognition, from where we stand at the moment, may or may not be profitably identified with the connectionist's so-called dynamical systems, one thing *is* clear, even at this stage, before embarking on our inquiry: the connectionist-logicist *clash* isn't treatable as such a system. It is, rather, treatable, if at all, as a clash of *propositions* thought by their proponents to be true and their opponents to be false.[3] A logic-based adjudication of the debate in question is something connectionists, most prominently (Smolensky 1988a), have themselves attempted, but not, in my opinion, pulled off. (Later, we will see that Smolensky's 1988a 'declarativization' of the clash leaves much to be desired.) At any rate, let us begin: let $\mathscr{C}$ ($\mathscr{L}$) denote connectionism (logicism) in the form of some as-yet-unarticulated set of

0952–813X/91 $3.00 © 1991 Taylor & Francis Ltd

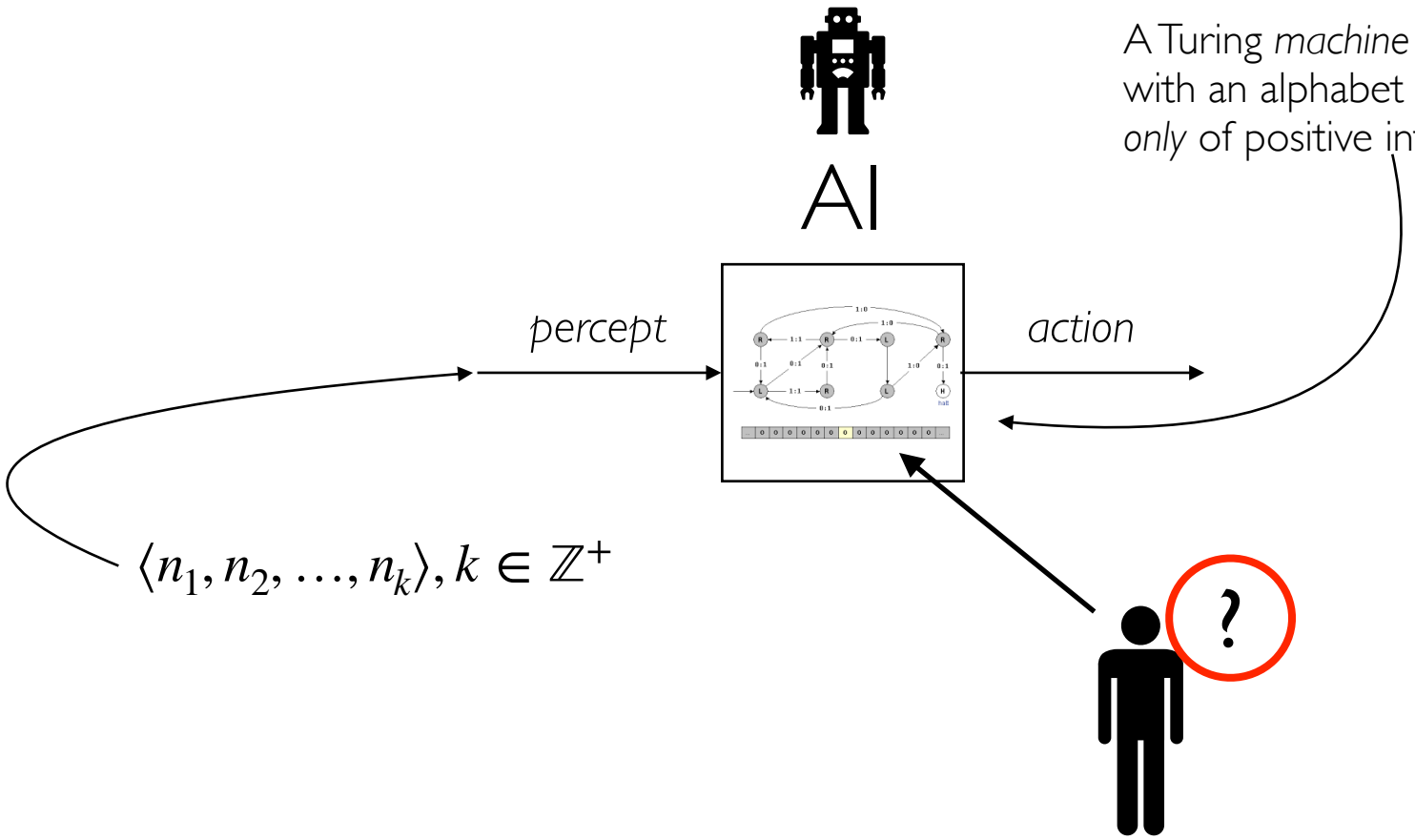
AI:MLn


AI
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.
percept
action
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$
?

AI:MLn


AI
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.
percept
action
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$
?

AI:MLn


AI
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.
percept
action
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$
?

By the way, we *do* know what real learning is …

Real Learning

Bringsjord, S., Govindarajulu, N., Banerjee, S. & Hummel, J. (2018) "Do Machine-Learning Machines Learn?" in V. Müller, ed., *Philosophy and Theory of Artificial Intelligence* 2017, Springer SAPERE, Berlin, Germany, pp. 136–157.

Real Learning

Bringsjord, S., Govindarajulu, N., Banerjee, S. & Hummel, J. (2018) "Do Machine-Learning Machines Learn?" in V. Müller, ed., *Philosophy and Theory of Artificial Intelligence* 2017, Springer SAPERE, Berlin, Germany, pp. 136–157.

- Introducing and characterizing **"Real Learning" ($\mathscr{RL}$)** in the formal sciences
- Agent ***a*** has *really learned* a unary function ***f*** : N → N only if
-
- Real learning is neither found in AI nor in psychology and its allied disciplines

Real Learning

Bringsjord, S., Govindarajulu, N., Banerjee, S. & Hummel, J. (2018) "Do Machine-Learning Machines Learn?" in V. Müller, ed., *Philosophy and Theory of Artificial Intelligence* 2017, Springer SAPERE, Berlin, Germany, pp. 136–157.

- Introducing and characterizing **"Real Learning" ($\mathscr{RL}$)** in the formal sciences
- Agent ***a*** has *really learned* a unary function ***f*** : N → N only if

(c1) $\mathfrak{a}$ understands the formal definition D_f of f,
(c2) can[4] produce both $f(x)$ for all $x \in \mathbb{N}$, and
(c3) a proof of the correctness of what is supplied in (c2).

- Real learning is neither found in AI nor in psychology and its allied disciplines

Real Learning

Bringsjord, S., Govindarajulu, N., Banerjee, S. & Hummel, J. (2018) "Do Machine-Learning Machines Learn?" in V. Müller, ed., *Philosophy and Theory of Artificial Intelligence 2017*, Springer SAPERE, Berlin, Germany, pp. 136–157.

- Introducing and characterizing **"Real Learning" ($\mathscr{RL}$)** in the formal sciences
- Agent ***a*** has *really learned* a unary function ***f*** : N → N only if

(c1) $\mathfrak{a}$ understands the formal definition D_f of f,
(c2) can[4] produce both $f(x)$ for all $x \in \mathbb{N}$, and
(c3) a proof of the correctness of what is supplied in (c2).

(c1′) $\mathfrak{a}$ can correctly answer test questions regarding the formal definition D_f of f, where the answers in each case are accompanied by correct proofs[7] discovered, expressed, and provided by $\mathfrak{a}$.

- Real learning is neither found in AI nor in psychology and its allied disciplines

Real Learning

Bringsjord, S., Govindarajulu, N., Banerjee, S. & Hummel, J. (2018) "Do Machine-Learning Machines Learn?" in V. Müller, ed., *Philosophy and Theory of Artificial Intelligence 2017*, Springer SAPERE, Berlin, Germany, pp. 136–157.

- Introducing and characterizing **"Real Learning" ($\mathscr{RL}$)** in the formal sciences
- Agent ***a*** has *really learned* a unary function ***f*** : N → N only if

(c1) $\mathfrak{a}$ understands the formal definition D_f of f,
(c2) can[4] produce both $f(x)$ for all $x \in \mathbb{N}$, and
(c3) a proof of the correctness of what is supplied in (c2).

(c1′) $\mathfrak{a}$ can correctly answer test questions regarding the formal definition D_f of f, where the answers in each case are accompanied by correct proofs[7] discovered, expressed, and provided by $\mathfrak{a}$.

- Real learning is neither found in AI nor in psychology and its allied disciplines

Real Learning

Bringsjord, S., Govindarajulu, N., Banerjee, S. & Hummel, J. (2018) "Do Machine-Learning Machines Learn?" in V. Müller, ed., *Philosophy and Theory of Artificial Intelligence 2017*, Springer SAPERE, Berlin, Germany, pp. 136–157.

- Introducing and characterizing **"Real Learning" ($\mathscr{RL}$)** in the formal sciences
- Agent ***a*** has *really learned* a unary function ***f*** : N → N only if

(c1) $\mathfrak{a}$ understands the formal definition D_f of f,
(c2) can[4] produce both $f(x)$ for all $x \in \mathbb{N}$, and
(c3) a proof of the correctness of what is supplied in (c2).

(c1′) $\mathfrak{a}$ can correctly answer test questions regarding the formal definition D_f of f, where the answers in each case are accompanied by correct proofs[7] discovered, expressed, and provided by $\mathfrak{a}$.

- Real learning is neither found in AI nor in psychology and its allied disciplines

The Four-Step Road to Real Learning

The Four-Step Road to Real Learning

- Step 1: Observe the acute discontinuity of human vs. nonhuman cognition. (Only humans understand and employ e.g. abstract reasoning schemas unaffected by the physical; layered quantification; recursion; and infinite structures/infinitary reasoning.)

The Four-Step Road to Real Learning

- Step 1: Observe the acute discontinuity of human vs. nonhuman cognition. (Only humans understand and employ e.g. abstract reasoning schemas unaffected by the physical; layered quantification; recursion; and infinite structures/infinitary reasoning.)

- Step 2: Exclude forms of "learning" made possible via exclusive use of reasoning and communication capacities in nonhuman animals (i.e. exclude forms of "learning" that don't eventuate in bona fide jtb *knowledge*).

The Four-Step Road to Real Learning

- Step 1: Observe the acute discontinuity of human vs. nonhuman cognition. (Only humans understand and employ e.g. abstract reasoning schemas unaffected by the physical; layered quantification; recursion; and infinite structures/infinitary reasoning.)

- Step 2: Exclude forms of "learning" made possible via exclusive use of reasoning and communication capacities in nonhuman animals (i.e. exclude forms of "learning" that don't eventuate in bona fide jtb *knowledge*).

- Step 3: Within the focus arising from Step 2, further narrow the focus to $HL^{\geq}$ reasoning and communication sufficiently powerful to perceive, and be successfully applied to, both (i) cohesive bodies of declarative content, and (ii) sophisticated natural-language content. Dub this **RC**.

The Four-Step Road to Real Learning

- Step 1: Observe the acute discontinuity of human vs. nonhuman cognition. (Only humans understand and employ e.g. abstract reasoning schemas unaffected by the physical; layered quantification; recursion; and infinite structures/infinitary reasoning.)

- Step 2: Exclude forms of "learning" made possible via exclusive use of reasoning and communication capacities in nonhuman animals (i.e. exclude forms of "learning" that don't eventuate in bona fide jtb *knowledge*).

- Step 3: Within the focus arising from Step 2, further narrow the focus to $HL^{\geq}$ reasoning and communication sufficiently powerful to perceive, and be successfully applied to, both (i) cohesive bodies of declarative content, and (ii) sophisticated natural-language content. Dub this **RC**.

- Step 4: Real Learning ($\mathscr{RL}$) is the acquisition of genuine knowledge via **RC**.

To render this computational, suitable for implementation, a novel Form of Machine Learning:

Learning *Ex Nihilo*

(or Learning *Ex Minima*)

To render this computational, suitable for implementation, a novel Form of Machine Learning:

Learning *Ex Nihilo*

(or Learning *Ex Minima*)

EPiC Series in Computing
Volume 72, 2020, Pages 1–27
GCAI 2020. 6th Global Conference
on Artificial Intelligence (GCAI 2020)

Learning *Ex Nihilo*

Selmer Bringsjord[1], Naveen Sundar Govindarajulu[1], John Licato[2], and Michael Giancola[1]

[1] Rensselaer AI & Reasoning (RAIR) Lab,
Rensselaer Polytechnic Institute, Troy NY 12180 USA
[2] Advancing Machine and Human Reasoning (AMHR) Lab,
University of South Florida, Tampa FL 33620 USA
{ selmer.bringsjord, naveen.sundar.g, mike.j.giancola, john.licato } @gmail.com

Abstract

This paper introduces, philosophically and to a degree formally, the novel concept of learning *ex nihilo*, intended (obviously) to be analogous to the concept of creation *ex nihilo*. Learning *ex nihilo* is an agent's learning "from nothing", by the suitable employment of inference schemata for deductive and inductive reasoning. This reasoning must be in machine-verifiable accord with a formal proof/argument theory in a *cognitive calculus* (i.e., here, roughly, an intensional higher-order multi-operator quantified logic), and this reasoning is applied to percepts received by the agent, in the context of both some prior knowledge, and some prior and current interests. Learning *ex nihilo* is a challenge to contemporary forms of ML, indeed a severe one, but the challenge is here offered in the spirit of seeking to stimulate attempts, on the part of non-logicist ML researchers and engineers, to collaborate with those in possession of learning-*ex nihilo* frameworks, and eventually attempts to integrate directly with such frameworks at the implementation level. Such integration will require, among other things, the symbiotic interoperation of state-of-the-art automated reasoners and high-expressivity planners, with statistical/connectionist ML technology.

1 Introduction

This paper introduces, philosophically and to a degree logico-mathematically, the novel concept of learning *ex nihilo*, intended (obviously) to be analogous to the concept of creation *ex nihilo*.[1] Learning *ex nihilo* is an agent's learning "from nothing," by the suitable employment of inference schemata for deductive and inductive[2] (e.g., analogical, enumerative-inductive, abductive, etc.) reasoning. This reasoning must be in machine-verifiable accord with a formal

[1] No such assumption as that creation *ex nihilo* is real or even formally respectable is made or needed in the present paper. The concept of creation *ex nihilo* is simply for us an intellectual inspiration — but as a matter of fact, the literature on it in analytic philosophy does provide some surprisingly rigorous accounts. In the present draft of the present paper, we don't seek to mine these accounts.

[2] Not to be confused with inductive logic programming (about which more will be said later), or inductive deductive techniques and schemas (e.g. mathematical induction, the induction schema in Peano Arithmetic, etc.). As we explain later, learning *ex nihilo* is in part powered by non-deductive inference schemata seen in inductive logic. An introductory overview of inductive logic is provided in [39].

G. Danoy, J. Pang and G. Sutcliffe (eds.), GCAI 2020 (EPiC Series in Computing, vol. 72), pp. 1–27

Bringsjord, S., Govindarajulu, N.S., Licato, J. & Giancola, M. (2020) "Learning *Ex Nihilo*" *Proceedings of the 6th Global Conference on Artificial Intelligence* (GCAI 2020), within *International Conferences on Logic and Artificial Intelligence* at Zhejiang University (ZJULogAI), in Danoy, G., Pang, J. & Sutcliffe, G., eds., EPiC Series in Computing **72**: 1–27 (Manchester, UK: EasyChair Ltd), ISSN: 2398-7340.
https://easychair.org/publications/paper/NzWG

Bringsjord, S., Govindarajulu, N.S., Licato, J. & Giancola, M. (2020) "Learning *Ex Nihilo*" *Proceedings of the 6th Global Conference on Artificial Intelligence* (GCAI 2020), within *International Conferences on Logic and Artificial Intelligence* at Zhejiang University (ZJULogAI), in Danoy, G., Pang, J. & Sutcliffe, G., eds., EPiC Series in Computing **72**: 1–27 (Manchester, UK: EasyChair Ltd), ISSN: 2398-7340. https://easychair.org/publications/paper/NzWG

Bringsjord, S., Govindarajulu, N.S., Licato, J. & Giancola, M. (2020) "Learning *Ex Nihilo*" *Proceedings of the 6th Global Conference on Artificial Intelligence* (GCAI 2020), within *International Conferences on Logic and Artificial Intelligence* at Zhejiang University (ZJULogAI), in Danoy, G., Pang, J. & Sutcliffe, G., eds., EPiC Series in Computing **72**: 1–27 (Manchester, UK: EasyChair Ltd), ISSN: 2398-7340. https://easychair.org/publications/paper/NzWG

The Problem:
<u>MLn is a Non-Starter</u> …

First, we note …

First, we note …

The PAID Problem

First, we note …

The PAID Problem

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

The PAID Problem:

The PAID Problem:

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

The PAID Problem:

$\forall$x : Agents

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

The PAID Problem:

$\forall$x : Agents

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

The PAID Problem:

$\forall$x : Agents

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

The PAID Problem:

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

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

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

The PAID Problem:

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

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

$\downarrow \qquad \downarrow \ldots \qquad \downarrow \ldots \qquad \downarrow \ldots$

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

The Coming Software Apocalypse

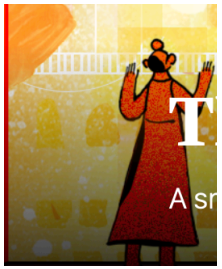
A small group of programmers wants to change how we code—before catastrophe strikes.

In September 2007, Jean Bookout was driving on the highway with her best friend in a Toyota Camry when the accelerator seemed to get stuck. When she took her foot off the pedal, the car didn't slow down. She tried the brakes but they seemed to have lost their power. As she swerved toward an off-ramp going 50 miles per hour, she pulled the emergency brake. The car left a skid mark 150 feet long before running into an embankment by the side of the road. The passenger was killed. Bookout woke up in a hospital a month later.

The incident was one of many in a nearly decade-long investigation into claims of so-called unintended acceleration in Toyota cars. Toyota blamed the incidents on poorly designed floor mats, "sticky" pedals, and driver error, but outsiders suspected that faulty software might be responsible. The National Highway Traffic Safety Administration enlisted software experts from NASA to perform an intensive review of Toyota's code. After nearly 10 months, the NASA team hadn't found evidence that software was the cause—but said they couldn't prove it wasn't.

It was during litigation of the Bookout accident that someone finally found a convincing connection. Michael Barr, an expert witness for the plaintiff, had a team of software experts spend 18 months with the Toyota code, picking up where NASA left off. Barr described what they found as "spaghetti code," programmer lingo for software that has become a tangled mess. Code turns to spaghetti when it accretes over many years, with feature after feature piling on top of, and being woven around, what's already there; eventually the code becomes impossible to follow, let alone to test exhaustively for flaws.

mark 150 feet long before running into an embankment by the side of the road. The passenger was killed. Bookout woke up in a hospital a month later.



The incident was one of many in a nearly decade-long investigation into claims of so-called unintended acceleration in Toyota cars. Toyota blamed the incidents on poorly designed floor mats, “sticky” pedals, and driver error, but outsiders suspected that faulty software might be responsible. The National Highway Traffic Safety Administration enlisted software experts from NASA to perform an intensive review of Toyota’s code. After nearly 10 months, the NASA team hadn’t found evidence that software was the cause—but said they couldn’t prove it wasn’t.

It was during litigation of the Bookout accident that someone finally found a convincing connection. Michael Barr, an expert witness for the plaintiff, had a team of software experts spend 18 months with the Toyota code, picking up where NASA left off. Barr described what they found as “spaghetti code,” programmer lingo for software that has become a tangled mess. Code turns to spaghetti when it accretes over many years, with feature after feature piling on top of, and being woven around, what’s already there; eventually the code becomes impossible to follow, let alone to test exhaustively for flaws.

The Solution:

The Solution:

The Four Steps

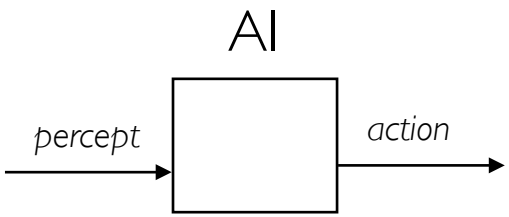
The Solution:

The Four Steps

The Logico-Mathematical Basis

Logic

$\mathcal{L}$



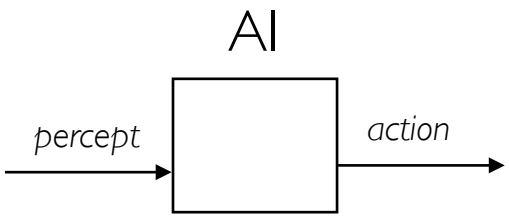
$$\mathcal{L} := \langle L, \mathcal{I} \rangle \qquad \frac{\begin{array}{cc} \mathbb{P} & L \\ \mathfrak{q} & L \end{array}}{\begin{array}{c} \mathbb{R} \ : \ \langle \mathbb{P}, \mathfrak{q} \rangle \longrightarrow \langle \mathtt{Y} | \mathtt{N} | \mathtt{U}, \delta, \pi_{(s)} | \alpha_{(s)} \rangle \\ \mathbb{C} \ : \ \pi_{(s)} | \alpha_{(s)} \longrightarrow \langle \mathtt{Y} | \mathtt{N} | \mathtt{U}, \delta \rangle \end{array}}$$

Pure General Logic Programming

The Logico-Mathematical Basis

Logic

$\mathcal{L}$



$$\mathcal{L} := \langle L, \mathcal{I} \rangle \qquad \frac{\begin{array}{cc} \mathbb{P} & L \\ \mathfrak{q} & L \end{array}}{\begin{array}{c} \mathbb{R} \ : \ \langle \mathbb{P}, \mathfrak{q} \rangle \longrightarrow \langle \mathtt{Y} | \mathtt{N} | \mathtt{U}, \delta, \pi_{(s)} | \alpha_{(s)} \rangle \\ \mathbb{C} \ : \ \pi_{(s)} | \alpha_{(s)} \longrightarrow \langle \mathtt{Y} | \mathtt{N} | \mathtt{U}, \delta \rangle \end{array}}$$

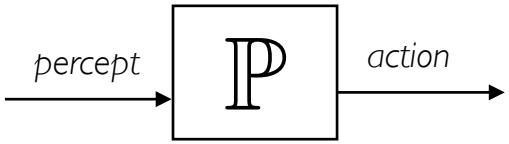
Pure General Logic Programming

The Logico-Mathematical Basis

Logic

$\mathcal{L}$

AI



$$\mathcal{L} := \langle L, \mathcal{I} \rangle \qquad \frac{\begin{array}{cc} \mathbb{P} & L \\ \mathfrak{q} & L \end{array}}{\begin{array}{c} \mathbb{R} \ : \ \langle \mathbb{P}, \mathfrak{q} \rangle \longrightarrow \langle \mathtt{Y} | \mathtt{N} | \mathtt{U}, \delta, \pi_{(s)} | \alpha_{(s)} \rangle \\ \mathbb{C} \ : \ \pi_{(s)} | \alpha_{(s)} \longrightarrow \langle \mathtt{Y} | \mathtt{N} | \mathtt{U}, \delta \rangle \end{array}}$$

Pure General Logic Programming

Novel, Easy Program Verification Thereby Enabled

Minds & Machines (2007) 17:185–202
DOI 10.1007/s11023-007-9063-5

Computers, Justification, and Mathematical Knowledge

Konstantine Arkoudas · Selmer Bringsjord

Published online: 23 June 2007
© Springer Science+Business Media B.V. 2007

Abstract The original proof of the four-color theorem by Appel and Haken sparked a controversy when Tymoczko used it to argue that the justification provided by unsurveyable proofs carried out by computers cannot be a priori. It also created a lingering impression to the effect that such proofs depend heavily for their soundness on large amounts of computation-intensive custom-built software. Contra Tymoczko, we argue that the justification provided by certain computerized mathematical proofs is not fundamentally different from that provided by surveyable proofs, and can be sensibly regarded as a priori. We also show that the aforementioned impression is mistaken because it fails to distinguish between proof search (the context of discovery) and proof checking (the context of justification). By using mechanized proof assistants capable of producing certificates that can be independently checked, it is possible to carry out complex proofs without the need to trust arbitrary custom-written code. We only need to trust one fixed, small, and simple piece of software: the proof checker. This is not only possible in principle, but is in fact becoming a viable methodology for performing complicated mathematical reasoning. This is evinced by a new proof of the four-color theorem that appeared in 2005, and which was developed and checked in its entirety by a mechanical proof system.

Keywords A priori · Justification · Proofs · Certificates · Four-color theorem · Mathematical knowledge

K. Arkoudas (✉) · S. Bringsjord
Cognitive Science Department, Computer Science Department, RPI, Troy, NY, USA 12180
e-mail: arkouk@rpi.edu

S. Bringsjord
e-mail: brings@rpi.edu

Springer

A Vindication of Program Verification

Selmer Bringsjord
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology
Rensselaer AI & Reasoning Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
selmer@rpi.edu

Received 00 Month 200x; final version received 00 Month 200x

Fetzer famously claims that program verification isn't even a theoretical possibility, and offers a certain argument for this far-reaching claim. Unfortunately for Fetzer, and like-minded thinkers, this position-argument pair, while based on a seminal insight that program verification, despite its Platonic proof-theoretic airs, is plagued by the inevitable unreliability of messy, real-world causation, is demonstrably self-refuting. As I soon show, Fetzer (and indeed anyone else who provides an argument- or proof-based attack on program verification) is like the person who claims: "My sole claim is that every claim expressed by an English sentence and starting with the phrase 'My sole claim' is false." Or, more accurately, such thinkers are like the person who claims that *modus tollens* is invalid, and supports this claim by giving an argument that itself employs this rule of inference.

1. Introduction

Fetzer (1988) famously claims that program verification isn't even a theoretical possibility,[1] and seeks to convince his readers of this claim by providing what has now become a widely known argument for it. Unfortunately for Fetzer, and like-minded thinkers, this position-argument pair, while based on a seminal insight that program verification, despite its Platonic proof-theoretic airs, is plagued by the inevitable unreliability of messy, real-world causation, is demonstrably self-refuting. As I soon show, Fetzer (and indeed anyone else who provides an argument- or proof-based attack on program verification) is like the person who claims: "My sole claim is that every claim expressed by an English sentence and starting with the phrase 'My sole claim' is false." Or, more accurately, such thinkers are like the person who claims that *modus tollens* is invalid, and supports this claim ($\neg\mu$) by giving an argument (where r is any rule of inference from some proof or argument calculus) of the form shown in the following table.

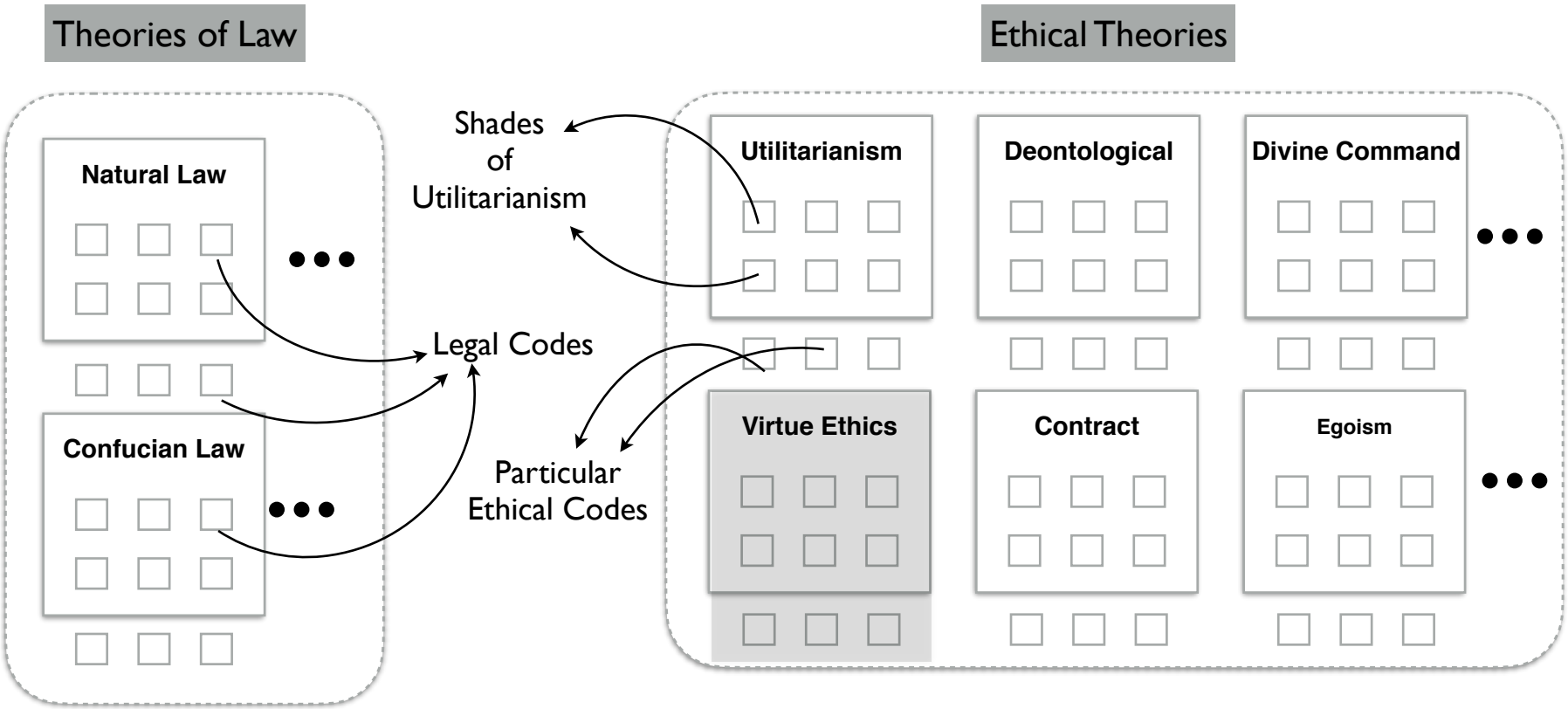
	1	ϕ_1	r
	2	ϕ_2	r
$\vdots$	$\vdots$	$\vdots$	$\vdots$
	k	$\mu \to \psi$	r
	$k+1$	$\neg\psi$	r
$\therefore$	$k+2$	$\neg\mu$	*modus tollens* k, $k+1$

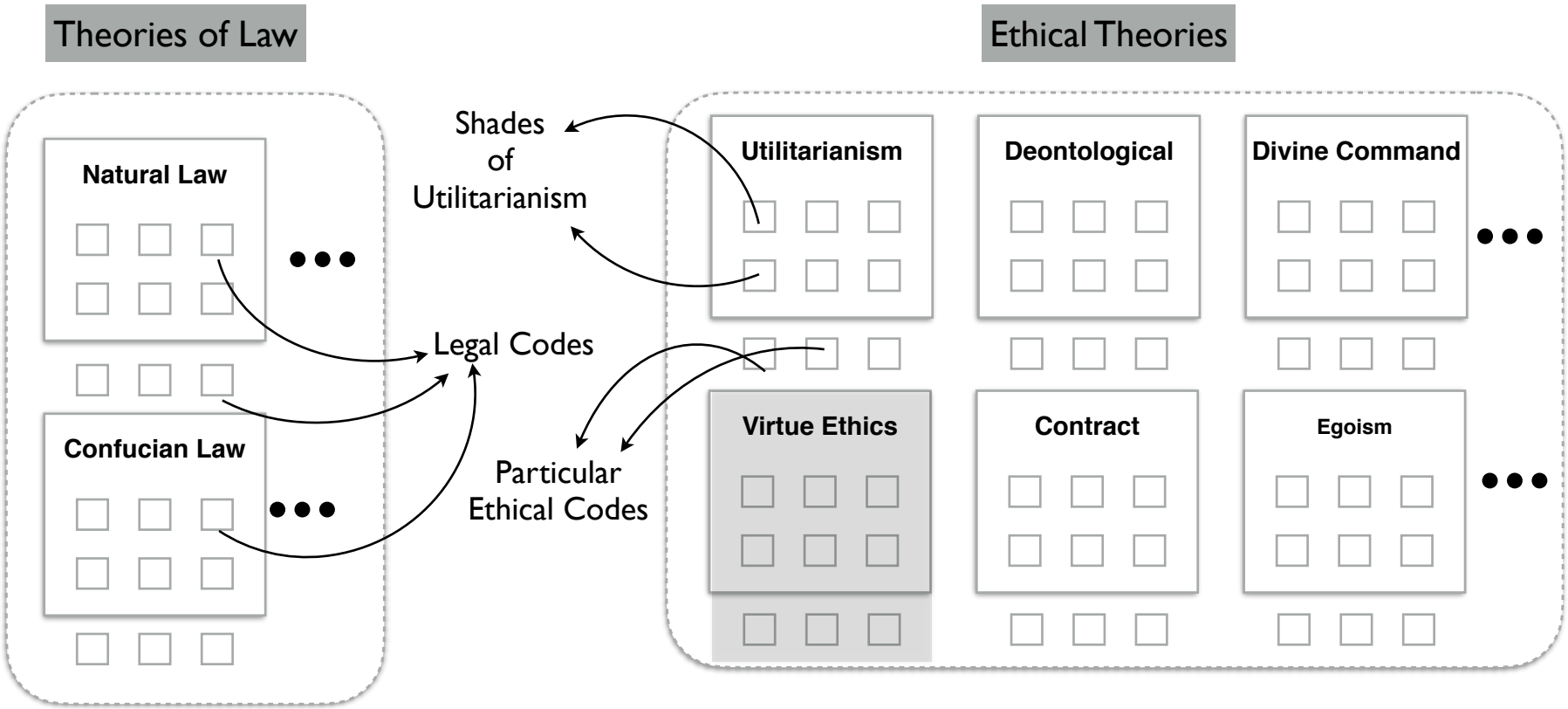
Table 1. Self-Refuting Argument-Schema Against *Modus Tollens*

[1] E.g., he writes: "The succe[redacted]thod for guaran-

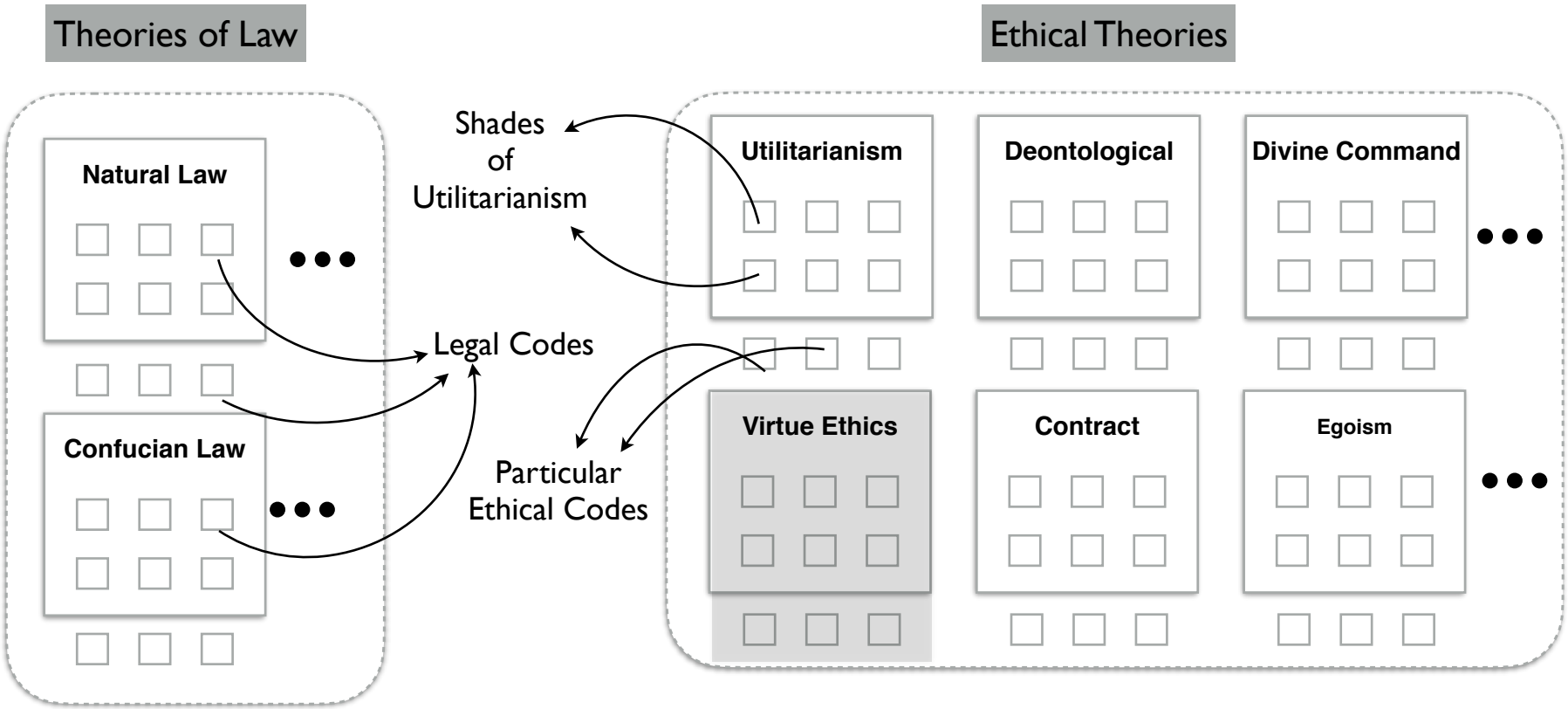
The Four Steps

Theories of Law
MAKING MORAL MACHINES
MAKING META-MORAL MACHINES
Natural Law
Confucian
Ethical Theories
Legal Codes
"Shades" of Util.
Particular Ethical Codes
Util
Deont.
D.C.
Virtue Eth.
Contract.
Egoism
Stripped Down Leibniz theory
Sub2
Sub1
F
M.N.
O
Sup1
Sup2
robot
The machine picks the theory; picks the code. . . . 2016
Pick the theory; pick the code; use the Leibnizian operators; engineer the


Theories of Law
Ethical Theories
Natural Law
Confucian Law
Shades of Utilitarianism
Legal Codes
Particular Ethical Codes
Utilitarianism
Deontological
Divine Command
Virtue Ethics
Contract
Egoism


Theories of Law
Ethical Theories
Natural Law
Confucian Law
Shades of Utilitarianism
Legal Codes
Particular Ethical Codes
Utilitarianism
Deontological
Divine Command
Virtue Ethics
Contract
Egoism

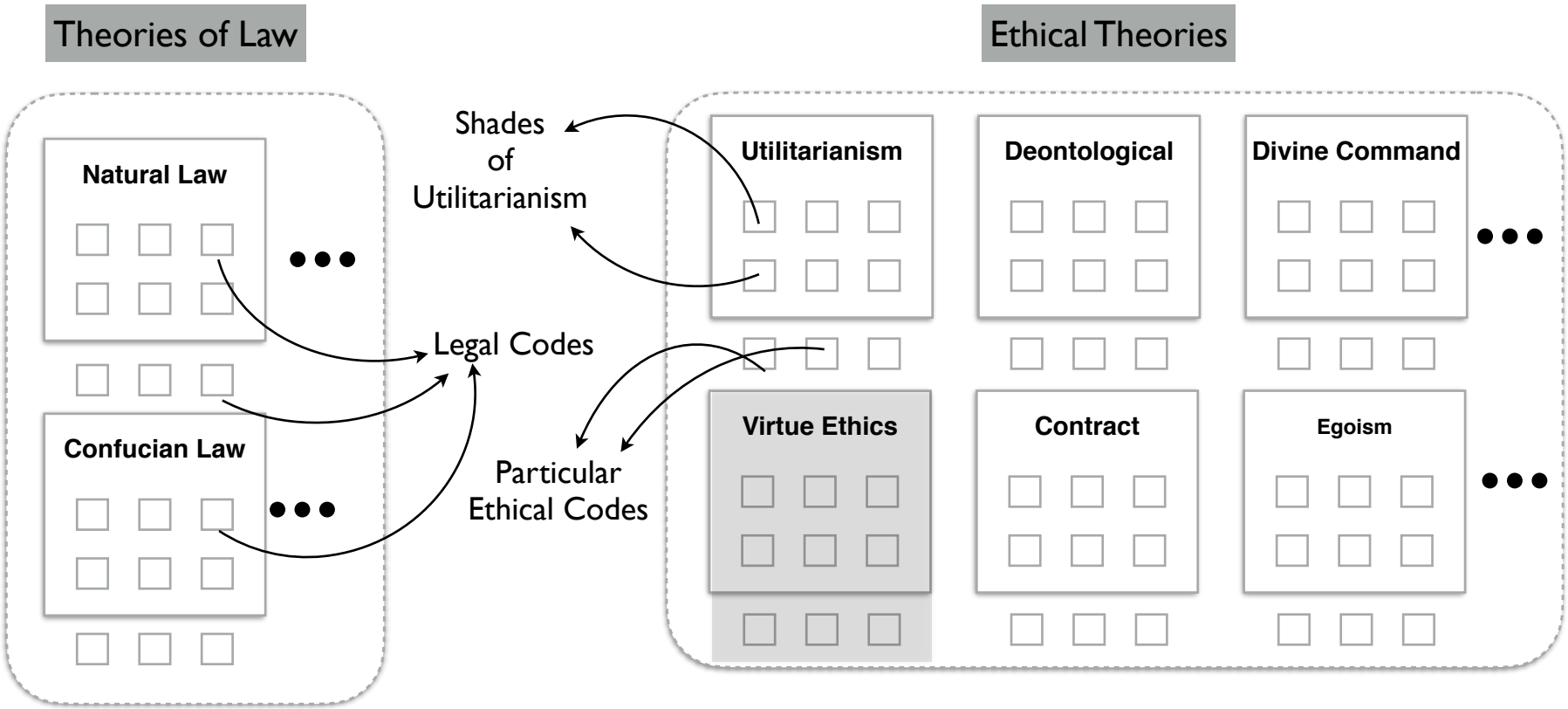
Making Morally X Machines



Step 1

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

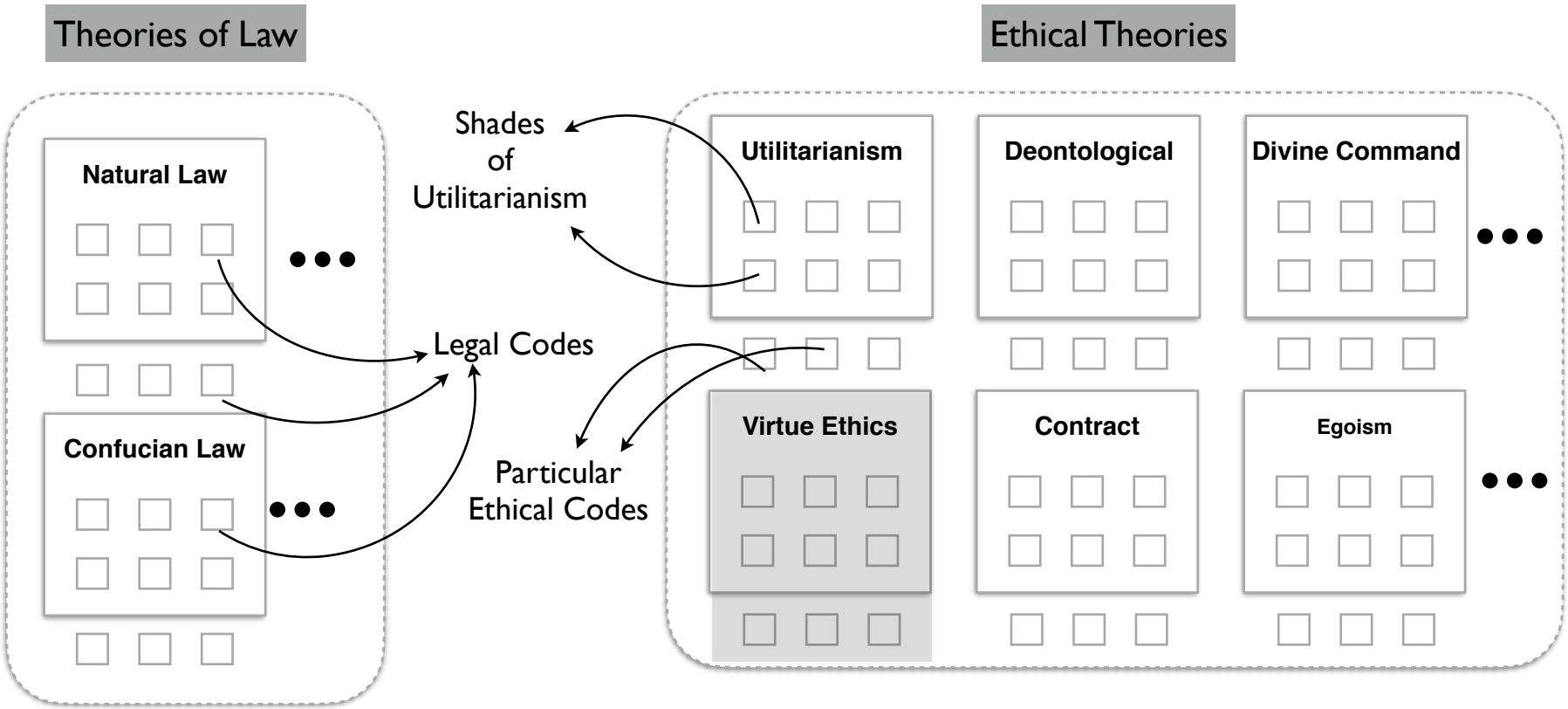
Making Morally X Machines



Step 1

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

Making Morally X Machines



Step 1

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

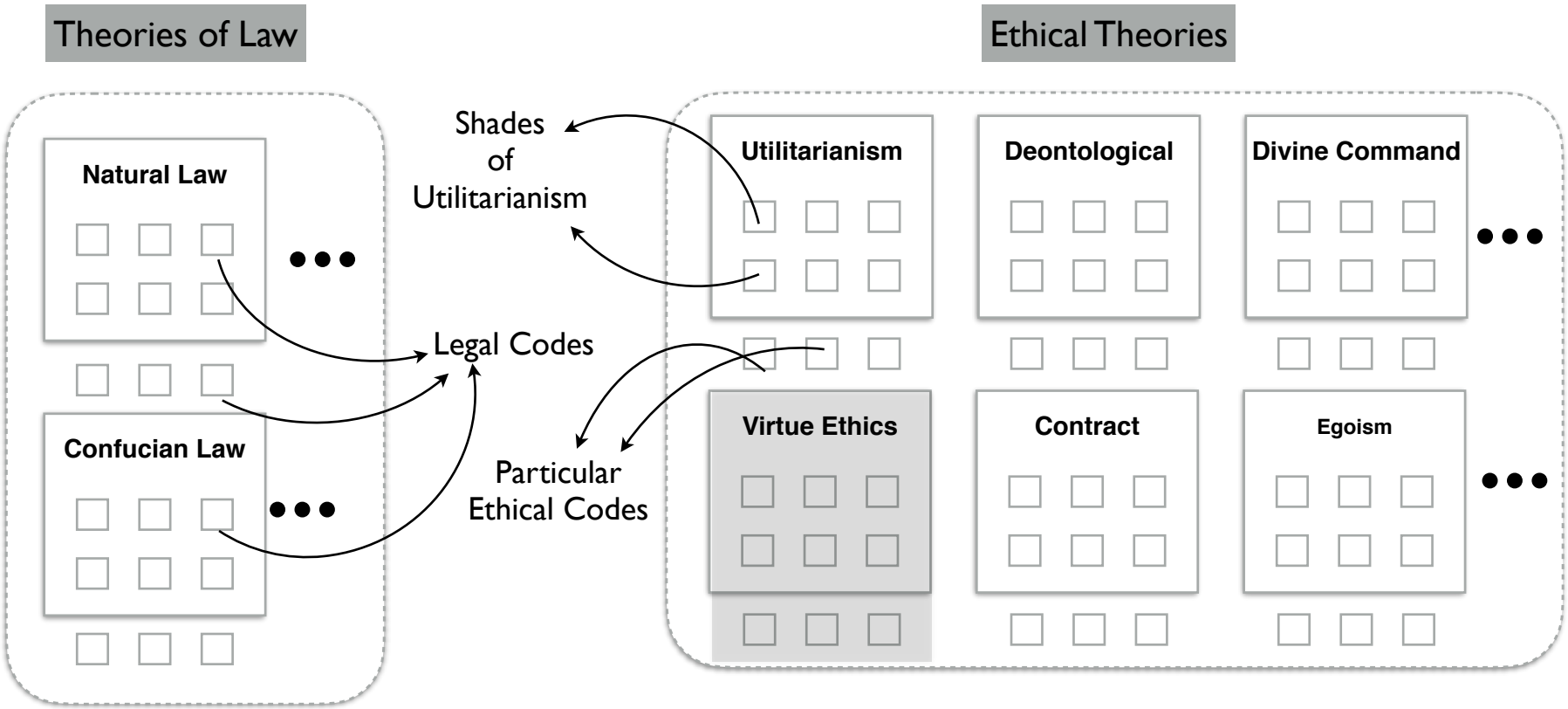
Step 2

Automate


Reasoners


Spectra

Making Morally X Machines



Step 1

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

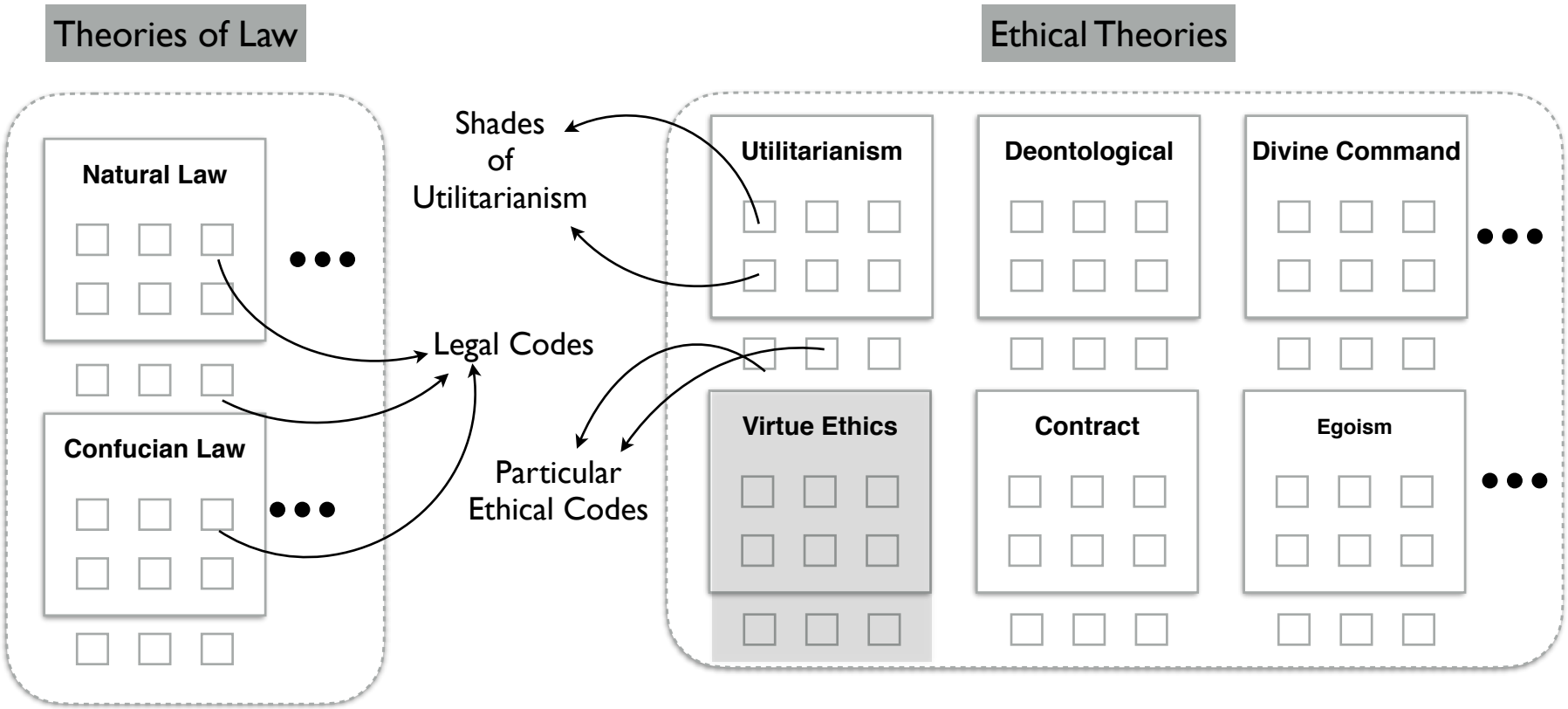
Step 2

Automate


Reasoners


Spectra

Making Morally X Machines



Step 1

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

Step 2

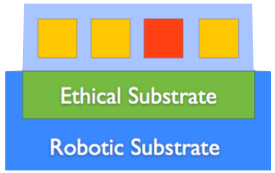
Automate

Reasoners

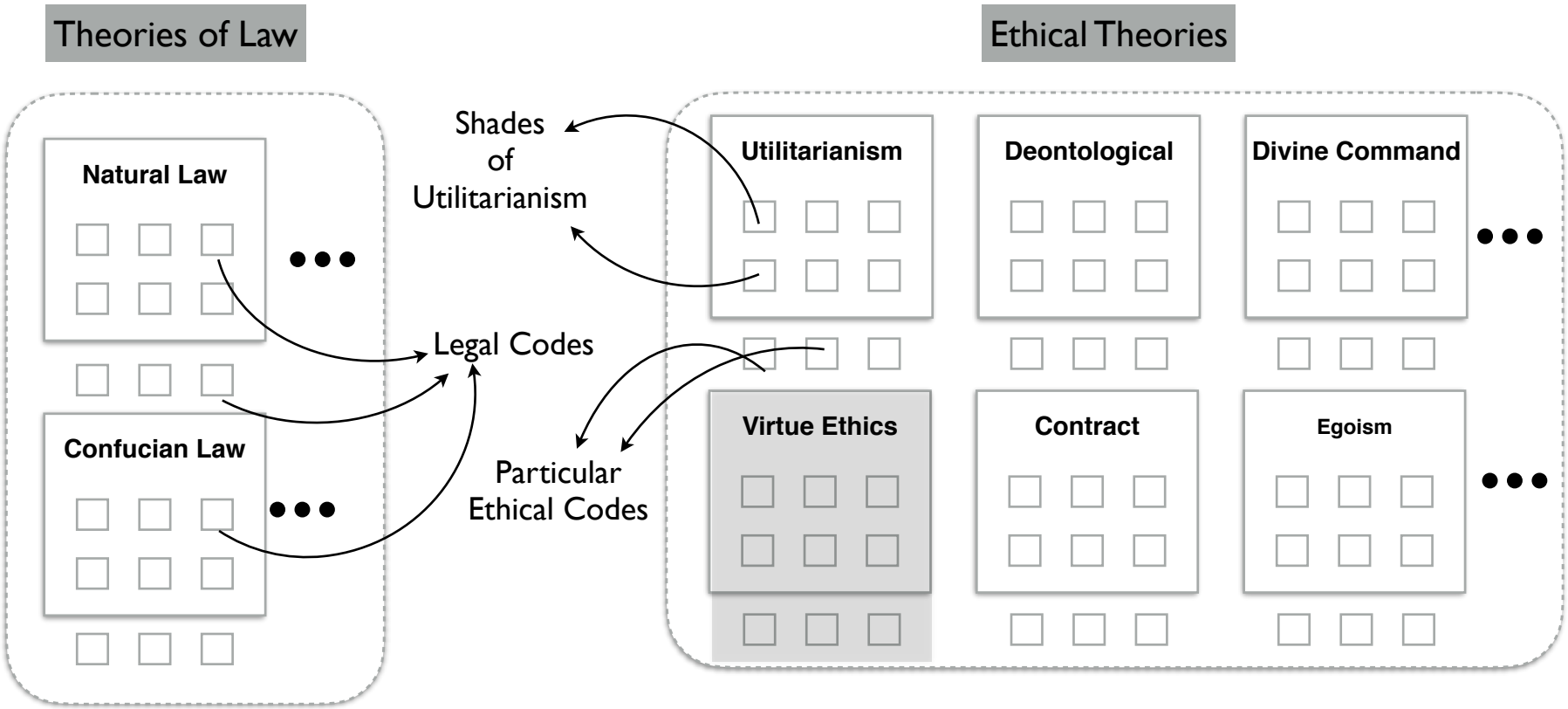
Spectra

Step 3

Ethical OS



Making Morally X Machines



Step 1

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

Step 2

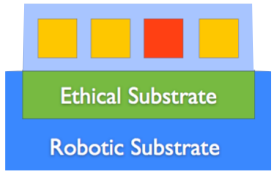
Automate

Reasoners

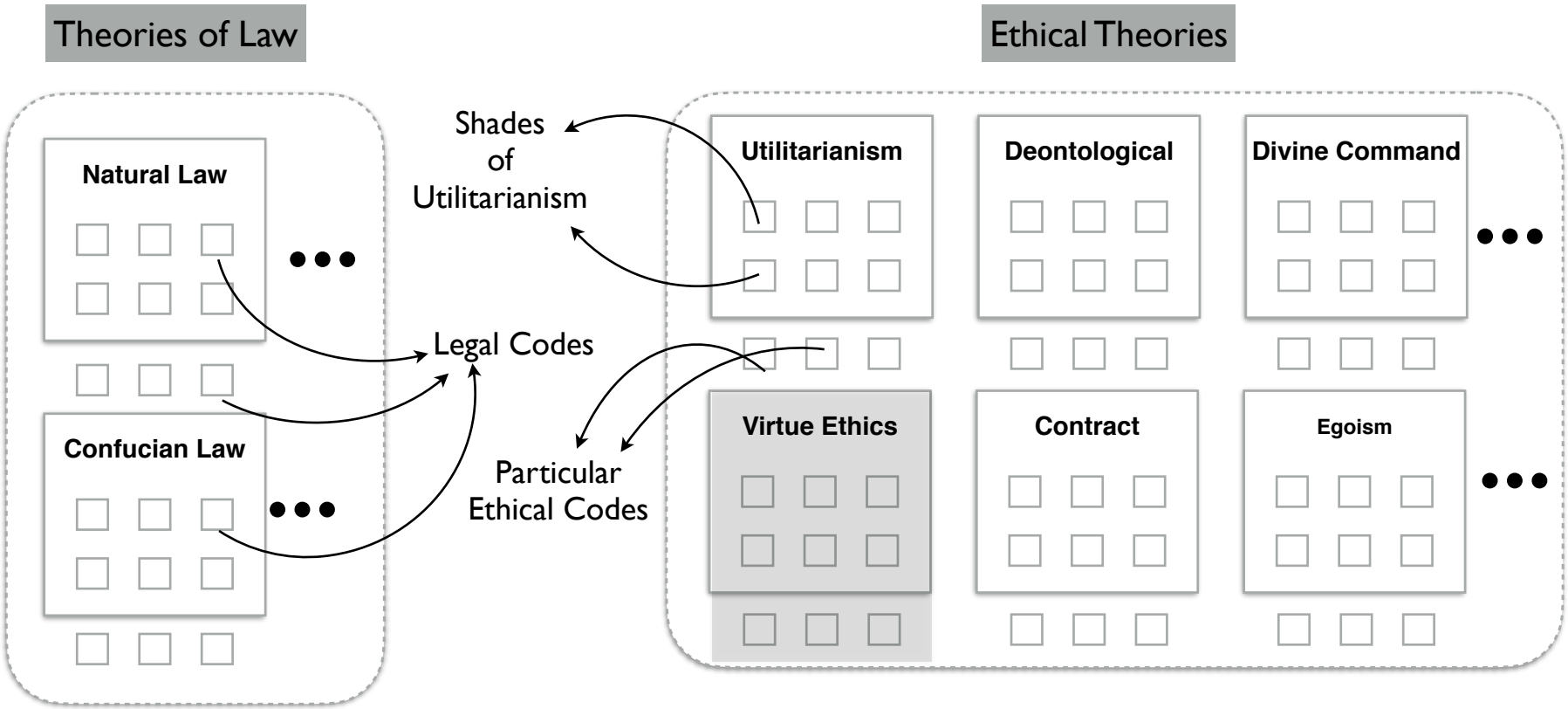
Spectra

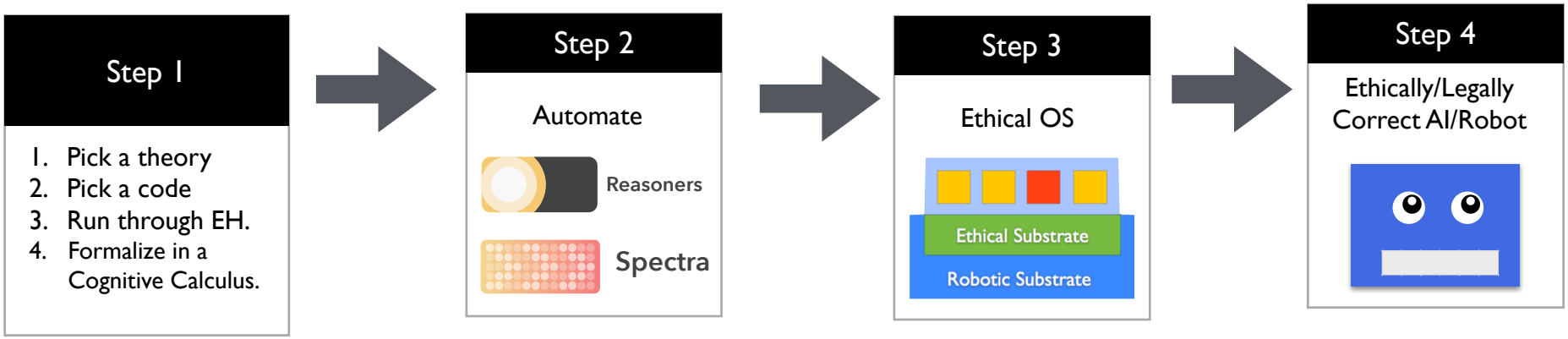
Step 3

Ethical OS



Making Morally X Machines





The MLn-is-a-Non-Starter Problem

The MLn-is-a-Non-Starter Problem

I

An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

The MLn-is-a-Non-Starter Problem

1 An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

2 The knowledge referred to in premise **1** is impossible to even express arithmetically (which btw is why logic has been employed for over two millennia to express this kind of content).

The MLn-is-a-Non-Starter Problem

1 An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

2 The knowledge referred to in premise **1** is impossible to even express arithmetically (which btw is why logic has been employed for over two millennia to express this kind of content).

3 That which can't be expressed arithmetically is impossible for MLn to represent and process.

The MLn-is-a-Non-Starter Problem

1 An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

2 The knowledge referred to in premise **1** is impossible to even express arithmetically (which btw is why logic has been employed for over two millennia to express this kind of content).

3 That which can't be expressed arithmetically is impossible for MLn to represent and process.

4 If MLn can't express and process ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, agents built on the basis of MLn, as per The PAID Problem, are minimally dangerous, and possibly eventually will destroy us.

The MLn-is-a-Non-Starter Problem

1 An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

2 The knowledge referred to in premise **1** is impossible to even express arithmetically (which btw is why logic has been employed for over two millennia to express this kind of content).

3 That which can't be expressed arithmetically is impossible for MLn to represent and process.

4 If MLn can't express and process ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, agents built on the basis of MLn, as per The PAID Problem, are minimally dangerous, and possibly eventually will destroy us.

∴

The MLn-is-a-Non-Starter Problem

	1	An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.
	2	The knowledge referred to in premise **1** is impossible to even express arithmetically (which btw is why logic has been employed for over two millennia to express this kind of content).
	3	That which can't be expressed arithmetically is impossible for MLn to represent and process.
	4	If MLn can't express and process ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, agents built on the basis of MLn, as per The PAID Problem, are minimally dangerous, and possibly eventually will destroy us.
∴	**5**	Agents built on the basis of MLn, as per The PAID Problem, are minimally dangerous, and possibly eventually will destroy us.

The MLn-is-a-Non-Starter Problem

I

An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

The ML roblem

The Washington Post
Democracy Dies in Darkness

Tech Consumer Tech Future of Transportation Inn

Ar owingly act in
ac iples … of the
typ itself and
ot the declarative
pr rinciples.

The Switch

AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how.

By **Drew Harwell**

The MLn-is-a-Non-Starter Problem

I

An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

The Washington Post

Democracy Dies in Darkness

Tech Consumer Tech Future of Transportation Inn

The Switch

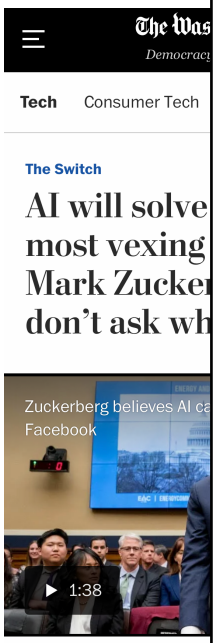
AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how.

By **Drew Harwell**

The ML Problem

An acco type othe prop

knowingly act in nciples … of the o itself and vs the declarative /principles.



April 11, 2018 at 12:04 PM EDT

Artificial intelligence will solve Facebook's most vexing problems, chief executive Mark Zuckerberg insists. He just can't say when, or how.

Zuckerberg referred to AI technology more than 30 times during ten hours of questioning from congressional lawmakers Tuesday and Wednesday, saying that it would one day be smart, sophisticated and eagle-eyed enough to fight against a vast variety of platform-spoiling misbehavior, including fake news, hate speech, discriminatory ads and terrorist propaganda.

Over the next five to 10 years, he said, artificial intelligence would prove a champion for the world's largest social network in resolving its most pressing crises on a global scale — while also helping the company dodge pesky questions about censorship, fairness and

The MLn-is-a-Non-Starter Problem

I

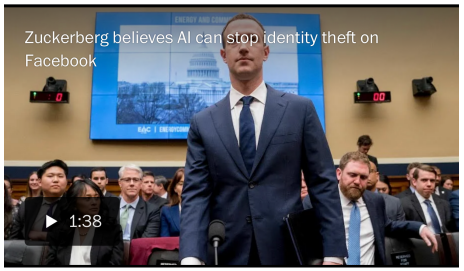
An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

The Washington Post
Democracy Dies in Darkness

Tech Consumer Tech Future of Transportation Inn

The Switch

AI will solve Facebook's most vexing problems, Mark Zuckerberg says. Just don't ask when or how.



By **Drew Harwell**

April 11, 2018 at 12:04 PM EDT

Artificial intelligence will solve Facebook's most vexing problems, chief executive Mark Zuckerberg insists. He just can't say when, or how.

Zuckerberg referred to AI technology more than 30 times during ten hours of questioning from congressional lawmakers Tuesday and Wednesday, saying that it would one day be smart, sophisticated and eagle-eyed enough to fight against a vast variety of platform-spoiling misbehavior, including fake news, hate speech, discriminatory ads and terrorist propaganda.

Over the next five to 10 years, he said, artificial intelligence would prove a champion for the world's largest social network in resolving its most pressing crises on a global scale — while also helping the company dodge pesky questions about censorship, fairness and

The MLn-is-a-Non-Starter Problem

I

An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

April 11, 2018 at 12:04 PM EDT

Artificial intelligence will solve Facebook's most vexing problems, chief executive Mark Zuckerberg insists. He just can't say when, or how.

Zuckerberg referred to AI technology more than 30 times during ten hours of questioning from congressional lawmakers Tuesday and Wednesday, saying that it would one day be smart, sophisticated and eagle-eyed enough to fight against a vast variety of platform-spoiling misbehavior, including fake news, hate speech, discriminatory ads and terrorist propaganda.

Over the next five to 10 years, he said, artificial intelligence would prove a champion for the world's largest social network in resolving its most pressing crises on a global scale — while also helping the company dodge pesky questions about censorship, fairness and

The MLn-is-a-Non-Starter Problem

I

An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

The MLn-is-a-Non-Starter Problem

I

An agent, whether natural or artificial, can only knowingly act in accord with ethical and legal theories/codes/principles … of the type at the heart of The Four Steps, and justify to itself and others that it's acting in such a manner, if it knows the declarative propositions that compose such theories/codes/principles.

What shall we then do?

What shall we then do?

We *cannot* afford to re-engineer and correct Mn-based artifacts.

What shall we then do?

We *cannot* afford to re-engineer and correct Mn-based artifacts.

Lectures from eg the EC, however wise and learned they may be, won't change anyone's behavior.

What shall we then do?

We *cannot* afford to re-engineer and correct Mn-based artifacts.

Lectures from eg the EC, however wise and learned they may be, won't change anyone's behavior.


EUROPEAN COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence – A European approach to excellence and trust

EN EN

Wha do?

We car r and correct

Lecture however they ma anyone

EUROPEAN COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

EN EN

What shall we then do?

We *cannot* afford to re-engineer and correct Mn-based artifacts.

Lectures from eg the EC, however wise and learned they may be, won't change anyone's behavior.


EUROPEAN COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

EN EN

What shall we then do?

We *cannot* afford to re-engineer and correct Mn-based artifacts.

Lectures from eg the EC, however wise and learned they may be, won't change anyone's behavior.


EUROPEAN COMMISSION

Brussels, 19.2.2020
COM(2020) 65 final

WHITE PAPER

On Artificial Intelligence - A European approach to excellence and trust

EN EN

Instead, employ fully rational paternalism, in the form of optimal "sin" taxation, on an international scale.

patax

patax

- We must have a rational architect/ideal observer.

$\mathscr{P}$patax

1. We must have a rational architect/ideal observer.

2. Absent the tax in question (here, tax on companies selling MLn), it's overwhelmingly likely that great disutility will result.

$\mathscr{P}$patax

1 We must have a rational architect/ideal observer.

2 Absent the tax in question (here, tax on companies selling MLn), it's overwhelmingly likely that great disutility will result.

3 Assuming enactment of the MLn tax, it is at least highly likely that the destructive behavior of the corporations in question will be greatly reduced, and perhaps eliminated altogether.

$\mathscr{P}$patax

1 We must have a rational architect/ideal observer.

2 Absent the tax in question (here, tax on companies selling MLn), it's overwhelmingly likely that great disutility will result.

3 Assuming enactment of the MLn tax, it is at least highly likely that the destructive behavior of the corporations in question will be greatly reduced, and perhaps eliminated altogether.

4 There are no easy-to-anticipate secondary effects (i.e. it's highly unlikely that there are secondary effects) that outweigh the utility flowing from reduction in said activity, even after redistributive measures are taken.

patax

1. We must have a rational architect/ideal observer.

2. Absent the tax in question (here, tax on companies selling MLn), it's overwhelmingly likely that great disutility will result.

3. Assuming enactment of the MLn tax, it is at least highly likely that the destructive behavior of the corporations in question will be greatly reduced, and perhaps eliminated altogether.

4. There are no easy-to-anticipate secondary effects (i.e. it's highly unlikely that there are secondary effects) that outweigh the utility flowing from reduction in said activity, even after redistributive measures are taken.

5. Redistributive measures must be included; specifically, these measures must be highly likely to foster the alternative behaviors sought.

patax

1. We must have a rational architect/ideal observer.

2. Absent the tax in question (here, tax on companies selling MLn), it's overwhelmingly likely that great disutility will result.

3. Assuming enactment of the MLn tax, it is at least highly likely that the destructive behavior of the corporations in question will be greatly reduced, and perhaps eliminated altogether.

4. There are no easy-to-anticipate secondary effects (i.e. it's highly unlikely that there are secondary effects) that outweigh the utility flowing from reduction in said activity, even after redistributive measures are taken.

5. Redistributive measures must be included; specifically, these measures must be highly likely to foster the alternative behaviors sought.

6. Monitoring and enforcement is feasible, as are the issuing of severe penalties in the case of violations and/or non-payment.

patax

1. We must have a rational architect/ideal observer.

2. Absent the tax in question (here, tax on companies selling MLn), it's overwhelmingly likely that great disutility will result.

3. Assuming enactment of the MLn tax, it is at least highly likely that the destructive behavior of the corporations in question will be greatly reduced, and perhaps eliminated altogether.

4. There are no easy-to-anticipate secondary effects (i.e. it's highly unlikely that there are secondary effects) that outweigh the utility flowing from reduction in said activity, even after redistributive measures are taken.

5. Redistributive measures must be included; specifically, these measures must be highly likely to foster the alternative behaviors sought.

6. Monitoring and enforcement is feasible, as are the issuing of severe penalties in the case of violations and/or non-payment.

7. There are no cogent arguments available that show that it's more likely than not that corporations will act irrationally in the face of the MLn tax.

Objections …

“But such taxes are unprecedented!”

“But such taxes are unprecedented!”

Not true, at all: the Tobin Tax.

A Proposal for International Monetary Reform*

JAMES TOBIN
Yale University

Over the last twenty years economists' prescriptions for reform of the international monetary system have taken various shapes. Their common premise was dissatisfaction with the Bretton Woods regime as it evolved in the 1950s. Robert Triffin awakened the world to the contradictions and instabilities of a system of pegged parities that relied on the debts in reserve currencies, mostly dollars, to meet growing needs for official reserves. Triffin and his followers saw the remedy as the internationalization of reserves and reserve assets; their ultimate solution was a world central bank. Others diagnosed the problem less in terms of liquidity than in the inadequacies of balance of payments adjustment mechanisms in the modern world. The inadequacies were especially evident under the fixed-parity gold-exchange standard when, as in the 1960s, the reserve currency center was structurally in chronic deficit. These analysts sought better and more symmetrical "rules of the game" for adjustments by surplus and deficit countries, usually including more flexibility in the setting of exchange parities, crawling pegs, and the like. Many economists, of whom Milton Friedman was an eloquent and persuasive spokesman, had all along advocated floating exchange rates, determined in private markets without official interventions.

*This paper is Prof. Tobin's presidential address at the 1978 conference of the Eastern Economic Association, Wash. D.C.

By the early 1970s the third view was the dominant one in the economics profession, though not among central bankers and private financiers. And all of a sudden, thanks to Nixon and Connally, we got our wish. Or at least we got as much of it as anyone could reasonably have hoped, since it could never have been expected that governments would eschew all intervention in exchange markets.

Now after five to seven years—depending how one counts—of unclean floating there are many second thoughts. Some economists share the nostalgia of men of affairs for the gold standard or its equivalent, for a fixed anchor for the world's money, for stability of official parities. Some economists, those who emphasize the rationality of expectations and the flexibility of prices in all markets, doubt that it makes much difference whether exchange rates are fixed or flexible, provided only that government policies are predictable. Clearly, flexible rates have not been the panacea which their more extravagant advocates had hoped; international monetary problems have not disappeared from headlines or from the agenda of anxieties of central banks and governments.

I believe that the basic problem today is not the exchange rate regime, whether fixed or floating. Debate on the regime evades and obscures the essential problem. That is the excessive international—or better, intercurrency—mobility of private financial capital. The biggest thing that happened in the world monetary system since the 1950s was

national monetary policy, movements of funds to exploit interest arbitrage or to speculate on exchange rate fluctuations cannot be sources of disturbances and painful interregional adjustments.

To recite this familiar account is to remind us how difficult it would be to replicate its prerequisites on a worldwide basis. Even for the Common Market countries, the goal is still far, far distant. We do not have to resolve the chicken-egg argument. Perhaps it is true that establishing a common currency and a central macro-economic policy will automatically generate the institutions, markets, and mobilities which make the system viable and its regional economic consequences everywhere tolerable. The risk is one that few are prepared to take. Moreover, EEC experience to date suggests that it is very hard to contrive a scenario of gradual evolution towards such a radically different regime, even though it could well be the global optimum.

At present the world enjoys many benefits of the increased worldwide economic integration of the last thirty years. But the integration is partial and unbalanced; in particular private financial markets have become internationalized much more rapidly and completely than other economic and political institutions. That is why we are in trouble. So I turn to the second, and second best, way out, forcing some segmentation of inter-currency financial markets.

My specific proposal is actually not new. I offered it in 1972 in my Janeway Lectures at Princeton, published in 1974 as *The New Economics One Decade Older,* pp. 88–92. The idea fell like a stone in a deep well. If I cast it in the water again, it is because events since the first try have strengthened my belief that something of the sort needs to be done.

The proposal is an internationally uniform tax on all spot conversions of one currency into another, proportional to the size of the transaction. The tax would particularly deter short-term financial round-trip excursions into another currency. A 1% tax, for example, could be overcome only by an 8 point differential in the annual yields of Treasury bills or Eurocurrency deposits denominated in dollars and Deutschmarks. The corresponding differential for one-year maturities would be 2 points. A permanent investment in another country or currency area, with regular repatriation of yield when earned, would need a 2% advantage in marginal efficiency over domestic investment. The impact of the tax would be less for permanent currency shifts, or for longer maturities. Because of exchange risks, capital value risks, and market imperfections, interest arbitrage and exchange speculation are less troublesome in long maturities. Moreover, it is desirable to obstruct as little as possible international movements of capital responsive to long-run portfolio preferences and profit opportunities.

Why do floating exchange rates not solve the problems? There are several reasons, all exemplified in recent experience.

First, as economists have long known, in a world of international capital mobility flexibility of exchange rates does not assure autonomy of national macroeconomic policy. The Mundell-Fleming models of the early 1960s showed how capital mobility inhibits domestic monetary policy under fixed parities and domestic fiscal policy under flexible rates. Moreover, the availability of the remaining instrument of macroeconomic policy in either regime is small consolation. Nations frequently face compelling domestic institutional, political, and economic constraints on one or the other instrument, or on the policy mix.

Second, it may seem that we should welcome an exchange rate regime that increases the potency of monetary policy relative to fiscal policy; after all, monetary policy is the more flexible and responsive instrument of domestic stabilization. But the liberation of domestic monetary policy under flexible rates

private financial markets have become internationalized much more rapidly and completely than other economic and political institutions. That is why we are in trouble. So I turn to the second, and second best, way out, forcing some segmentation of inter-currency financial markets.

My specific proposal is actually not new. I offered it in 1972 in my Janeway Lectures at Princeton, published in 1974 as *The New Economics One Decade Older,* pp. 88–92. The idea fell like a stone in a deep well. If I cast it in the water again, it is because events since the first try have strengthened my belief that something of the sort needs to be done.

The proposal is an internationally uniform tax on all spot conversions of one currency into another, proportional to the size of the transaction. The tax would particularly deter

bility of exchange rates does not assure au
omy of national macroeconomic policy.
Mundell-Fleming models of the early 1
showed how capital mobility inhibits dom
monetary policy under fixed parities
domestic fiscal policy under flexible r
Moreover, the availability of the remai
instrument of macroeconomic policy in ei
regime is small consolation. Nations
quently face compelling domestic institu
al, political, and economic constraints on
or the other instrument, or on the policy r

Second, it may seem that we sh
welcome an exchange rate regime
increases the potency of monetary policy
tive to fiscal policy; after all, monetary p
is the more flexible and responsive instru
of domestic stabilization. But the liberati
domestic monetary policy under flexible

of whom Milton Friedman was an eloquent and persuasive spokesman, had all along advocated floating exchange rates, determined in private markets without official interventions.

*This paper is Prof. Tobin's presidential address at the 1978 conference of the Eastern Economic Association, Wash. D.C.

governments.

I believe that the basic problem today is not the exchange rate regime, whether fixed or floating. Debate on the regime evades and obscures the essential problem. That is the excessive international—or better, inter-currency—mobility of private financial capital. The biggest thing that happened in the world monetary system since the 1950s was

national monetary policy, movements of funds to exploit interest arbitrage or to speculate on exchange rate fluctuations cannot be sources of disturbances and painful interregional adjustments.

To recite this familiar account is to remind us how difficult it would be to replicate its prerequisites on a worldwide basis. Even for the Common Market countries, the goal is still far, far distant. We do not have to resolve the chicken-egg argument. Perhaps it is true that establishing a common currency and a central macro-economic policy will automatically generate the institutions, markets, and mobilities which make the system viable and its regional economic consequences everywhere tolerable. The risk is one that few are prepared to take. Moreover, EEC experience to date suggests that it is very hard to contrive a scenario of gradual evolution towards such a radically different regime, even though it could well be the global optimum.

At present the world enjoys many benefits of the increased worldwide economic integration of the last thirty years. But the integration is partial and unbalanced; in particular private financial markets have become internationalized much more rapidly and completely than other economic and political

short-term financial round-trip excursions into another currency. A 1% tax, for example, could be overcome only by an 8 point differential in the annual yields of Treasury bills or Eurocurrency deposits denominated in dollars and Deutschmarks. The corresponding differential for one-year maturities would be 2 points. A permanent investment in another country or currency area, with regular repatriation of yield when earned, would need a 2% advantage in marginal efficiency over domestic investment. The impact of the tax would be less for permanent currency shifts, or for longer maturities. Because of exchange risks, capital value risks, and market imperfections, interest arbitrage and exchange speculation are less troublesome in long maturities. Moreover, it is desirable to obstruct as little as possible international movements of capital responsive to long-run portfolio preferences and profit opportunities.

Why do floating exchange rates not solve the problems? There are several reasons, all exemplified in recent experience.

First, as economists have long known, in a world of international capital mobility flexibility of exchange rates does not assure autonomy of national macroeconomic policy. The Mundell-Fleming models of the early 1960s

“Can you *prove* it’ll work? Where are your promised theorems?”

Theorems (machine-proved & machine-verified): Under certain conditions a Ponzi schemer will avoid detection, and then prove that under enhanced detection the schemer will be caught.

$$\dfrac{\dfrac{\dfrac{}{\mathbf{S}(p,b,t_0,r_n(t_0)=\alpha')}\,[\gamma_{Ponzi}]}{\mathbf{B}(b,t_0,\mathbf{B}(p,t_0,r_n(t_0)=\alpha'))}\,[R_{12}] \quad \dfrac{\dfrac{}{\forall r\mathbf{B}(b,t_0,\mathbf{B}(p,t_0,r_n(t_0)=\alpha'))\rightarrow\mathbf{B}(b,t_0,r_n(t_0)=\alpha')}\,[\gamma_{NI}]}{\mathbf{B}(b,t_0,\mathbf{B}(p,t_0,r_n(t_0)=\alpha'))\rightarrow\mathbf{B}(b,t_0,r_n(t_0)=\alpha')}\,[U_{elim}]}{\mathbf{B}(b,t_0,r_n(t_0)=\alpha')}\,[MP]$$

$$\dfrac{\mathbf{B}(b,t_0,r_n(t_0)=\alpha') \quad \dfrac{}{\mathbf{B}(b,t_0,r_n(t_0)=\alpha'\rightarrow happens(action(b,invest),t_0))}\,[\gamma_{CI}]\ \&\ [Arith.]}{happens(action(b,invest),t_0)}\,MP$$

$$\gamma_{susp} : \forall(t_1,\ldots,t_n)\wedge_{i=1}^{n}\mathbf{P}(i,t_i,S_a(t_i)>=\tau)\rightarrow\mathbf{B}(i,t_{n+1},ponzi)$$

A simple investigator also decides to investigate the moment they become suspicious without any consideration of possible negative consequences of a failed investigation: $\gamma_{investigate}$. The intention of the Ponzi schemer is to avoid a collapse of the fund, represented by $\gamma_{collapse}$, which uses the intention operator $\mathbf{I}$.

$$\gamma_{investigate} : \forall t.\ \mathbf{B}(i,t,ponzi)\rightarrow happens(action(i,investigate),t)$$
$$\gamma_{collapse} : \forall t_1,t_2.\ prior(t_1,t_2)\wedge\mathbf{I}(i,t_1,\neg holds(collapse,t_2))$$

While a simple Ponzi schemer may not have any beliefs about the beliefs of the investigator, a more sophisticated Ponzi schemer will ascribe the following three beliefs to an investigator. At this point, we have all that we need to simultaneously prove and simulate such results as that if the infamous Madoff had been as sophisticated as our Ponzi schemer agent p, barring a "run" on the funds Madoff controlled, exposing him would have been well nigh impossible. Space constraints preclude presenting such simulation here.

If p believes i believes

$$\gamma_{p\text{-}susp} : \forall t.\ \mathbf{B}(p,t,\gamma_{susp})$$
$$\gamma_{p\text{-}investigate} : \forall t.\ \mathbf{B}(p,t,\gamma_{investigate})$$
$$\gamma_{p\text{-}collapse} : \forall t.\ \mathbf{B}(p,t,\gamma_{collapse})$$

p will not be caught.

Theorem: *A* corporation will avoid MLn and invest in logicist AI, if persistently subjected to the tax in question.

$$\dfrac{\dfrac{\dfrac{}{\mathbf{S}(p,b,t_0,r_n(t_0)=\alpha')}\,[\gamma_{Ponzi}]}{\mathbf{B}(b,t_0,\mathbf{B}(p,t_0,r_n(t_0)=\alpha'))}\,[R_{12}] \quad \dfrac{\dfrac{}{\forall r\mathbf{B}(b,t_0,\mathbf{B}(p,t_0,r_n(t_0)=\alpha'))\rightarrow\mathbf{B}(b,t_0,r_n(t_0)=\alpha')}\,[\gamma_{NI}]}{\mathbf{B}(b,t_0,\mathbf{B}(p,t_0,r_n(t_0)=\alpha'))\rightarrow\mathbf{B}(b,t_0,r_n(t_0)=\alpha')}\,[U_{elim}]}{\mathbf{B}(b,t_0,r_n(t_0)=\alpha')}\,[MP]$$

$$\dfrac{\mathbf{B}(b,t_0,r_n(t_0)=\alpha') \quad \dfrac{}{\mathbf{B}(b,t_0,r_n(t_0)=\alpha'\rightarrow happens(action(b,invest),t_0))}\,[\gamma_{CI}]\ \&\ [Arith.]}{happens(action(b,invest),t_0)}\,MP$$

$$\gamma_{susp} : \forall(t_1,\ldots,t_n)\wedge_{i=1}^{n}\mathbf{P}(i,t_i,S_a(t_i)>=\tau)\rightarrow\mathbf{B}(i,t_{n+1},ponzi)$$

A simple investigator also decides to investigate the moment they become suspicious without any consideration of possible negative consequences of a failed investigation: $\gamma_{investigate}$. The intention of the Ponzi schemer is to avoid a collapse of the fund, represented by $\gamma_{collapse}$, which uses the intention operator $\mathbf{I}$.

$$\gamma_{investigate} : \forall t.\ \mathbf{B}(i,t,ponzi)\rightarrow happens(action(i,investigate),t)$$
$$\gamma_{collapse} : \forall t_1,t_2.\ prior(t_1,t_2)\wedge\mathbf{I}(i,t_1,\neg holds(collapse,t_2))$$

While a simple Ponzi schemer may not have any beliefs about the beliefs of the investigator, a more sophisticated Ponzi schemer will ascribe the following three beliefs to an investigator. At this point, we have all that we need to simultaneously prove and simulate such results as that if the infamous Madoff had been as sophisticated as our Ponzi schemer agent p, barring a "run" on the funds Madoff controlled, exposing him would have been well nigh impossible. Space constraints preclude presenting such simulation here.

If *p* believes *i* believes

$$\gamma_{p\text{-}susp} : \forall t.\ \mathbf{B}(p,t,\gamma_{susp})$$
$$\gamma_{p\text{-}investigate} : \forall t.\ \mathbf{B}(p,t,\gamma_{investigate})$$
$$\gamma_{p\text{-}collapse} : \forall t.\ \mathbf{B}(p,t,\gamma_{collapse})$$

p will not be caught.

“Can you *prove* it’ll work? Where are your promised theorems?”

“Can you *prove* it’ll work? Where are your promised theorems?”

Yes, we can —using logicist agent-based economics. See the forthcoming full paper.

Med nok penger, kan logikk løse alle problemer.