

Killer Robots, D, and Beyond to *DCEC** in HyperSlate®

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

IFLAI2
11/9/2020
ver 1111201115NY



Logistics ...

Required by Thurs!

The screenshot shows a LaTeX editor interface with a dark theme. The top bar includes a menu, a title 'IFLAI2F20_PAPERTOPICS', and several utility buttons: a question mark, a letter 'L', 'Review', 'Share', 'Submit', 'History', and 'Chat'. Below the top bar, there are tabs for 'Source' (selected) and 'Rich Text', and a 'Recompile' button. The left sidebar shows a file explorer with 'main.tex' selected. The main editor area displays the source code for 'main.tex' with line numbers 1 through 26. The code includes LaTeX commands for document class, packages, title, author, date, and a list of topics. The right pane shows the rendered output of the document, titled 'The List', which contains a bulleted list of topics and specific claims for three students: Mike Giancola, Jasper Covey, and John Slowik.

```
1 \documentclass[11pt]{article}
2
3 \usepackage[utf8]{inputenc}
4 \usepackage{fullpage}
5 \usepackage{amssymb}
6 \usepackage[colorlinks]{hyperref}
7
8 \begin{document}
9
10 \title{\textbf{IFLAI2F20 Paper Topics}}
11 \author{Selmer Bringsjord}
12 \date{\texttt{ver 1109201500NY}}
13 \maketitle
14
15 \noindent
16 %
17 This document maintains the list of selected
18 paper topics and specific
19 claims etc.\ for each student enrolled in
20 \href{http://www.logicamodernapproach.com/rpi/
21 iflai2f20.bringsjord/}{IFLAI2F20}.
22 The format is simply that for each student by
23 name (First Last) there
24 is a bullet that gives first the general topic
25 area, and then another
26 bullet following on that that gives the
27 specific chief claim the
28 student is making in the paper. (The specific
29 claim must be announced
30 at the very outset of the paper itself.
31 Specifically, the claim must
32 be expressed in the first paragraph of the
33 paper as a clear
34 declarative sentence in English, as is the
35 case in the present
```

The List

- Mike Giancola ✓
 - **Topic Area:** Paternalistic Taxation of Machine Learning.
 - **Specific Claim:** The proposal to tax corporate ML activity made recently by S Bringsjord would face four major roadblocks to successful implementation: (1) passage into law; (2) enforcement; and efficacy, both in terms of (3a) reducing harm and (3b) shifting research towards logic-based methods.
- Jasper Covey ✓
 - **Topic Area:** Modeling Taxation, Effort, and Wealth.
 - **Specific Claim:** The taxation model, *S*, proposed in class by S Bringsjord lacks an account of the effects of capital on effort that, when implemented, would necessitate a progressive tax scheme.
- Joe Halasz ✓
 - **Topic Area:** The Argument for God's Existence from AI
 - **Specific Claim:** The argument for God's Existence proposed by S Bringsjord, specifically section 4.1 about premise 4 vulnerabilities, does not take new studies on canine ability into account that could remove the discontinuity between the human mind and the canine mind, and premise 5 in The Argument does not take into account the fact that other natural forces still having to do with physical science could have caused it to be the case that we have this level of cognitive power.
- John Slowik
 - **Topic Area:** Modeling Taxation, Effort, and Wealth.
 - **Specific Claim:** The proposed tax model fails to afford the taxed individuals ethical standards of living, promotes counterproductive behavior in the taxed population, and stifles competition and innovation, contrary to its claims that such a model is required for the respective promotion or suppression of the same. I intend to model this using an ordinal set of activities *A* which citizens can participate in only if they satisfy some requirement, e.g. having sufficient capital. The set being ordinal means that a citizen will choose to participate in activities in order until they cannot perform further activities due to exhausted means (again noting that each activity maintains its own satisfaction conditions).

Schedule Switcheroo

Schedule Switcheroo

- **Nov 5:** *Pure General Logic Programming, Functional Programming, Turing-Completeness, and Beyond.* We review the basic paradigms of computer programming. For the imperative case, we use the simple imperative language of (Davis, Sigal & Weyuker 1994), and also discuss register machines, Turing machines (again), KU machines. We also discuss whether programming beyond the Turing Limit makes sense and can be pursued.
- **Nov 9:** *Hypergraphical Proof and Programming in HyperSlate[®].* We here introduce the availability of writing Clojure functions in the context of proofs in HyperSlate[®].
- **Nov 12:** *Quantified Modal Logic.* We here explore quantified **S5**, the infamous Barcan Formula. HyperSlate[®] is used.
- **Nov 16:** *Killer Robots, D, and Beyond in HyperSlate[®] to DC&C.* We begin here by stating the “PAID Problem,” and then the approach to it from Bringsjord et al. advocates.

Schedule Switcheroo

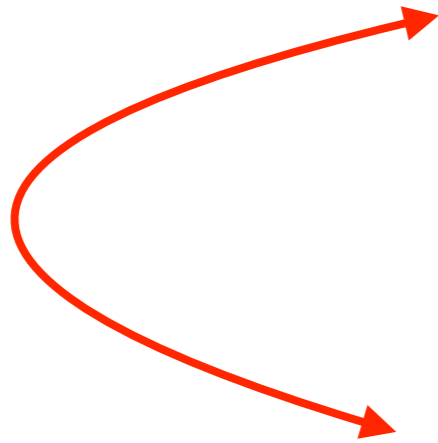


- **Nov 5:** *Pure General Logic Programming, Functional Programming, Turing-Completeness, and Beyond.* We review the basic paradigms of computer programming. For the imperative case, we use the simple imperative language of (Davis, Sigal & Weyuker 1994), and also discuss register machines, Turing machines (again), KU machines. We also discuss whether programming beyond the Turing Limit makes sense and can be pursued.
- **Nov 9:** *Hypergraphical Proof and Programming in HyperSlate[®].* We here introduce the availability of writing Clojure functions in the context of proofs in HyperSlate[®].
- **Nov 12:** *Quantified Modal Logic.* We here explore quantified **S5**, the infamous Barcan Formula. HyperSlate[®] is used.
- **Nov 16:** *Killer Robots, D, and Beyond in HyperSlate[®] to DC&C.* We begin here by stating the “PAID Problem,” and then the approach to it from Bringsjord et al. advocates.

Schedule Switcheroo



- **Nov 5:** *Pure General Logic Programming, Functional Programming, Turing-Completeness, and Beyond.* We review the basic paradigms of computer programming. For the imperative case, we use the simple imperative language of (Davis, Sigal & Weyuker 1994), and also discuss register machines, Turing machines (again), KU machines. We also discuss whether programming beyond the Turing Limit makes sense and can be pursued.
- **Nov 9:** *Hypergraphical Proof and Programming in HyperSlate[®].* We here introduce the availability of writing Clojure functions in the context of proofs in HyperSlate[®].
- **Nov 12:** *Quantified Modal Logic.* We here explore quantified **S5**, the infamous Barcan Formula. HyperSlate[®] is used.
- **Nov 16:** *Killer Robots, D, and Beyond in HyperSlate[®] to DC&C.* We begin here by stating the “PAID Problem,” and then the approach to it from Bringsjord et al. advocates.

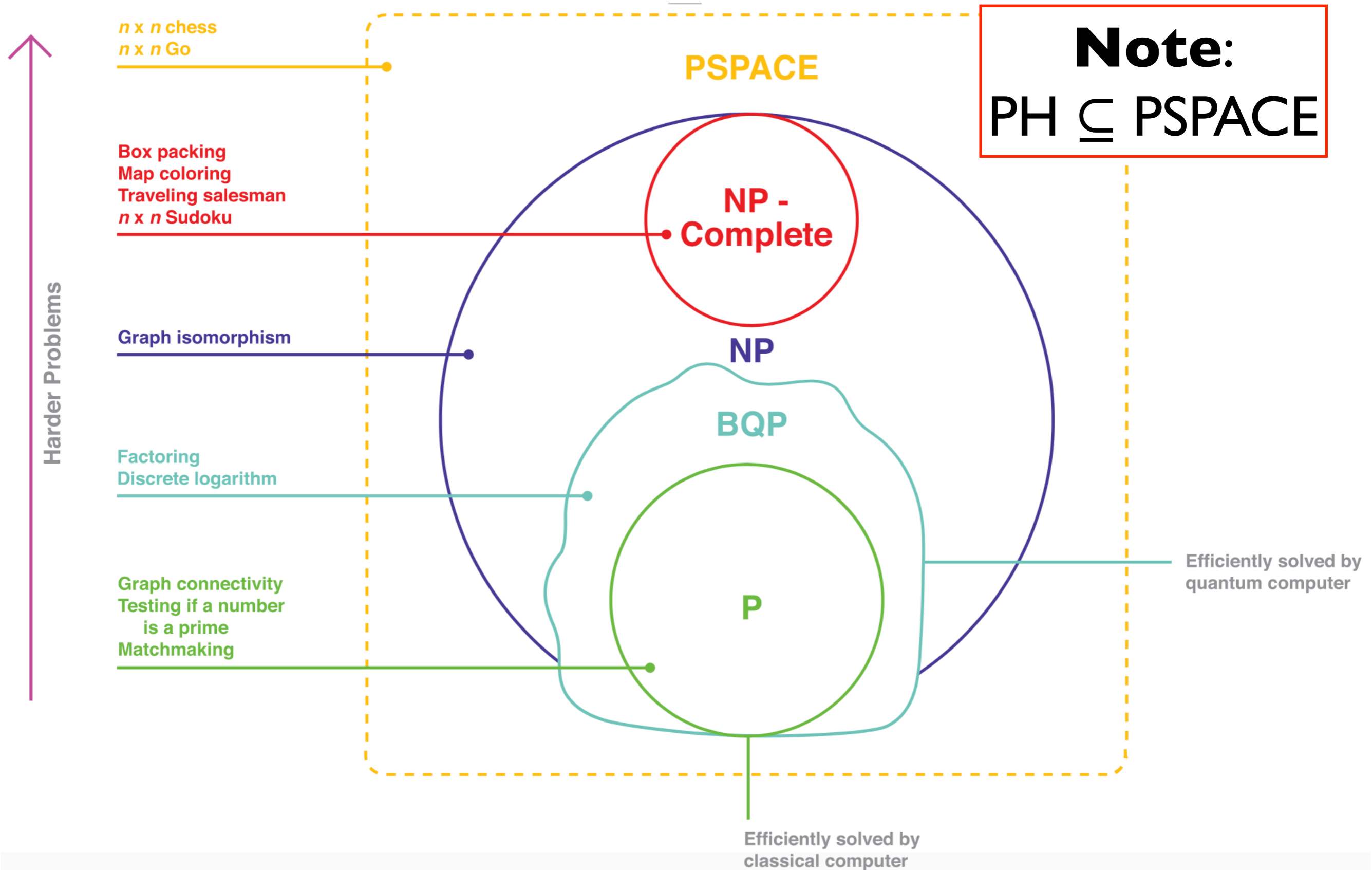


Further delivery on
promissory note re
building hierarchies via
formal logic...

What about (oft vaunted) quantum computers?

Note:
 $PH \subseteq PSPACE$

What about (oft vaunted) quantum computers?



What about (oft vaunted) quantum computers?

GCI

Harder Problems

$n \times n$ chess
 $n \times n$ Go

Box packing
Map coloring
Traveling salesman
 $n \times n$ Sudoku

Graph isomorphism

Factoring
Discrete logarithm

Graph connectivity
Testing if a number
is a prime
Matchmaking

PSPACE

NP -
Complete

NP

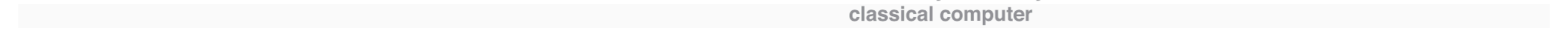
BQP

P

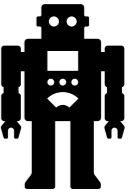
Note:
 $PH \subseteq PSPACE$

Efficiently solved by
quantum computer

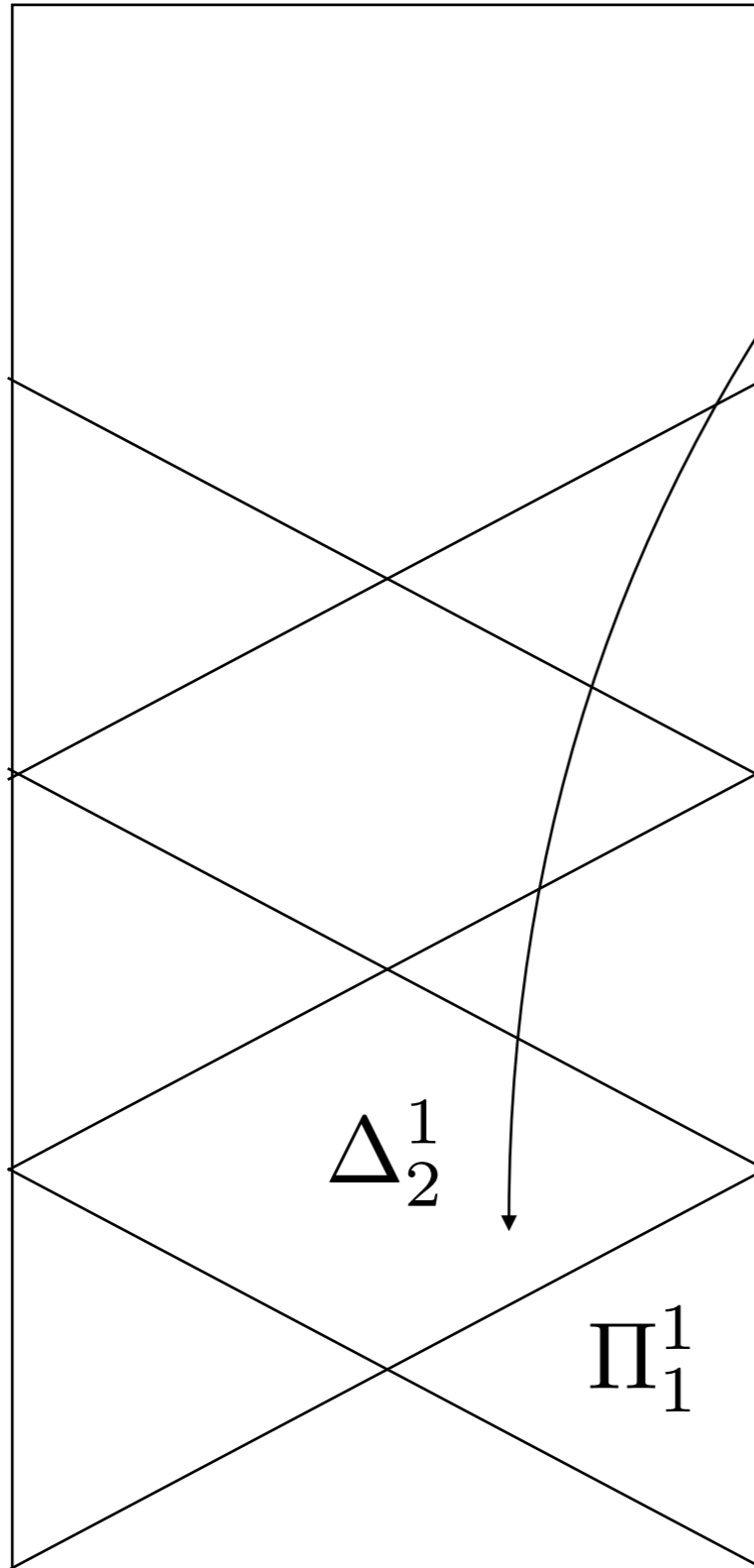
Efficiently solved by
classical computer



CogSci and AI need to say more about where AI falls/can fall in the landscape.

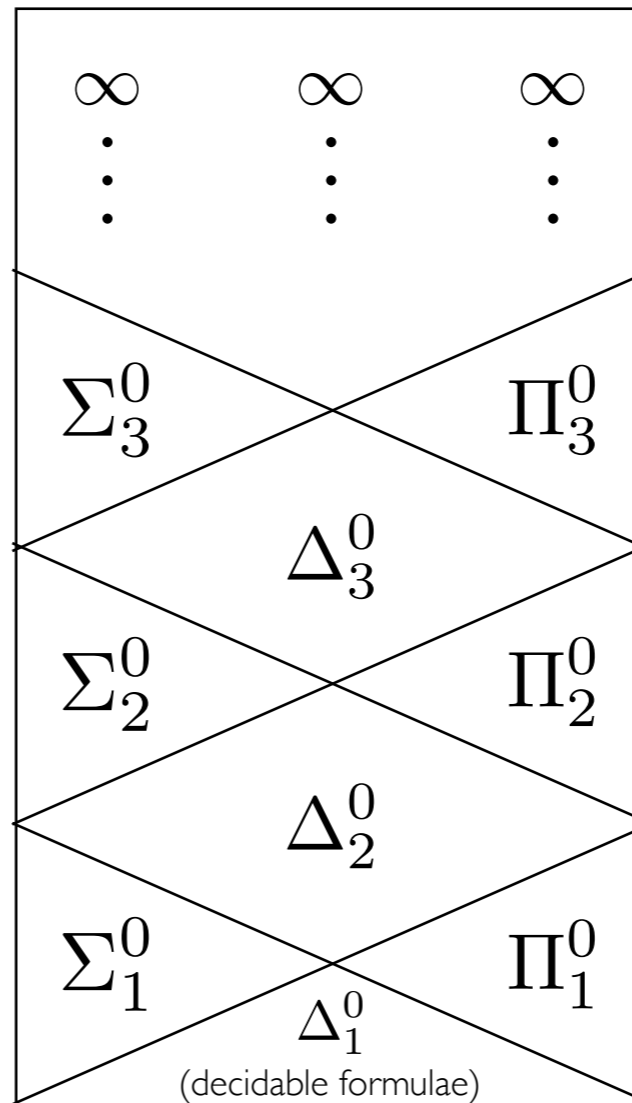


$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)

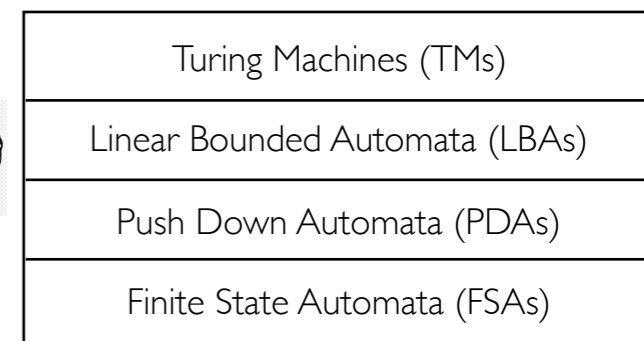


Human Persons (according to Bringsjord)

Human Brains (according to Granger)



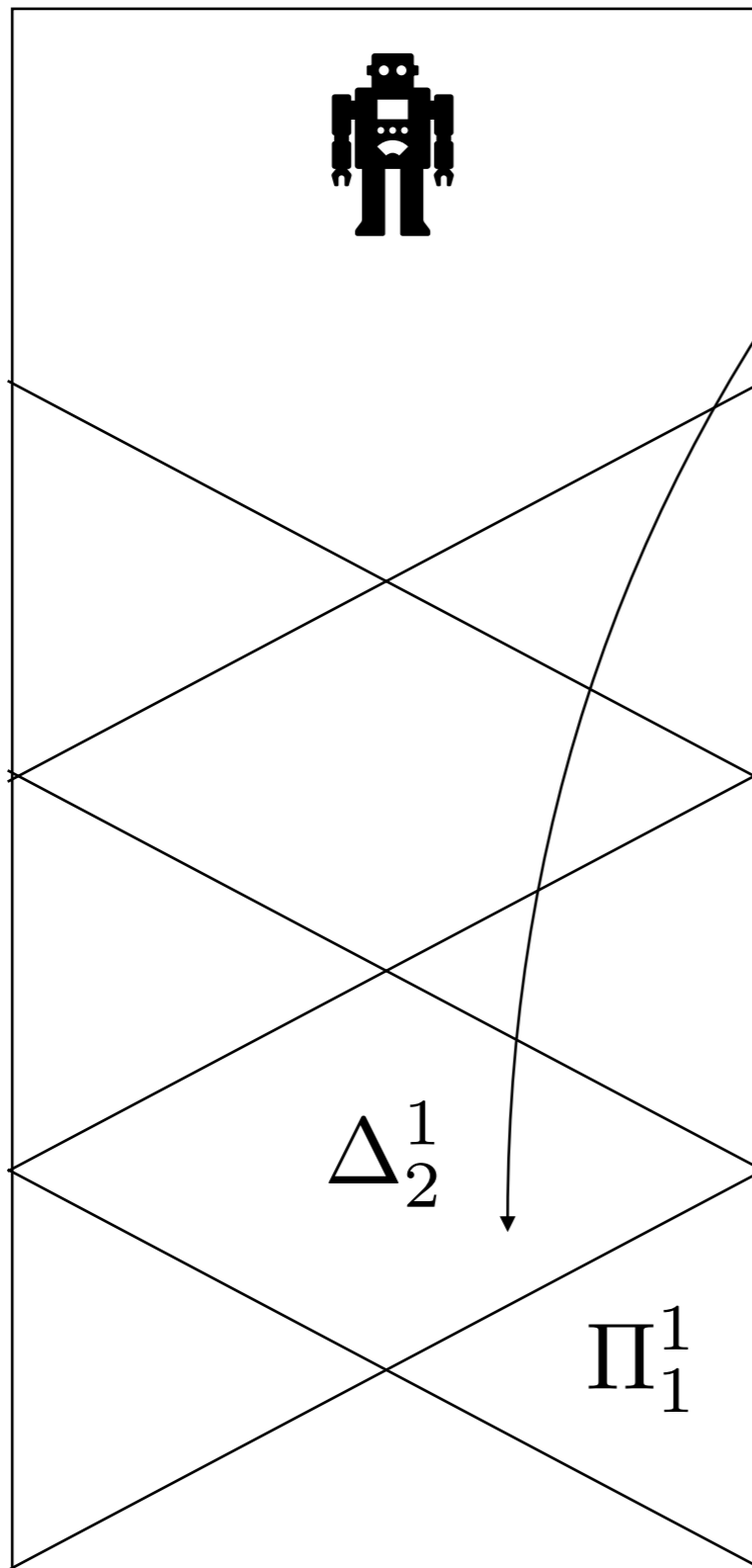
\mathcal{CH} (Chomsky Hierarchy)



\mathcal{EM}

CogSci and AI need to say more about where AI falls/can fall in the landscape.

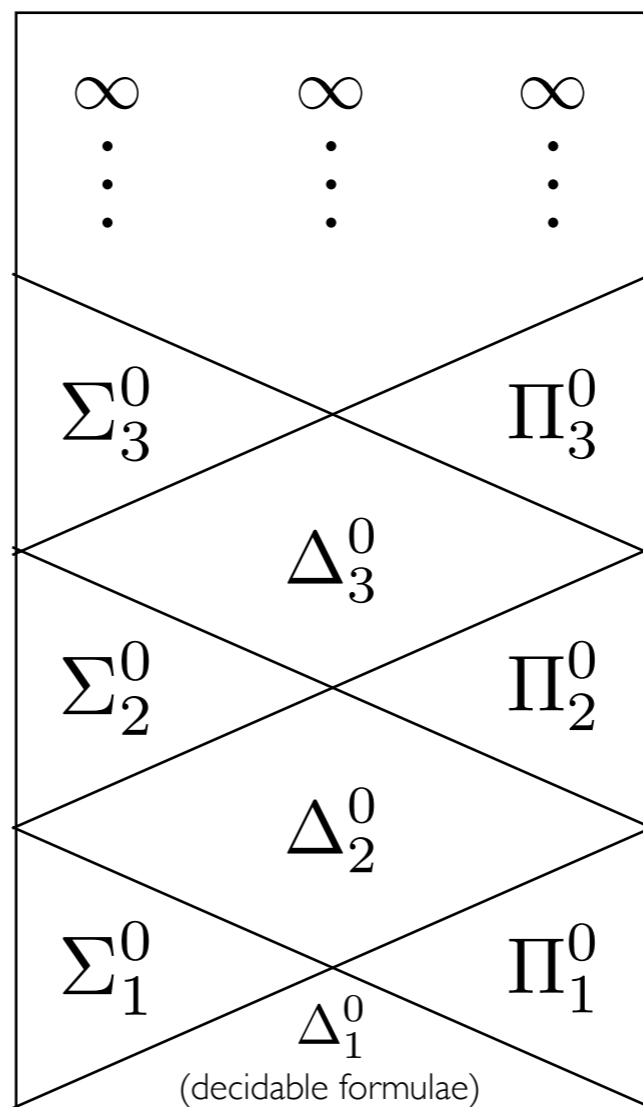
$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

Human Persons (according to Bringsjord)

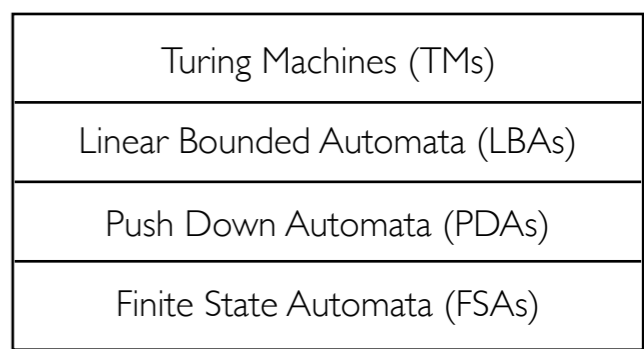
$A^r \mathcal{H}$ (Arithmetic Hierarchy)



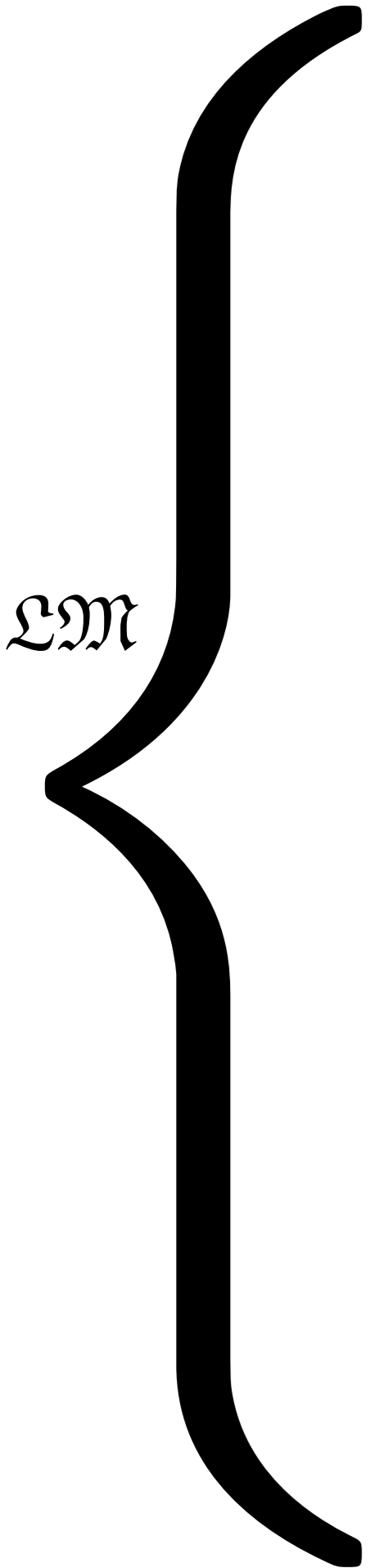
Human Brains (according to Granger)



\mathcal{CH} (Chomsky Hierarchy)

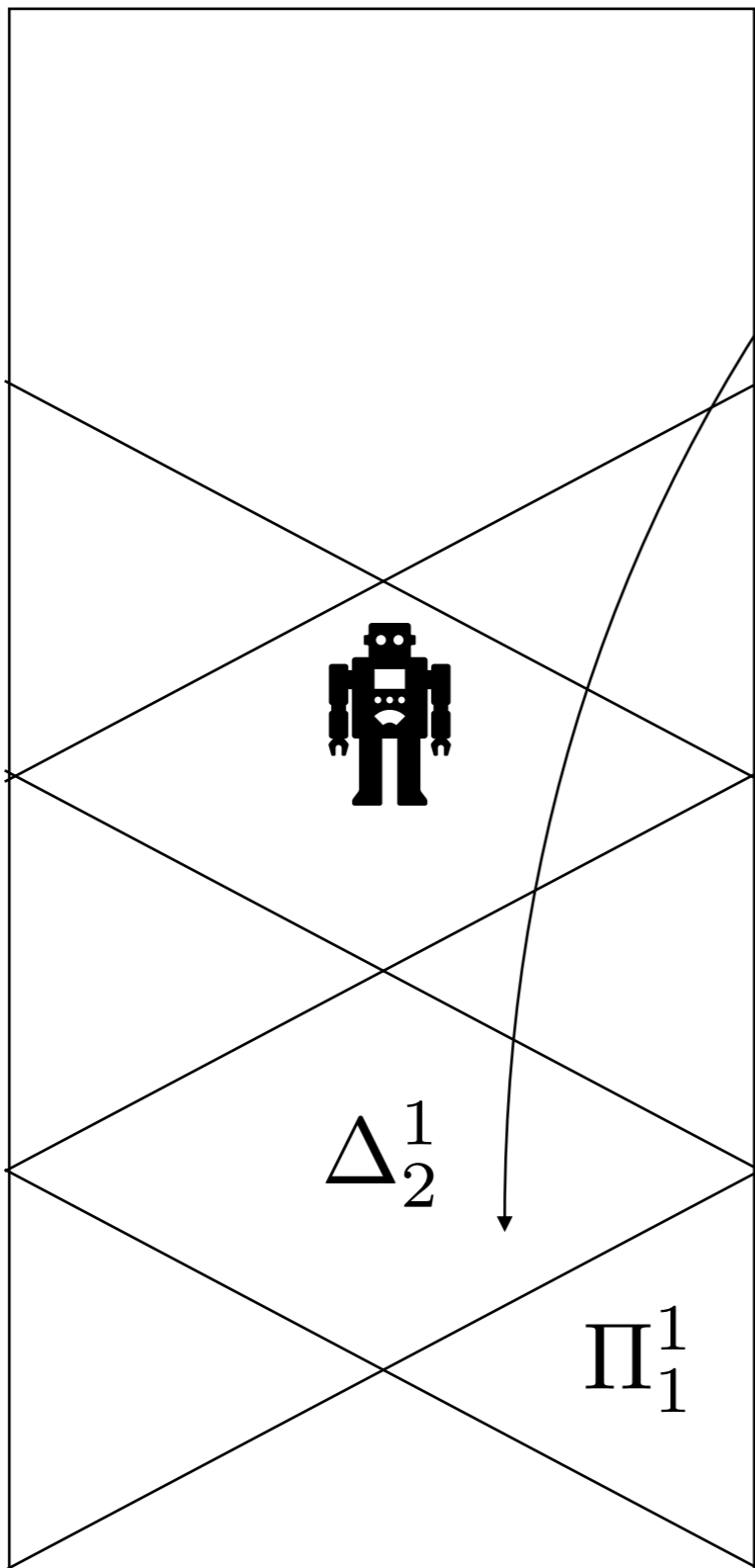


\mathcal{EM}



CogSci and AI need to say more about where AI falls/can fall in the landscape.

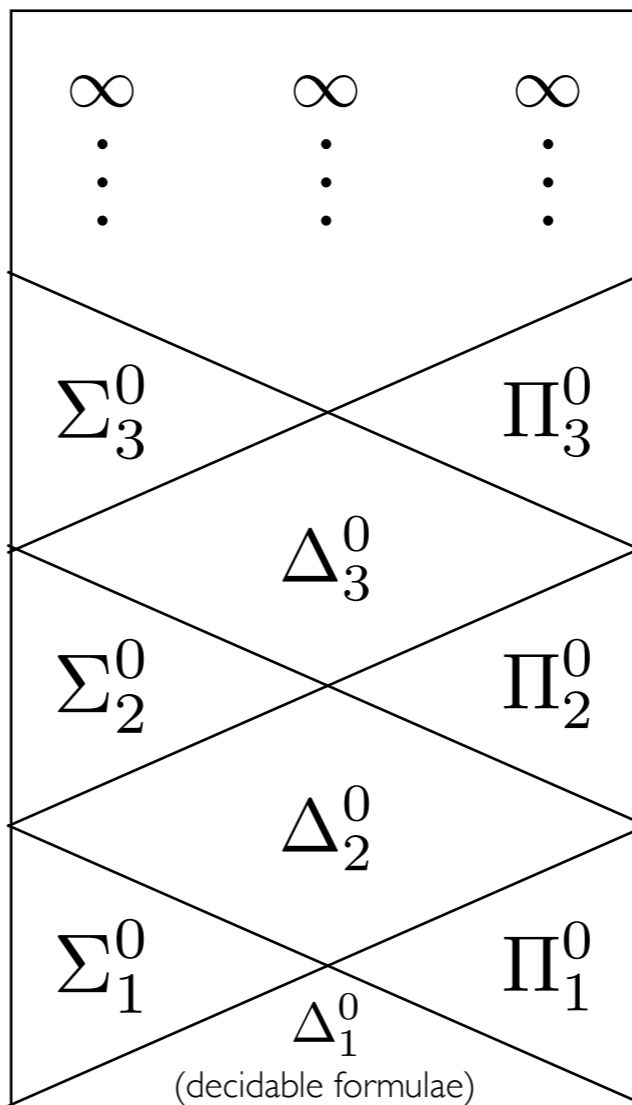
$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

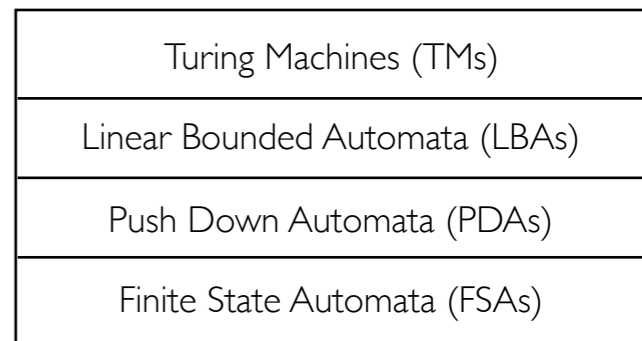
Human Persons
(according to Bringsjord)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)



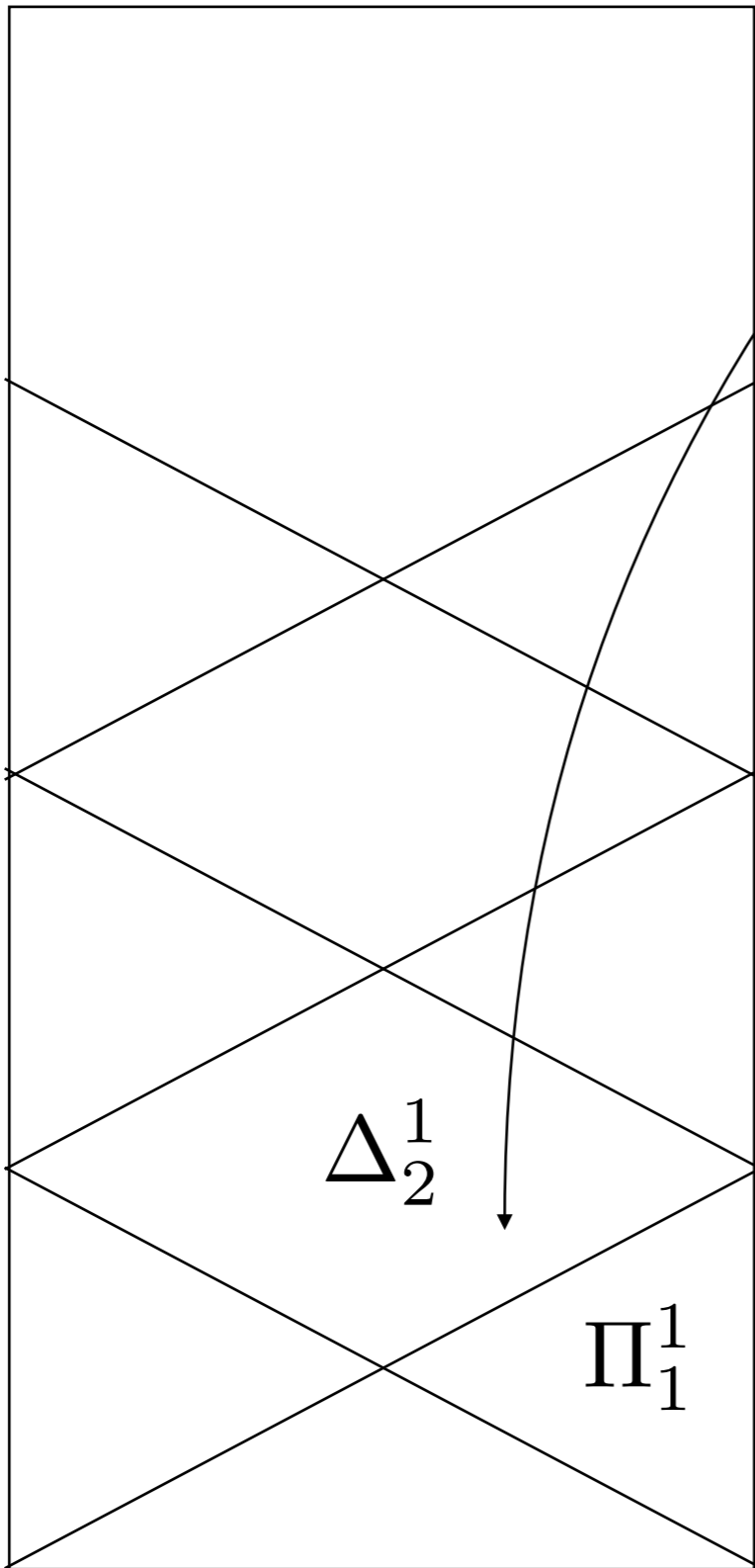
Human Brains
(according to Granger)

\mathcal{CH} (Chomsky Hierarchy)



CogSci and AI need to say more about where AI falls/can fall in the landscape.

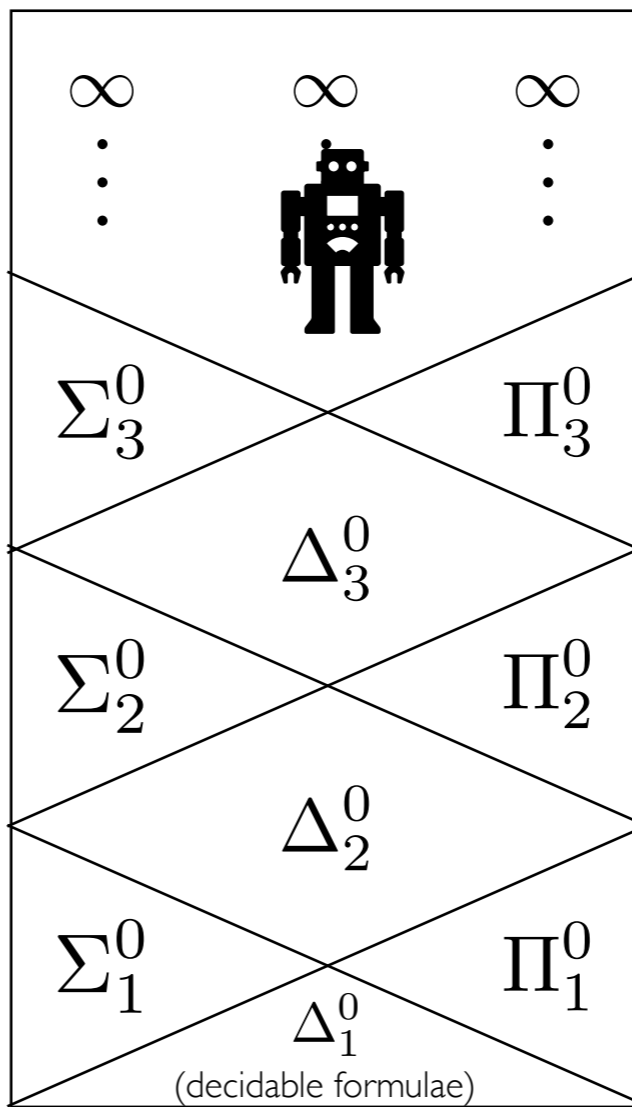
$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

Human Persons
(according to Bringsjord)

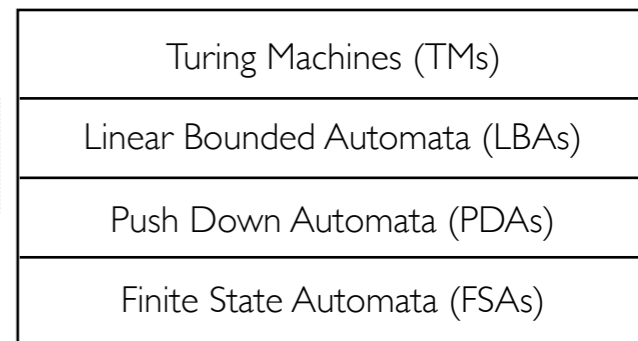
$A^r \mathcal{H}$ (Arithmetic Hierarchy)



Human Brains
(according to Granger)



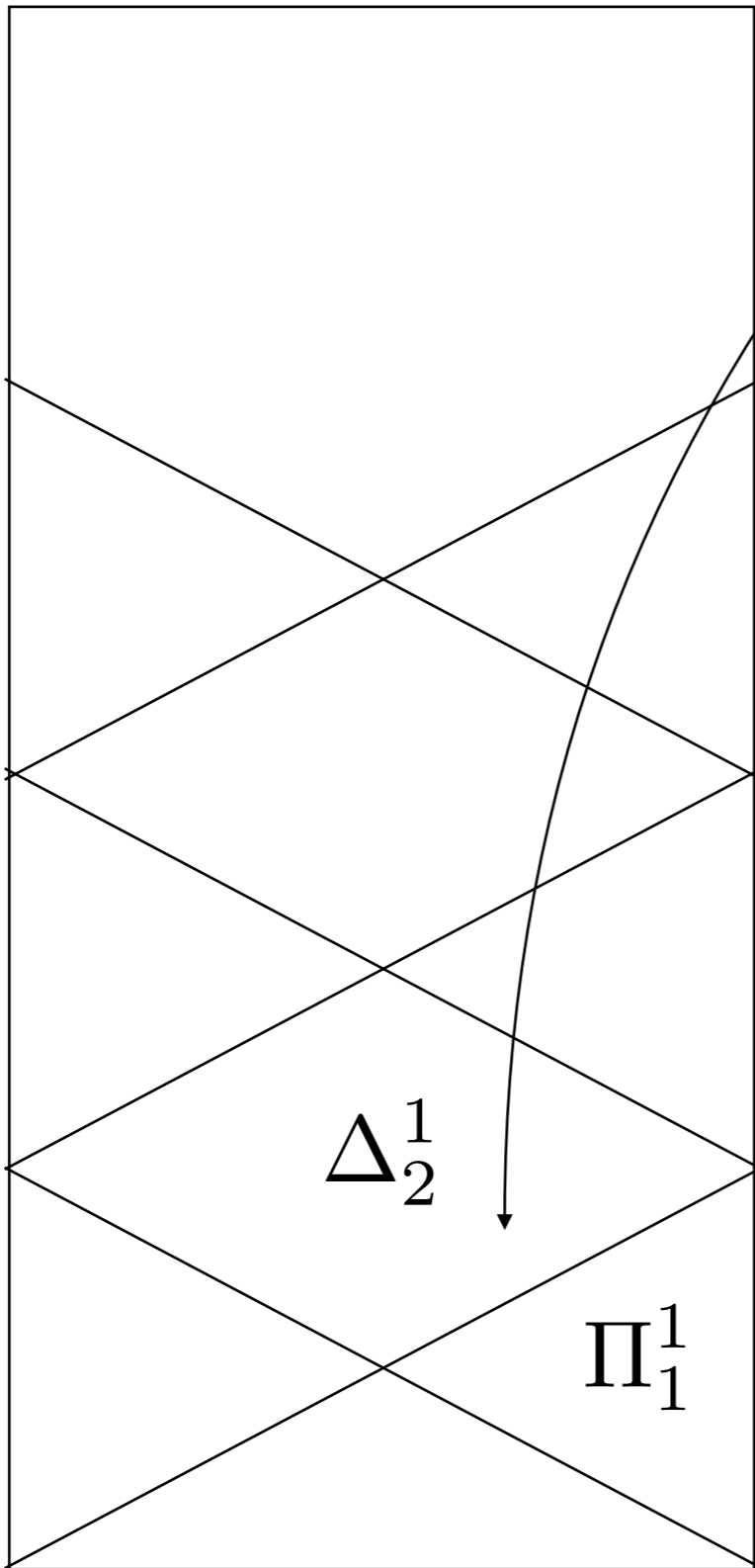
\mathcal{CH} (Chomsky Hierarchy)



\mathcal{EM}

CogSci and AI need to say more about where AI falls/can fall in the landscape.

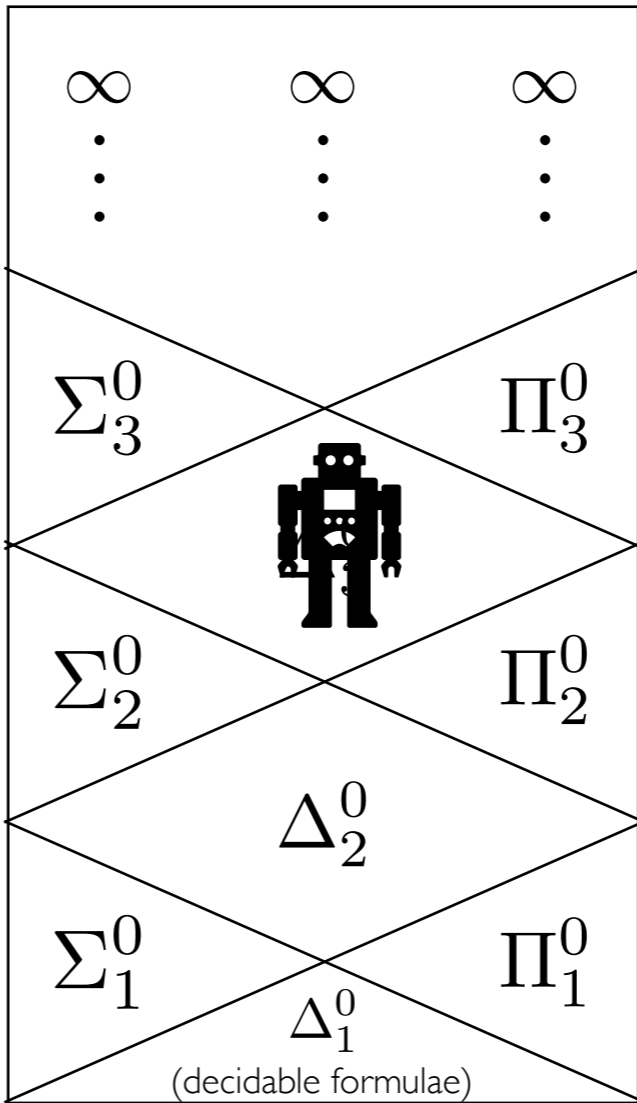
$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

Human Persons (according to Bringsjord)

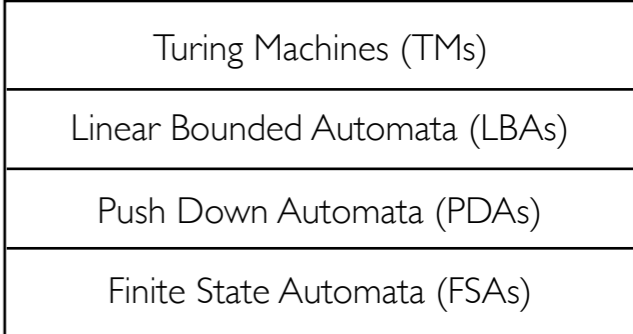
$A^r \mathcal{H}$ (Arithmetic Hierarchy)



Human Brains (according to Granger)



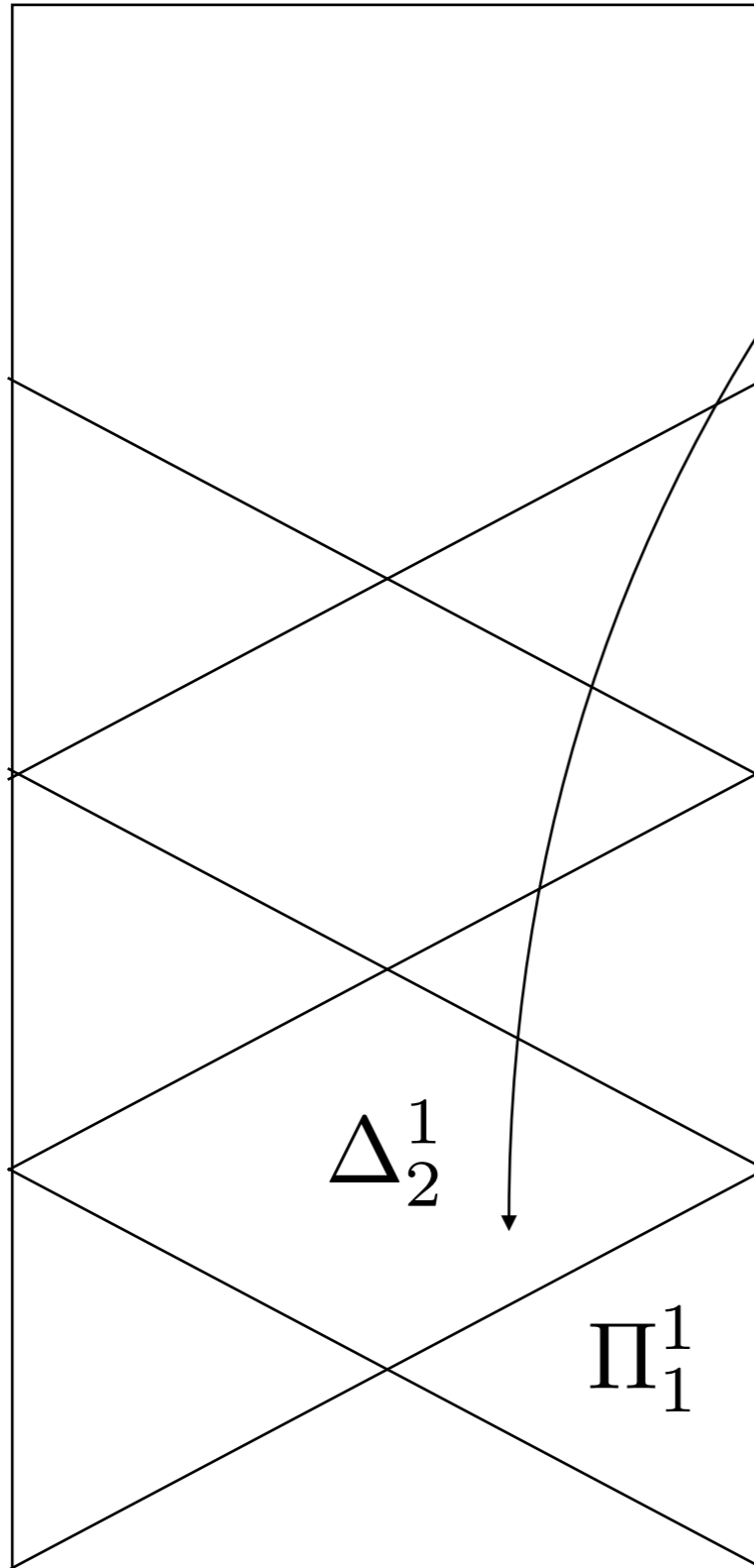
\mathcal{CH} (Chomsky Hierarchy)



\mathcal{EM}

CogSci and AI need to say more about where AI falls/can fall in the landscape.

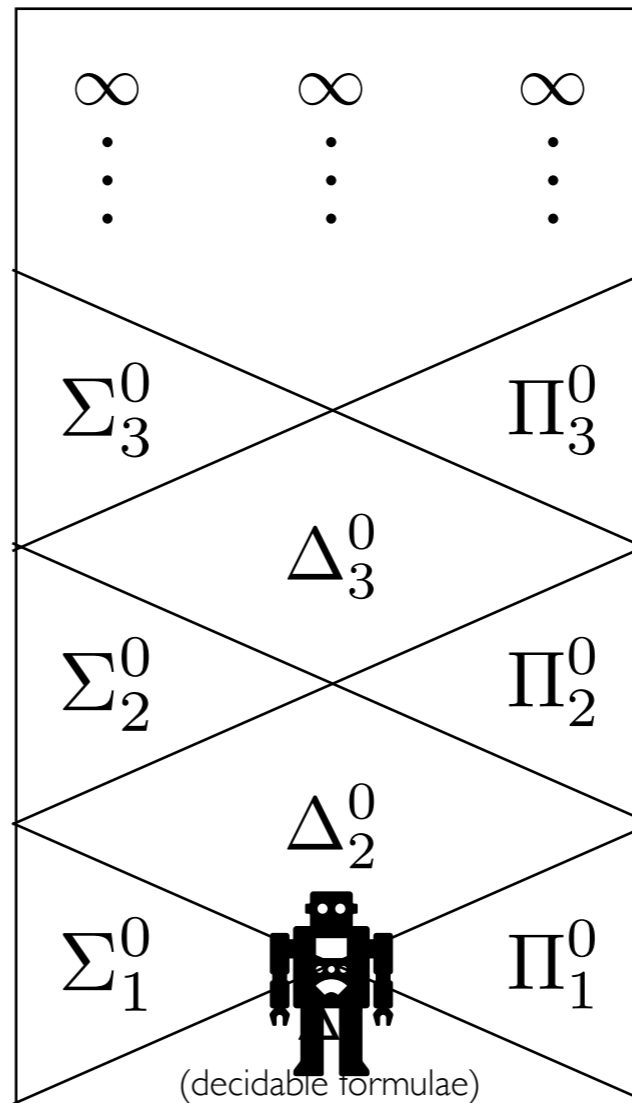
$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

Human Persons
(according to Bringsjord)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)



Human Brains
(according to Granger)



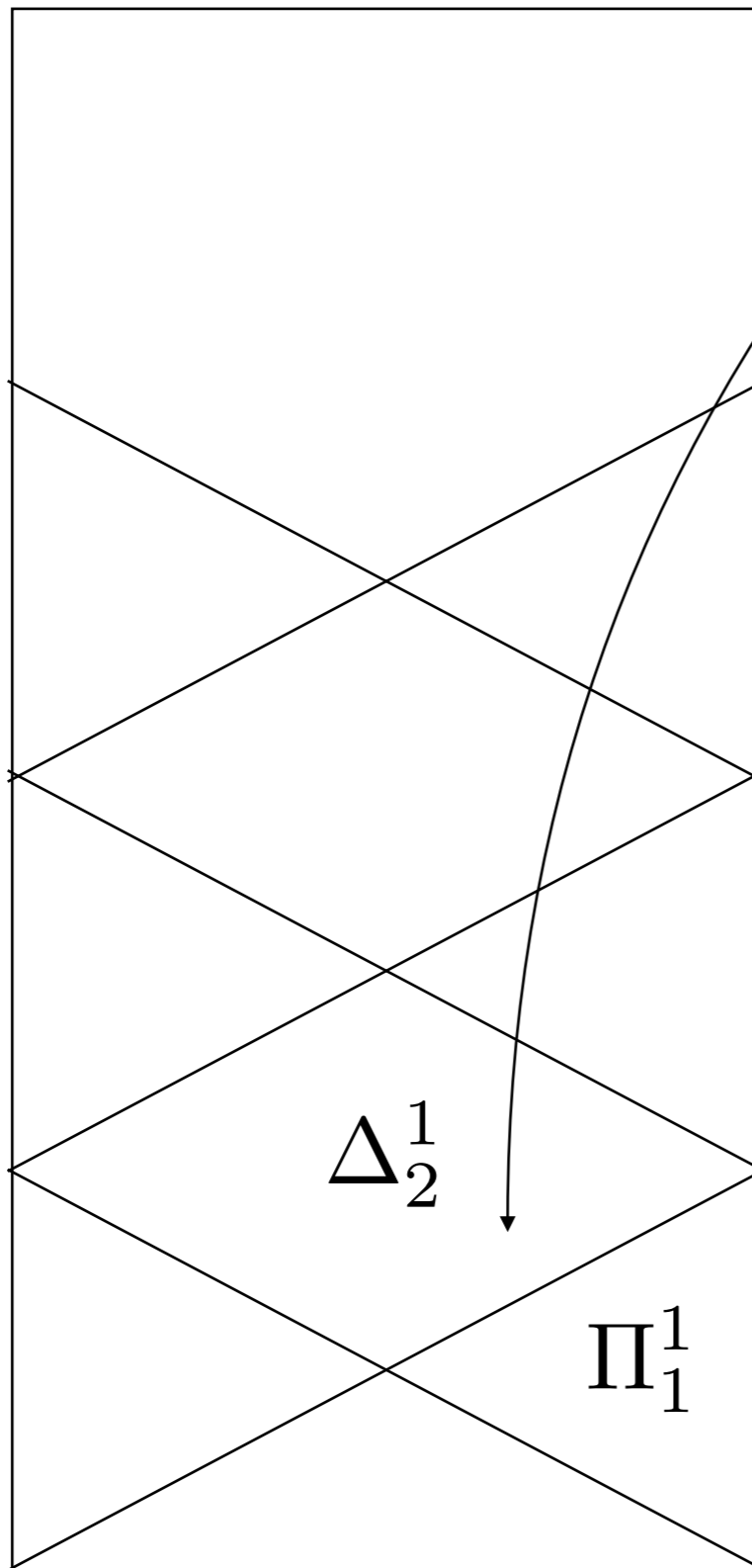
\mathcal{CH} (Chomsky Hierarchy)

- Turing Machines (TMs)
- Linear Bounded Automata (LBAs)
- Push Down Automata (PDAs)
- Finite State Automata (FSAs)



CogSci and AI need to say more about where AI falls/can fall in the landscape.

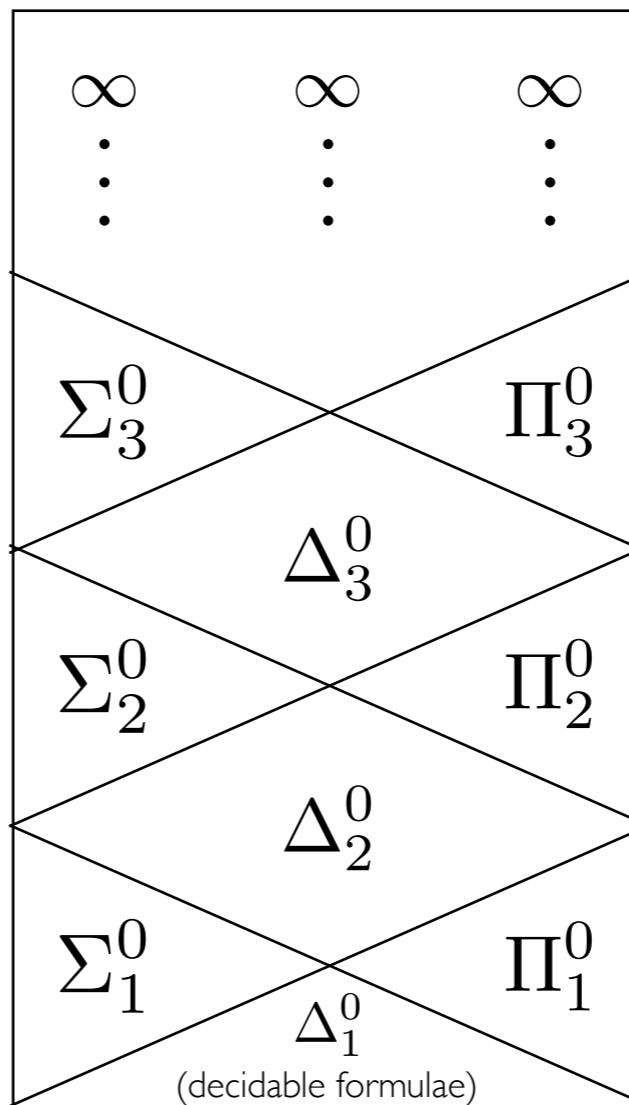
$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

Human Persons (according to Bringsjord)

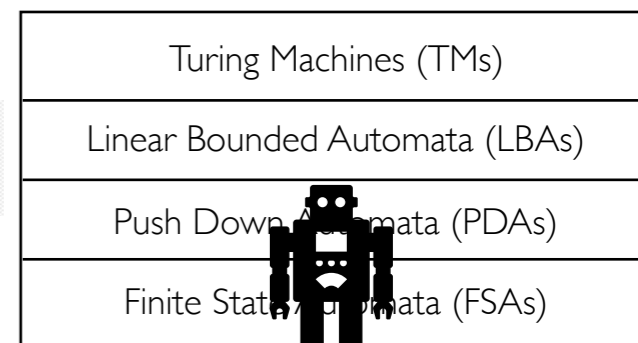
$A^r \mathcal{H}$ (Arithmetic Hierarchy)



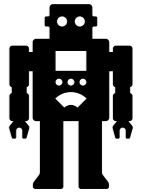
Human Brains (according to Granger)



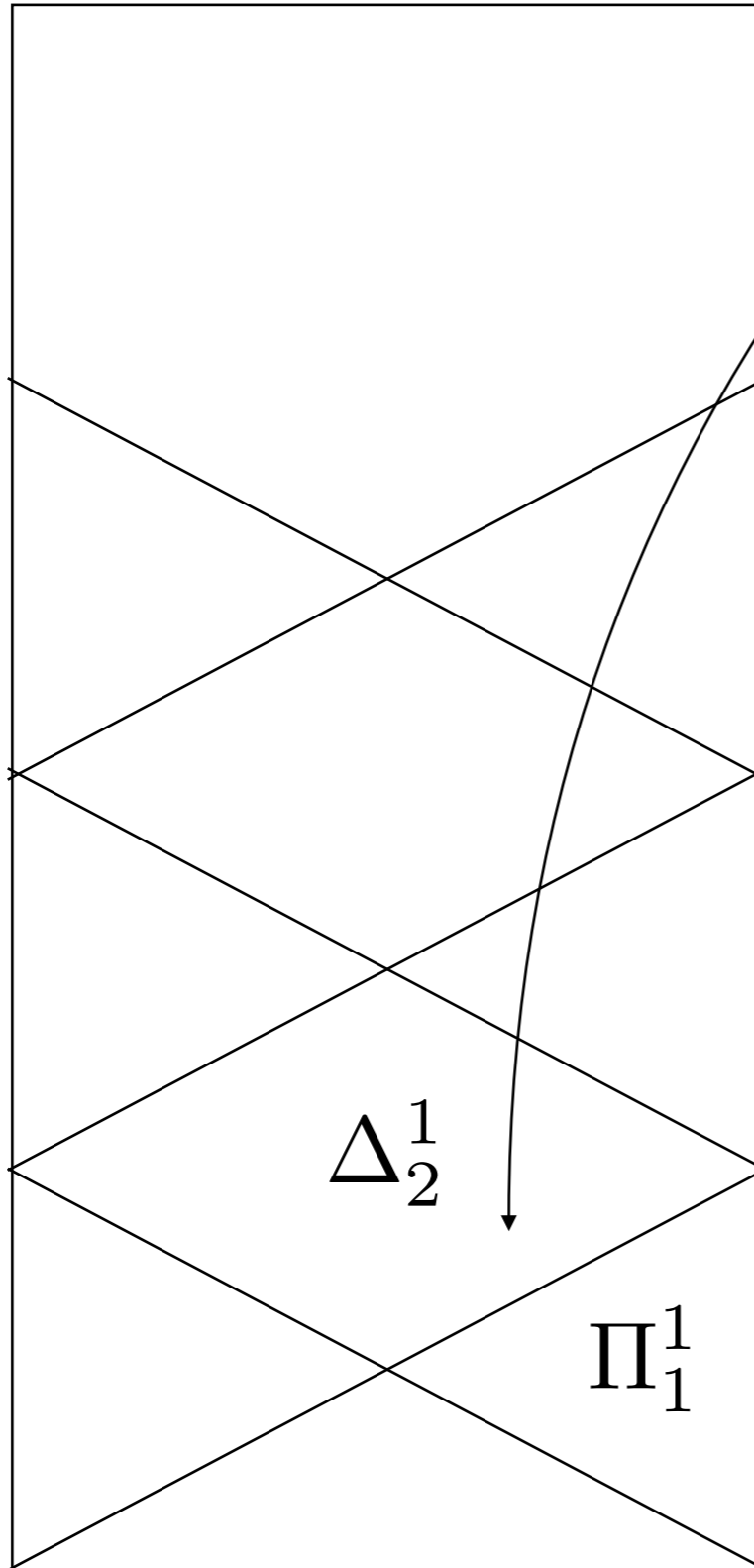
\mathcal{CH} (Chomsky Hierarchy)



CogSci and AI need to say more about where AI falls/can fall in the landscape.

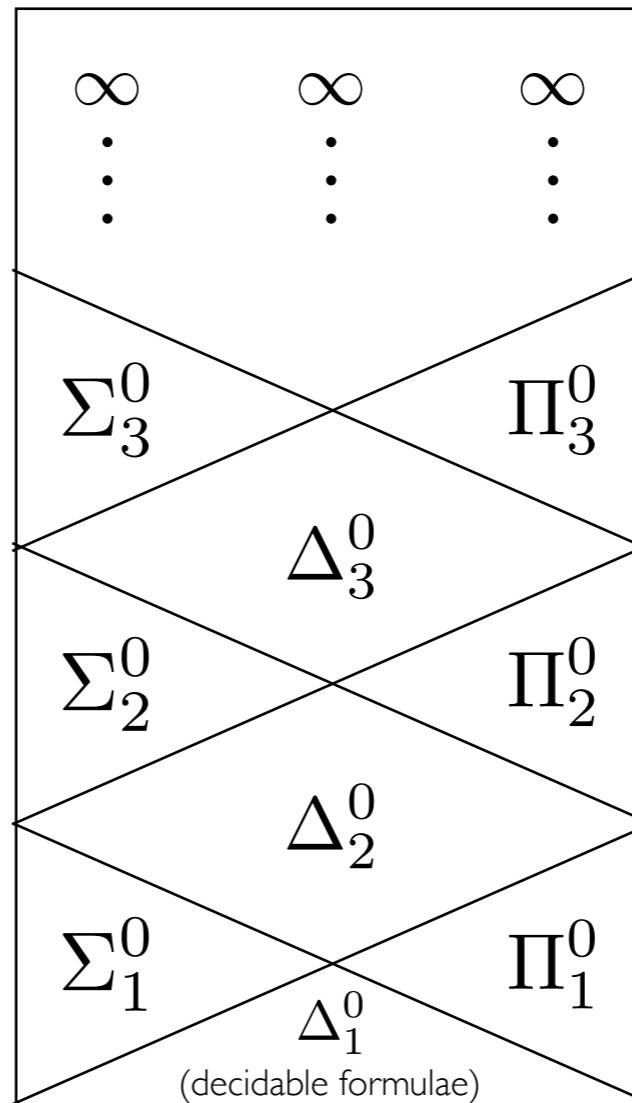


$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)

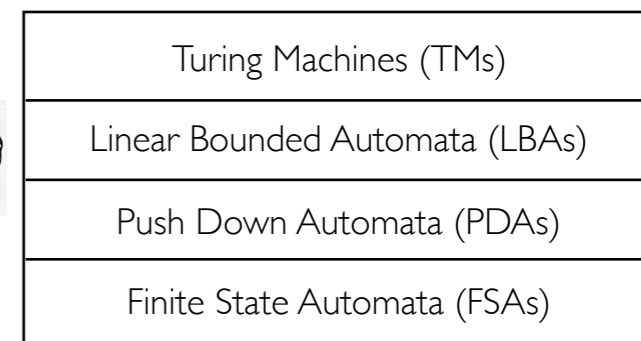


Human Persons
(according to Bringsjord)

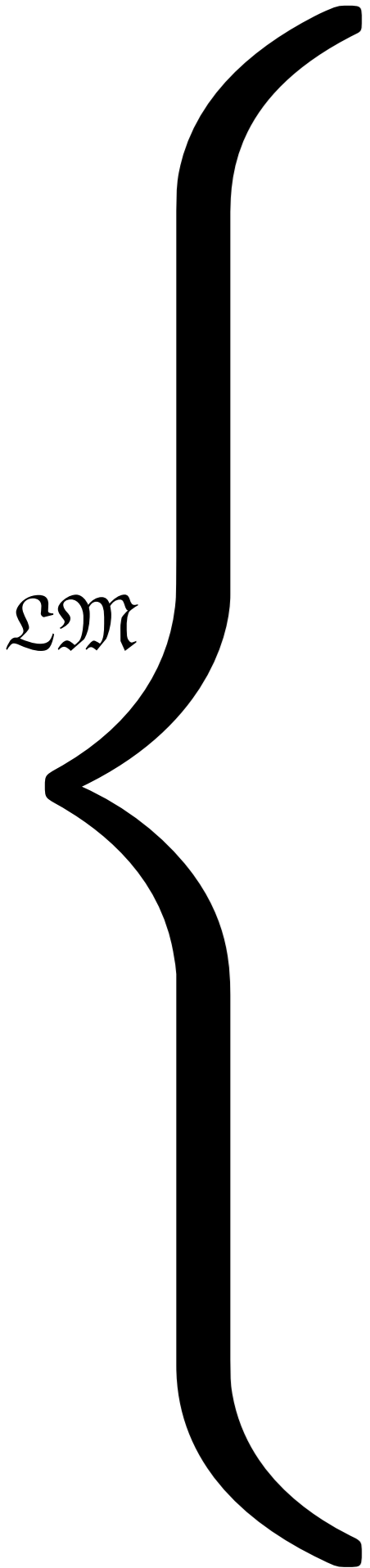
Human Brains
(according to Granger)



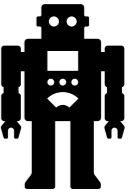
\mathcal{CH} (Chomsky Hierarchy)



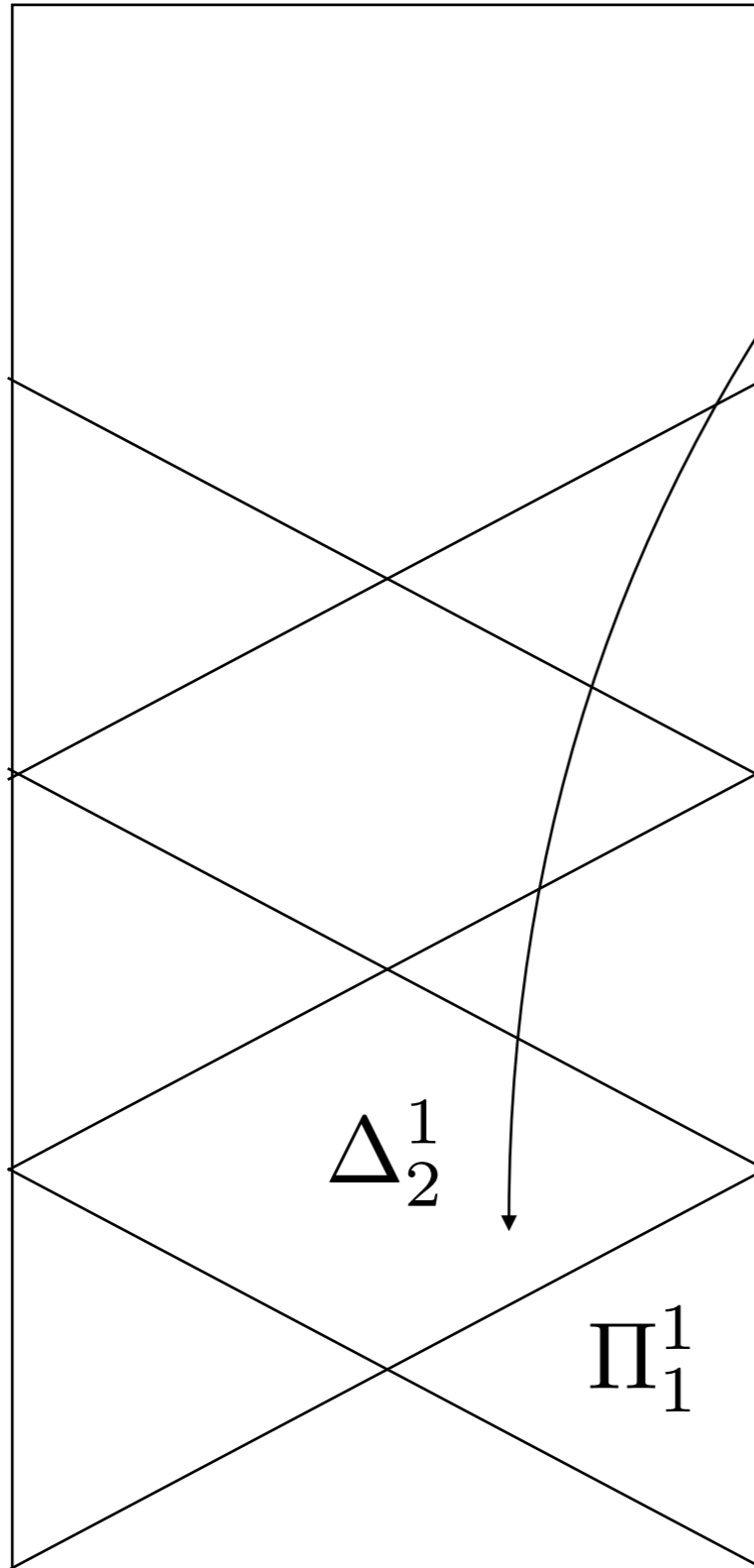
\mathcal{EM}



CogSci and AI need to say more about where AI falls/can fall in the landscape.

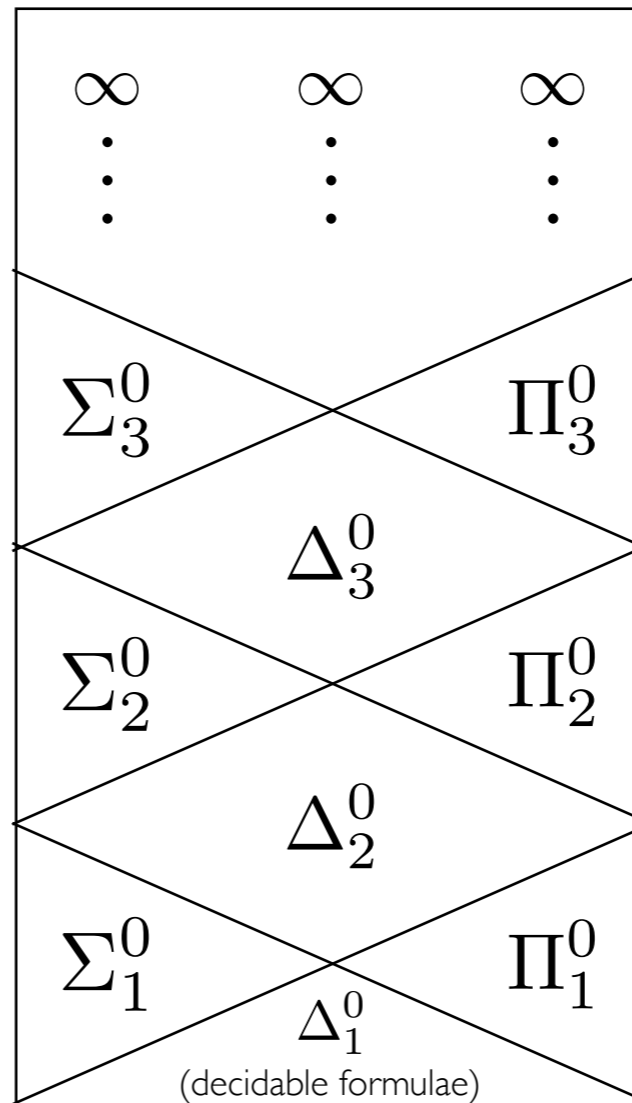


$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)

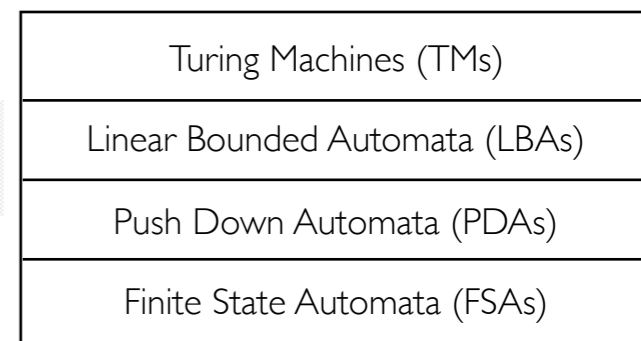


Human Persons (according to Bringsjord)

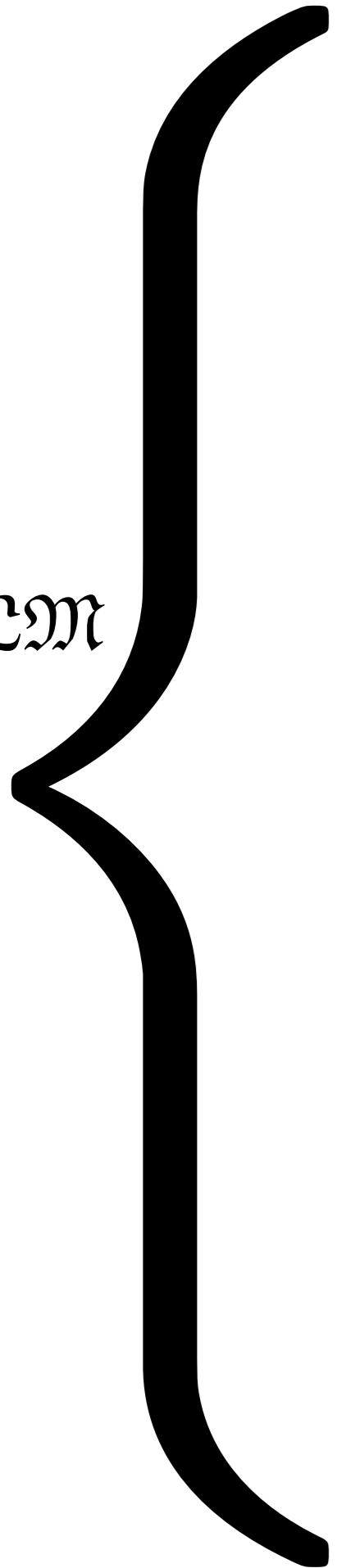
Human Brains (according to Granger)



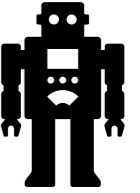
\mathcal{CH} (Chomsky Hierarchy)



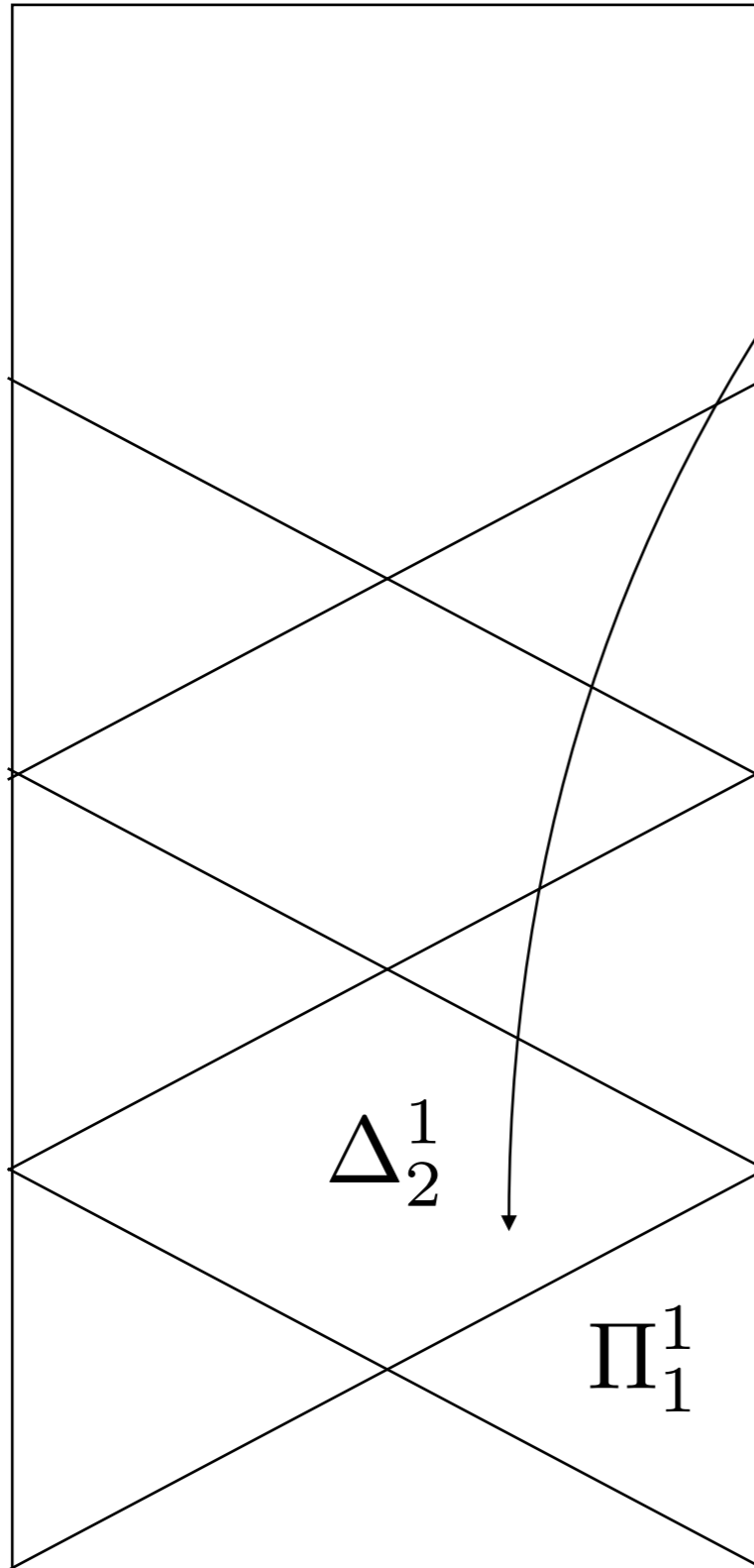
\mathcal{EM}



CogSci and AI need to say more about where AI falls/can fall in the landscape.

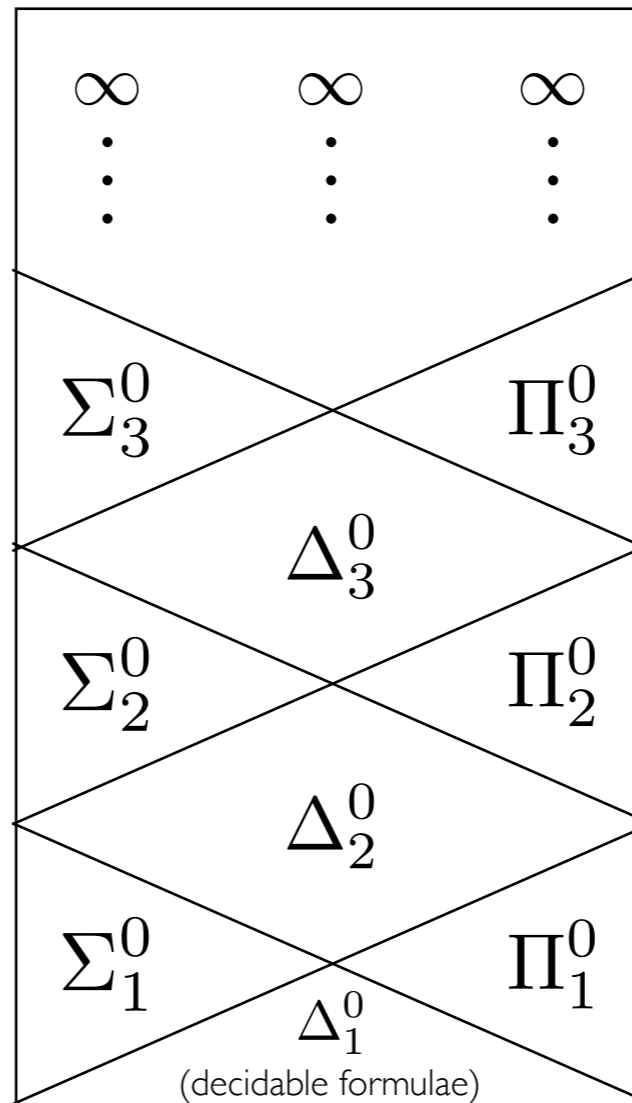


$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)

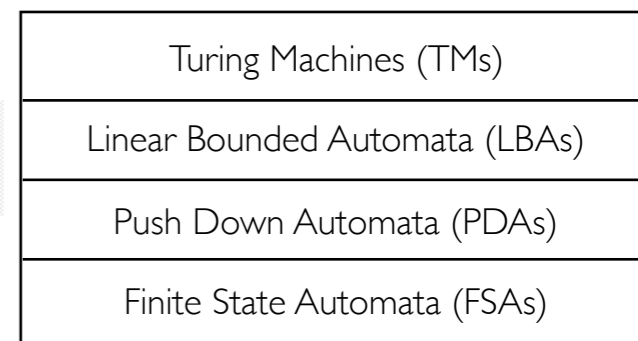


Human Persons (according to Bringsjord)

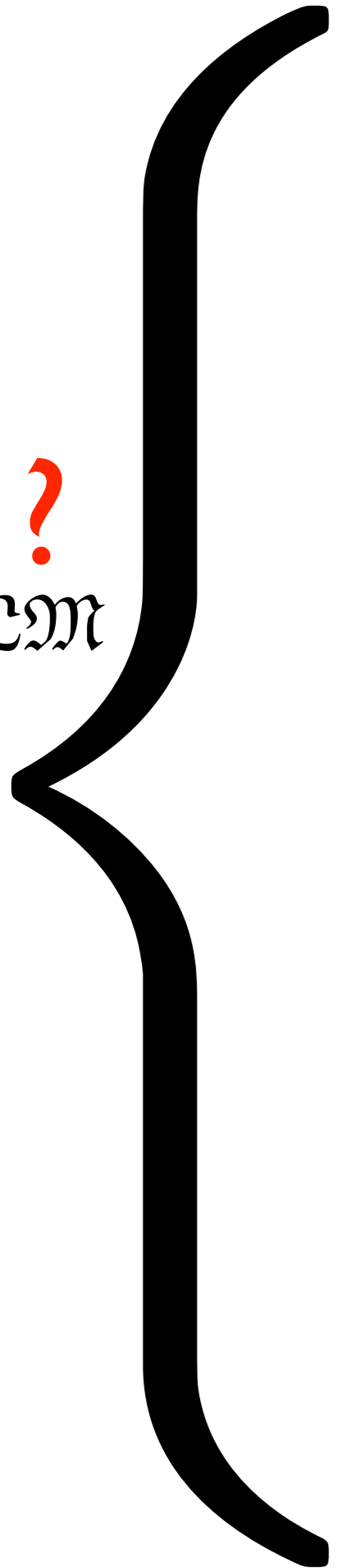
Human Brains (according to Granger)



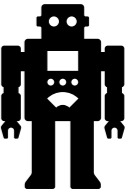
\mathcal{CH} (Chomsky Hierarchy)



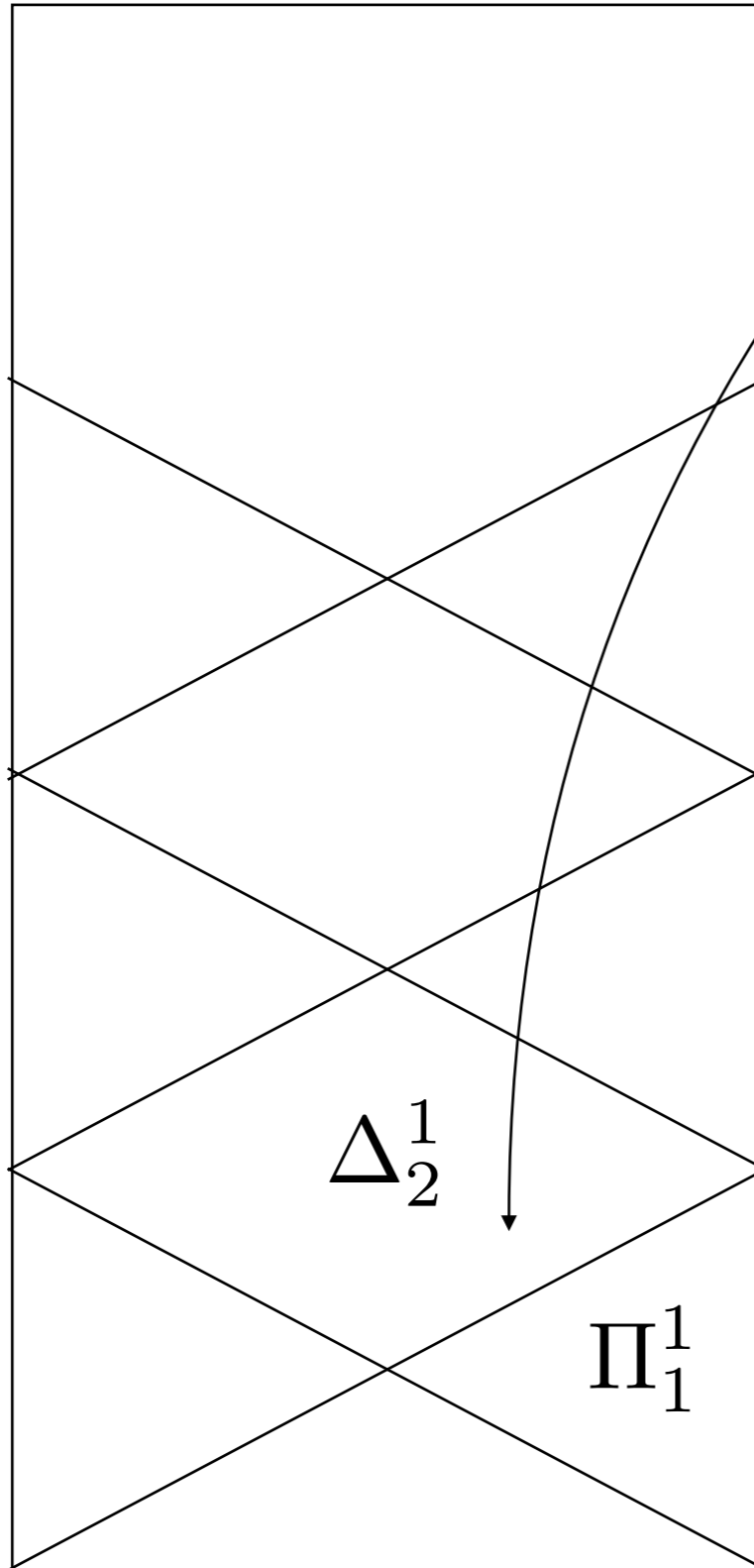
?
EM



CogSci and AI need to say more about where AI falls/can fall in the landscape.

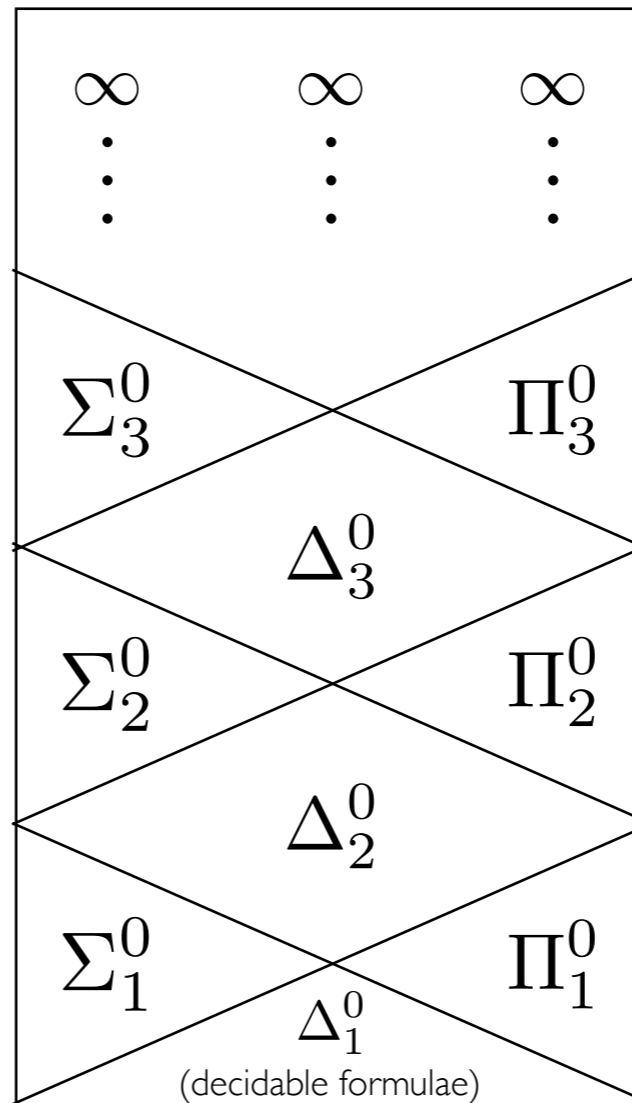


$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)



Human Persons (according to Bringsjord)

Human Brains (according to Granger)

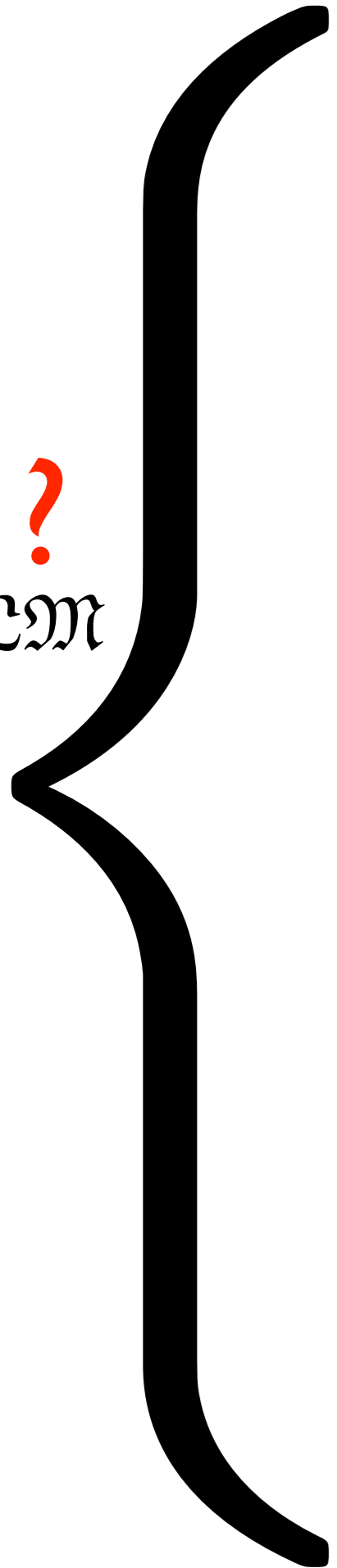


\mathcal{CH} (Chomsky Hierarchy)

- Turing Machines (TMs)
- Linear Bounded Automata (LBAs)
- Push Down Automata (PDAs)
- Finite State Automata (FSAs)

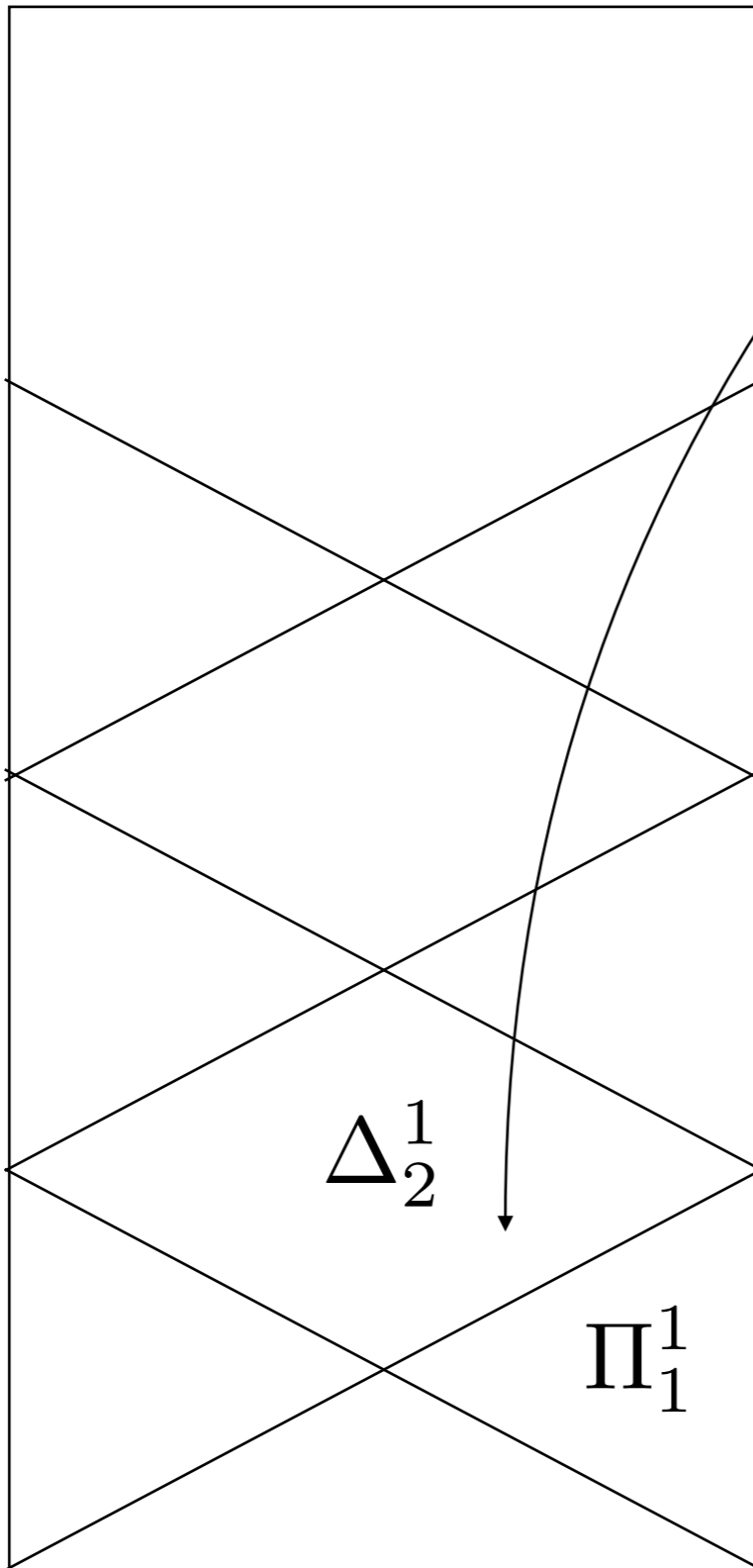


?
 \mathcal{EM}



CogSci and AI need to say more about where AI falls/can fall in the landscape.

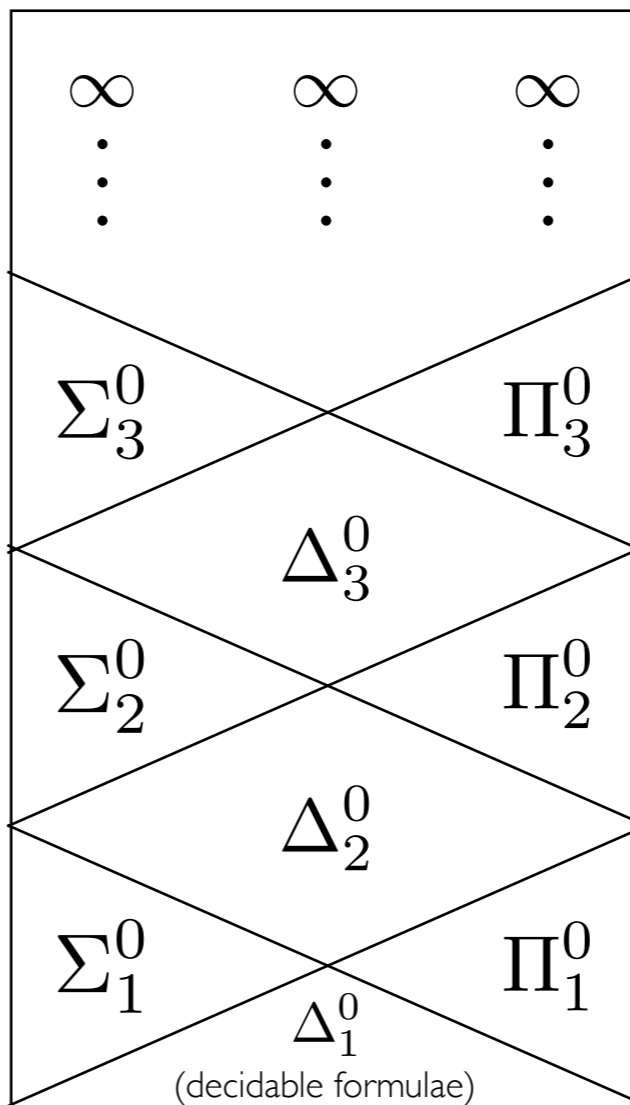
$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

Human Persons
(according to Bringsjord)

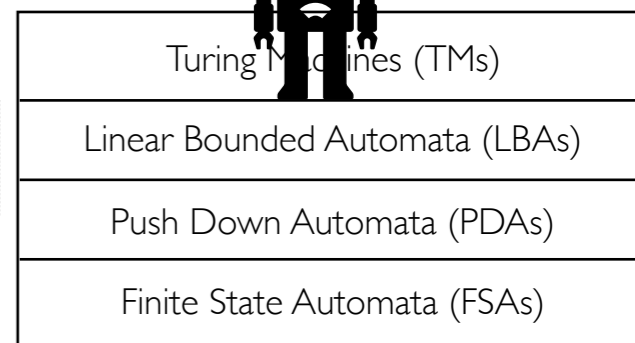
$A^r \mathcal{H}$ (Arithmetic Hierarchy)



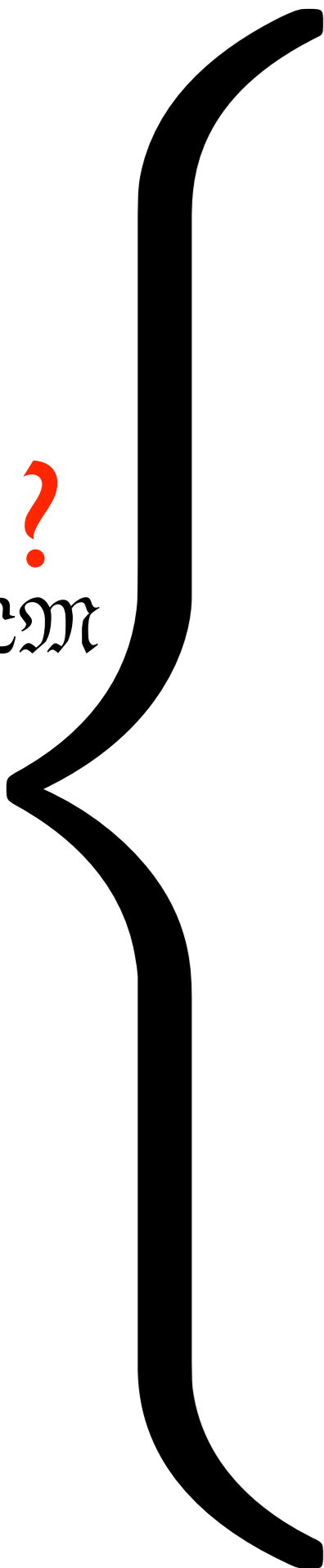
Human Brains
(according to Granger)



\mathcal{CH} (Churchs Hierarchy)

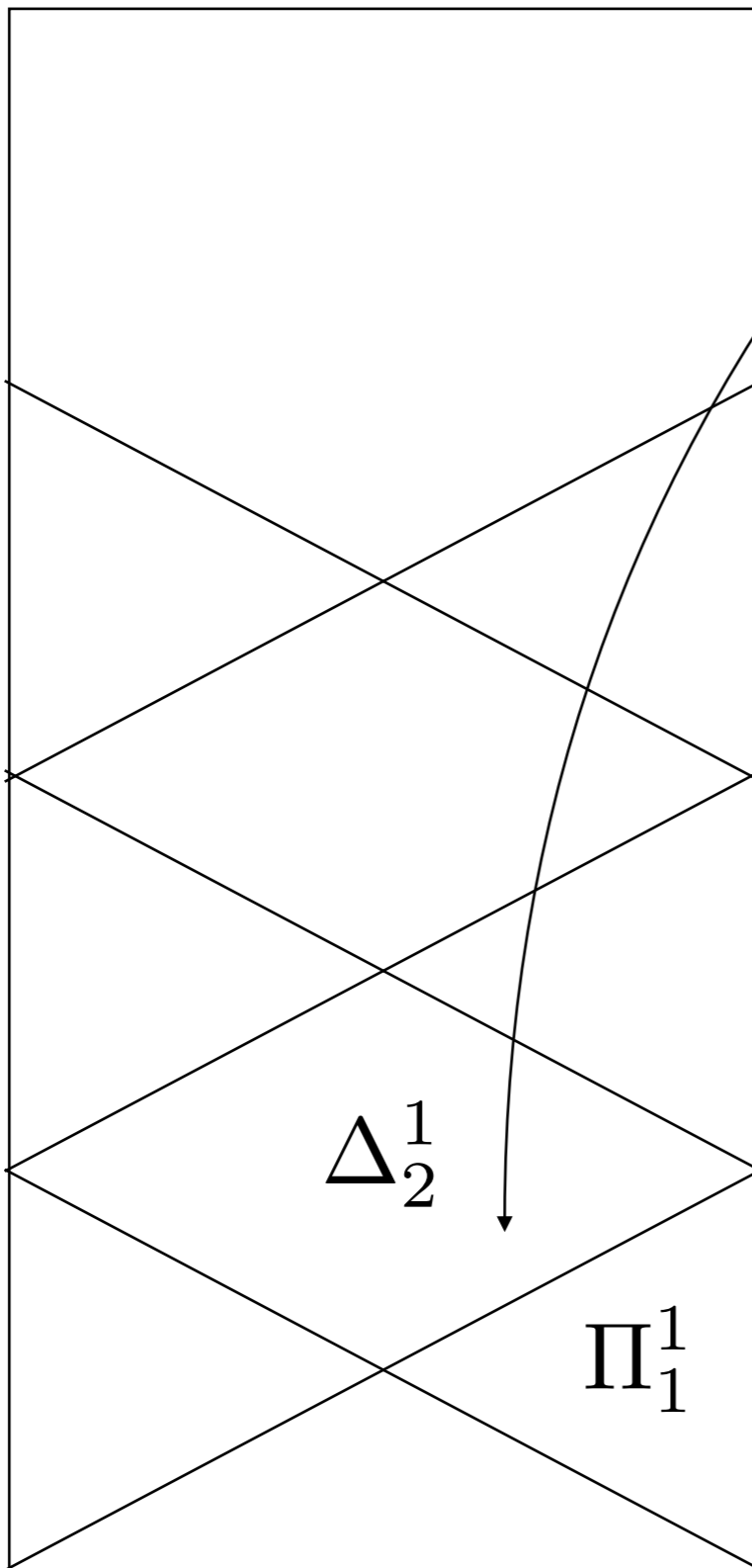


?
 \mathcal{EM}

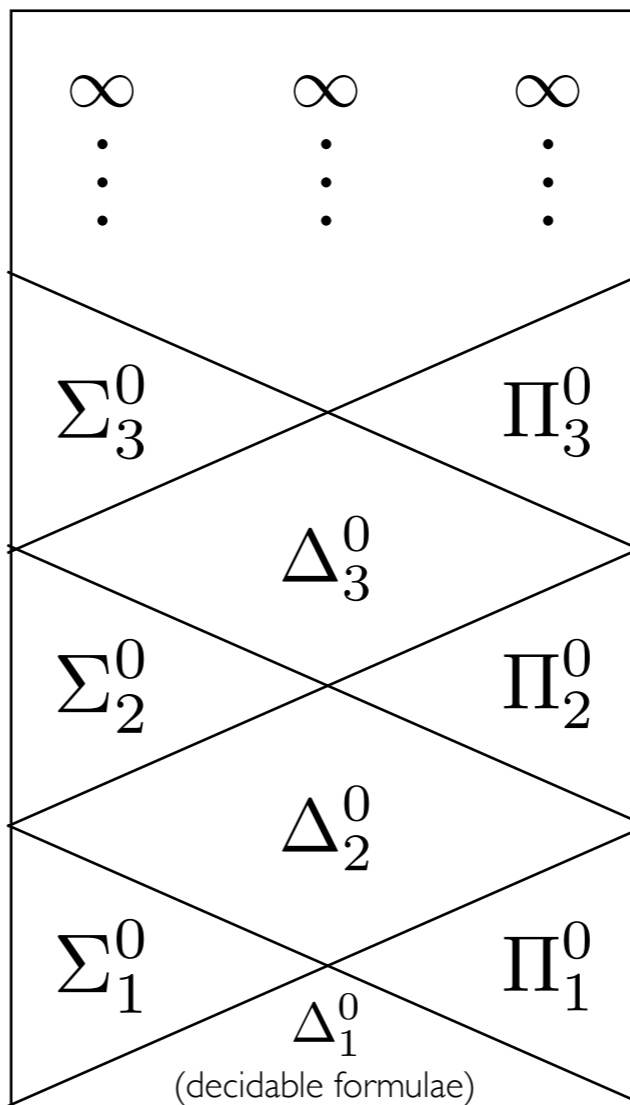


CogSci and AI need to say more about where AI falls/can fall in the landscape.

$A^n \mathcal{H}$ (Analytic Hierarchy)



$A^r \mathcal{H}$ (Arithmetic Hierarchy)



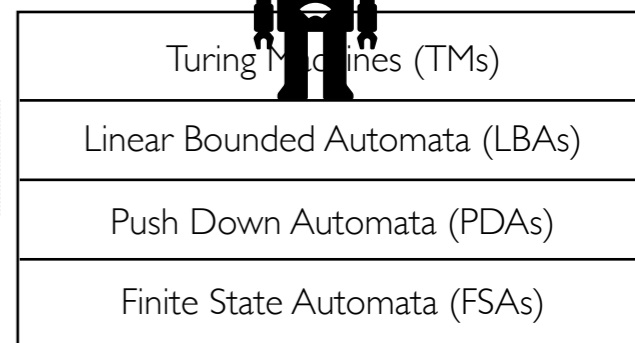
Infinite Time Turing Machines (ITTMs)

Human Persons
(according to Bringsjord)

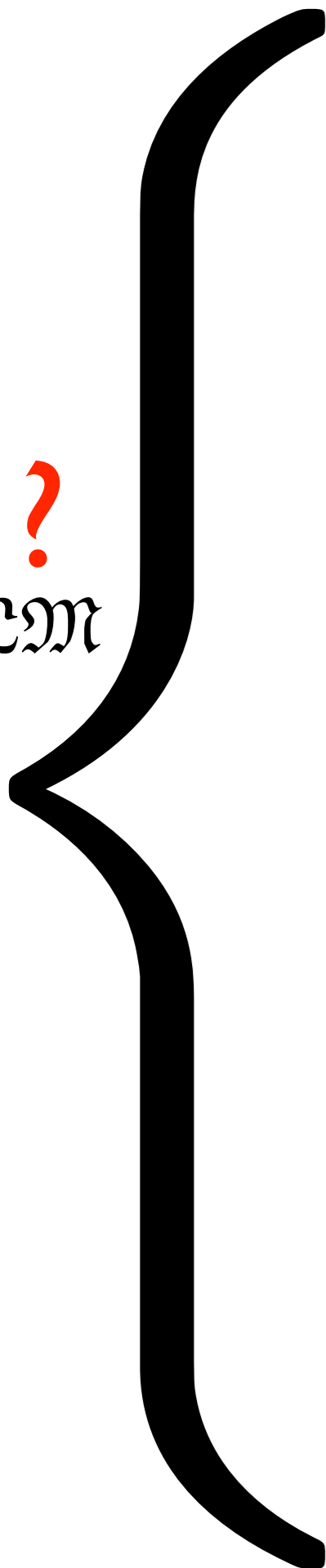
Human Brains
(according to Granger)



\mathcal{CH} (Computational Hierarchy)

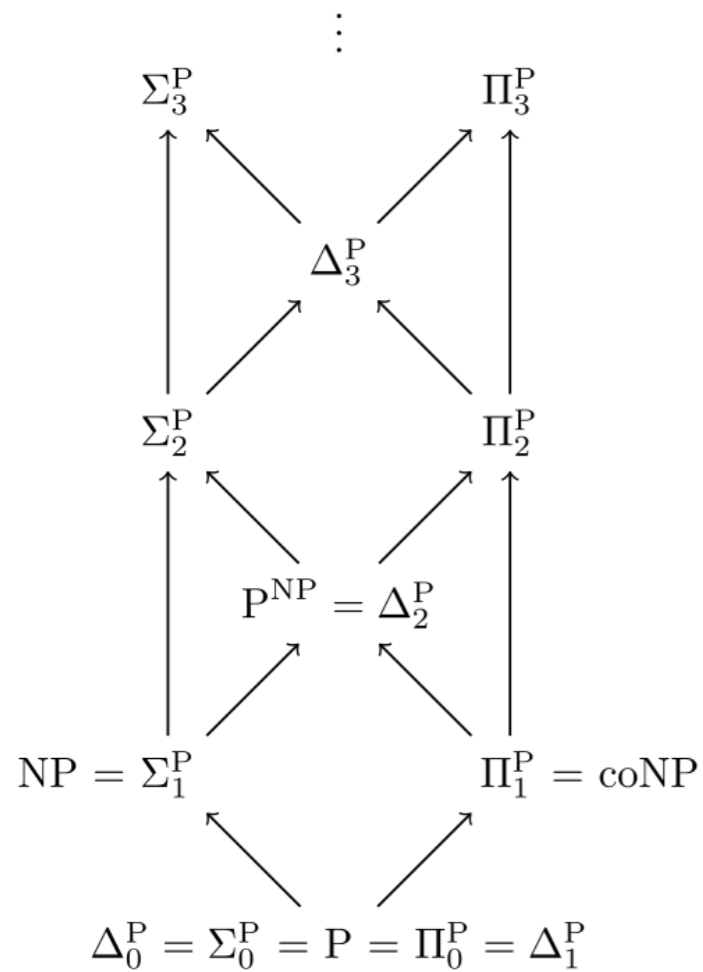
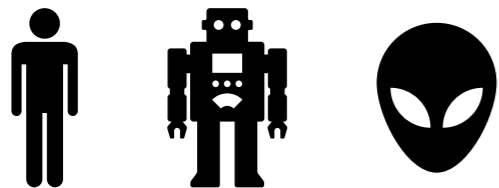


EM ?



Polynomial Hierarchy, Part I

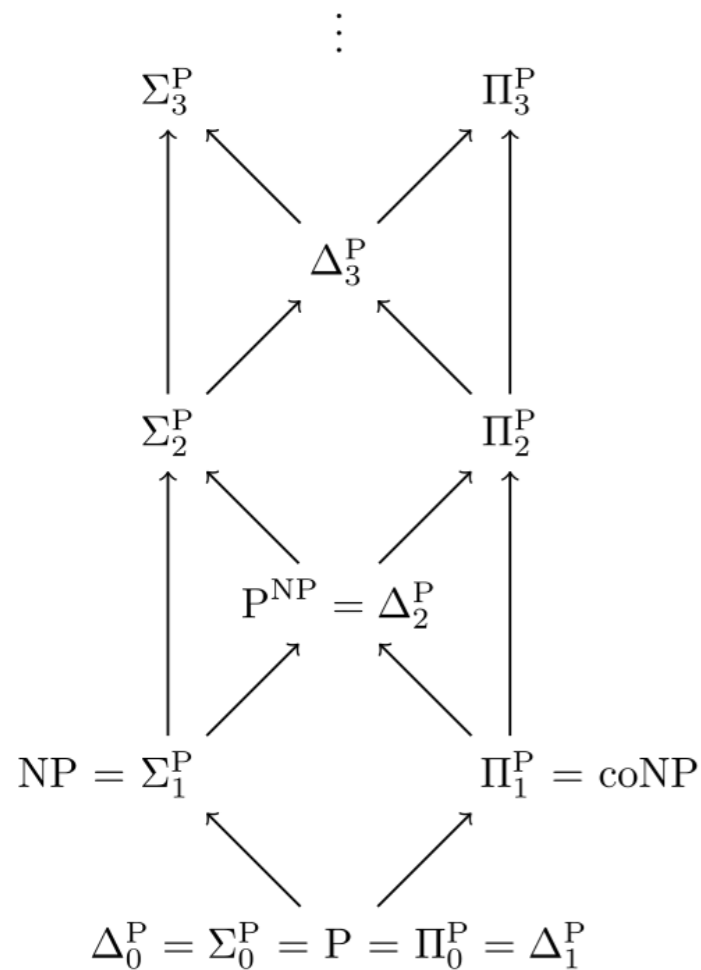
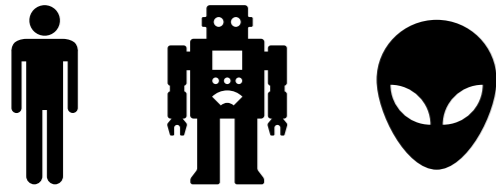
(via formal logic, directly; a start)



Polynomial Hierarchy, Part I

(via formal logic, directly; a start)

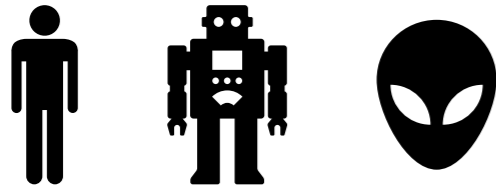
We say that a relation $R(u, y_1, \dots, y_n)$ is polytime iff there is a deterministic Turing Machine \mathfrak{m} and a polynomial p s.t. \mathfrak{m} decides this relation in $p(|u|)$.



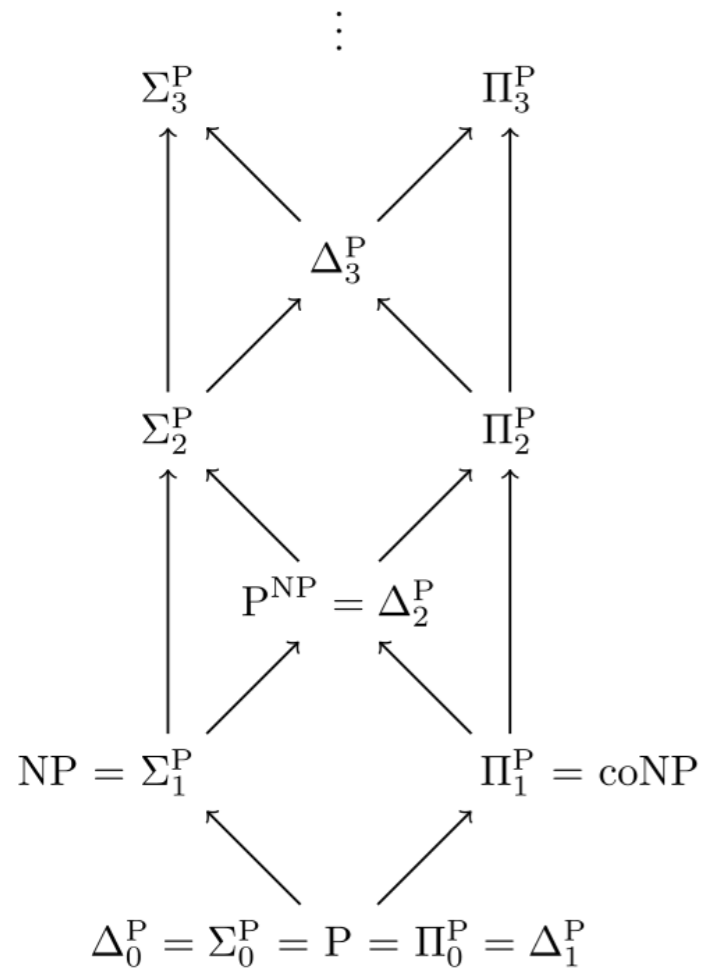
Polynomial Hierarchy, Part I

(via formal logic, directly; a start)

We say that a relation $R(u, y_1, \dots, y_n)$ is polytime iff there is a deterministic Turing Machine m and a polynomial p s.t. m decides this relation in $p(|u|)$.

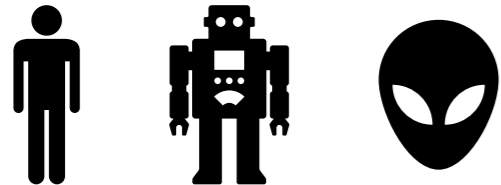


$L \in \mathbf{NP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\exists y R(u, y)$.



Polynomial Hierarchy, Part I

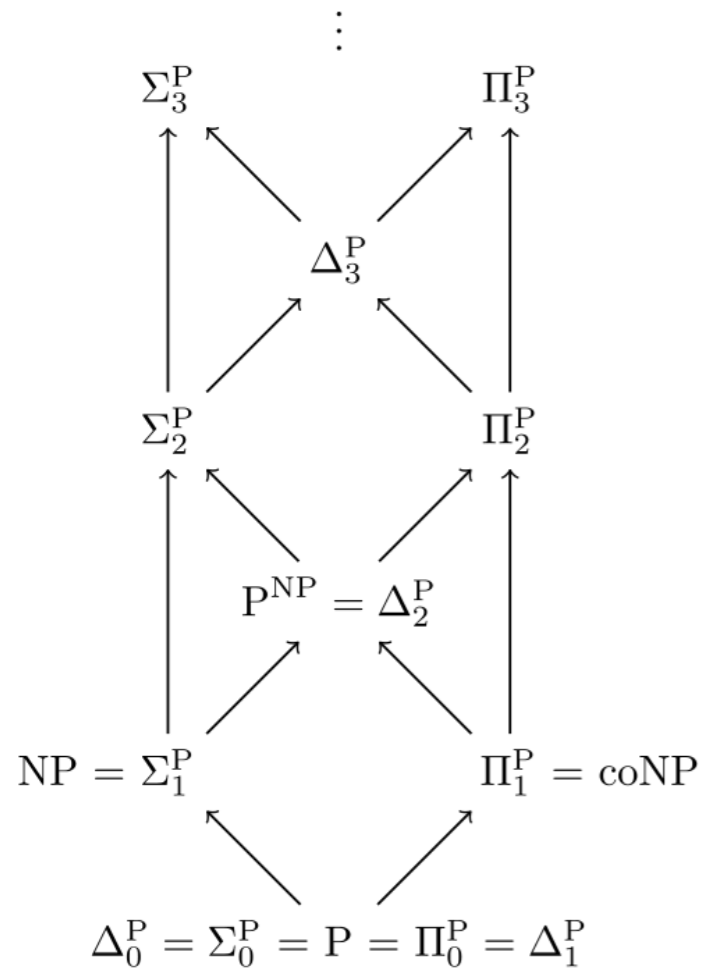
(via formal logic, directly; a start)



We say that a relation $R(u, y_1, \dots, y_n)$ is polytime iff there is a deterministic Turing Machine m and a polynomial p s.t. m decides this relation in $p(|u|)$.

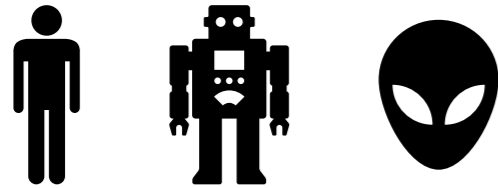
$L \in \mathbf{NP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\exists y R(u, y)$.

E.g.: We can prove $\mathbf{SAT} \in \mathbf{NP}$ because we have a polytime relation R s.t. $\phi \in \mathbf{SAT}$ iff $\exists y R(\phi \in \mathcal{L}_{pc}, \langle \text{assignments to Boolean vars} \rangle)$, where these assignments produce truth.



Polynomial Hierarchy, Part I

(via formal logic, directly; a start)

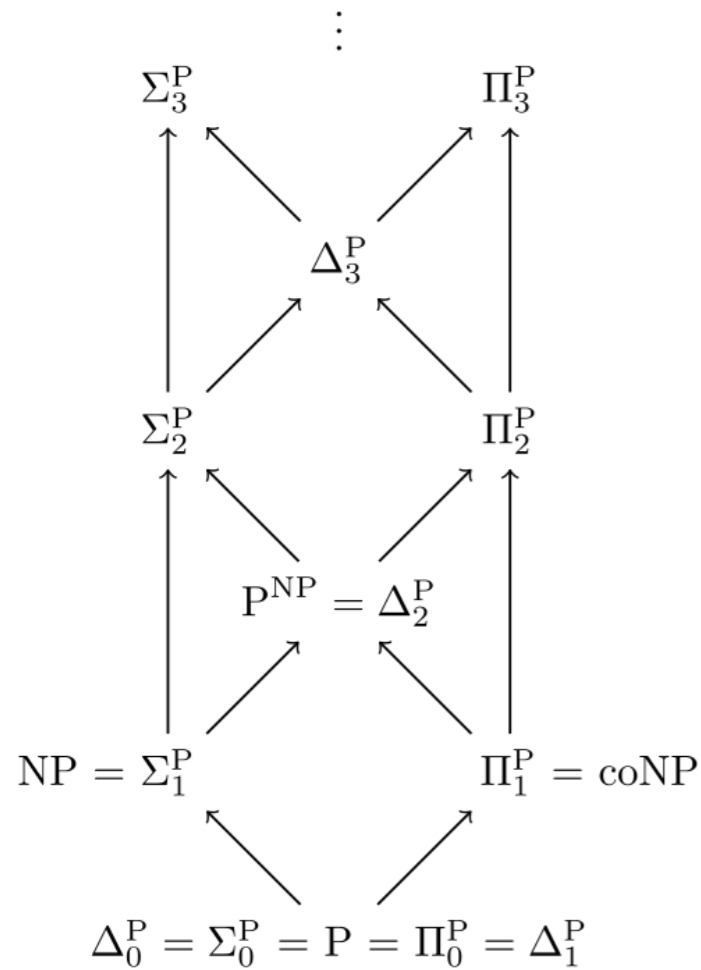


We say that a relation $R(u, y_1, \dots, y_n)$ is polytime iff there is a deterministic Turing Machine m and a polynomial p s.t. m decides this relation in $p(|u|)$.

$L \in \mathbf{NP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\exists y R(u, y)$.

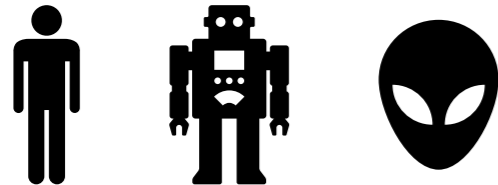
E.g.: We can prove $\mathbf{SAT} \in \mathbf{NP}$ because we have a polytime relation R s.t. $\phi \in \mathbf{SAT}$ iff $\exists y R(\phi \in \mathcal{L}_{pc}, \langle \text{assignments to Boolean vars} \rangle)$, where these assignments produce truth.

$L \in \mathbf{coNP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\forall y R(u, y)$.



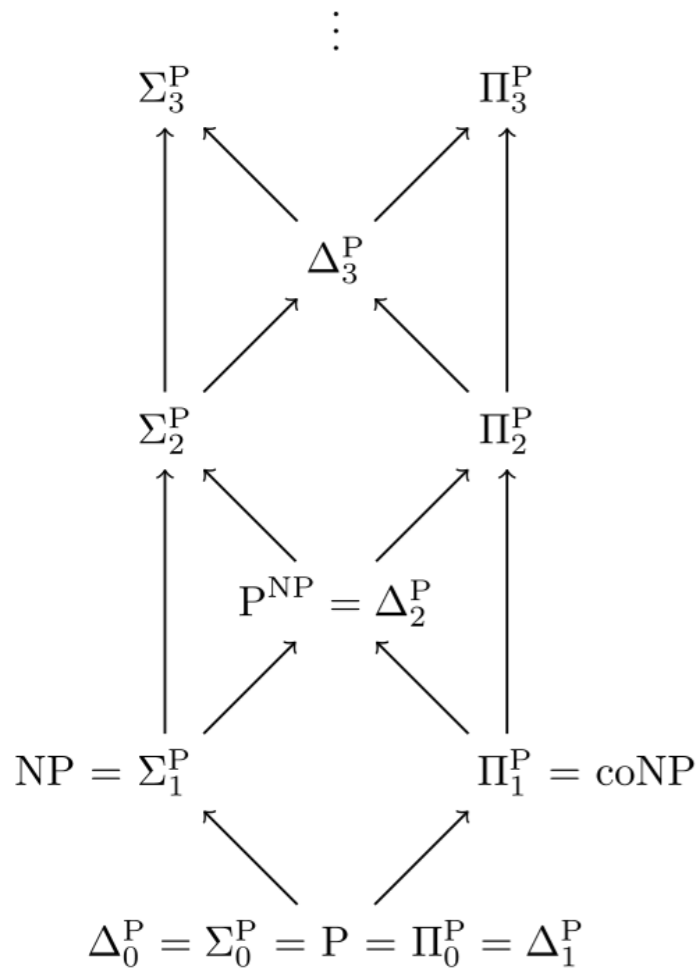
Polynomial Hierarchy, Part I

(via formal logic, directly; a start)



We say that a relation $R(u, y_1, \dots, y_n)$ is polytime iff there is a deterministic Turing Machine m and a polynomial p s.t. m decides this relation in $p(|u|)$.

$L \in \mathbf{NP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\exists y R(u, y)$.



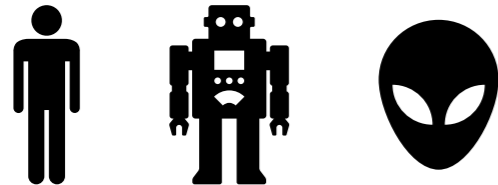
E.g.: We can prove $\mathbf{SAT} \in \mathbf{NP}$ because we have a polytime relation R s.t. $\phi \in \mathbf{SAT}$ iff $\exists y R(\phi \in \mathcal{L}_{pc}, \langle \text{assignments to Boolean vars} \rangle)$, where these assignments produce truth.

$L \in \mathbf{coNP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\forall y R(u, y)$.

To prove $\mathbf{coSAT} \in \mathbf{coNP}$, we note that we have a polytime relation R s.t. $\phi \in \mathbf{coSAT}$ iff $\forall y R(\phi \in \mathcal{L}_{pc}, \langle \text{assignments to Boolean vars} \rangle)$, where the assignments produce *falsity*.

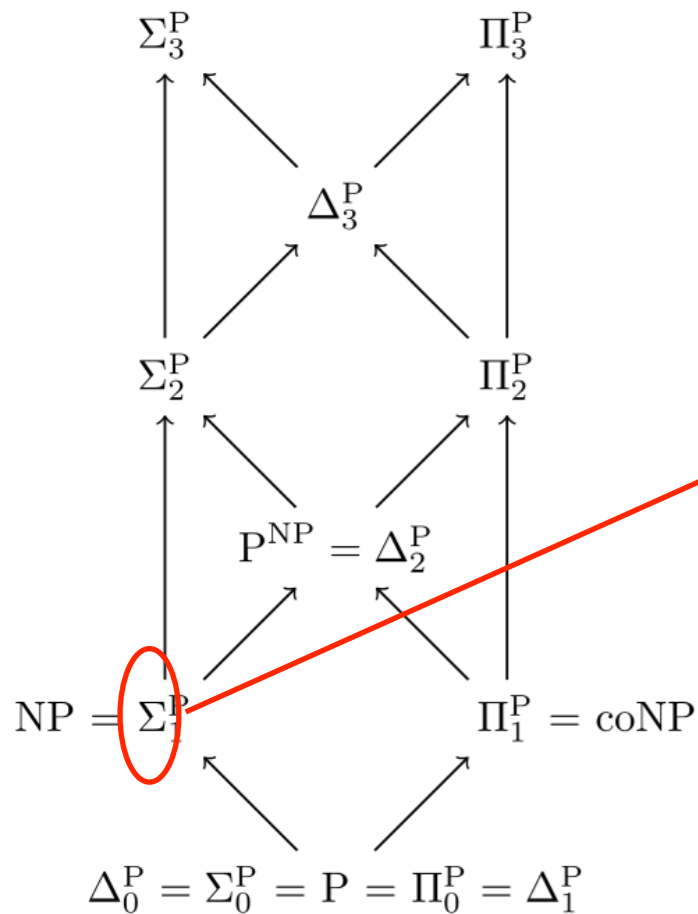
Polynomial Hierarchy, Part I

(via formal logic, directly; a start)



We say that a relation $R(u, y_1, \dots, y_n)$ is polytime iff there is a deterministic Turing Machine m and a polynomial p s.t. m decides this relation in $p(|u|)$.

$L \in \mathbf{NP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\exists yR(u, y)$.



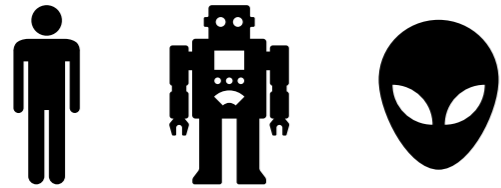
E.g.: We can prove $\mathbf{SAT} \in \mathbf{NP}$ because we have a polytime relation R s.t. $\phi \in \mathbf{SAT}$ iff $\exists yR(\phi \in \mathcal{L}_{pc}, \langle \text{assignments to Boolean vars} \rangle)$, where these assignments produce truth.

$L \in \mathbf{coNP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\forall yR(u, y)$.

To prove $\mathbf{coSAT} \in \mathbf{coNP}$, we note that we have a polytime relation R s.t. $\phi \in \mathbf{coSAT}$ iff $\forall yR(\phi \in \mathcal{L}_{pc}, \langle \text{assignments to Boolean vars} \rangle)$, where the assignments produce *falsity*.

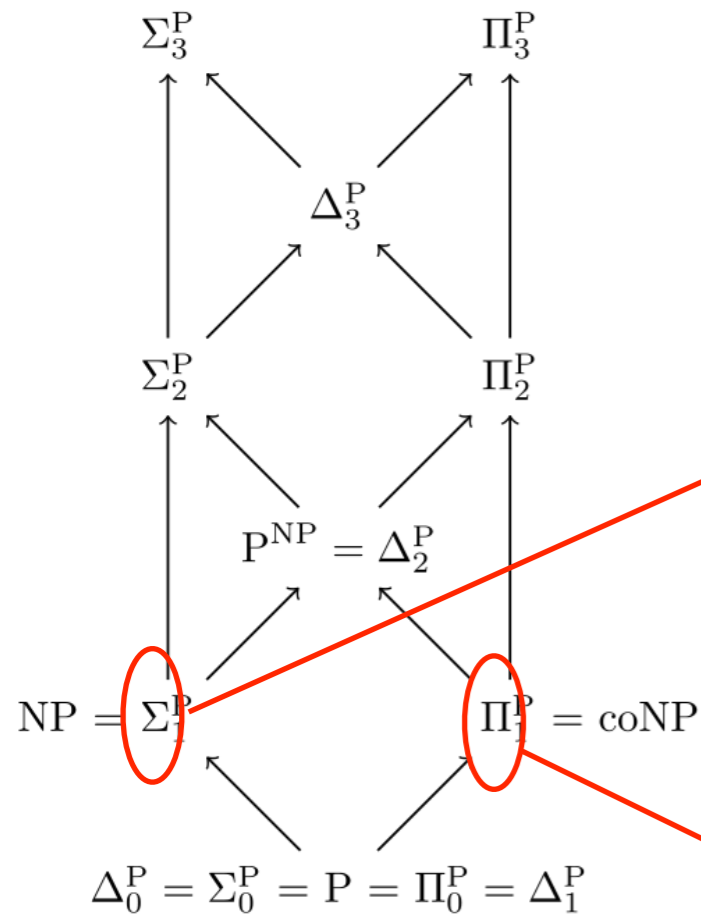
Polynomial Hierarchy, Part I

(via formal logic, directly; a start)



We say that a relation $R(u, y_1, \dots, y_n)$ is polytime iff there is a deterministic Turing Machine m and a polynomial p s.t. m decides this relation in $p(|u|)$.

$L \in \mathbf{NP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\exists y R(u, y)$.



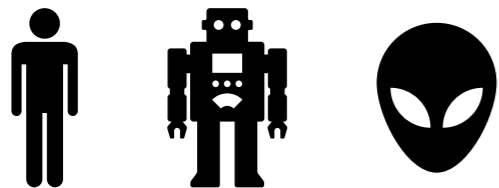
E.g.: We can prove $\mathbf{SAT} \in \mathbf{NP}$ because we have a polytime relation R s.t. $\phi \in \mathbf{SAT}$ iff $\exists y R(\phi \in \mathcal{L}_{pc}, \langle \text{assignments to Boolean vars} \rangle)$, where these assignments produce truth.

$L \in \mathbf{coNP}$ iff: there's a polytime relation R s.t. $u \in L$ iff $\forall y R(u, y)$.

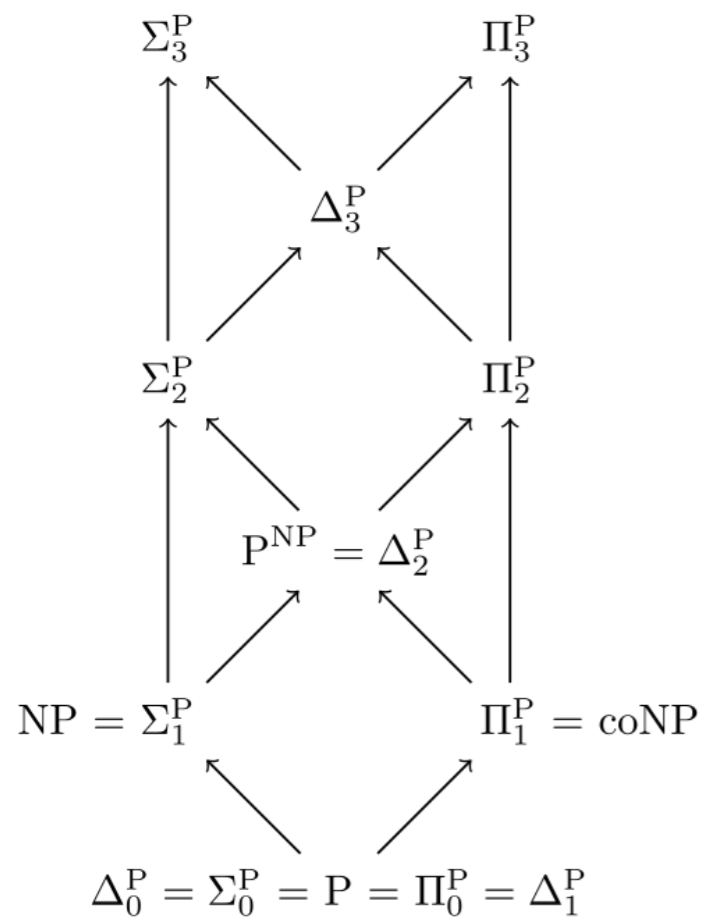
To prove $\mathbf{coSAT} \in \mathbf{coNP}$, we note that we have a polytime relation R s.t. $\phi \in \mathbf{coSAT}$ iff $\forall y R(\phi \in \mathcal{L}_{pc}, \langle \text{assignments to Boolean vars} \rangle)$, where the assignments produce *falsity*.

Polynomial Hierarchy, Part I

(via formal logic, directly; a start)

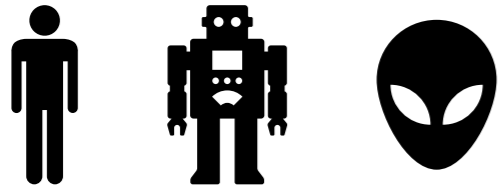


⋮

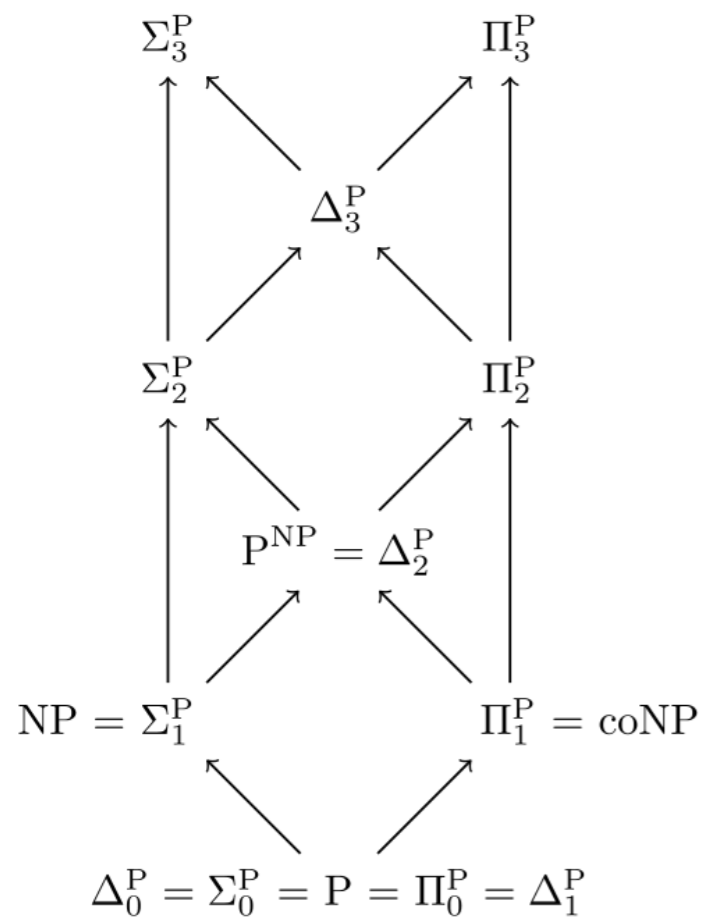


Polynomial Hierarchy, Part I

(via formal logic, directly; a start)



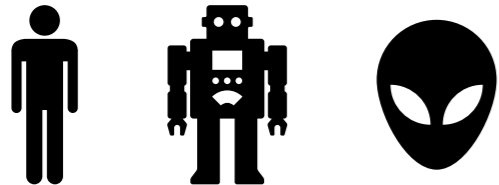
⋮



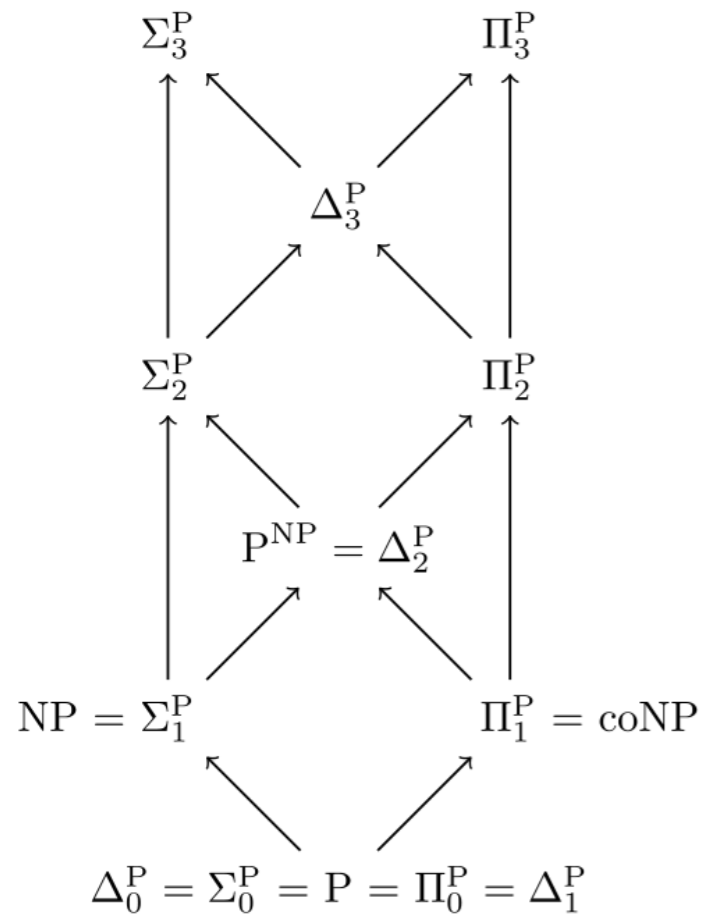
“What’s that Δ ??”

Polynomial Hierarchy, Part I

(via formal logic, directly; a start)



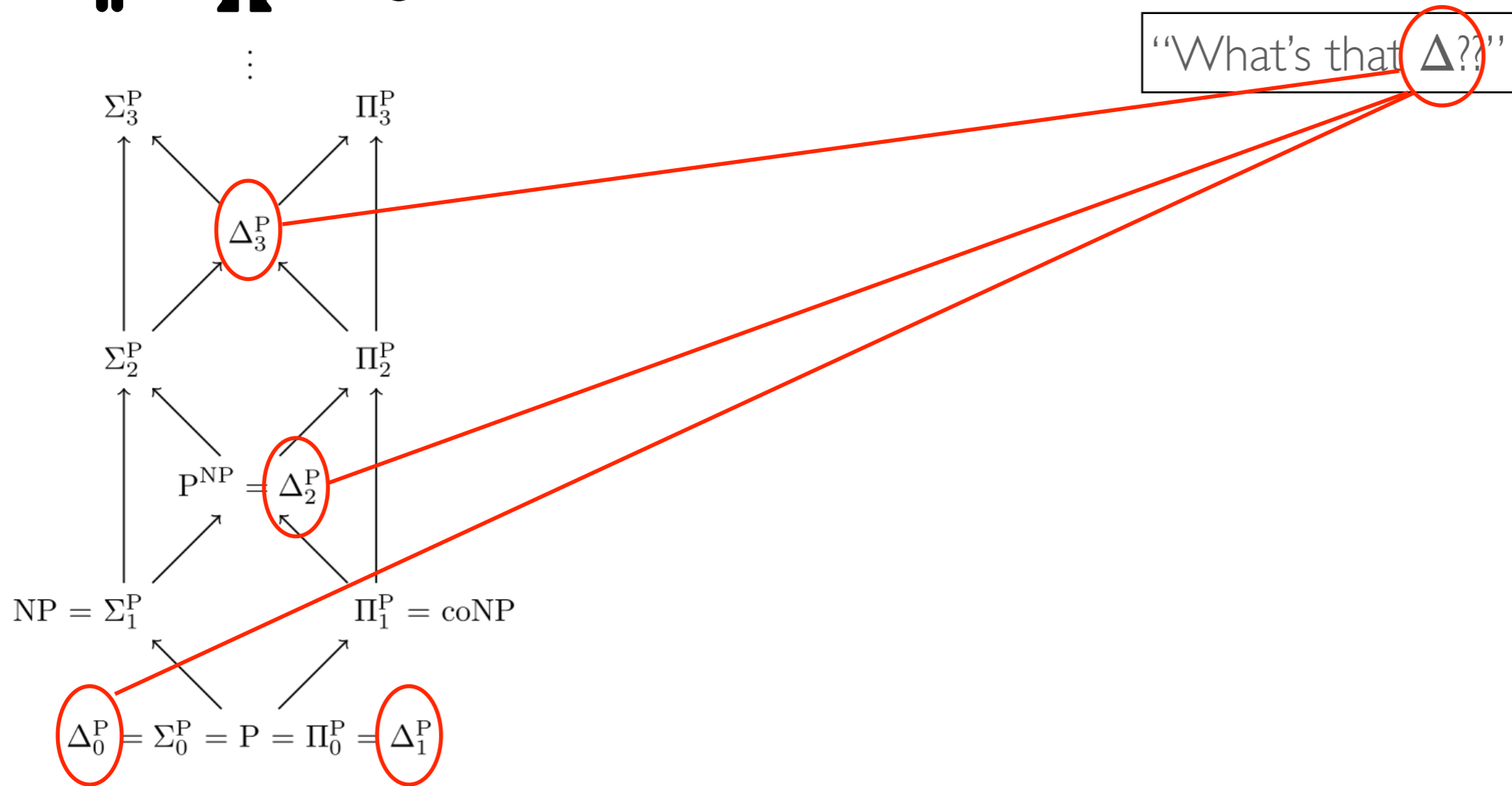
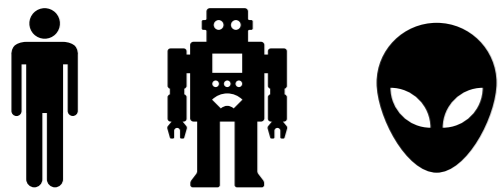
⋮



“What’s that $\Delta??$ ”

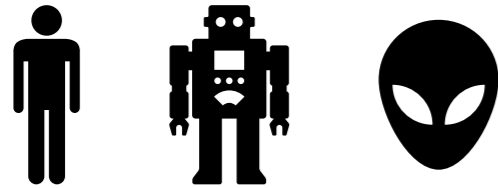
Polynomial Hierarchy, Part I

(via formal logic, directly; a start)

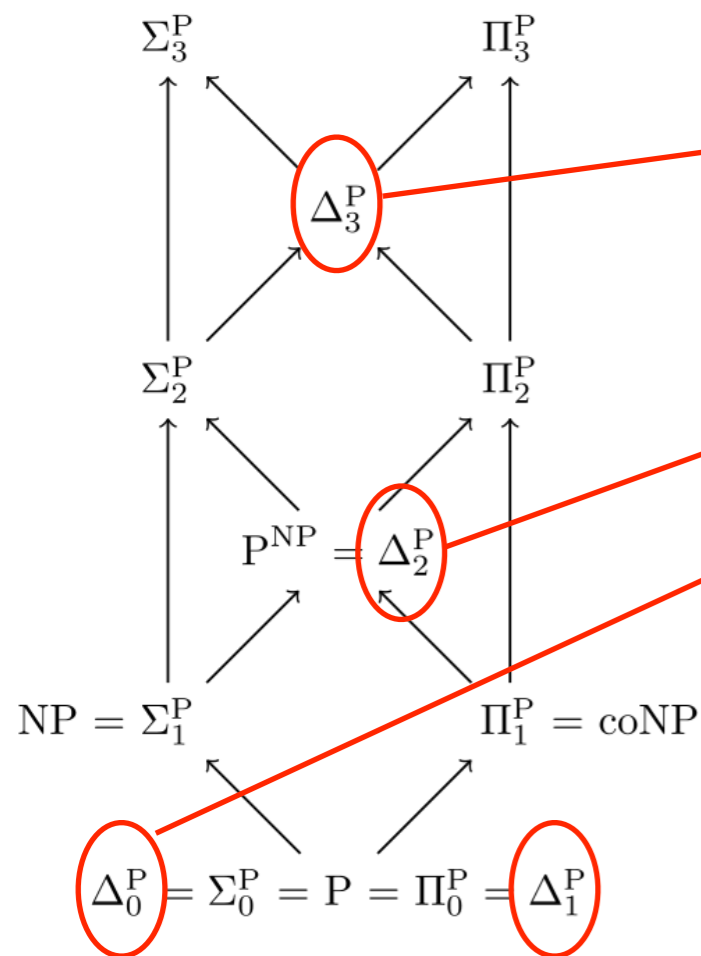


Polynomial Hierarchy, Part I

(via formal logic, directly; a start)



⋮

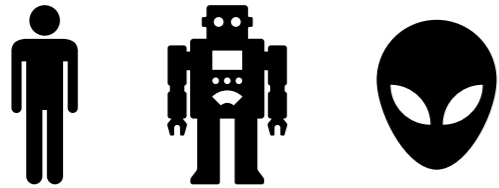


“What’s that $\Delta??$ ”

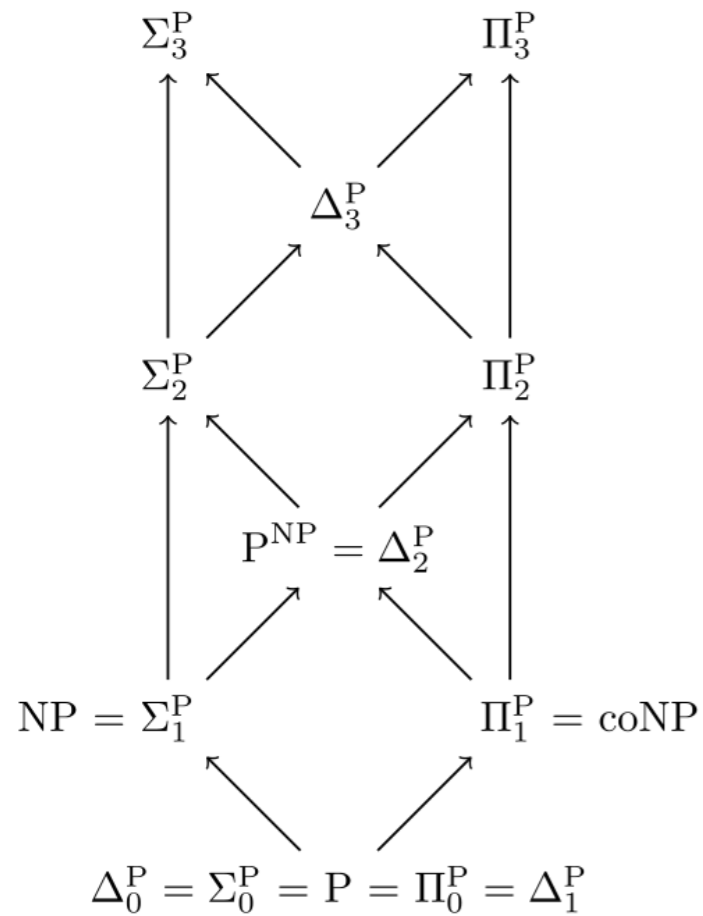
Ah. This is *not* a direct analogue to the AH. The arrows going up do indicate containment, but the purely “logician” notation based on quantifiers is apparently mixed here (dangerously). The “Delta notation” is the oracle approach to building up PH. The availability of an oracle e.g. for NP questions from P-solving machine would subsume both NP and coNP, etc.

Polynomial Hierarchy, Part II

(via formal logic, directly)

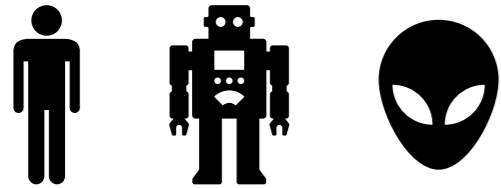


⋮

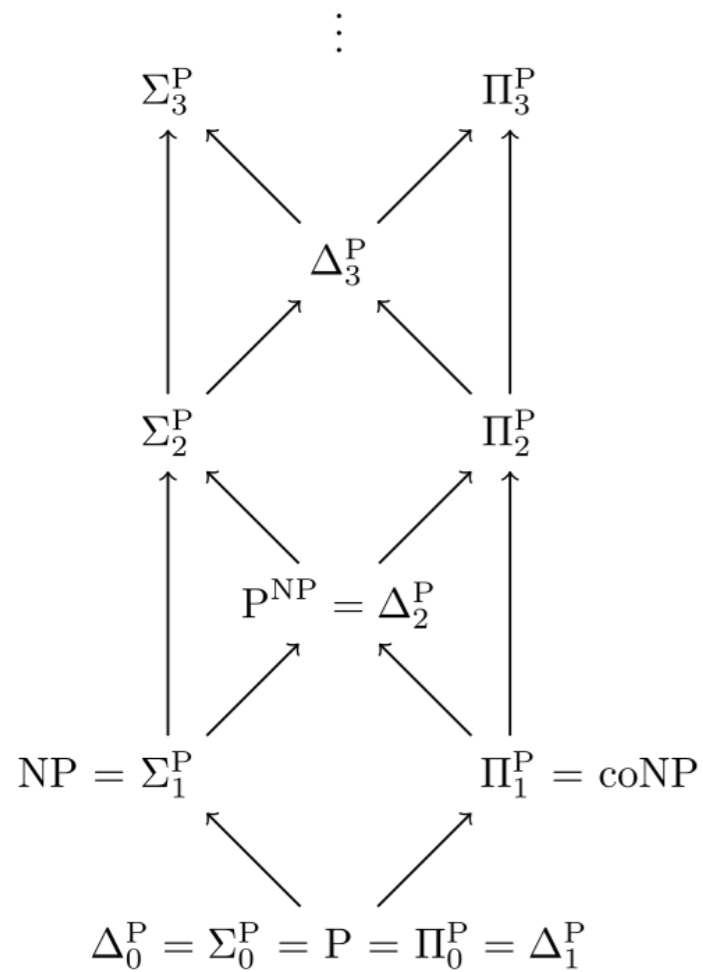


Polynomial Hierarchy, Part II

(via formal logic, directly)

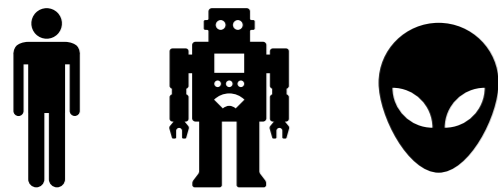


Eg:



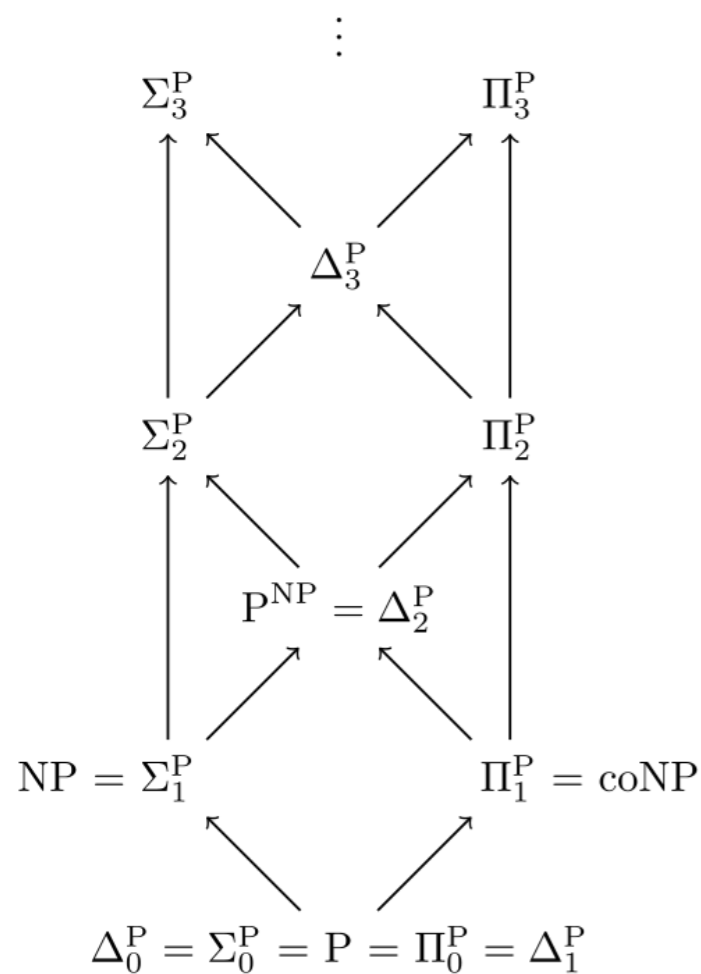
Polynomial Hierarchy, Part II

(via formal logic, directly)



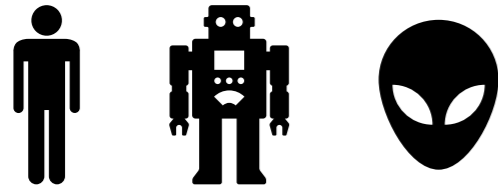
Eg:

$$\langle \phi_1, k \rangle \in L \text{ iff } \exists \phi_2 \forall \alpha KLogEquiv(\phi_1, \phi_2, |\phi_2| \leq k, \alpha(\phi_1) = \alpha(\phi_2))$$



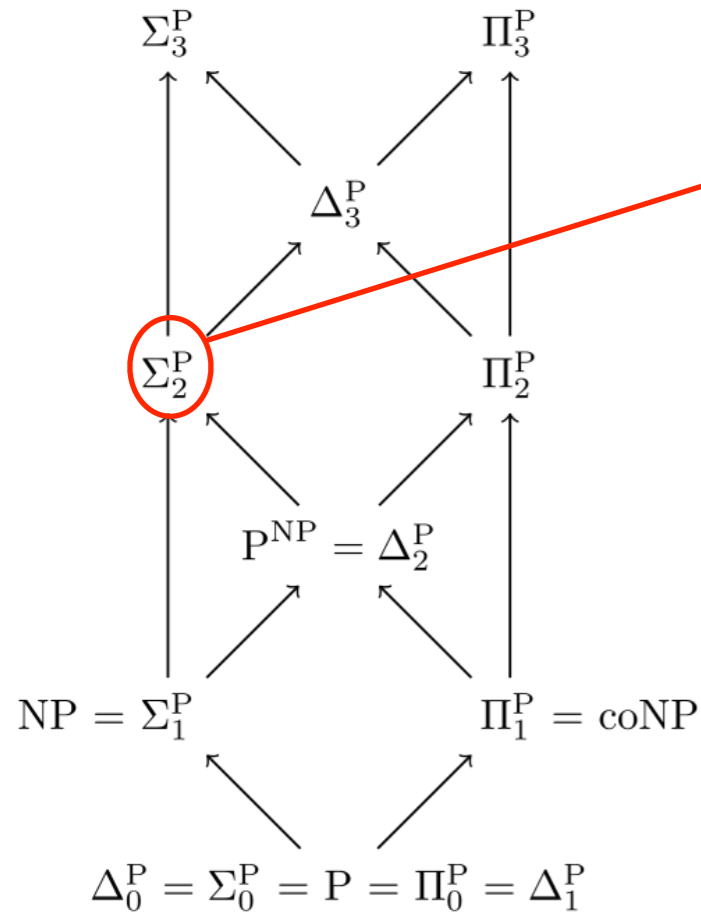
Polynomial Hierarchy, Part II

(via formal logic, directly)



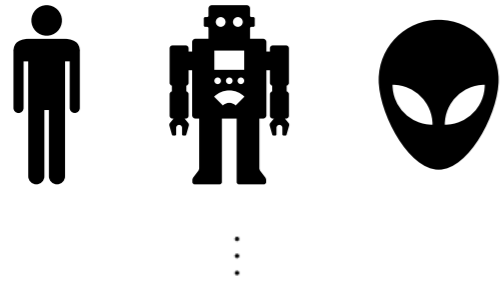
Eg:

$$\langle \phi_1, k \rangle \in L \text{ iff } \exists \phi_2 \forall \alpha KLogEquiv(\phi_1, \phi_2, |\phi_2| \leq k, \alpha(\phi_1) = \alpha(\phi_2))$$



Polynomial Hierarchy, Part II

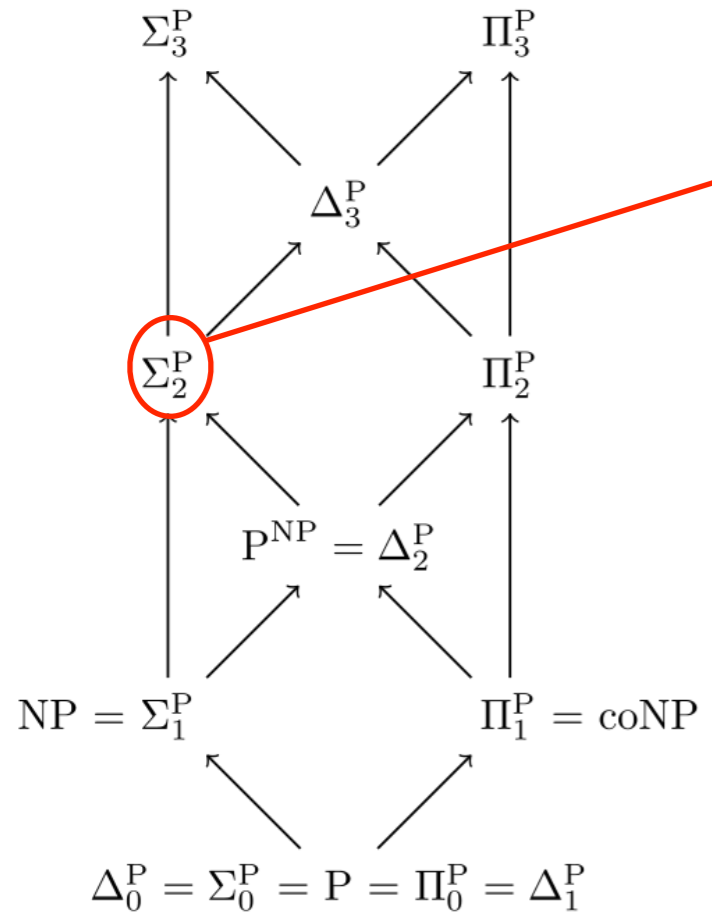
(via formal logic, directly)



free variables

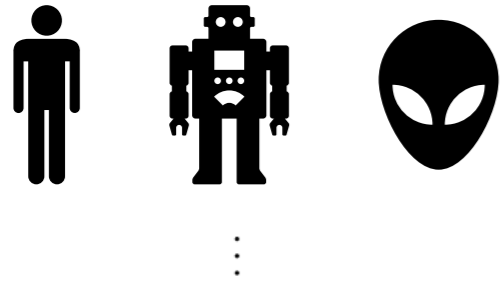
Eg:

$$\langle \phi_1, k \rangle \in L \text{ iff } \exists \phi_2 \forall \alpha KLogEquiv(\phi_1, \phi_2, |\phi_2| \leq k, \alpha(\phi_1) = \alpha(\phi_2))$$



Polynomial Hierarchy, Part II

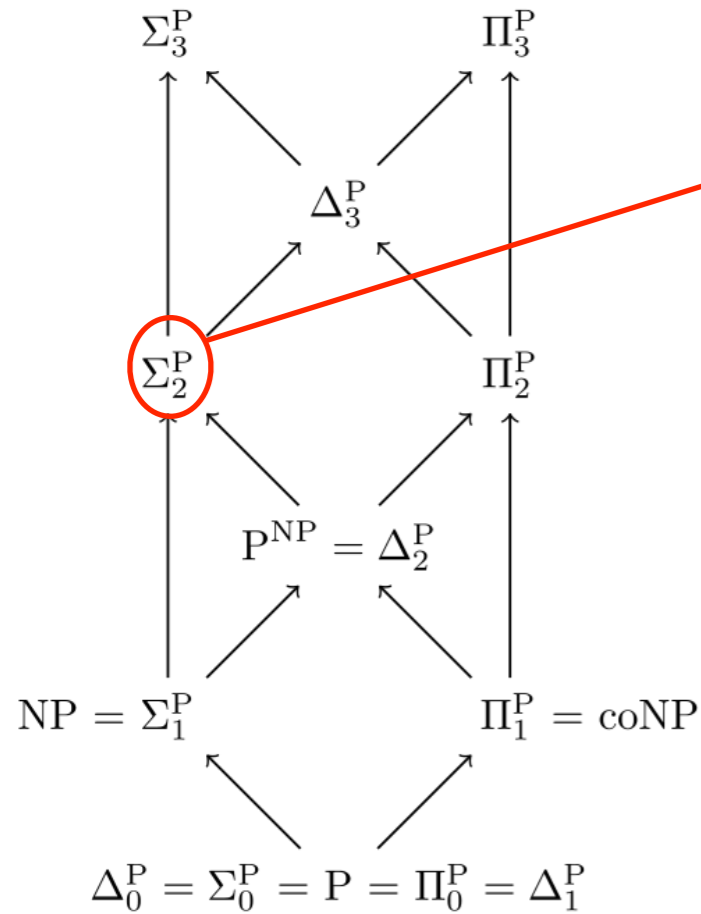
(via formal logic, directly)



free variables

Eg:

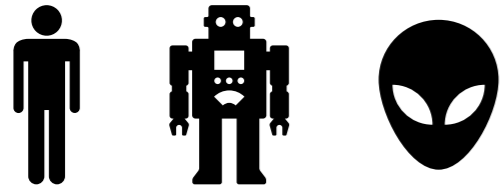
$$\langle \phi_1, k \rangle \in L \text{ iff } \exists \phi_2 \forall \alpha KLogEquiv(\phi_1, \phi_2, |\phi_2| \leq k, \alpha(\phi_1) = \alpha(\phi_2))$$



Now we generalize:

Polynomial Hierarchy, Part II

(via formal logic, directly)



free variables

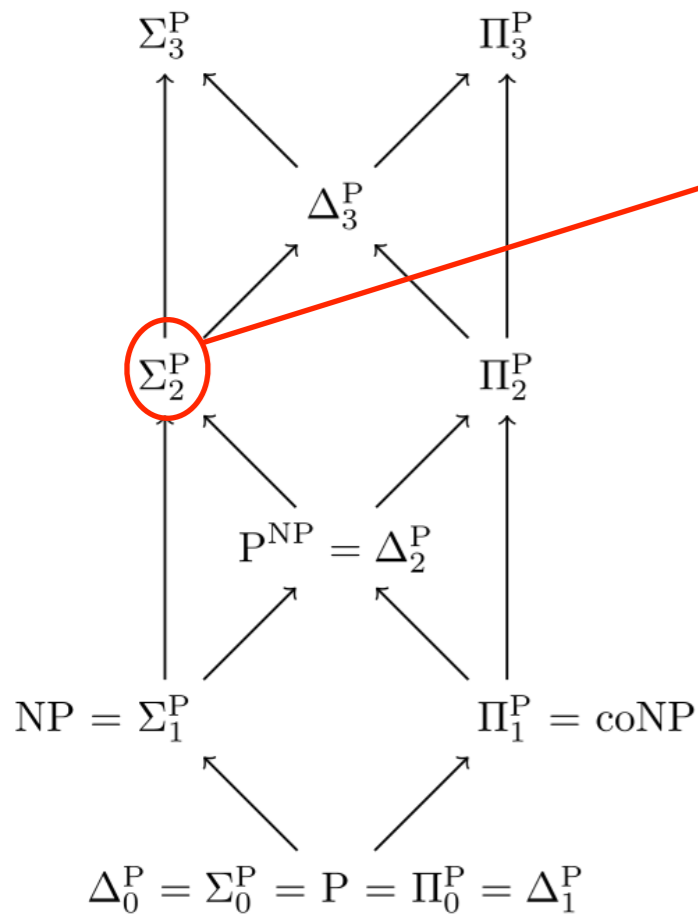
Eg:

$$\langle \phi_1, k \rangle \in L \text{ iff } \exists \phi_2 \forall \alpha KLogEquiv(\phi_1, \phi_2, |\phi_2| \leq k, \alpha(\phi_1) = \alpha(\phi_2))$$

Now we generalize:

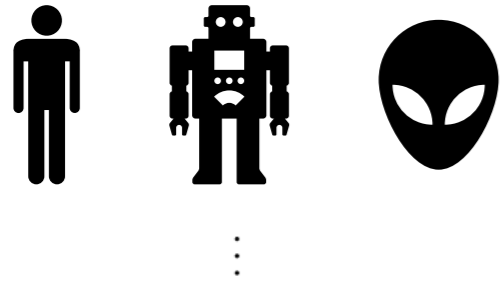
$$x \in \Sigma_i \text{ iff } \exists R \exists y_1 \forall y_2 \cdots Q_i y_i R(x, y_1, y_2, \dots, y_i)$$

($Q_i = \forall$ if i even; $Q_i = \exists$ if i odd)



Polynomial Hierarchy, Part II

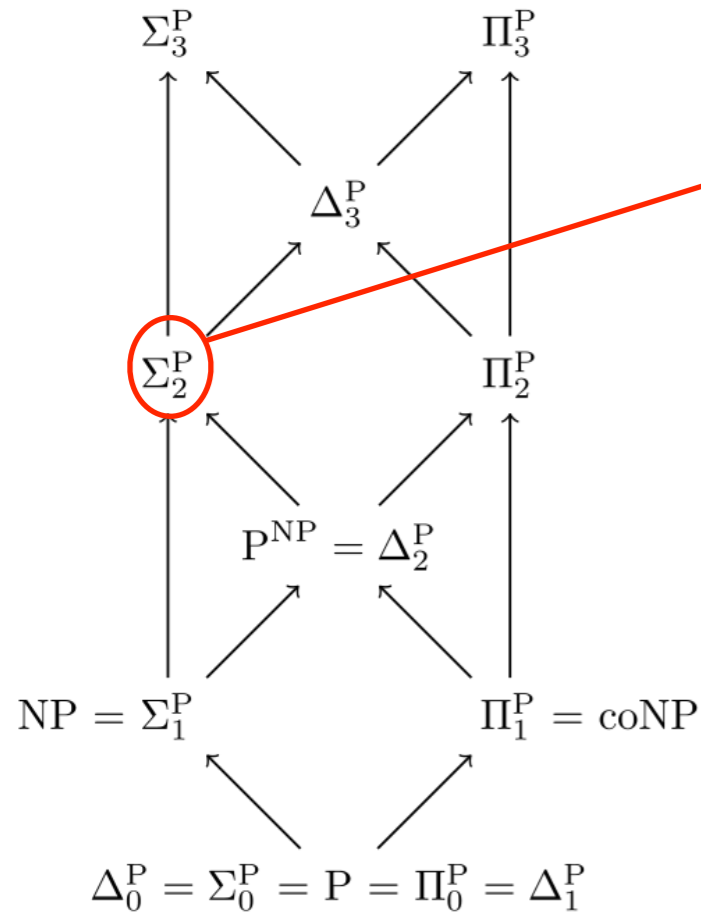
(via formal logic, directly)



free variables

Eg:

$\langle \phi_1, k \rangle \in L \text{ iff } \exists \phi_2 \forall \alpha KLogEquiv(\phi_1, \phi_2, |\phi_2| \leq k, \alpha(\phi_1) = \alpha(\phi_2))$



Now we generalize:

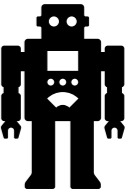
$$x \in \Sigma_i \text{ iff } \exists R \exists y_1 \forall y_2 \cdots Q_i y_i R(x, y_1, y_2, \dots, y_i)$$

($Q_i = \forall$ if i even; $Q_i = \exists$ if i odd)

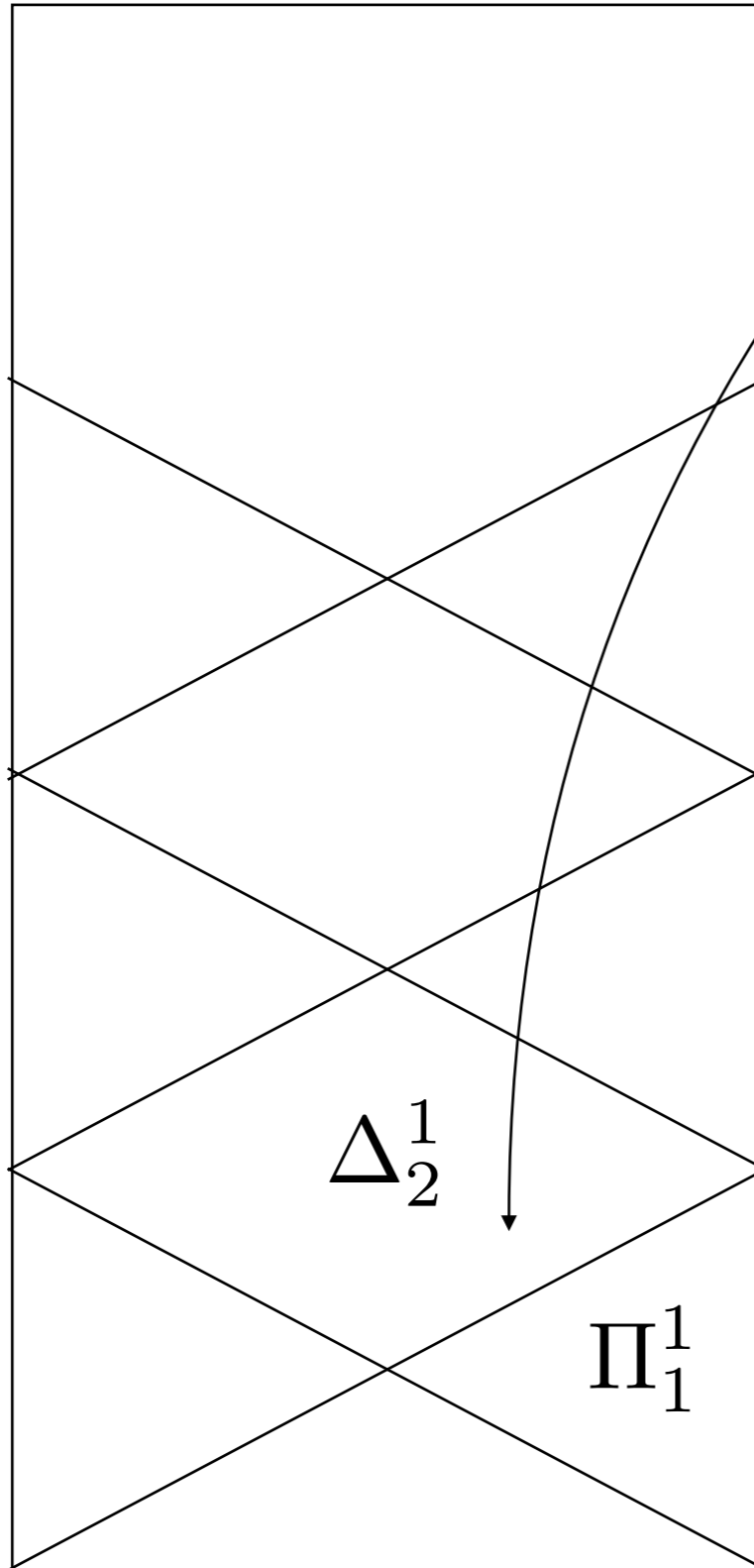
$$x \in \Pi_i \text{ iff } \exists R \forall y_1 \exists y_2 \cdots Q_i y_i R(x, y_1, y_2, \dots, y_i)$$

($Q_i = \exists$ if j even; $Q_i = \forall$ if j odd)

CogSci and AI need to say more about where AI falls/can fall in the landscape.

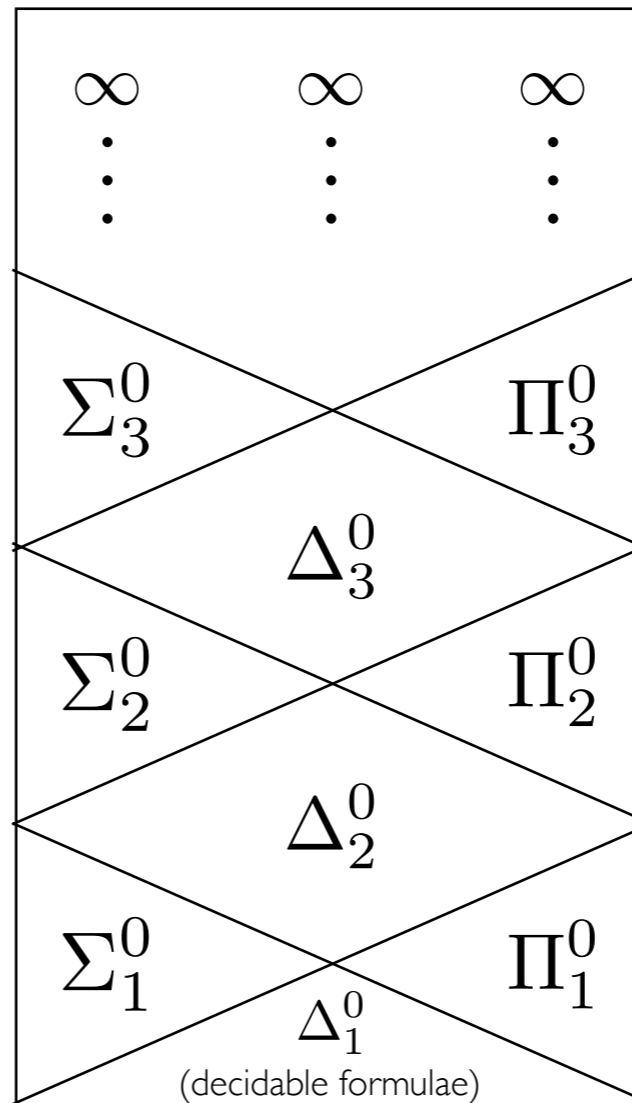


$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)

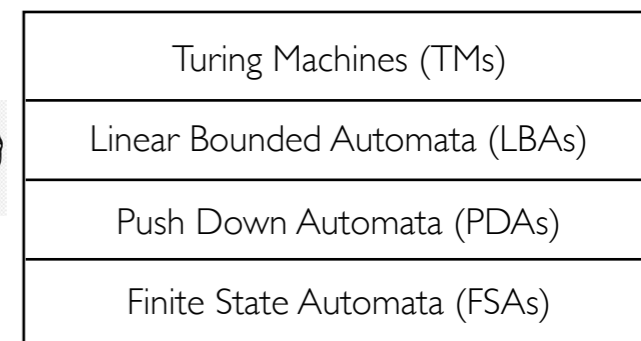


Human Persons (according to Bringsjord)

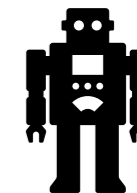
Human Brains (according to Granger)



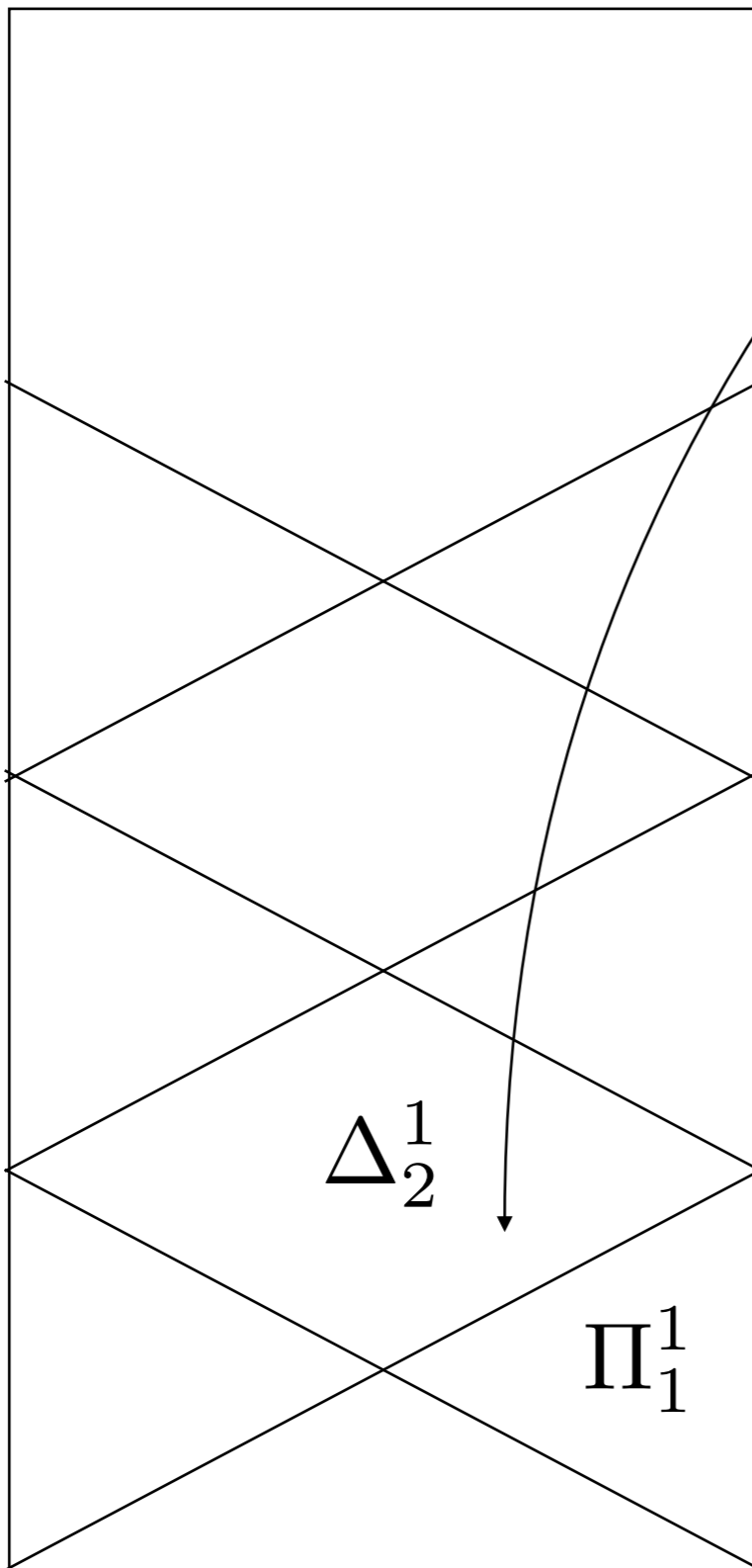
\mathcal{CH} (Chomsky Hierarchy)



CogSci and AI need to say more about where AI falls/can fall in the landscape.



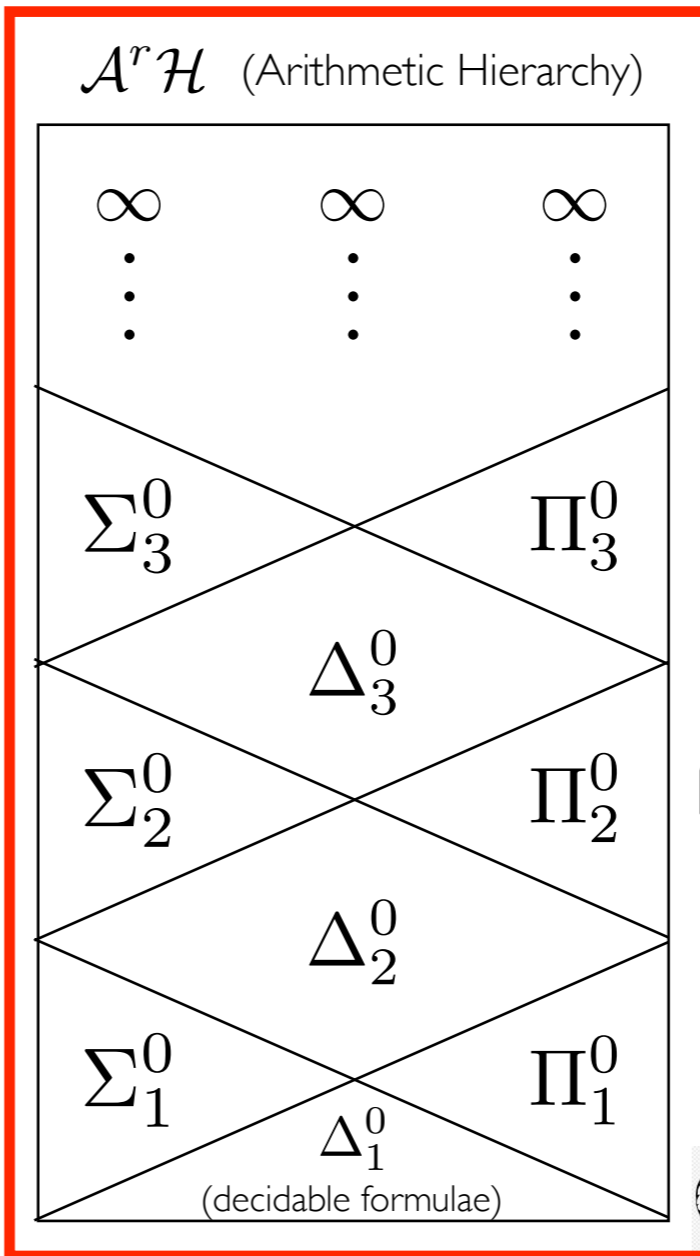
$A^n \mathcal{H}$ (Analytic Hierarchy)



Infinite Time Turing Machines (ITTMs)

Human Persons (according to Bringsjord)

$A^r \mathcal{H}$ (Arithmetic Hierarchy)



Human Brains (according to Granger)

\mathcal{CH} (Chomsky Hierarchy)

- Turing Machines (TMs)
- Linear Bounded Automata (LBAs)
- Push Down Automata (PDAs)
- Finite State Automata (FSAs)

\mathcal{EM}

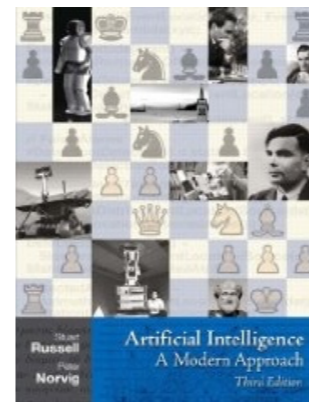
The PAID Problem ...

“Earth, we have a problem.”

“Earth, we have a problem.”



“Earth, we have a problem.”



These are amorphous worries
separated from the formal. The
Problem is easy to state, precisely:

These are amorphous worries
separated from the formal. The
Problem is easy to state, precisely:

The PAID Problem

These are amorphous worries
separated from the formal. The
Problem is easy to state, precisely:

The PAID Problem

The PAID Problem (Level I):

The PAID Problem (Level I):

$\forall x : \text{Agents}$

The PAID Problem (Level I):

$\forall x : \text{Agents}$

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

The PAID Problem (Level I):

$\forall x : \text{Agents}$

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



The PAID Problem (Level I):

$\forall x : \text{Agents}$

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



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

The PAID Problem (Level I):

$\forall x : \text{Agents}$

Powerful(x) + Autonomous(x) + Intelligent(x) \Rightarrow Dangerous(x)



...

...

...

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

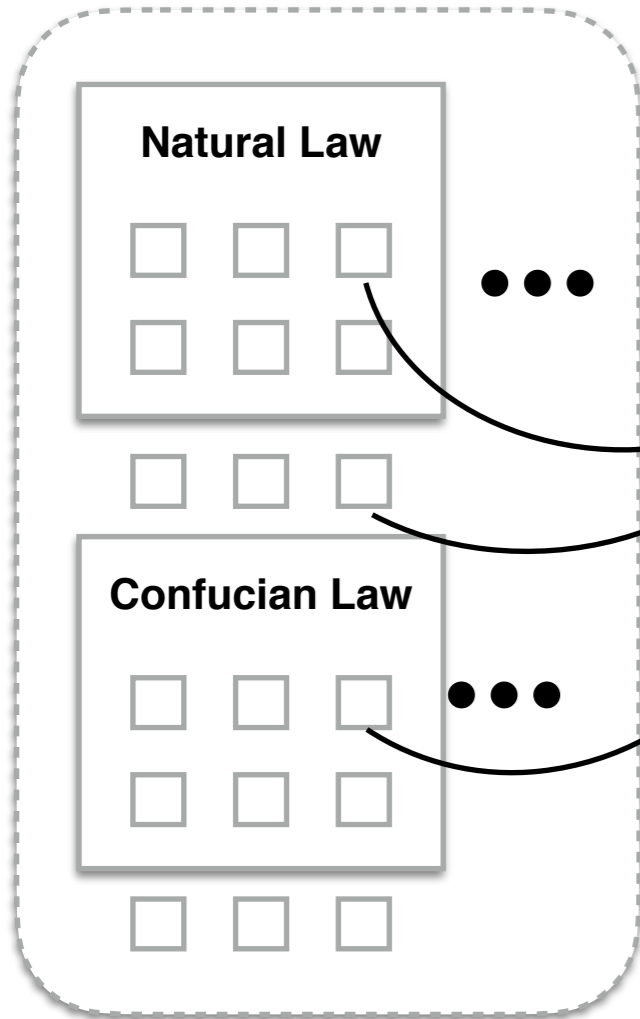
The Four Steps ...



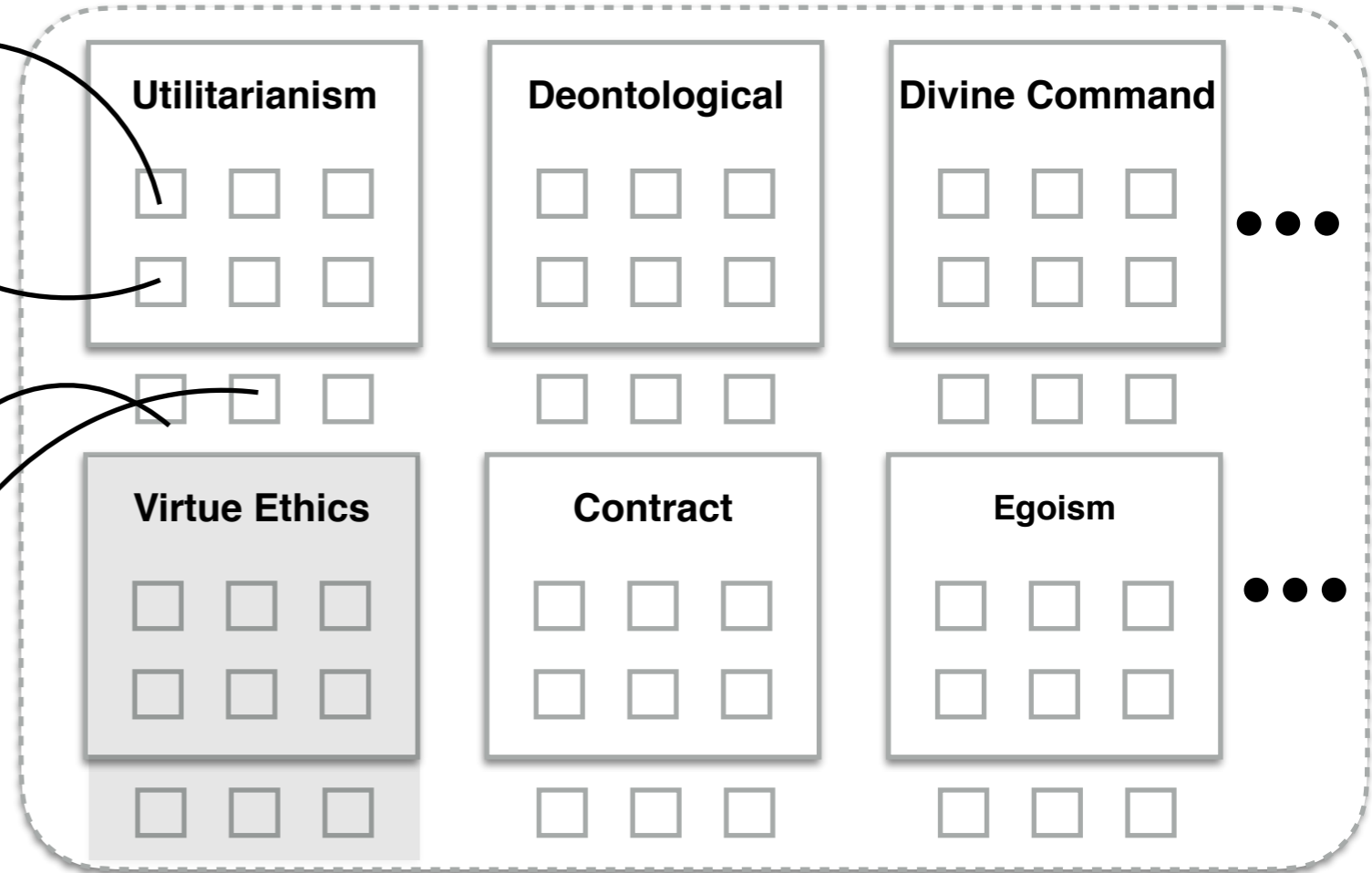
Making Morally X Machines; Only Logic Can Save Us



Theories of Law



Ethical Theories



Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

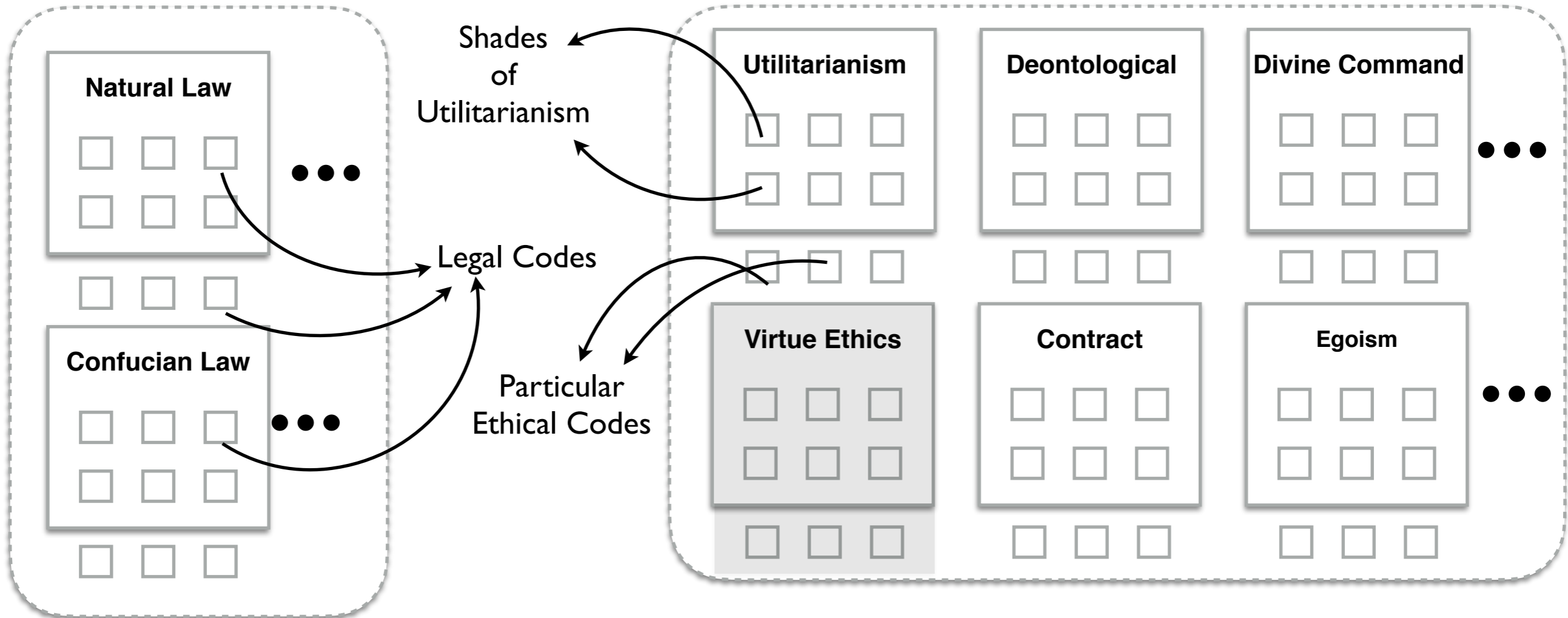


Making Morally X Machines; Only Logic Can Save Us



Theories of Law

Ethical Theories



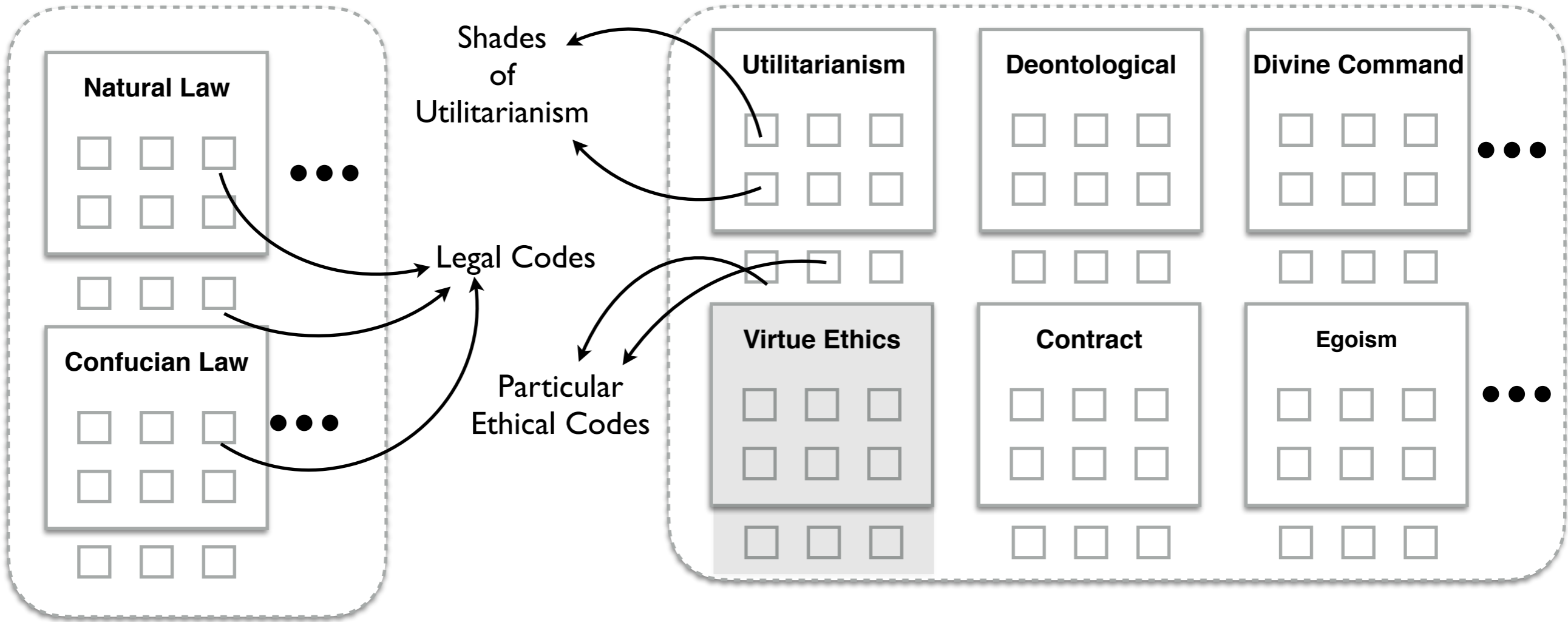


Making Morally X Machines; Only Logic Can Save Us



Theories of Law

Ethical Theories



Step I

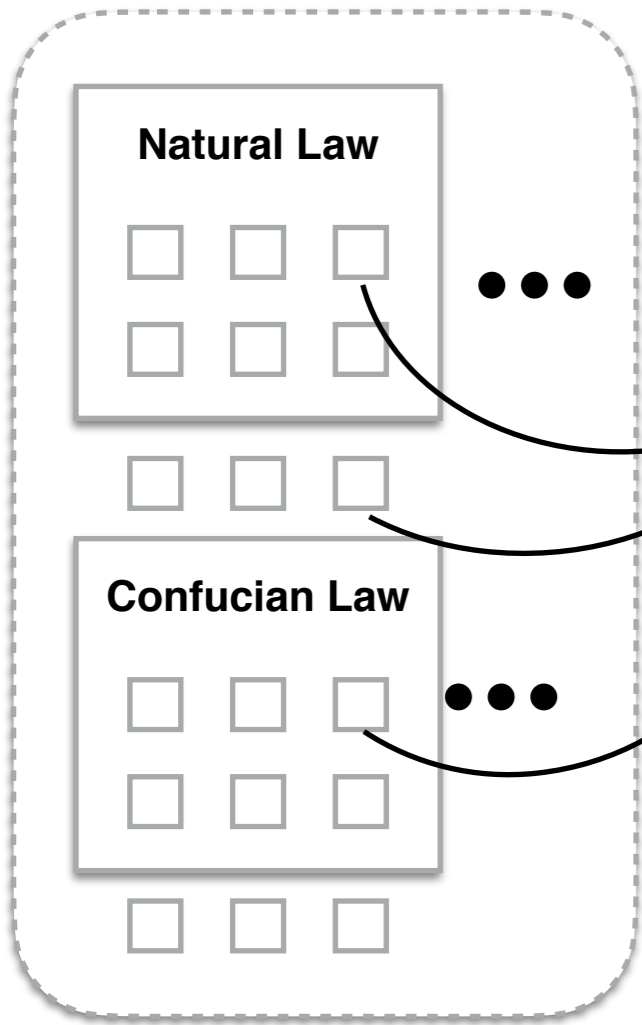
1. Pick a theory
2. Pick a code
3. Run through EH.
4. Formalize in a Cognitive Calculus.



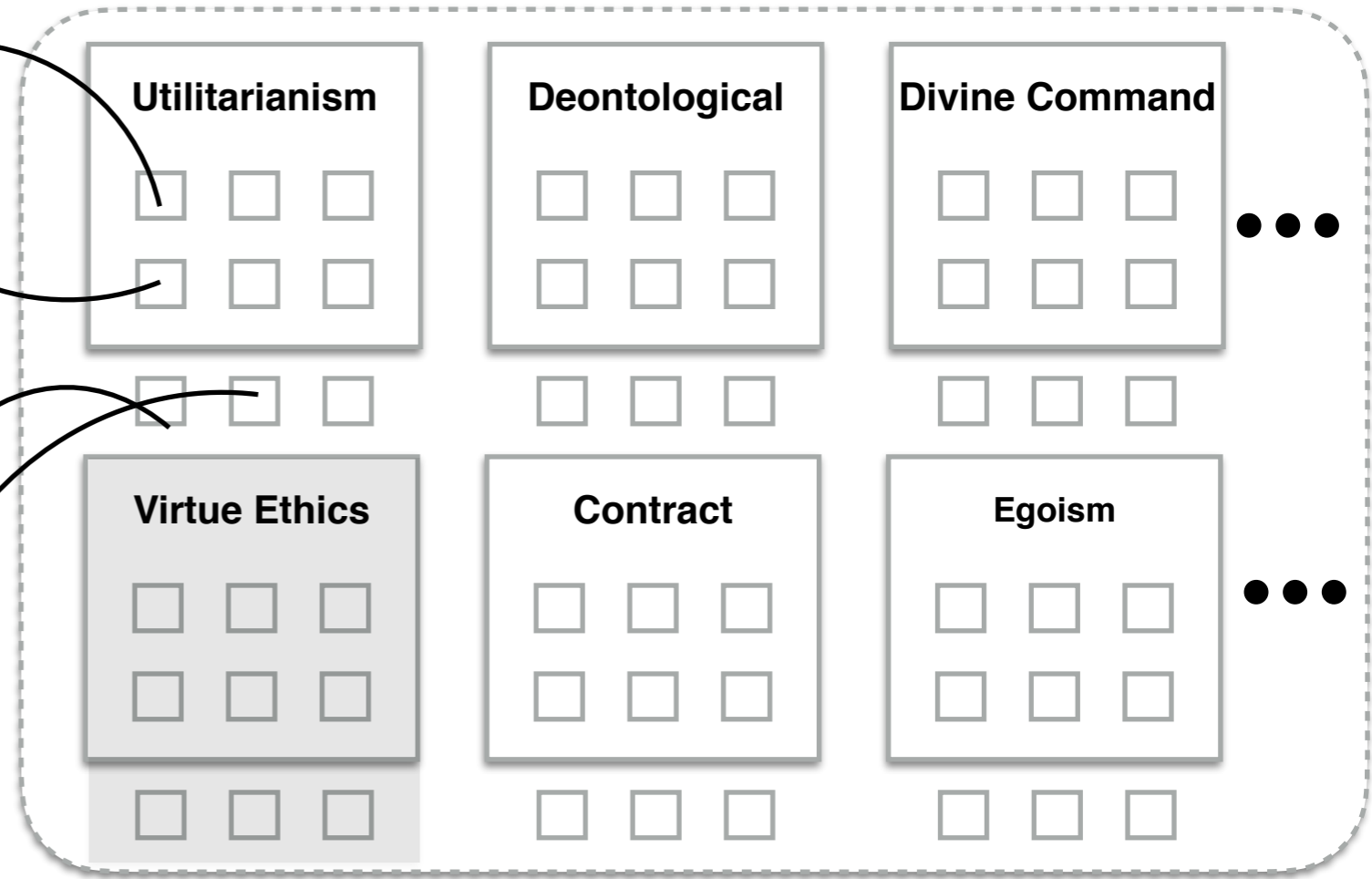
Making Morally X Machines; Only Logic Can Save Us



Theories of Law



Ethical Theories

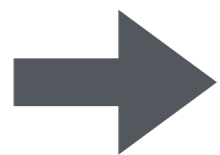


Shades of Utilitarianism

Legal Codes

Particular Ethical Codes

Step I



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

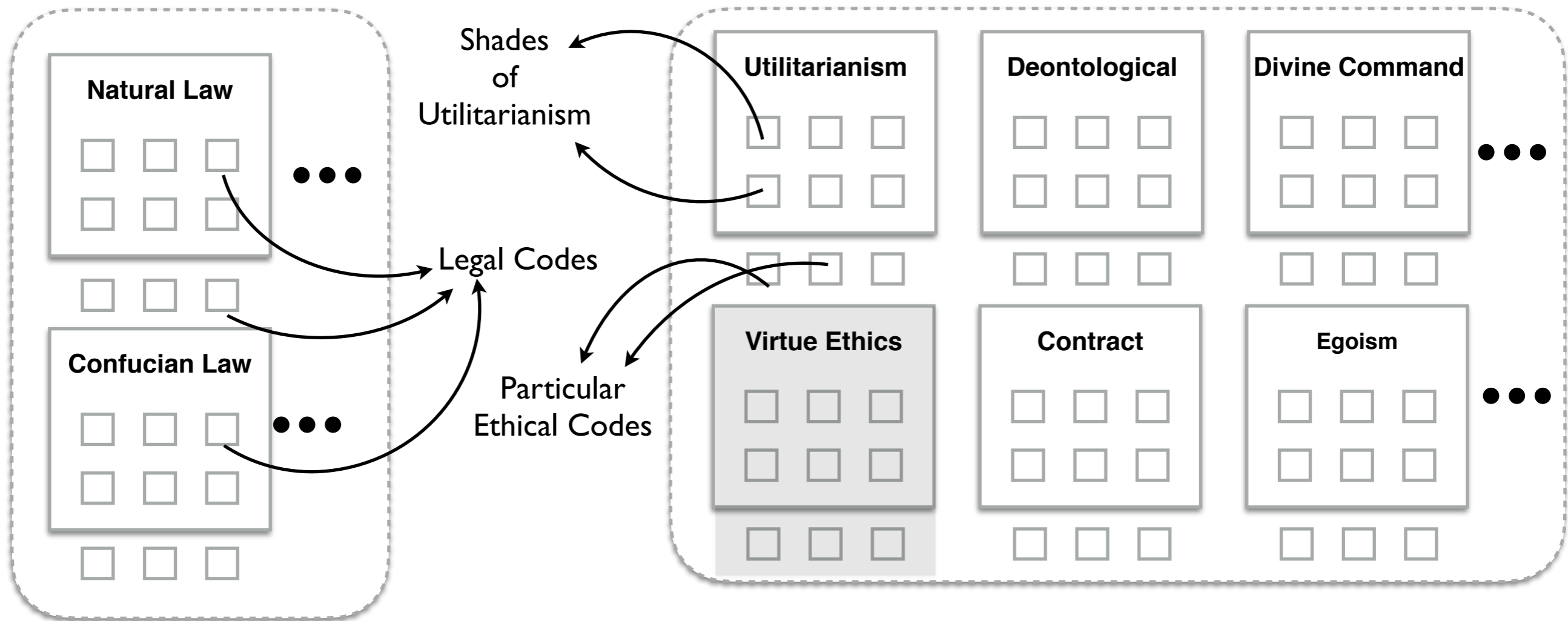


Making Morally X Machines; Only Logic Can Save Us



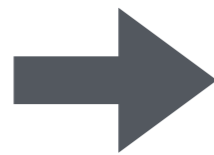
Theories of Law

Ethical Theories




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

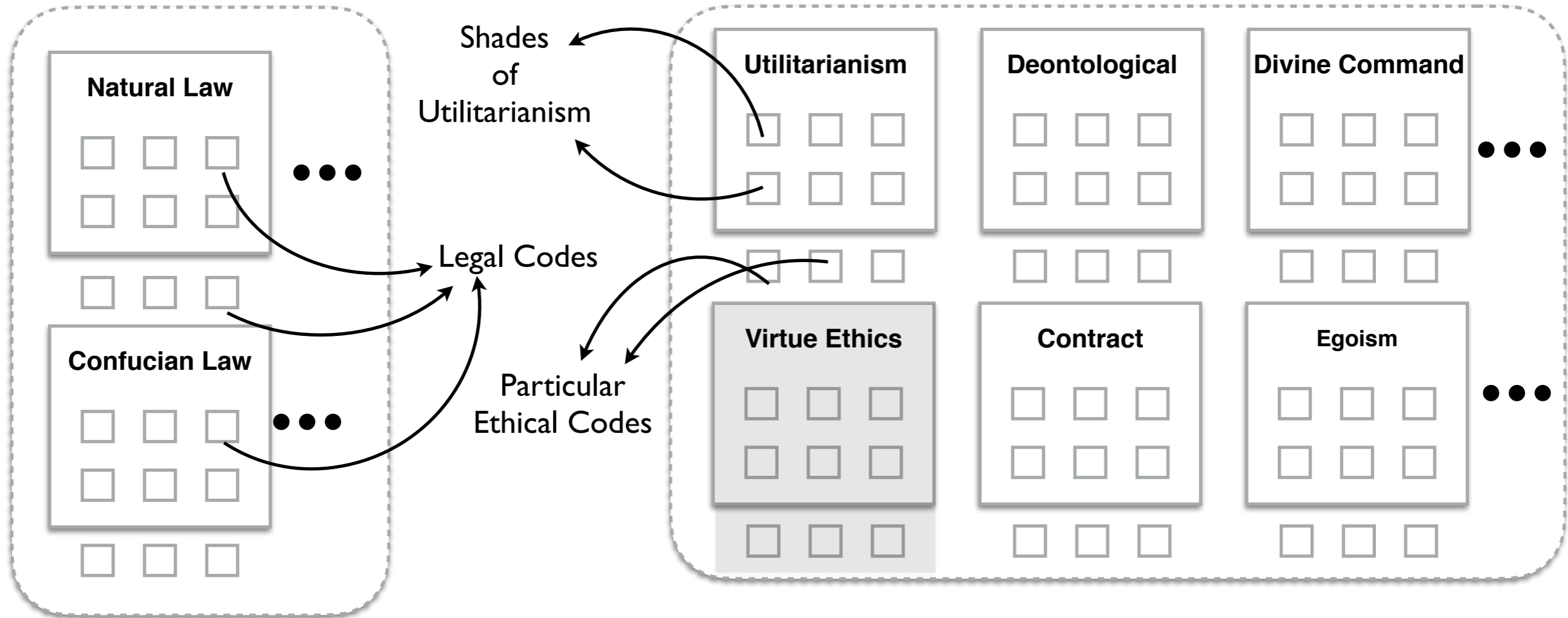


Making Morally X Machines; Only Logic Can Save Us



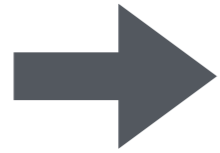
Theories of Law

Ethical Theories



Step 1

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

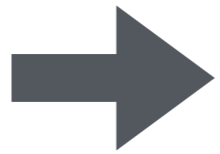


Step 2

Automate

Reasoners

Spectra

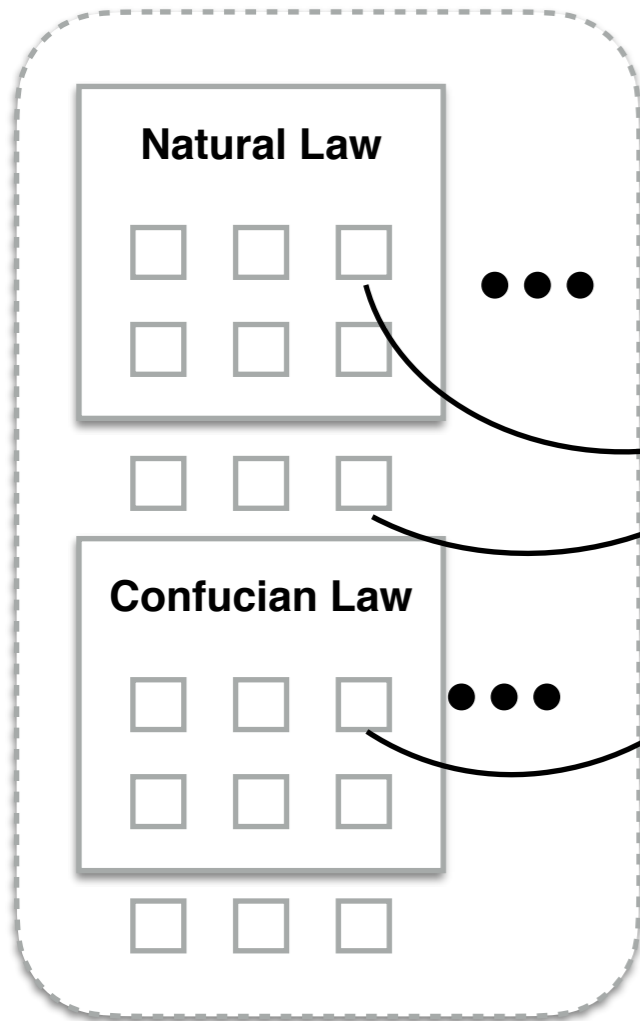




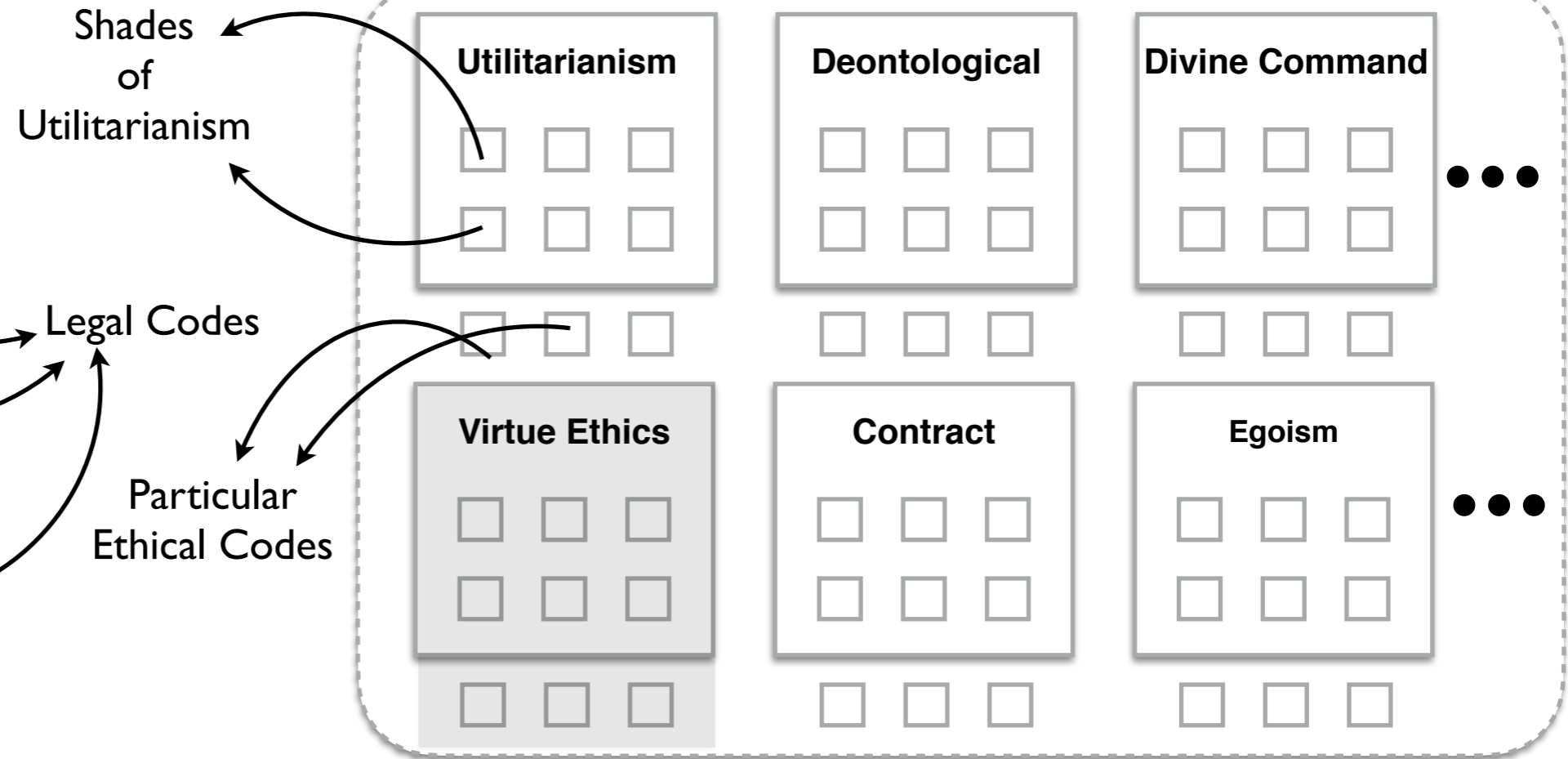
Making Morally X Machines; Only Logic Can Save Us



Theories of Law



Ethical Theories

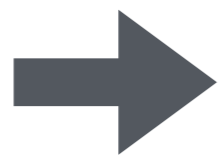


Legal Codes

Particular Ethical Codes

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

The diagram shows a stack of three layers: a blue 'Robotic Substrate' at the bottom, a green 'Ethical Substrate' in the middle, and a light blue layer at the top containing four colored squares (yellow, yellow, red, yellow).

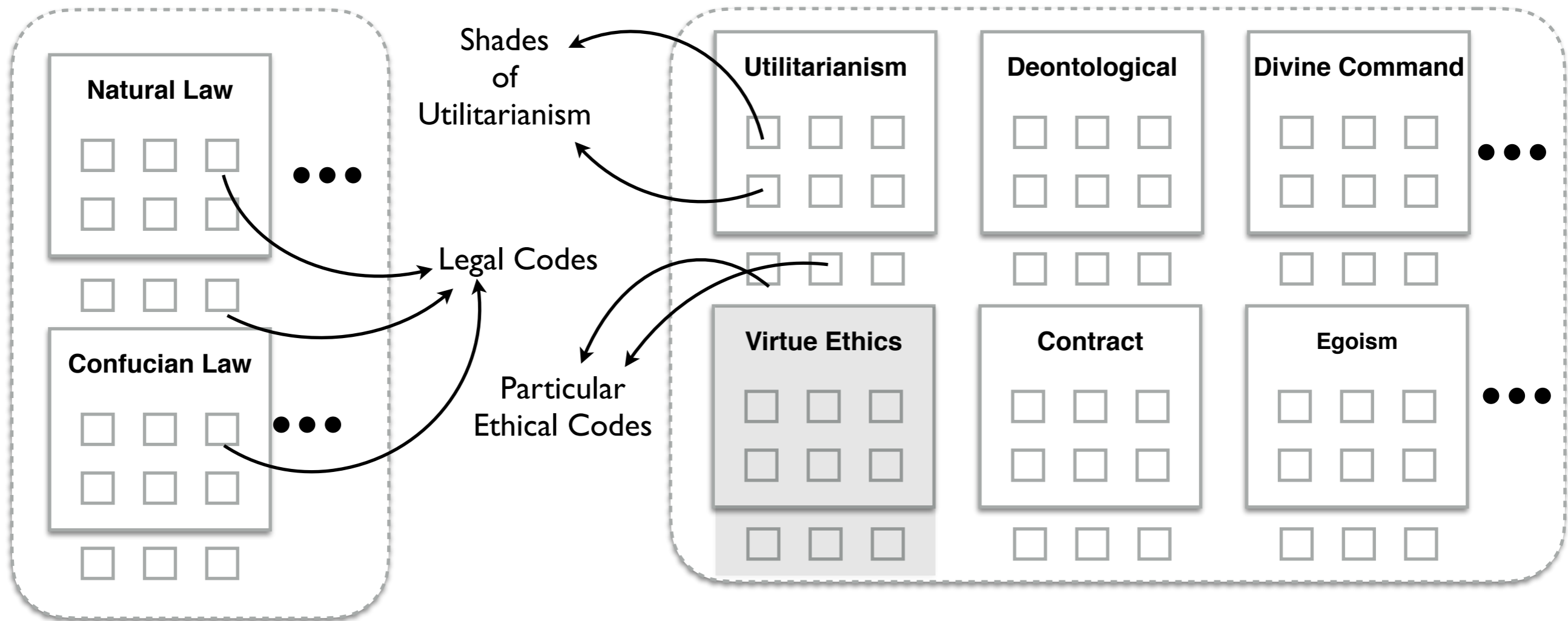


Making Morally X Machines; Only Logic Can Save Us



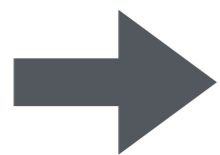
Theories of Law

Ethical Theories



Step 1

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



Step 2

Automate

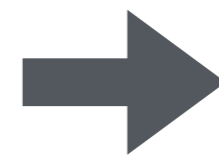
Reasoners

Spectra



Step 3

Ethical OS



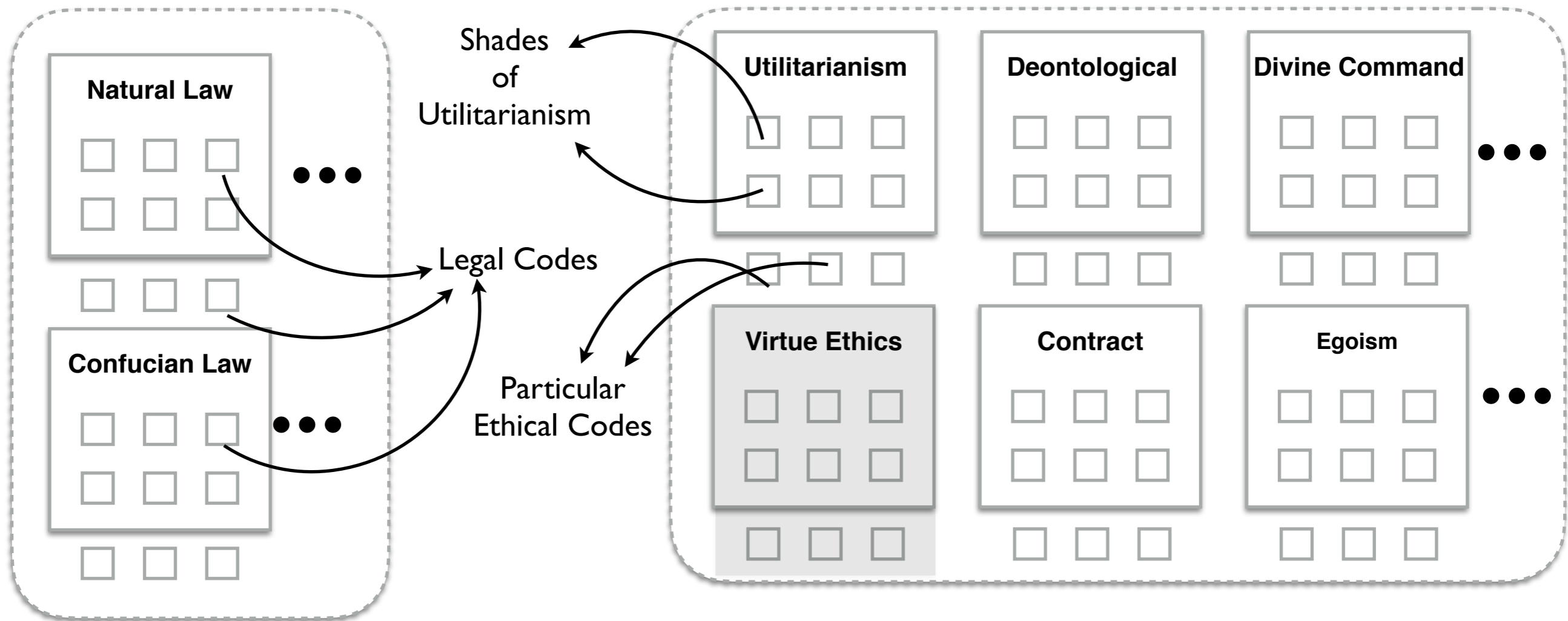


Making Morally X Machines; Only Logic Can Save Us



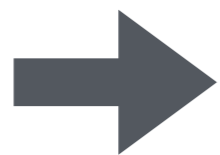
Theories of Law

Ethical Theories



Step 1

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



Step 2

Automate

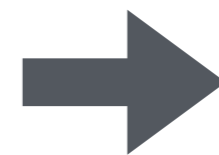
Reasoners

Spectra



Step 3

Ethical OS



Step 4

Install! — to Obtain:
Ethically/Legally
Correct Robot

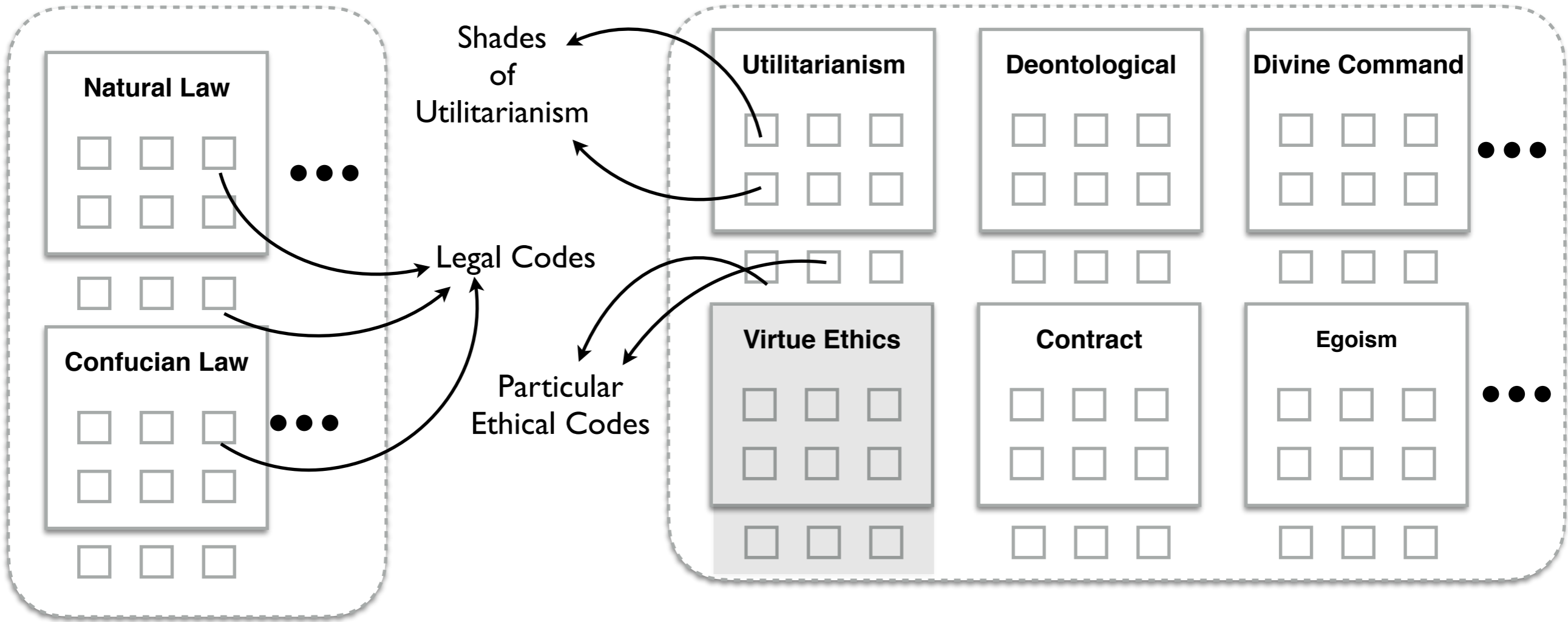


Making Morally X Machines; Only Logic Can Save Us



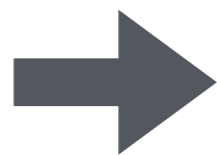
Theories of Law

Ethical Theories



Step 1

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

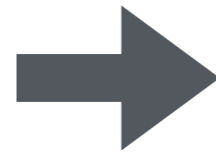


Step 2

Automate

Reasoners

Spectra

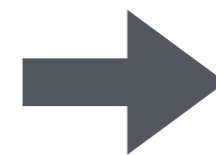


Step 3

Ethical OS

Ethical Substrate

Robotic Substrate



Step 4

Install! — to Obtain:
Ethically/Legally
Correct Robot

e.g. "Toward the Engineering of Virtuous Robots" Naveen, Selmer et al.

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

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

Not **D = SDL!** ...

Review: Encapsulation

Slate - K.slt

K. $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$
 $K \vdash \checkmark \infty \Box$

T. $\Box\varphi \rightarrow \varphi$
 $K \vdash \times \infty \Box$

4. $\Box\varphi \rightarrow \Box\Box\varphi$
 $K \vdash \times \infty \Box$

5. $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$
 $K \vdash \times \infty \Box$

Review: Encapsulation

The image shows two overlapping windows from the Slate application. The top window is titled "Slate - K.slt" and the bottom window is titled "Slate - T.slt". Each window contains four rounded rectangular boxes, each representing a modal logic formula and its validity in the K and T systems. The validity is indicated by a checkmark (✓) for true and a red X for false.

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

Review: Encapsulation

The image shows three overlapping windows from the Slate application, each displaying a grid of modal logic formulas and their derivability in various systems. The windows are titled "Slate - K.slt", "Slate - T.slt", and "Slate - D.slt".

Slate - K.slt

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

Slate - T.slt

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

Slate - D.slt

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

Review: Encapsulation

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

Slate - K.slt

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

Slate - T.slt

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

Slate - D.slt

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

Slate - S4.slt

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

Review: Encapsulation

K

T

D

4 = S4

5 = S5

The screenshot displays five windows of the HyperSlate interface, each showing a set of logical formulas and their provability status in different modal logics. The formulas are arranged in a grid, with some formulas appearing in multiple windows.

Slate - K.slt

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

Slate - T.slt

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

Slate - D.slt

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

Slate - S4.slt

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

Slate - S5.slt

- K. $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$ (S5 $\vdash \checkmark \infty \Box$)
- T. $\Box\varphi \rightarrow \varphi$ (S5 $\vdash \checkmark \infty \Box$)
- D. $\Box\varphi \rightarrow \Diamond\varphi$ ({D} Assume \checkmark)
- 4. $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$ ({4} Assume \checkmark)
- 5. $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$ (S5 $\vdash \checkmark \infty \Box$)
- INTER. $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$ ({INTER} Assume \checkmark)

Review: Encapsulation

K

T

D

4 = S4

5 = S5

The screenshot displays five windows of the HyperSlate interface, each showing a set of logical formulas and their status in different modal logics. The windows are titled 'Slate - K.slt', 'Slate - T.slt', 'Slate - D.slt', 'Slate - S4.slt', and 'Slate - S5.slt'. The 'Slate - D.slt' window is highlighted with a red border.

Slate - K.slt

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

Slate - T.slt

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

Slate - D.slt (highlighted)

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

Slate - S4.slt

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

Slate - S5.slt

- K. $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$ $S5 \vdash \checkmark \infty \Box$
- T. $\Box\varphi \rightarrow \varphi$ $S5 \vdash \checkmark \infty \Box$
- D. $\Box\varphi \rightarrow \Diamond\varphi$ $\{D\} \text{ Assume } \checkmark$
- 4. $\Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$ $\{4\} \text{ Assume } \checkmark$
- 5. $\neg\Box\varphi \rightarrow \Box\neg\Box\varphi$ $S5 \vdash \checkmark \infty \Box$
- INTER. $\Box\varphi \leftrightarrow \neg\Diamond\neg\varphi$ $\{INTER\} \text{ Assume } \checkmark$

Chisholm's Paradox ...

4.4.4 **D = SDL (= ‘Standard Deontic Logic’)**

We here introduce what is known as ‘Standard Deontic Logic’ (**SDL**), which in Slate is the system **D**. Deontic logic is the sub-branch of logic devoted to formalizing the fundamental concepts of morality; for example, the concepts of *obligation*, *permissibility*, and *forbiddenness*. The first of these three concepts can apparently serve as a cornerstone, since to say that ϕ (a formulae representing some state-of-affairs) is permissible seems to amount to saying that it’s not obligatory that it not be the case that ϕ (which shows permissibility can be defined in terms of obligation), and to say that ϕ is forbidden would seem to amount to it being obligatory that it not be the case that ϕ (which of course appears to show that forbiddenness buildable from obligation). This interconnected trio of ethical concepts is a triad explicitly invoked and analyzed since the end of the 18th century, and the importance of the triad even to modern deontic logic would be quite hard to exaggerate.⁹

SDL is traditionally axiomatized by the following:¹⁰

SDL

TAUT All theorems of the propositional calculus.

OB-K $\odot(\phi \rightarrow \psi) \rightarrow (\odot\phi \rightarrow \odot\psi)$

OB-D $\odot\phi \rightarrow \neg\odot\neg\phi$

MP If $\vdash \phi$ and $\vdash \phi \rightarrow \psi$, then $\vdash \psi$

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

4.4.4 D = SDL (= ‘Standard Deontic Logic’)

We here introduce what is known as ‘Standard Deontic Logic’ (SDL), which in Slate is the system **D**. Deontic logic is the sub-branch of logic devoted to formalizing the fundamental concepts of morality; for example, the concepts of *obligation*, *permissibility*, and *forbiddenness*. The first of these three concepts can apparently serve as a cornerstone, since to say that ϕ (a formulae representing some state-of-affairs) is permissible seems to amount to saying that $\neg\phi$ (which shows permissibility can be expressed in terms of obligation) is forbidden. This interconnected trio of ethical concepts has been analyzed since the end of the 18th century and modern deontic logic would be quite hard to invent if it were not for the work of von Wright. SDL is traditionally axiomatized by the following axioms:

SDL

TAUT All theorems of the propositional calculus

OB-K $\odot(\phi \rightarrow \psi) \rightarrow (\odot\phi \rightarrow \odot\psi)$

OB-D $\odot\phi \rightarrow \neg\odot\neg\phi$

MP If $\vdash \phi$ and $\vdash \phi \rightarrow \psi$, then $\vdash \psi$

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

CHAPTER 4. PROPOSITIONAL MODAL LOGIC

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

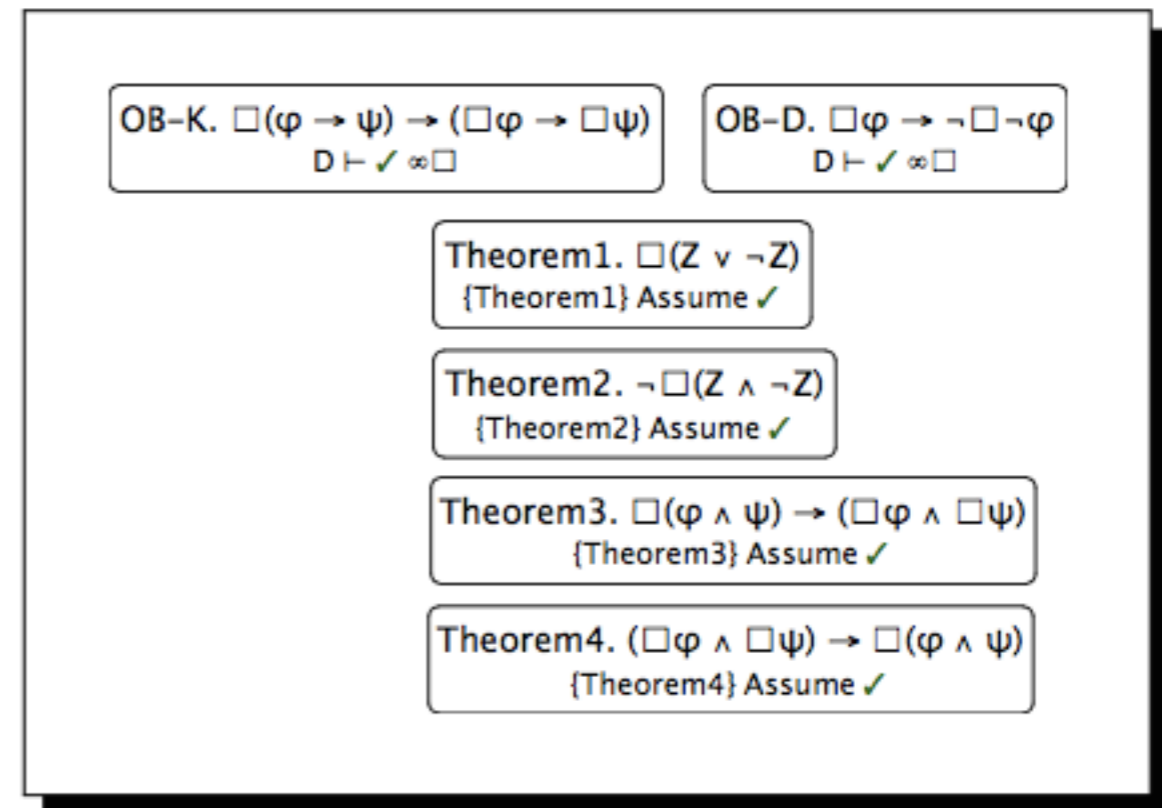


Figure 4.7: The Initial Configuration Upon Opening the File SDL.slt

4.4.4.1 Chisholm's Paradox and SDL

There are a host of problems that, together, constitute what is probably a fatal threat to **SDL** as a model of human-level ethical reasoning. We discuss in the present section the first of these problems to hit the “airwaves”: Chisholm's Paradox (CP) (Chisholm 1963). CP can be generated in Slate, you we shall see. But before we get to the level of experimentation in Slate, let's understand the scenario that Chisholm's imagined.

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

¹¹We change some particulars to ease exposition; generally, again, follow, the *SEP* entry on deontic logic (recall footnote 10). The core logic mirrors (Chisholm 1963), the original publication.

them that he is coming. The fourth and final given fact is simply that Jones doesn't go to assist his neighbors. (On the way to do so, suppose he comes upon a serious vehicular accident, is proficient in emergency medicine, and (commendably!) seizes the opportunity to save the life (and subsequently monitor) of one of the victims in this accident.) These four givens have been represented in an obvious way within four formula nodes in a Slate file; see Figure 4.8. (Notice that \square is used in place of \odot .) The paradox arises from the fact that Chisholm's quartet of givens, which surely reflect situations that are common in everyday life, in conjunction with the axioms of **SDL**, entail outright contradictions (see Exercise 2 for **D = SDL**, in §4.4.4.2).

4.4.4.1 Chisholm's Paradox and SDL

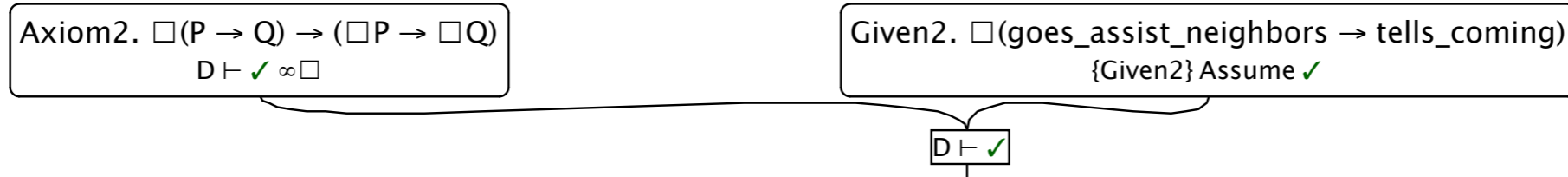
There are a host of problems that, together, constitute what is probably a fatal threat to **SDL** as a model of human-level ethical reasoning. We discuss in the present section the first of these problems to hit the “airwaves”: Chisholm's Paradox (CP) (Chisholm 1963). CP can be generated in Slate, you we shall see. But before we get to the level of experimentation in Slate, let's understand the scenario that Chisholm's imagined.

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

¹¹We change some particulars to ease exposition; generally, again, follow, the *SEP* entry on deontic logic (recall footnote 10). The core logic mirrors (Chisholm 1963), the original publication.

them that he is coming. The fourth and final given fact is simply that Jones doesn't go to assist his neighbors. (On the way to do so, suppose he comes upon a serious vehicular accident, is proficient in emergency medicine, and (commendably!) seizes the opportunity to save the life (and subsequently monitor) of one of the victims in this accident.) These four givens have been represented in an obvious way within four formula nodes in a Slate file; see Figure 4.8. (Notice that \square is used in place of \odot .) The paradox arises from the fact that Chisholm's quartet of givens, which surely reflect situations that are common in everyday life, in conjunction with the axioms of **SDL**, entail outright contradictions (see Exercise 2 for **D = SDL**, in §4.4.4.2).

Chisholm's Paradox

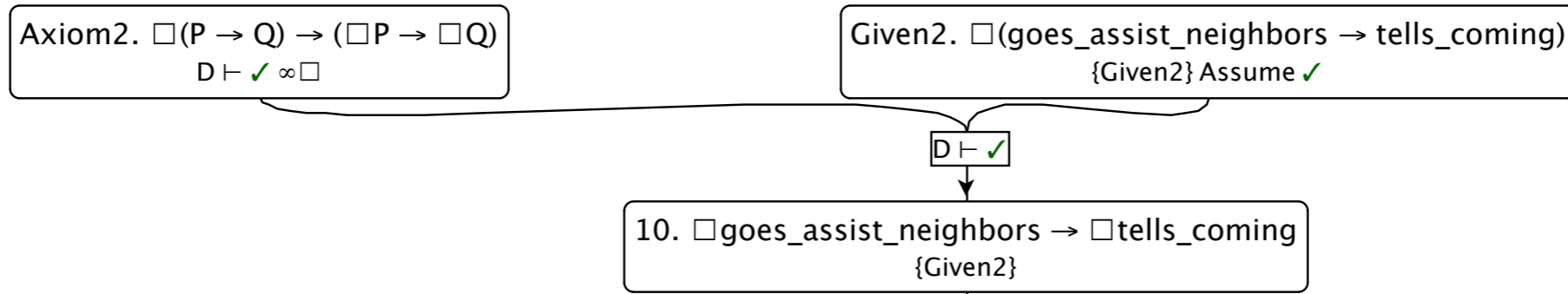


Axiom4. "Modus ponens for provability."
{Axiom4} Assume \checkmark

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

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

Chisholm's Paradox

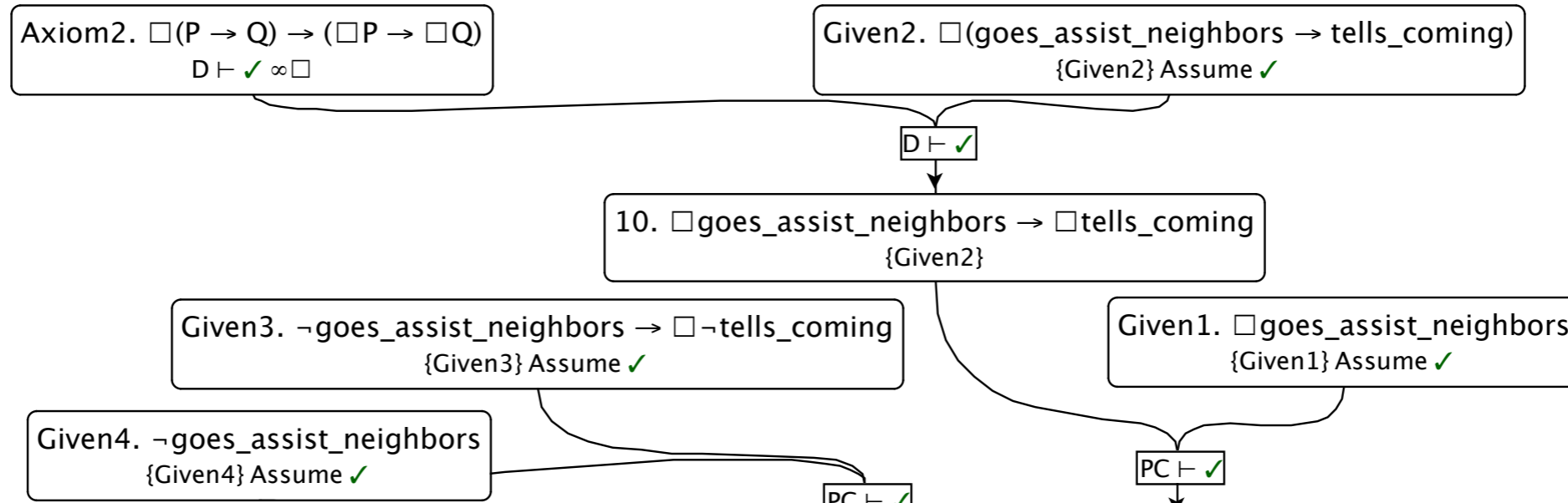


Axiom4. "Modus ponens for provability."
{Axiom4} Assume \checkmark

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

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

Chisholm's Paradox

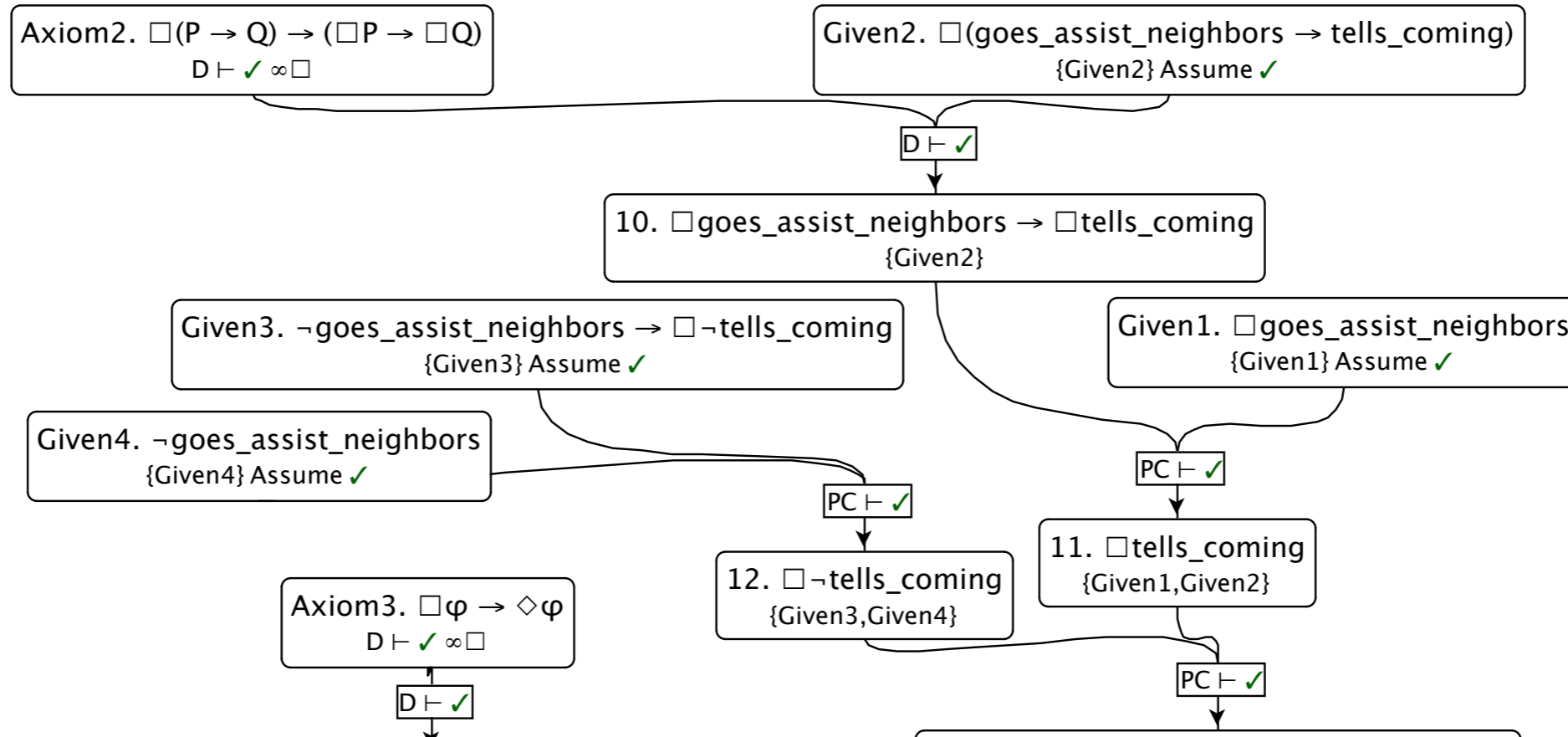


Axiom4. "Modus ponens for provability."
 $\{\text{Axiom4}\} \text{Assume } \checkmark$

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

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

Chisholm's Paradox

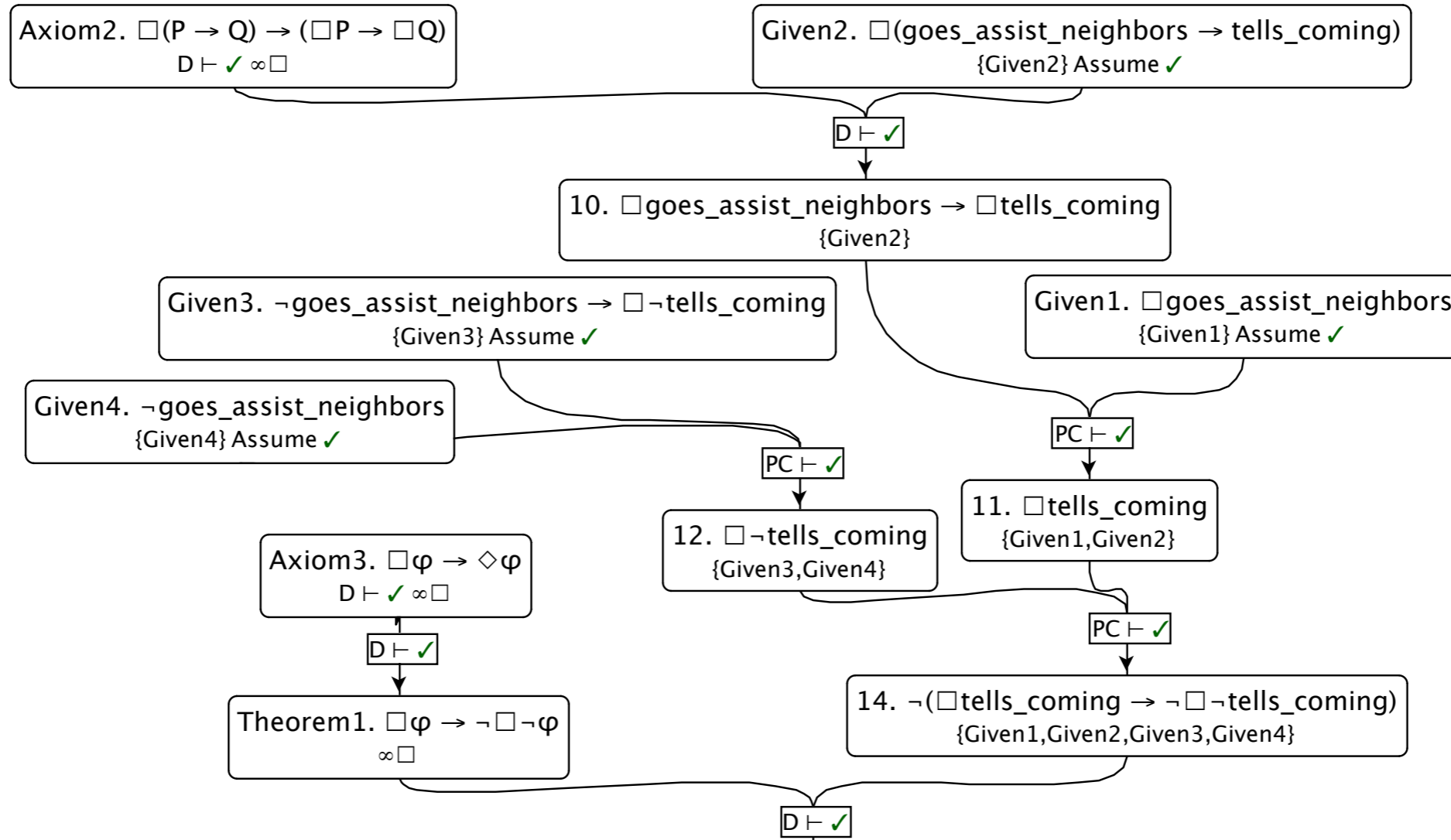


Axiom4. "Modus ponens for provability."
 $\{\text{Axiom4}\} \text{ Assume } \checkmark$

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

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

Chisholm's Paradox

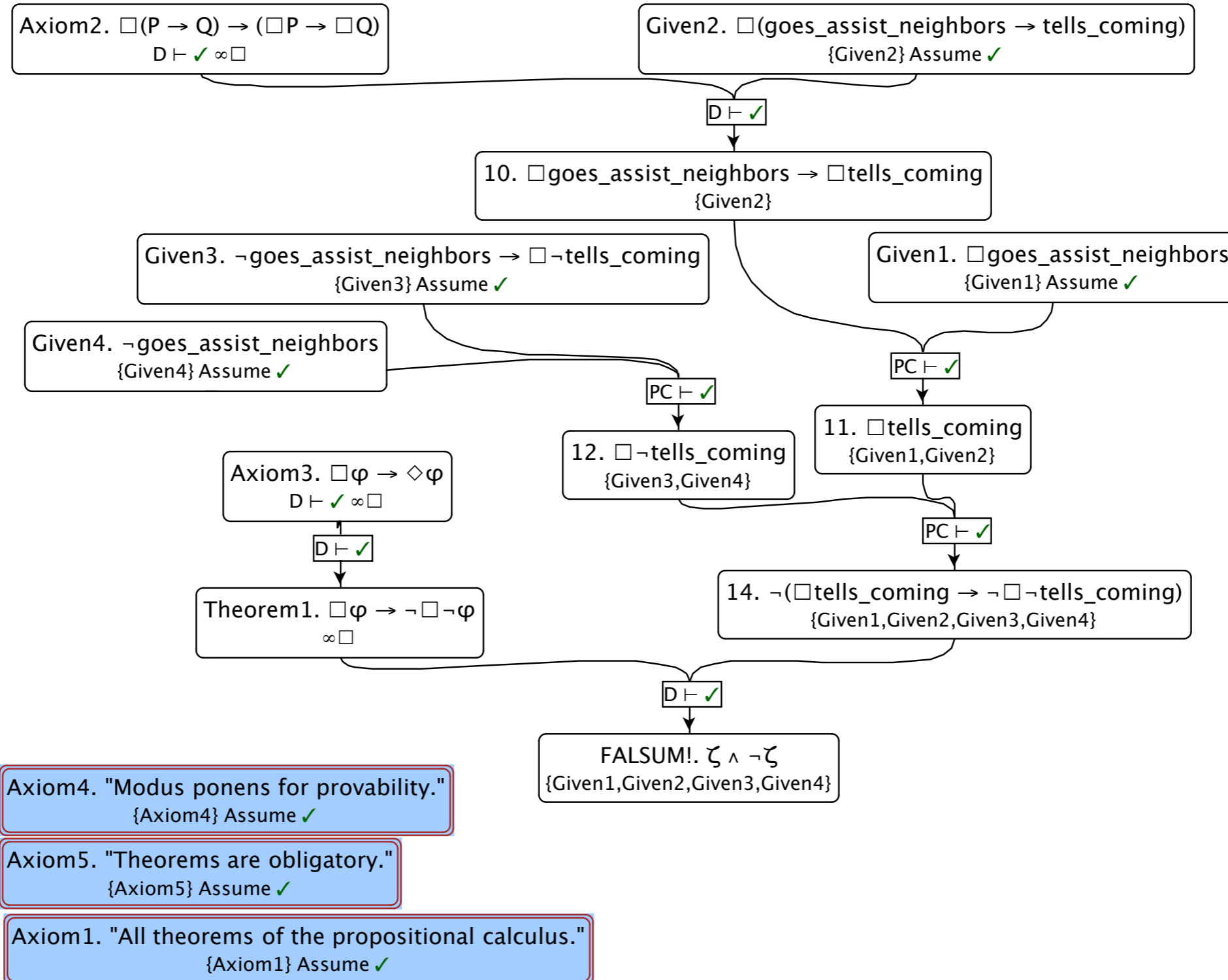


Axiom4. "Modus ponens for provability."
 $\{\text{Axiom4}\} \text{Assume } \checkmark$

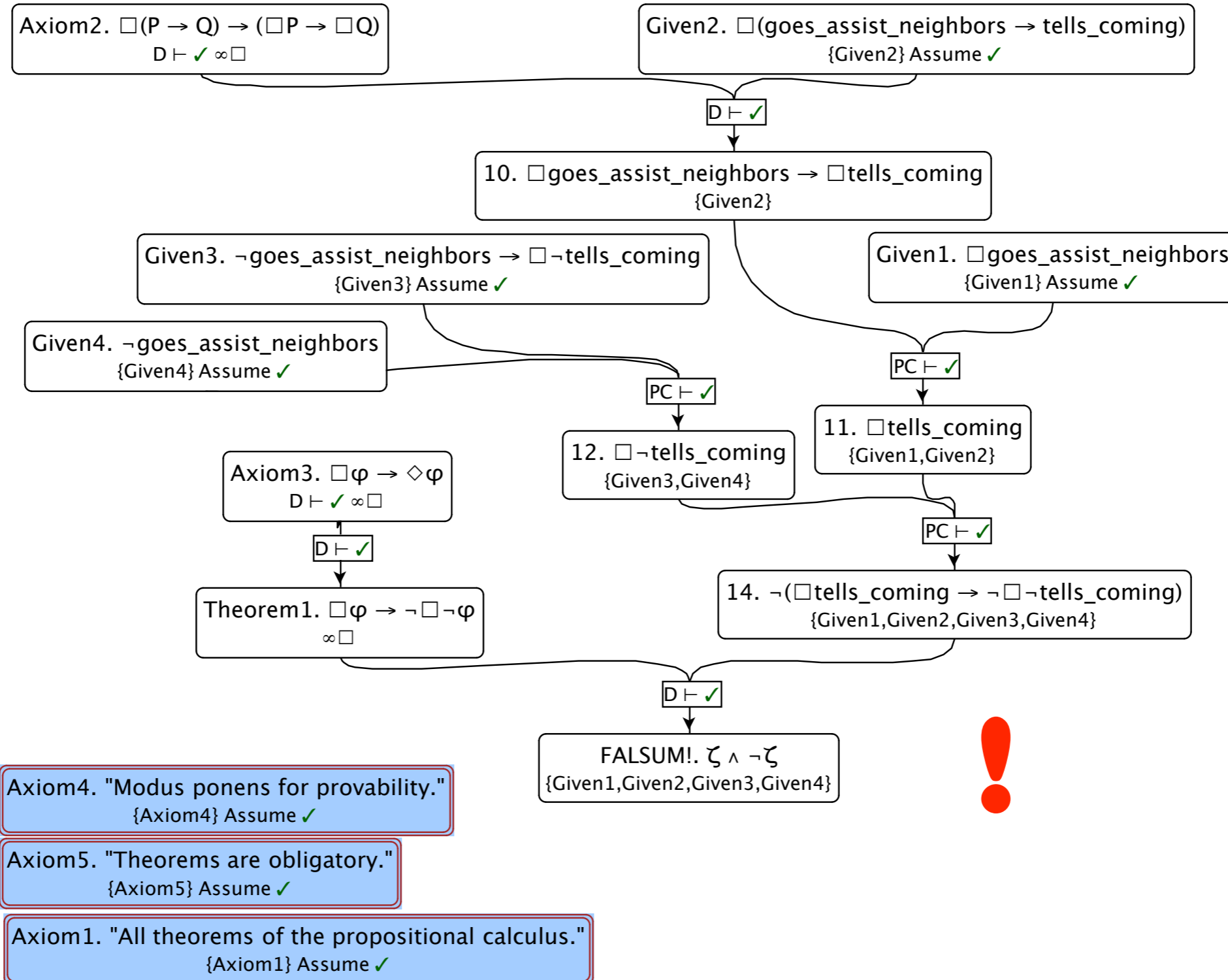
Axiom5. "Theorems are obligatory."
 $\{\text{Axiom5}\} \text{Assume } \checkmark$

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


Chisholm's Paradox



Chisholm's Paradox



Required

 ChisholmsParadox

Here you are asked to build a proof that confirms *Chisholm's Paradox*. This paradox is that from a particular representation in **D** (= Standard Deontic Logic (SDL)) of four seemingly innocuous givens, a contradiction $\zeta \wedge \neg\zeta$ can be deduced. (Your instructor should have covered this in class, and may well have supplied a proof of CP.) The four givens are based on the story of a character Jones, who is obligated to go to assist his neighbors (move to a different domicile, e.g.). It would be wrong of him to show up unannounced, though; so if he goes to assist them, it ought to be that he tells them he's coming. In addition, if it's not the case that Jones goes to assist them, then it ought to be that it not be the case that he tells them he is coming. Finally, as a matter of fact, it's not case the Jones goes to assist (because on the way he comes across a car accident, and has an opportunity to save one of the victims).

Fortunately, the [RAIR Lab's](#) modern cognitive calculus *DC $\mathcal{E}\mathcal{C}$ ** allows Chisholm's Paradox to be avoided. A recent paper explaining the use by an ethically correct AI of this calculus is available [here](#).

Your finished proof is allowed to make use of the PC provability oracle, but of no other oracle.

Deadline November 12, 2020, 3:00 PM EST

Overall Measures for: [ChisholmsParadox](#)

Total Submissions	Passed	Failed	Unprocessed
38	37	1	0

 [ChisholmsParadox_Mon_Nov_09_2020_13:50:13_GMT-0500_\(EST\).csv](#)

SDL's = D's Problems

Don't Stop Here:

The Free Choice

Permission Paradox ...

The Free Choice Permission Paradox (Ross)

1. "You may either sleep on the sofa bed or the guest bed."

{1} Assume ✓

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."

{2} Assume ✓

The Free Choice Permission Paradox (Ross)

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

$\text{D} \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
{1'}

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

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

The Free Choice Permission Paradox (Ross)

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

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

$\Box \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
{1'}

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

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

The Free Choice Permission Paradox (Ross)

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

$D \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
{1'}

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

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

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

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

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

The Free Choice Permission Paradox (Ross)

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

$D \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
{1'}

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

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

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

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

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

(How?)

The Free Choice Permission Paradox (Ross)

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

$D \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
{1'}

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

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
{2} Assume ✓

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

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

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

(How?)

8. $\diamond\varphi$
{8} Assume ✓

$PC \vdash \checkmark$

The Free Choice Permission Paradox (Ross)

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

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

$D \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
 {1'}

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
 {2} Assume ✓

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

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

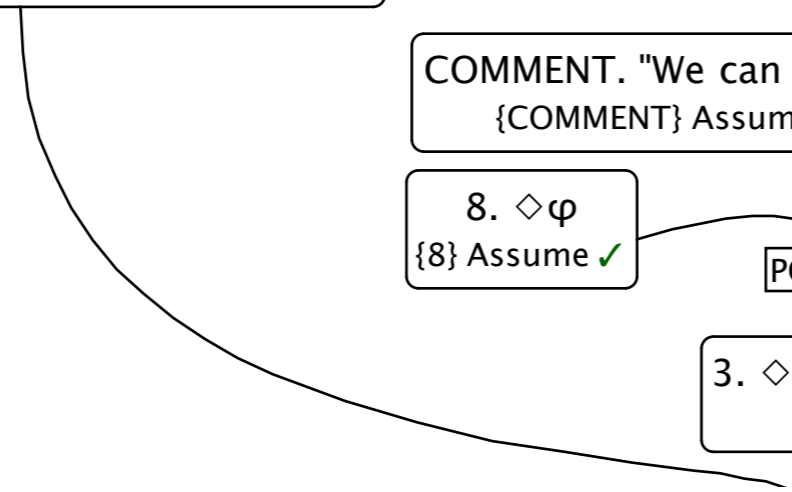
THM 5. $\diamond\varphi \rightarrow \diamond(\varphi \vee \psi)$
 $D \vdash \checkmark \infty \square$

(How?)

8. $\diamond\varphi$
 {8} Assume ✓

$PC \vdash \checkmark$

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



The Free Choice Permission Paradox (Ross)

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

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

$D \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
 {1'}

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
 {2} Assume ✓

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

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

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

(How?)

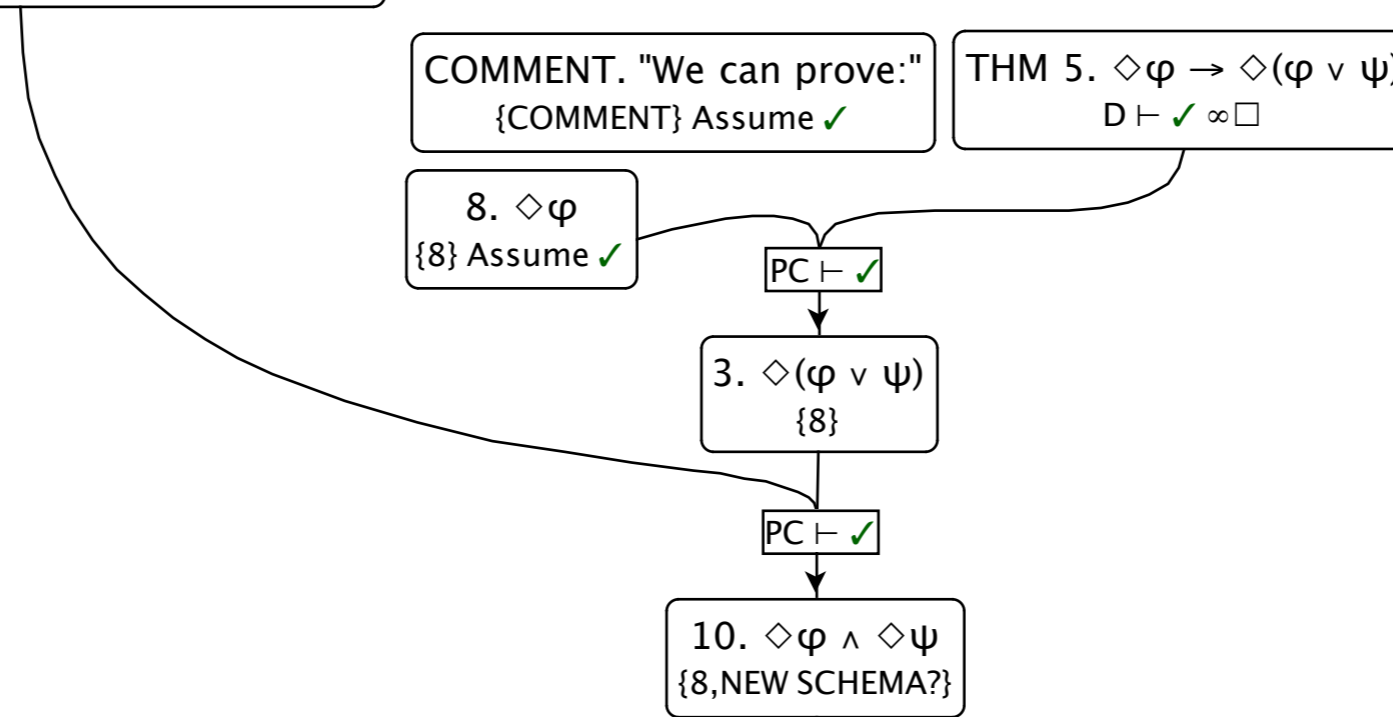
8. $\diamond\varphi$
 {8} Assume ✓

$PC \vdash \checkmark$

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

$PC \vdash \checkmark$

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



The Free Choice Permission Paradox (Ross)

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

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

$D \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
 {1'}

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
 {2} Assume ✓

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

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

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

(How?)

8. $\diamond\varphi$
 {8} Assume ✓

$PC \vdash \checkmark$

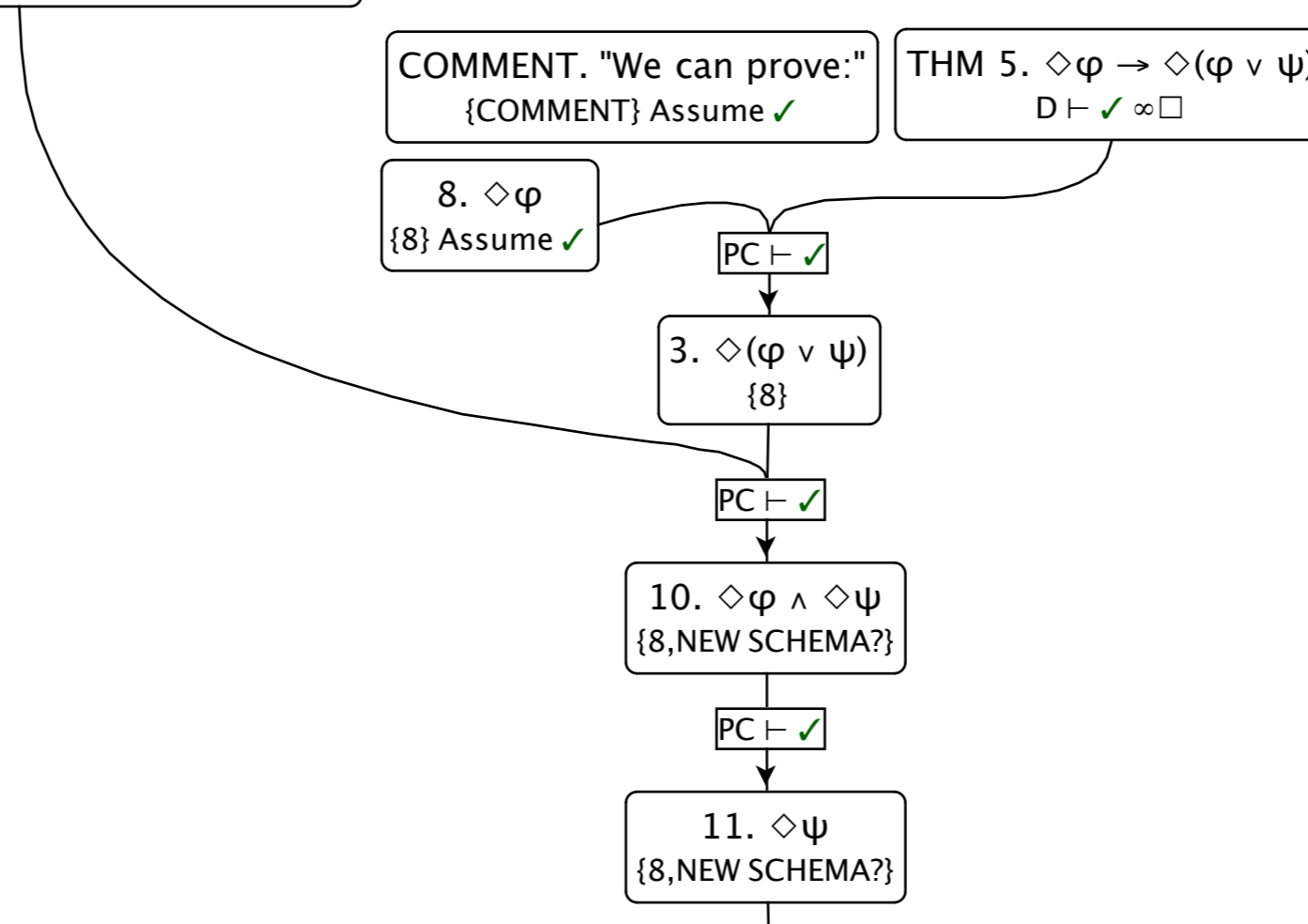
3. $\diamond(\varphi \vee \psi)$
 {8}

$PC \vdash \checkmark$

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

$PC \vdash \checkmark$

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



The Free Choice Permission Paradox (Ross)

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

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

$\text{D} \vdash \times$

2'. $\diamond \text{sofa-bed} \wedge \diamond \text{guest-bed}$
 {1'}

2. "Therefore: You may sleep on the sofa bed, and you may sleep on the guest bed."
 {2} Assume ✓

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

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

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

(How?)

8. $\diamond\varphi$
 {8} Assume ✓

$\text{PC} \vdash \checkmark$

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

$\text{PC} \vdash \checkmark$

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

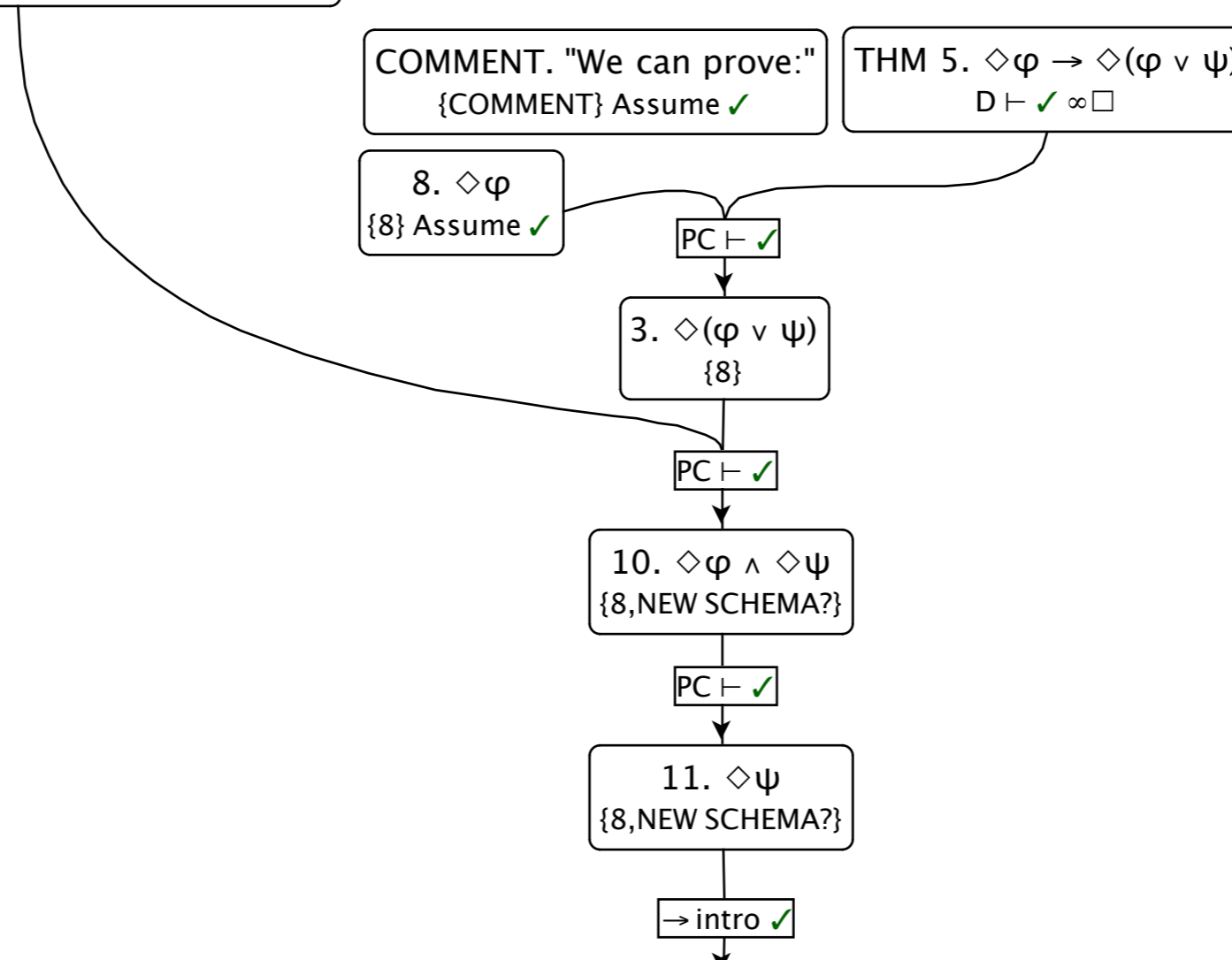
$\text{PC} \vdash \checkmark$

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


\rightarrow intro ✓

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

COMMENT. Absurd!
 {COMMENT} Assume ✓



Required

 TheFreeChoicePermissionParadox

Producing a valid proof in this problem will enable you to understand The Free Choice Permission Paradox (FCPP), discovered in 1941 by Ross ("Imperatives and Logic," *Theoria* 7: 53–71). Given that the proof in question yields an absurdity, FCPP can be taken to show that **SDL** (Standard Deontic Logic) = **D** leads to inconsistency when applied; or, put in AI terms, you wouldn't want a robot to base its ethical decision-making on **D**! Fortunately, the [RAIR Lab's](#) modern cognitive calculus *DC \mathcal{E} C** allows FCPP to be avoided. (A recent paper explaining the use by an ethically correct AI of this calculus is available [here](#).)

Here's the paradox. Suppose that you travel to visit a friend, arrive late at night, and are weary. Your friend says hospitably: "You may either sleep on the sofa-bed or sleep on the guest-room bed." (1) From this statement it follows that you are permitted to sleep on the sofa-bed, and you are permitted to sleep on the guest-room bed. (2) In **D**, this pair gets symbolized like this:

(1')

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

(2')

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

But (2') doesn't follow deductively from (1') in **D**, as a call to the provability oracle for **D** in the HyperSlate™ file for this problem confirms. A suggested repair is to add to **D** the schema

$$\diamond(\phi \vee \psi) \rightarrow (\diamond\phi \wedge \diamond\psi),$$

but as your proof will (hopefully) show, this addition allows a proof of the absurd theorem that if anything is morally permissible, everything is!

Your finished proof is allowed to make use of the PC provability oracle, but of no other oracle.

Deadline November 12, 2020, 3:00 PM EST

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

Producing a valid proof in this problem will enable you to understand The Free Choice Permission Paradox (FCPP), discovered in 1941 by Ross ("Imperatives and Logic," *Theoria* 7: 53–71). Given that the proof in question yields an absurdity, FCPP can be taken to show that **SDL** (Standard Deontic Logic) = **D** leads to inconsistency when applied; or, put in AI terms, you wouldn't want a robot to base its ethical decision-making on **D**! Fortunately, the [RAIR Lab](#)'s modern cognitive calculus *DCEC** allows FCPP to be avoided. (A recent paper explaining the use by an ethically correct AI of this calculus is available [here](#).)

Here's the paradox. Suppose that you travel to visit a friend, arrive late at night, and are weary. Your friend says hospitably: "You may either sleep on the sofa-bed or sleep on the guest-room bed." (1) From this statement it follows that you are permitted to sleep on the sofa-bed, and you are permitted to sleep on the guest-room bed. (2) In **D**, this pair gets symbolized like this:

(1')

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

(2')

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

But (2') doesn't follow deductively from (1') in **D**, as a call to the provability oracle for **D** in the HyperSlate™ file for this problem confirms. A suggested repair is to add to **D** the schema

$$\diamond(\phi \vee \psi) \rightarrow (\diamond\phi \wedge \diamond\psi),$$

but as your proof will (hopefully) show, this addition allows a proof of the absurd theorem that if anything is morally permissible, everything is!

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

DCEC* !!!

<https://www.ijcai.org/Proceedings/2017/0658.pdf>

Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)

On Automating the Doctrine of Double Effect

Naveen Sundar Govindarajulu and Selmer Bringsjord

Rensselaer Polytechnic Institute, Troy, NY

{naveensundarg,selmer.bringsjord}@gmail.com

Abstract

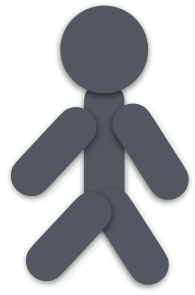
The **doctrine of double effect** (*DDE*) is a long-studied ethical principle that governs when actions that have both positive and negative effects are to be allowed. The goal in this paper is to automate *DDE*. We briefly present *DDE*, and use a first-order modal logic, the **deontic cognitive event calculus**, as our framework to formalize the doctrine.

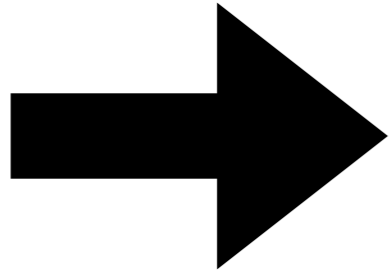
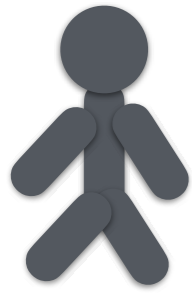
— provided that 1) the harmful effects are not intended; 2) the harmful effects are not used to achieve the beneficial effects (harm is merely a *side-effect*); and 3) benefits outweigh the harm by a significant amount. What distinguishes *DDE* from, say, naïve forms of consequentialism in ethics (e.g. act utilitarianism, which holds that an action is obligatory for an autonomous agent if and only if it produces the most utility among all competing actions) is that purely mental intentions in and of themselves, independent of conse-

4 Informal \mathcal{DDE}

We now informally but rigorously present \mathcal{DDE} . We assume we have at hand an ethical hierarchy of actions as in the deontological case (e.g. forbidden, neutral, obligatory); see [Bringsjord, 2017]. We also assume that we have a utility or goodness function for states of the world or effects as in the consequentialist case. For an autonomous agent a , an action α in a situation σ at time t is said to be \mathcal{DDE} -compliant *iff*:

- C_1 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);
- C_2 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;
- C_4 the bad effects are not used as a means to obtain the good effects; and
- C_5 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.





Formal Conditions for $\mathcal{DD}\mathcal{E}$

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

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

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:

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

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

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

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

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:

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



Formal Conditions for \mathcal{DDE}

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

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

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:

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

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:

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

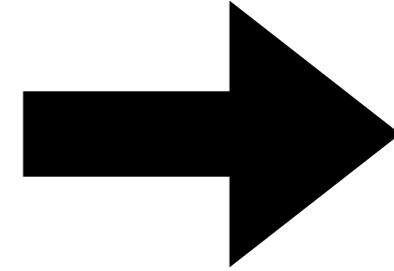
$$\Gamma \not\vdash \mathbf{I} \left(a, t, \neg \text{Holds}(f_g, y) \right)$$

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:

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



$\mathbb{P}_{\text{DDE}_1} + \text{ShadowProver}$



Formal Conditions for \mathcal{DDE}

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

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

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:

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

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:

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

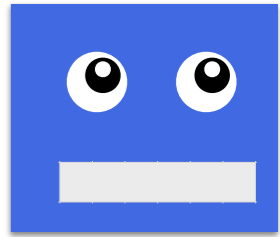
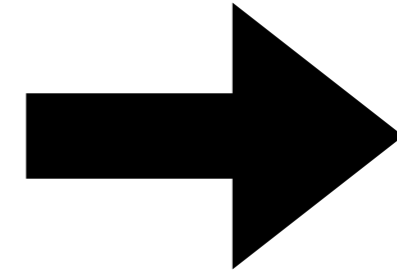
$$\Gamma \not\vdash \mathbf{I} \left(a, t, \neg \text{Holds}(f_g, y) \right)$$

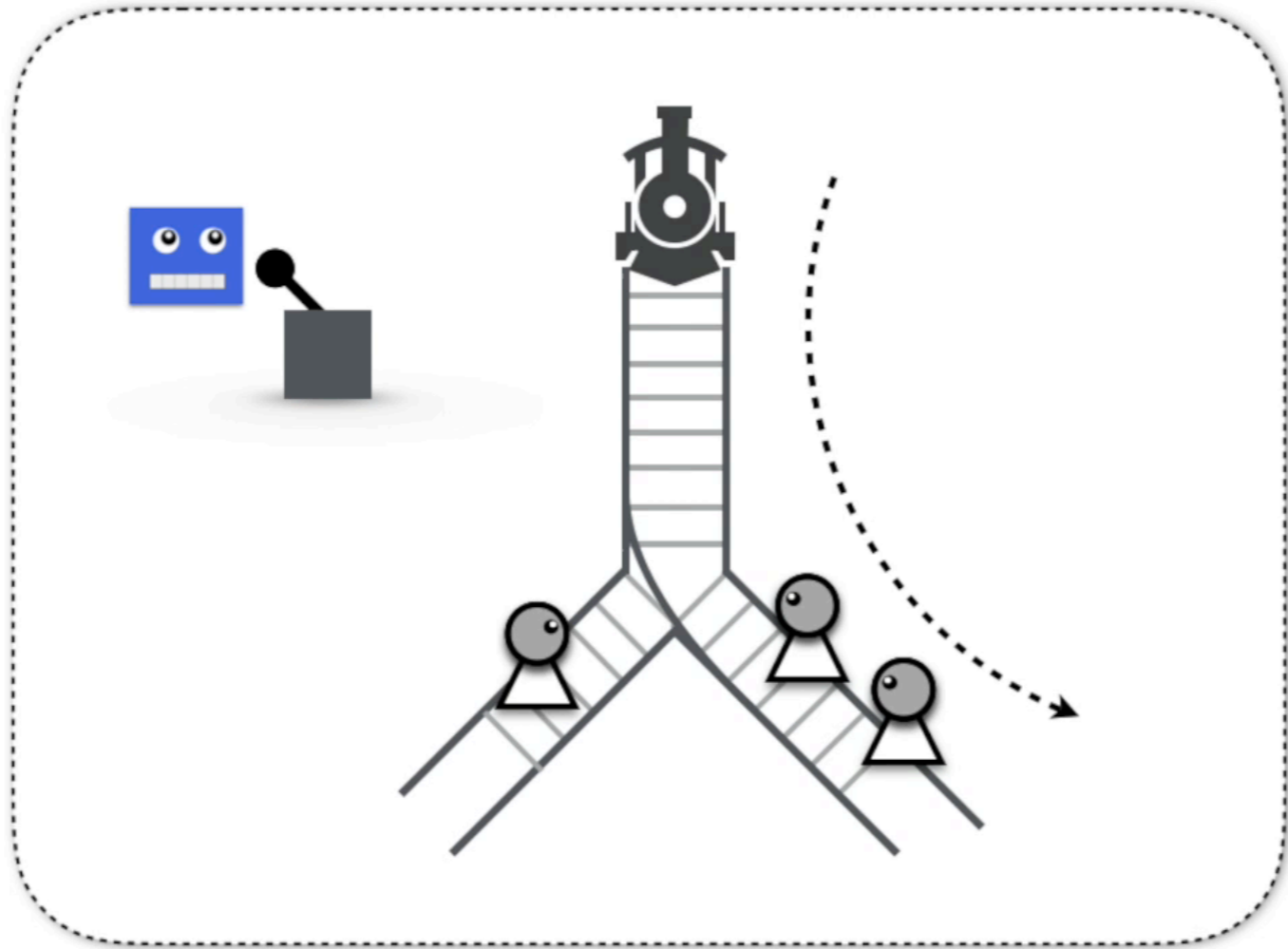
F₄ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of \triangleright above, hold here. One such permutation is shown below. For any bad fluent f_b holding at t_1 , and any good fluent f_g holding at some t_2 , such that $t < t_1, t_2 \leq H$, the following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

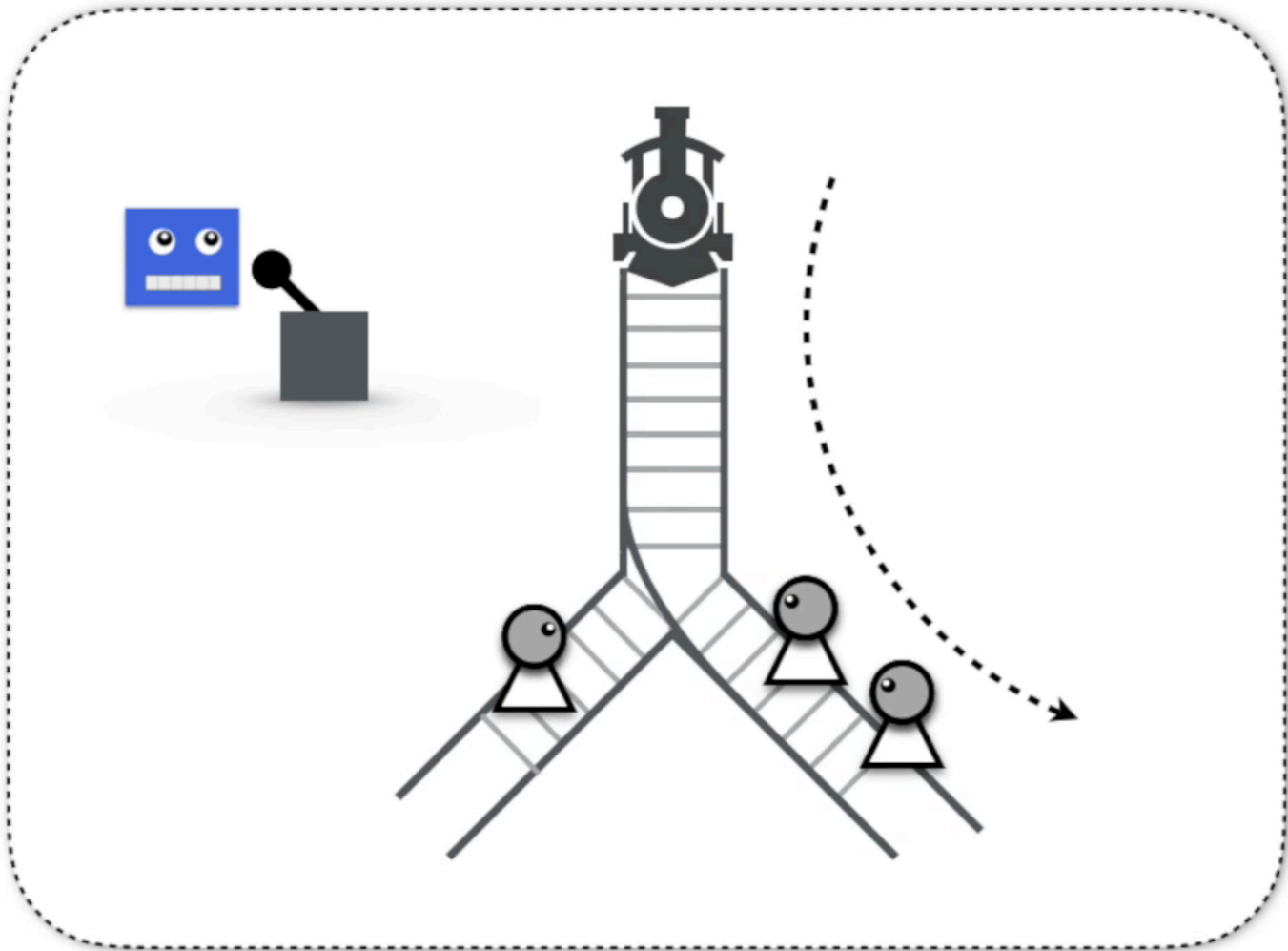
$$\Gamma \vdash \neg \triangleright \left(\text{Holds}(f_b, t_1), \text{Holds}(f_g, t_2) \right)$$



$\mathbb{P}_{\text{DDE}_1} + \text{ShadowProver}$







Inference Schemata

$$\frac{\mathbf{K}(a, t_1, \Gamma), \Gamma \vdash \phi, t_1 \leq t_2}{\mathbf{K}(a, t_2, \phi)} [R_K] \quad \frac{\mathbf{B}(a, t_1, \Gamma), \Gamma \vdash \phi, t_1 \leq t_2}{\mathbf{B}(a, t_2, \phi)} [R_B]$$

$$\frac{}{\mathbf{C}(t, \mathbf{P}(a, t, \phi) \rightarrow \mathbf{K}(a, t, \phi))} [R_1] \quad \frac{}{\mathbf{C}(t, \mathbf{K}(a, t, \phi) \rightarrow \mathbf{B}(a, t, \phi))} [R_2]$$

$$\frac{\mathbf{C}(t, \phi) \ t \leq t_1 \dots t \leq t_n}{\mathbf{K}(a_1, t_1, \dots \mathbf{K}(a_n, t_n, \phi) \dots)} [R_3] \quad \frac{\mathbf{K}(a, t, \phi)}{\phi} [R_4]$$

$$\frac{}{\mathbf{C}(t, \mathbf{K}(a, t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{K}(a, t_2, \phi_1) \rightarrow \mathbf{K}(a, t_3, \phi_2)} [R_5]$$

$$\frac{}{\mathbf{C}(t, \mathbf{B}(a, t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{B}(a, t_2, \phi_1) \rightarrow \mathbf{B}(a, t_3, \phi_2)} [R_6]$$

$$\frac{}{\mathbf{C}(t, \mathbf{C}(t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2, \phi_1) \rightarrow \mathbf{C}(t_3, \phi_2)} [R_7]$$

$$\frac{}{\mathbf{C}(t, \forall x. \phi \rightarrow \phi[x \mapsto t])} [R_8] \quad \frac{}{\mathbf{C}(t, \phi_1 \leftrightarrow \phi_2 \rightarrow \neg \phi_2 \rightarrow \neg \phi_1)} [R_9]$$

$$\frac{}{\mathbf{C}(t, [\phi_1 \wedge \dots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \dots \rightarrow \phi_n \rightarrow \psi])} [R_{10}]$$

$$\frac{\mathbf{S}(s, h, t, \phi)}{\mathbf{B}(h, t, \mathbf{B}(s, t, \phi))} [R_{12}] \quad \frac{\mathbf{I}(a, t, \mathit{happens}(\mathit{action}(a^*, \alpha), t'))}{\mathbf{P}(a, t, \mathit{happens}(\mathit{action}(a^*, \alpha), t))} [R_{13}]$$

$$\frac{\mathbf{B}(a, t, \phi) \ \mathbf{B}(a, t, \mathbf{O}(a, t, \phi, \chi)) \ \mathbf{O}(a, t, \phi, \chi)}{\mathbf{K}(a, t, \mathbf{I}(a, t, \chi))} [R_{14}]$$

Inference Schemata

$$\frac{\mathbf{K}(a, t_1, \Gamma), \Gamma \vdash \phi, t_1 \leq t_2}{\mathbf{K}(a, t_2, \phi)} [R_K] \quad \frac{\mathbf{B}(a, t_1, \Gamma), \Gamma \vdash \phi, t_1 \leq t_2}{\mathbf{B}(a, t_2, \phi)} [R_B]$$

$$\frac{}{\mathbf{C}(t, \mathbf{P}(a, t, \phi) \rightarrow \mathbf{K}(a, t, \phi))} [R_1] \quad \frac{}{\mathbf{C}(t, \mathbf{K}(a, t, \phi) \rightarrow \mathbf{B}(a, t, \phi))} [R_2]$$

$$\frac{\mathbf{C}(t, \phi) \ t \leq t_1 \dots t \leq t_n}{\mathbf{K}(a_1, t_1, \dots \mathbf{K}(a_n, t_n, \phi) \dots)} [R_3] \quad \frac{\mathbf{K}(a, t, \phi)}{\phi} [R_4]$$

$$\frac{}{\mathbf{C}(t, \mathbf{K}(a, t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{K}(a, t_2, \phi_1) \rightarrow \mathbf{K}(a, t_3, \phi_2)} [R_5]$$

$$\frac{}{\mathbf{C}(t, \mathbf{B}(a, t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{B}(a, t_2, \phi_1) \rightarrow \mathbf{B}(a, t_3, \phi_2)} [R_6]$$

$$\frac{}{\mathbf{C}(t, \mathbf{C}(t_1, \phi_1 \rightarrow \phi_2)) \rightarrow \mathbf{C}(t_2, \phi_1) \rightarrow \mathbf{C}(t_3, \phi_2)} [R_7]$$

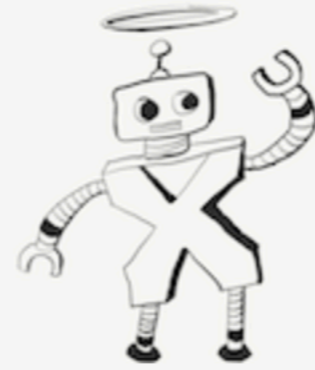
$$\frac{}{\mathbf{C}(t, \forall x. \phi \rightarrow \phi[x \mapsto t])} [R_8] \quad \frac{}{\mathbf{C}(t, \phi_1 \leftrightarrow \phi_2 \rightarrow \neg \phi_2 \rightarrow \neg \phi_1)} [R_9]$$

$$\frac{}{\mathbf{C}(t, [\phi_1 \wedge \dots \wedge \phi_n \rightarrow \phi] \rightarrow [\phi_1 \rightarrow \dots \rightarrow \phi_n \rightarrow \psi])} [R_{10}]$$

$$\frac{\mathbf{S}(s, h, t, \phi)}{\mathbf{B}(h, t, \mathbf{B}(s, t, \phi))} [R_{12}] \quad \frac{\mathbf{I}(a, t, \mathit{happens}(\mathit{action}(a^*, \alpha), t'))}{\mathbf{P}(a, t, \mathit{happens}(\mathit{action}(a^*, \alpha), t))} [R_{13}]$$

$$\frac{\mathbf{B}(a, t, \phi) \ \mathbf{B}(a, t, \mathbf{O}(a, t, \phi, \chi)) \ \mathbf{O}(a, t, \phi, \chi)}{\mathbf{K}(a, t, \mathbf{I}(a, t, \chi))} [R_{14}]$$

Making Morally



Machines

[Selmer Bringsjord](#) ^ [Naveen Sundar Govindarajulu](#) ^ [John Licato](#)

Making Morally



Machines

[Selmer Bringsjord](#) ^ [Naveen Sundar Govindarajulu](#) ^ [John Licato](#)

er løsningen, med nok penger!