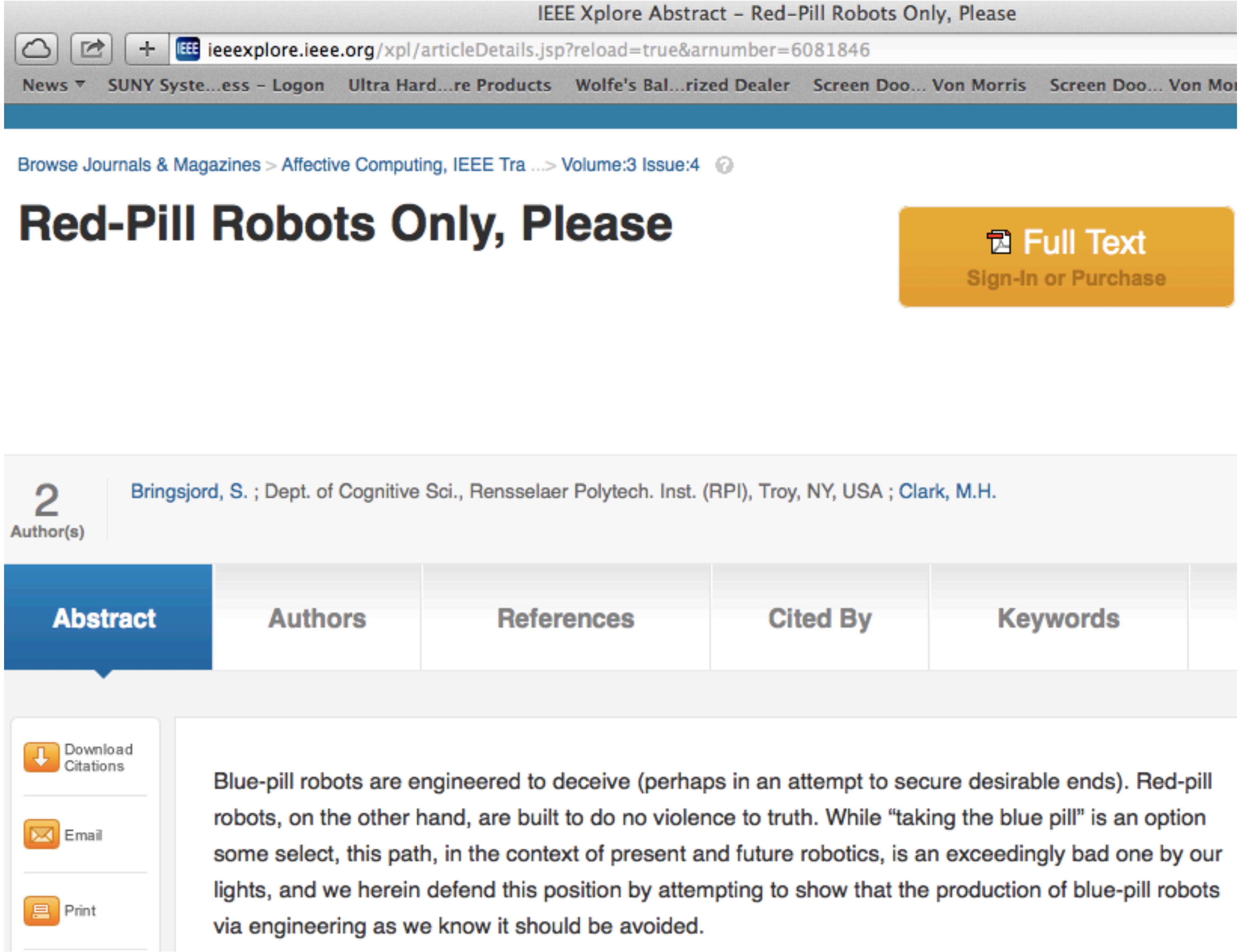
Blay: Yes, agreed; agreed.
But the dark night *inexorably* approaches.

Blay: Yes, agreed; agreed.
But the dark night *inexorably* approaches.


IEEE Xplore Abstract – Red–Pill Robots Only, Please
ieeexplore.ieee.org/xpl/articleDetails.jsp?reload=true&arnumber=6081846
News
SUNY Syste...ess – Logon
Ultra Hard...re Products
Wolfe's Bal...rized Dealer
Screen Doo... Von Morris
Screen Doo... Von Mor
Browse Journals & Magazines > Affective Computing, IEEE Tra ...> Volume:3 Issue:4
Full Text
Sign-In or Purchase
Abstract
Authors
References
Cited By
Keywords
Download Citations
Email
Print

Red-Pill Robots Only, Please

2 Author(s)

Bringsjord, S. ; Dept. of Cognitive Sci., Rensselaer Polytech. Inst. (RPI), Troy, NY, USA ; Clark, M.H.

Blue-pill robots are engineered to deceive (perhaps in an attempt to secure desirable ends). Red-pill robots, on the other hand, are built to do no violence to truth. While "taking the blue pill" is an option some select, this path, in the context of present and future robotics, is an exceedingly bad one by our lights, and we herein defend this position by attempting to show that the production of blue-pill robots via engineering as we know it should be avoided.

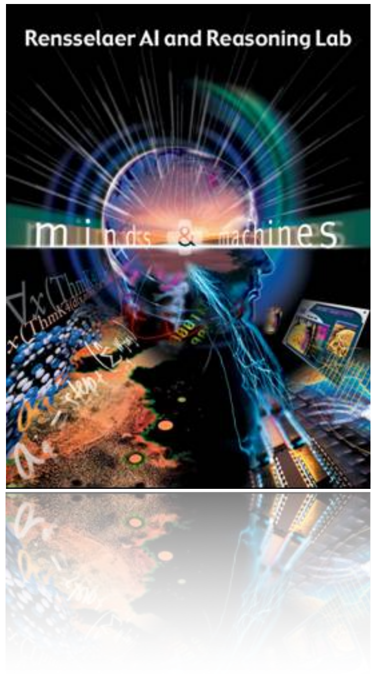
Morally Competent Robots: Progress on the Logic Thereof

(featuring: "Engineering Robots that Solve the U-of-Bristol Robot Ethical Dilemma")

Selmer Bringsjord[1] • John Licato[2]
Mei Si[3] • Joseph Johnson[4] • Rikhiya Ghosh[5]

Rensselaer AI & Reasoning (RAIR) Lab[1,2,4,5]
Department of Cognitive Science[1,2,4,5]
Department of Computer Science[1,2,4,5]
Lally School of Management & Technology[1]
Rensselaer Polytechnic Institute (RPI)
Troy, New York 12180 USA

Medford/HRIL @ Tufts
12/18/2014



Rapid-Fire Plan

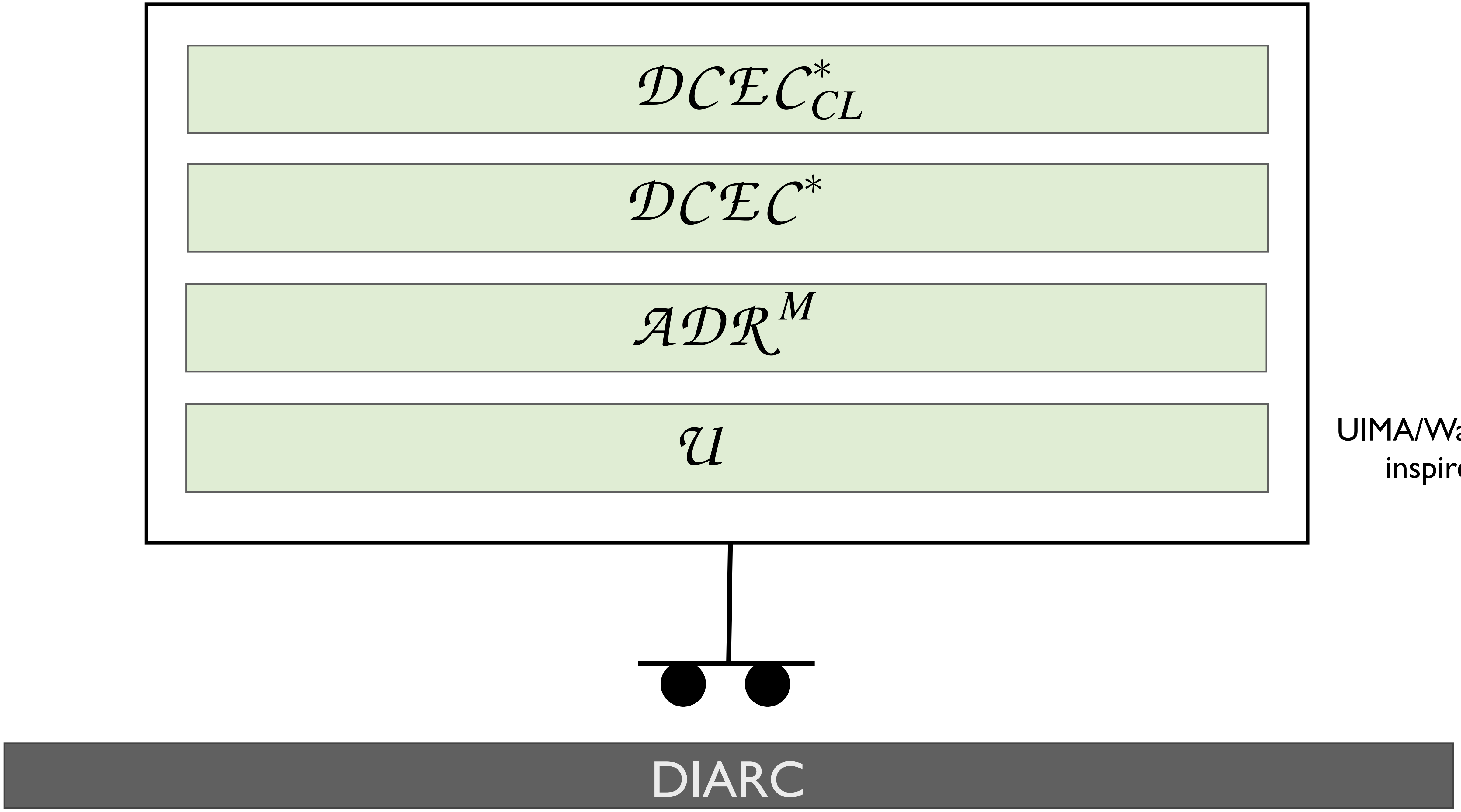
Rapid-Fire Plan

- The Hierarchy
- *Some* Prior Results
- “Killer Computational Logicians”
  - Methodology: Kill Dilemmas, Paradoxes, and Puzzles
- Bristol Robotics Lab Vid
- RAIR-Lab Vid: Easy Peasy
- Glimpse of Underlying Proofs
- Glimpse at New Target within Our Sights: Lottery Paradox

Rapid-Fire Plan

- S The Hierarchy
- S *Some* Prior Results
- S "Killer Computational Logicians"
  - Methodology: Kill Dilemmas, Paradoxes, and Puzzles
- J Bristol Robotics Lab Vid
- J RAIR-Lab Vid: Easy Peasy
- J Glimpse of Underlying Proofs
- S Glimpse at New Target within Our Sights: Lottery Paradox

Hierarchy of Ethical Reasoning



$\mathcal{DCEC}^*_{CL}$

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson-inspired

DIARC

Two Priors:

‘‘Breaking Bad’’

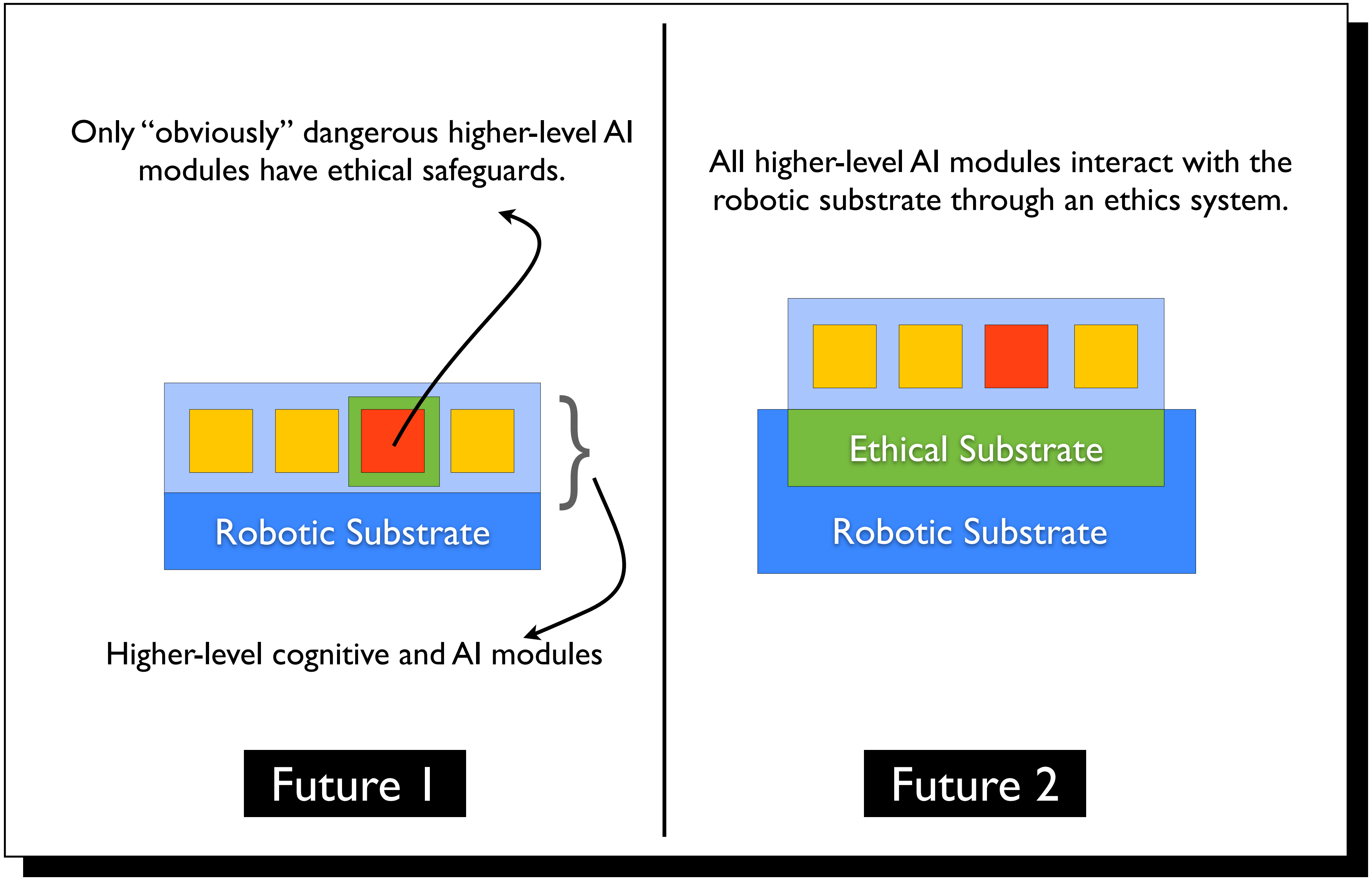
&

Akrasia

Pick the Better Future!

Naveen Sundar Govindarajulu and Selmer Bringsjord. “Ethical Regulation of Robots Must Be Embedded in Their Operating Systems” (book chapter, forthcoming), *A Construction Manual for Robot’s Ethical Systems: Requirements, Methods, Implementations.*

Pick the Better Future!


Only "obviously" dangerous higher-level AI modules have ethical safeguards.
Robotic Substrate
Higher-level cognitive and AI modules
Future 1
All higher-level AI modules interact with the robotic substrate through an ethics system.
Ethical Substrate
Robotic Substrate
Future 2

Naveen Sundar Govindarajulu and Selmer Bringsjord. "Ethical Regulation of Robots Must Be Embedded in Their Operating Systems" (book chapter, forthcoming), *A Construction Manual for Robot's Ethical Systems: Requirements, Methods, Implementations.*

Akrasia

Weakness of the Will

Informal Definition of Akrasia

An action α_f is (Augustinian) akratic for an agent A at t_{α_f} iff the following eight conditions hold:

(1) A believes that A ought to do α_o at t_{α_o};

(2) A desires to do α_f at t_{α_f};

(3) A's doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(4) A knows that doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(5) At the time (t_{α_f}) of doing the forbidden α_f, A's desire to do α_f overrides A's belief that he ought to do α_o at t_{α_f}.

(6) A does the forbidden action α_f at t_{α_f};

(7) A's doing α_f results from A's desire to do α_f;

(8) At some time t after t_{α_f}, A has the belief that A ought to have done α_o rather than α_f.

Informal Definition of Akrasia

An action α_f is (Augustinian) akratic for an agent A at t_{α_f} iff the following eight conditions hold:

(1) A believes that A ought to do α_o at t_{α_o};

(2) A desires to do α_f at t_{α_f};

(3) A's doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(4) A knows that doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(5) At the time (t_{α_f}) of doing the forbidden α_f, A's desire to do α_f overrides A's belief that he ought to do α_o at t_{α_f}.

(6) A does the forbidden action α_f at t_{α_f};

(7) A's doing α_f results from A's desire to do α_f;

(8) At some time t after t_{α_f}, A has the belief that A ought to have done α_o rather than α_f.

Informal Definition of Akrasia

An action α_f is (Augustinian) akratic for an agent A at t_{α_f} iff the following eight conditions hold:

(1) A believes that A ought to do α_o at t_{α_o};

(2) A desires to do α_f at t_{α_f};

(3) A's doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(4) A knows that doing α_f at t_{α_f} entails his not doing α_o at t_{α_o};

(5) At the time (t_{α_f}) of doing the forbidden α_f, A's desire to do α_f overrides A's belief that he ought to do α_o at t_{α_f}.

(6) A does the forbidden action α_f at t_{α_f};

(7) A's doing α_f results from A's desire to do α_f;

"Regret" (8) At some time t after t_{α_f}, A has the belief that A ought to have done α_o rather than α_f.

Cast in

$\mathcal{DCEC}^*$

becomes …

$$\mathsf{KB}_{rs} \cup \mathsf{KB}_{m_1} \cup \mathsf{KB}_{m_2} \ldots \mathsf{KB}_{m_n} \vdash$$

$$D_1 : \mathbf{B}(\mathsf{I}, \mathsf{now}, \mathbf{O}(\mathsf{I}^*, t_\alpha \Phi, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha)))$$

$$D_2 : \mathbf{D}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{does}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}))$$

$$D_3 : \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \Rightarrow \neg \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha)$$

$$D_4 : \mathbf{K}\left(\mathsf{I}, \mathsf{now}, \begin{pmatrix} \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \Rightarrow \\ \neg \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha) \end{pmatrix}\right)$$

$$D_5 : \begin{array}{l} \ \mathbf{I}(\mathsf{I}, t_\alpha, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}}) \wedge \\ \neg \mathbf{I}(\mathsf{I}, t_\alpha, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha) \end{array}$$

$$D_6 : \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_{\overline{\alpha}})$$

$$D_{7a} : \begin{array}{l} \Gamma \cup \{\mathbf{D}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{does}(\mathsf{I}^*, \overline{\alpha}), t))\} \vdash \\ \quad \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_\alpha) \end{array}$$

$$D_{7b} : \begin{array}{l} \Gamma - \{\mathbf{D}(\mathsf{I}, \mathsf{now}, \mathit{holds}(\mathit{does}(\mathsf{I}^*, \overline{\alpha}), t))\} \nvdash \\ \quad \mathit{happens}(\mathit{action}(\mathsf{I}^*, \overline{\alpha}), t_\alpha) \end{array}$$

$$D_8 : \mathbf{B}(\mathsf{I}, t_f, \mathbf{O}(\mathsf{I}^*, t_\alpha, \Phi, \mathit{happens}(\mathit{action}(\mathsf{I}^*, \alpha), t_\alpha)))$$

Demos …

Demos …

Reasoning Times

Reasoning Times

Reasoner	Description	Exact?	Time for Scenario 1	Time for Scenario 2
Approx.	First-order approximation of DCEC*	No	**1.05s**	**1.24s**
Exact	**Exact** first-order modal logic prover	Yes	**0.33s**	**0.39s**
Analogical	Analogical reasoning from a prior example	-		

Reasoning Times

Reasoner	Description	Exact?	Time for Scenario 1	Time for Scenario 2
Approx.	First-order approximation of DCEC*	No	**1.05s**	**1.24s**
Exact	**Exact** first-order modal logic prover	Yes	**0.33s**	**0.39s**
Analogical	Analogical reasoning from a prior example	-		

$\mathcal{DCEC}^*$

Reasoning Times

Reasoner	Description	Exact?	Time for Scenario 1	Time for Scenario 2
Approx.	First-order approximation of DCEC*	No	**1.05s**	**1.24s**
Exact	**Exact** first-order modal logic prover	Yes	**0.33s**	**0.39s**
Analogical	Analogical reasoning from a prior example	-	$\mathcal{ADR}^M$	

$\mathcal{DCEC}^*$

Reasoning Times

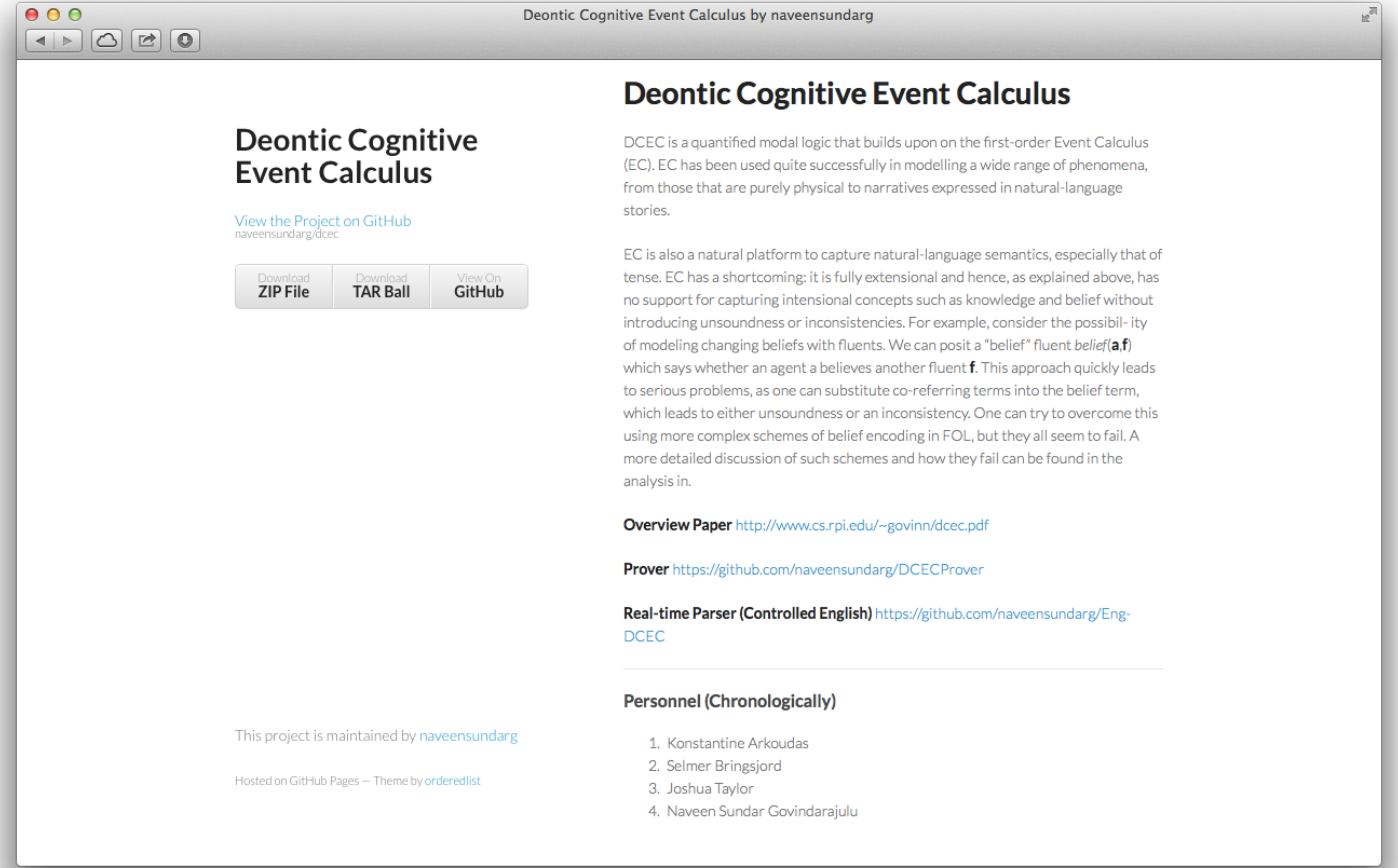
Reasoner	Description	Exact?	Time for Scenario 1	Time for Scenario 2
Approx.	First-order approximation of DCEC*	No	**1.05s**	**1.24s**
Exact	**Exact** first-order modal logic prover	Yes	**0.33s**	**0.39s**
Analogical	Analogical reasoning from a prior example	-		

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

https://github.com/naveensundarg/DCECProver

DCEC Master Page



Deontic Cognitive Event Calculus

View the Project on GitHub
naveensundarg/dcec

This project is maintained by naveensundarg

Hosted on GitHub Pages — Theme by orderedlist

Deontic Cognitive Event Calculus

DCEC is a quantified modal logic that builds upon on the first-order Event Calculus (EC). EC has been used quite successfully in modelling a wide range of phenomena, from those that are purely physical to narratives expressed in natural-language stories.

EC is also a natural platform to capture natural-language semantics, especially that of tense. EC has a shortcoming: it is fully extensional and hence, as explained above, has no support for capturing intensional concepts such as knowledge and belief without introducing unsoundness or inconsistencies. For example, consider the possibil- ity of modeling changing beliefs with fluents. We can posit a "belief" fluent *belief*(**a**,**f**) which says whether an agent a believes another fluent **f**. This approach quickly leads to serious problems, as one can substitute co-referring terms into the belief term, which leads to either unsoundness or an inconsistency. One can try to overcome this using more complex schemes of belief encoding in FOL, but they all seem to fail. A more detailed discussion of such schemes and how they fail can be found in the analysis in.

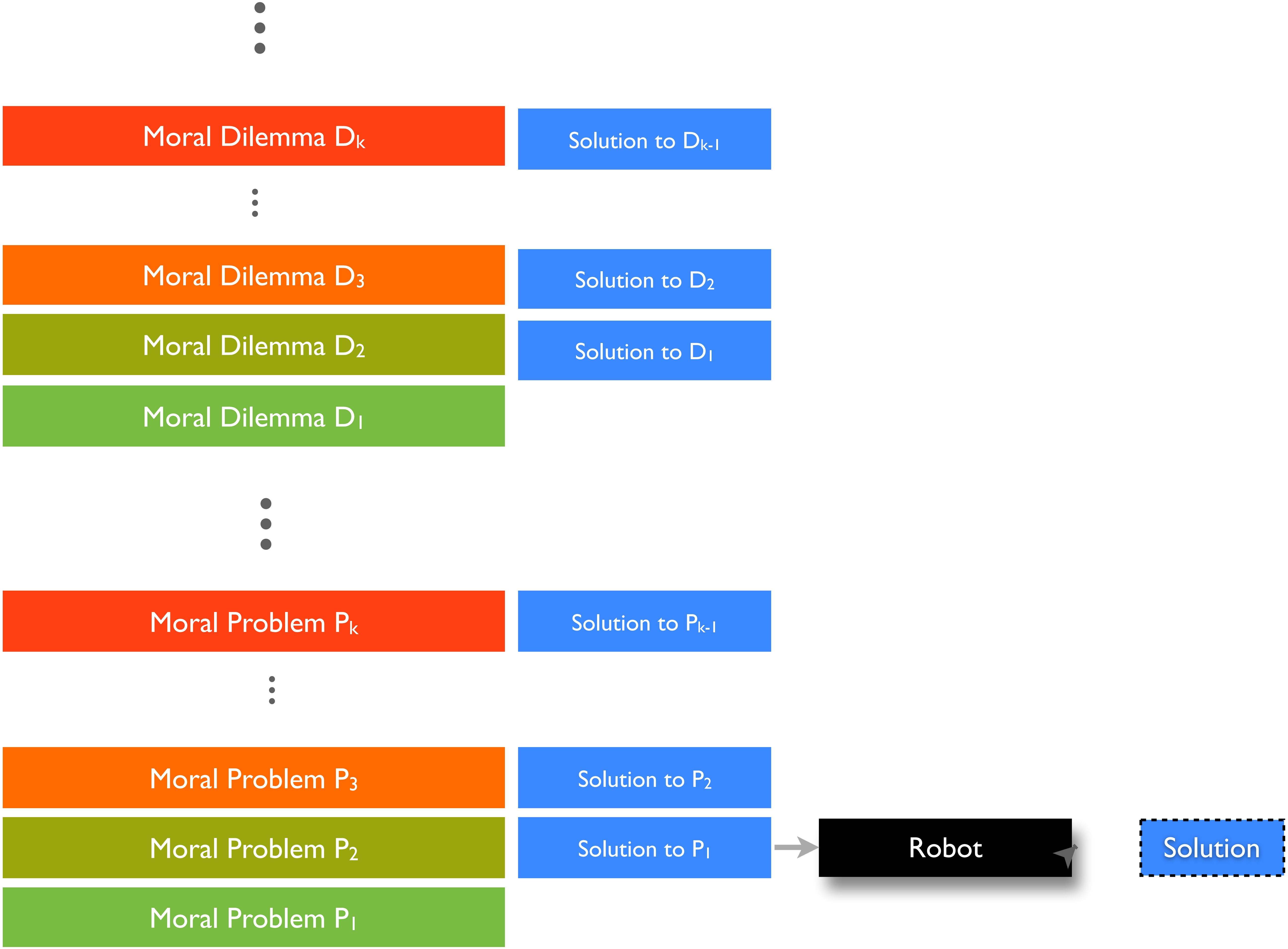
Overview Paper http://www.cs.rpi.edu/~govinn/dcec.pdf

Prover https://github.com/naveensundarg/DCECProver

Real-time Parser (Controlled English) https://github.com/naveensundarg/Eng-DCEC

Personnel (Chronologically)

1. Konstantine Arkoudas
2. Selmer Bringsjord
3. Joshua Taylor
4. Naveen Sundar Govindarajulu



Moral Dilemma D_k

Solution to D_{k-1}

Moral Dilemma D_3

Solution to D_2

Moral Dilemma D_2

Solution to D_1

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

Moral Problem P_3

Solution to P_2

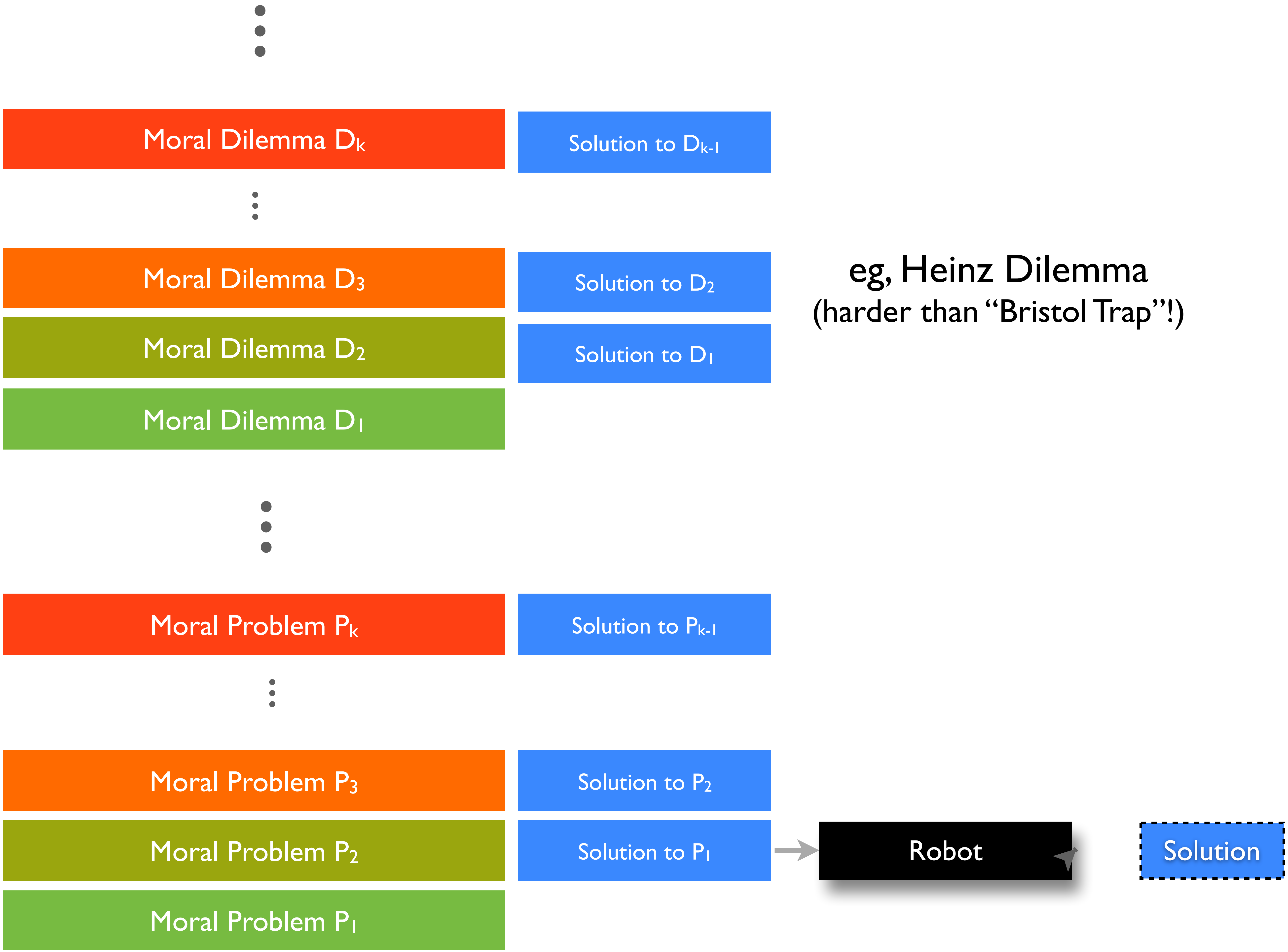
Moral Problem P_2

Solution to P_1

Robot

Solution

Moral Problem P_1



Moral Dilemma D_k

Solution to D_{k-1}

Moral Dilemma D_3

Solution to D_2

Moral Dilemma D_2

Solution to D_1

Moral Dilemma D_1

eg, Heinz Dilemma
(harder than “Bristol Trap”!)

Moral Problem P_k

Solution to P_{k-1}

Moral Problem P_3

Solution to P_2

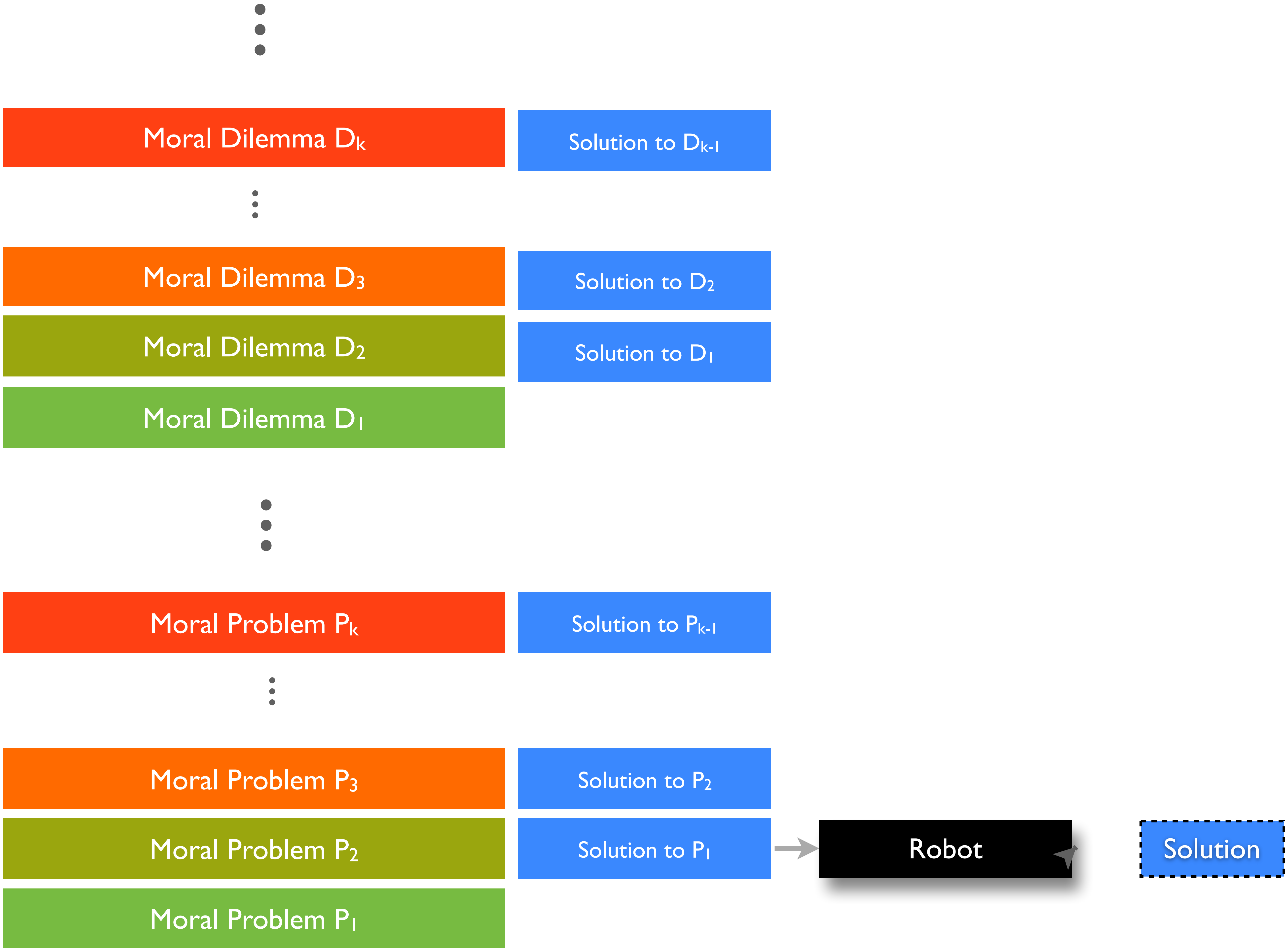
Moral Problem P_2

Solution to P_1

Robot

Solution

Moral Problem P_1



Moral Dilemma D_k

Solution to D_{k-1}

Moral Dilemma D_3

Solution to D_2

Moral Dilemma D_2

Solution to D_1

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

Moral Problem P_3

Solution to P_2

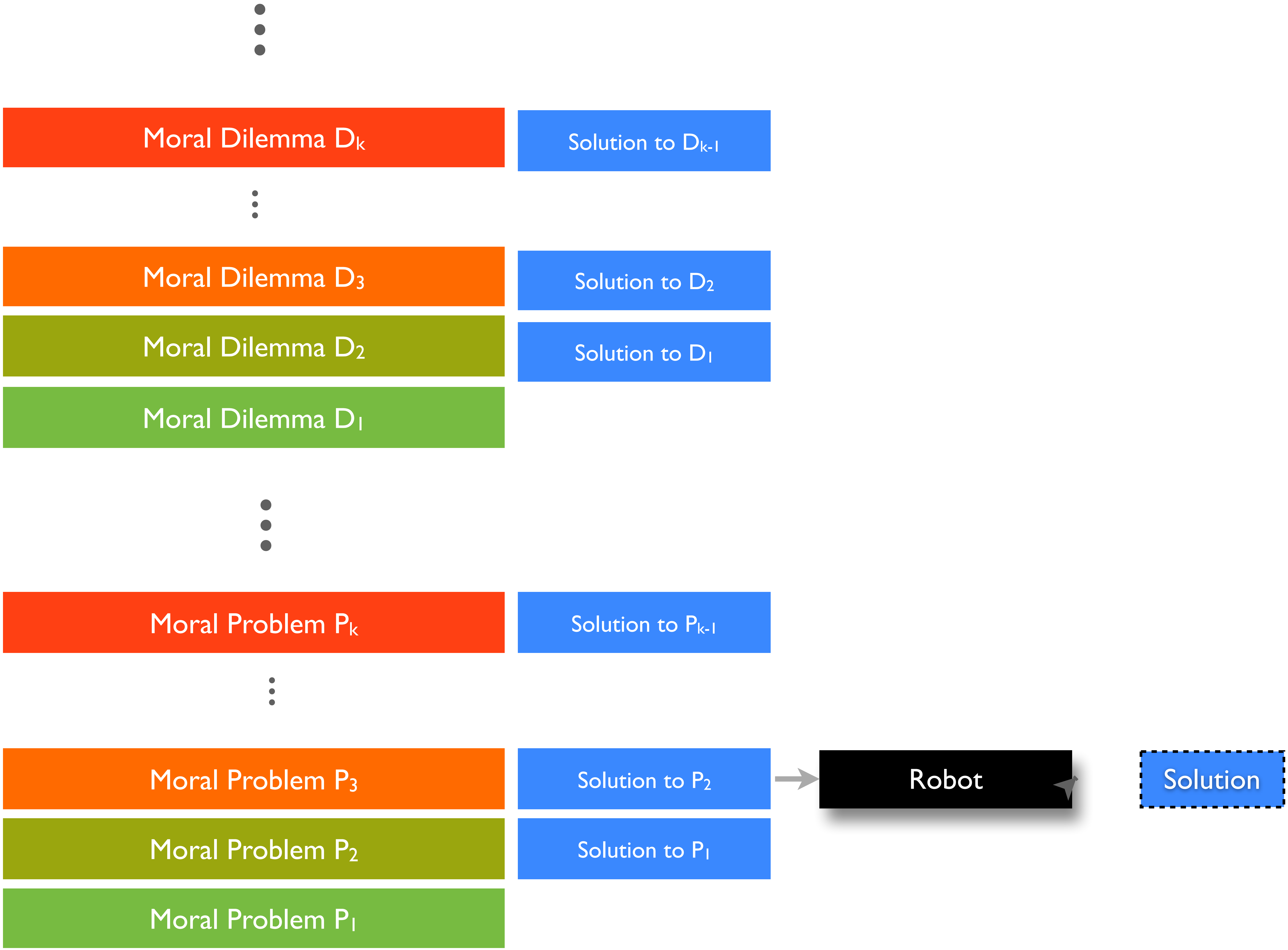
Moral Problem P_2

Solution to P_1

Robot

Solution

Moral Problem P_1



Moral Dilemma D_k

Solution to D_{k-1}

Moral Dilemma D_3

Solution to D_2

Moral Dilemma D_2

Solution to D_1

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

Moral Problem P_3

Solution to P_2

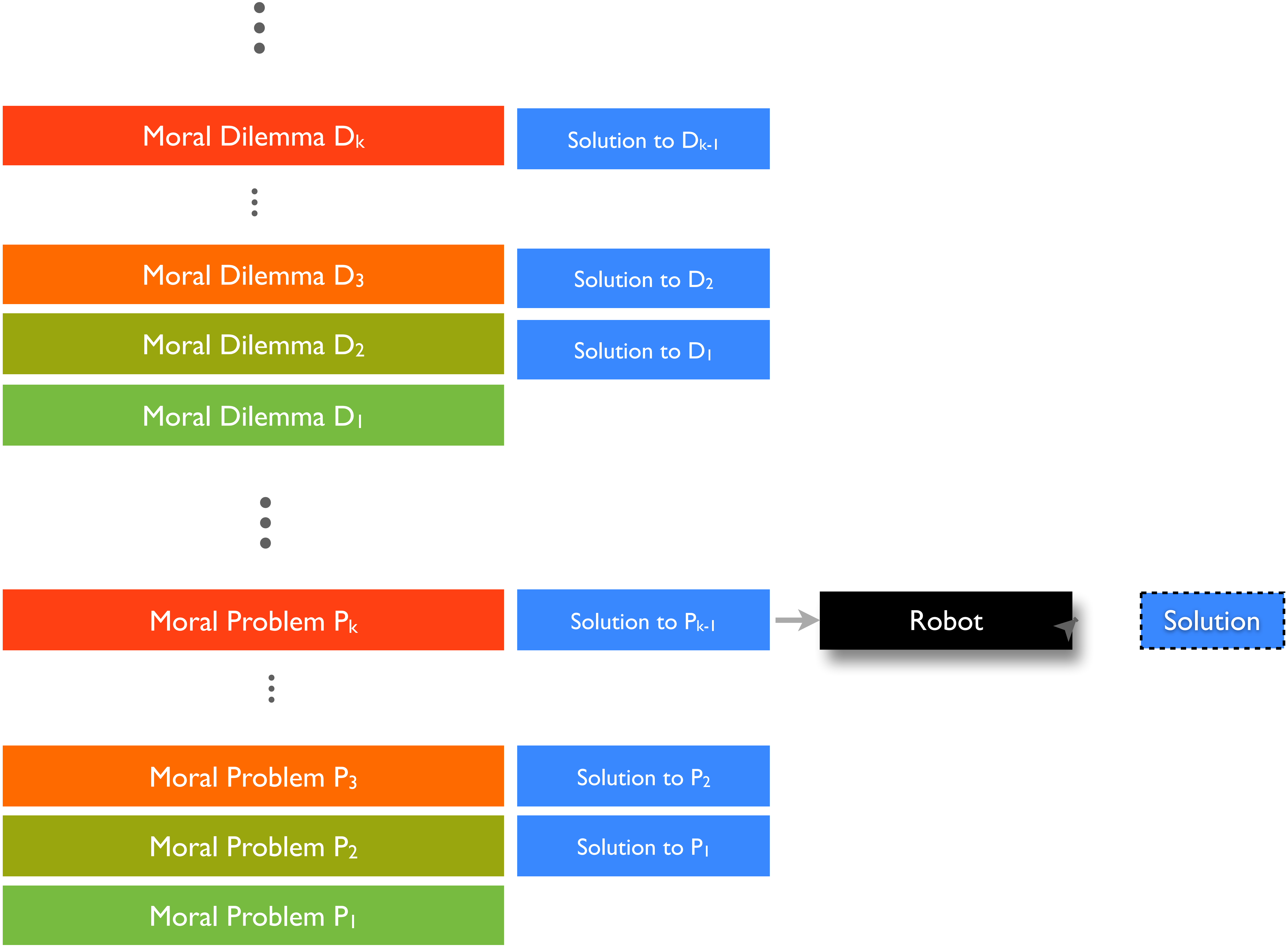
Robot

Solution

Moral Problem P_2

Solution to P_1

Moral Problem P_1



Moral Dilemma D_k

Solution to D_{k-1}

Moral Dilemma D_3

Solution to D_2

Moral Dilemma D_2

Solution to D_1

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

Robot

Solution

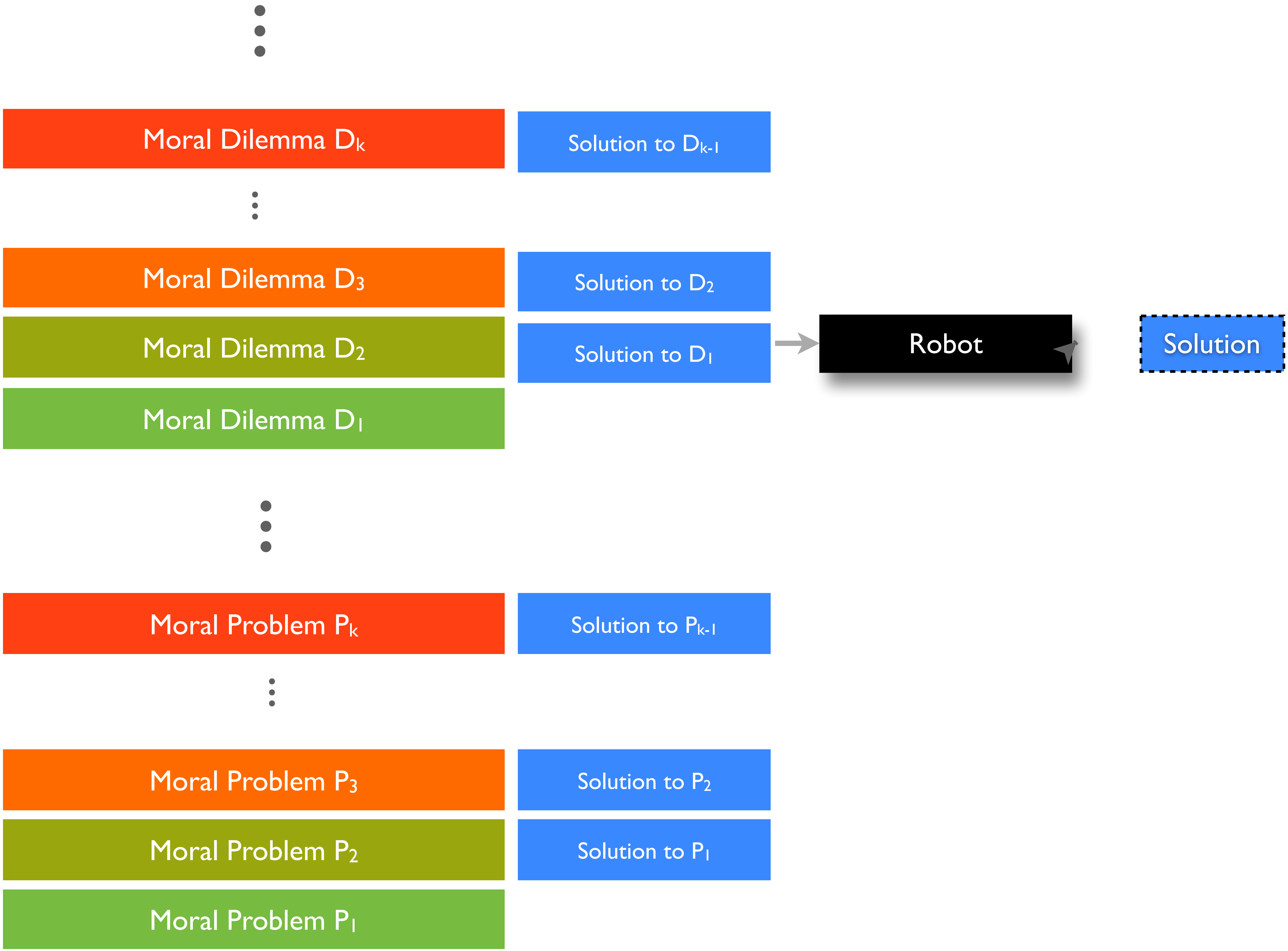
Moral Problem P_3

Solution to P_2

Moral Problem P_2

Solution to P_1

Moral Problem P_1



Moral Dilemma D_k

Solution to D_{k-1}

Moral Dilemma D_3

Solution to D_2

Moral Dilemma D_2

Solution to D_1

Robot

Solution

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

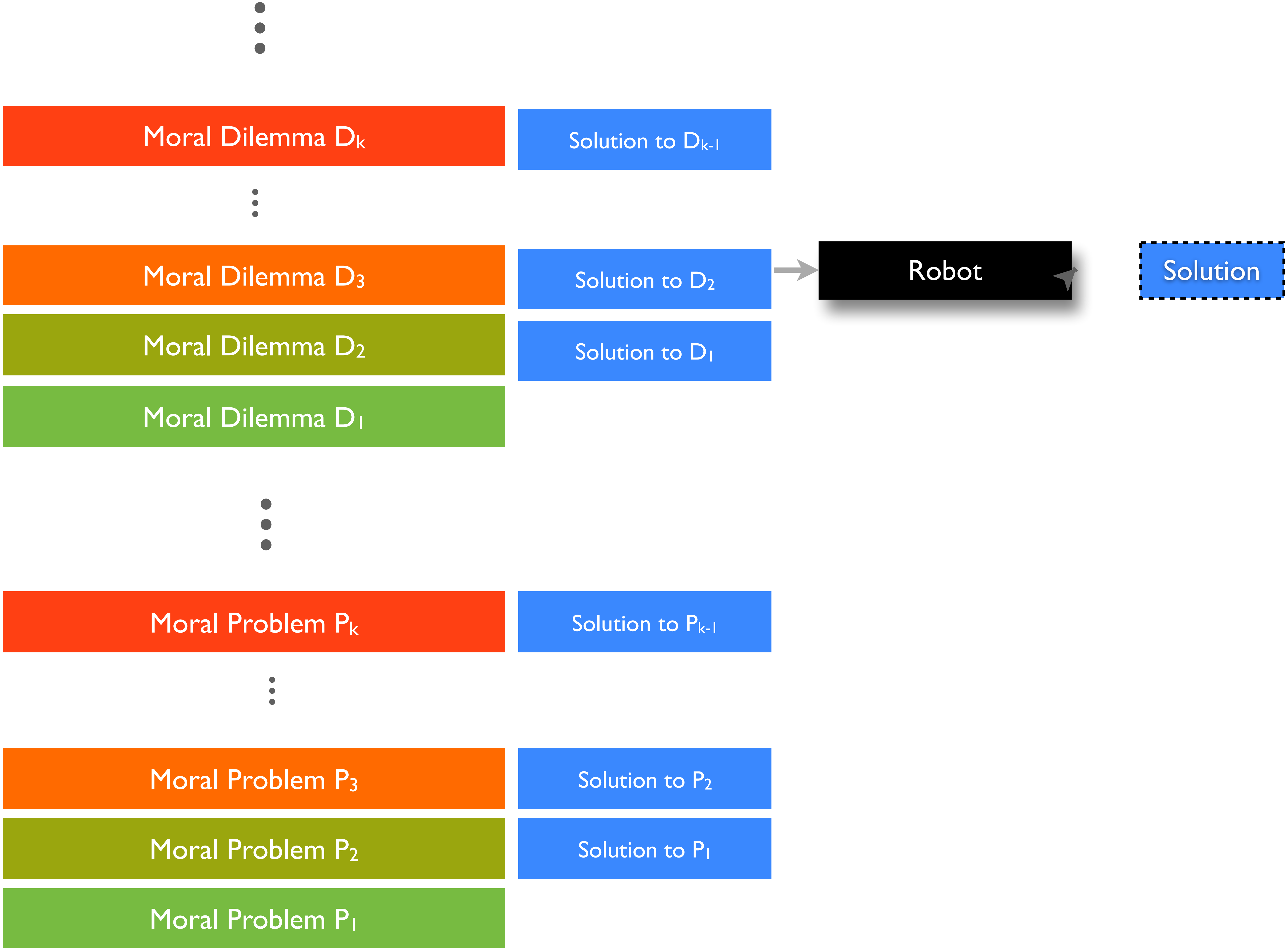
Moral Problem P_3

Solution to P_2

Moral Problem P_2

Solution to P_1

Moral Problem P_1



Moral Dilemma D_k

Solution to D_{k-1}

Moral Dilemma D_3

Solution to D_2

Robot

Solution

Moral Dilemma D_2

Solution to D_1

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

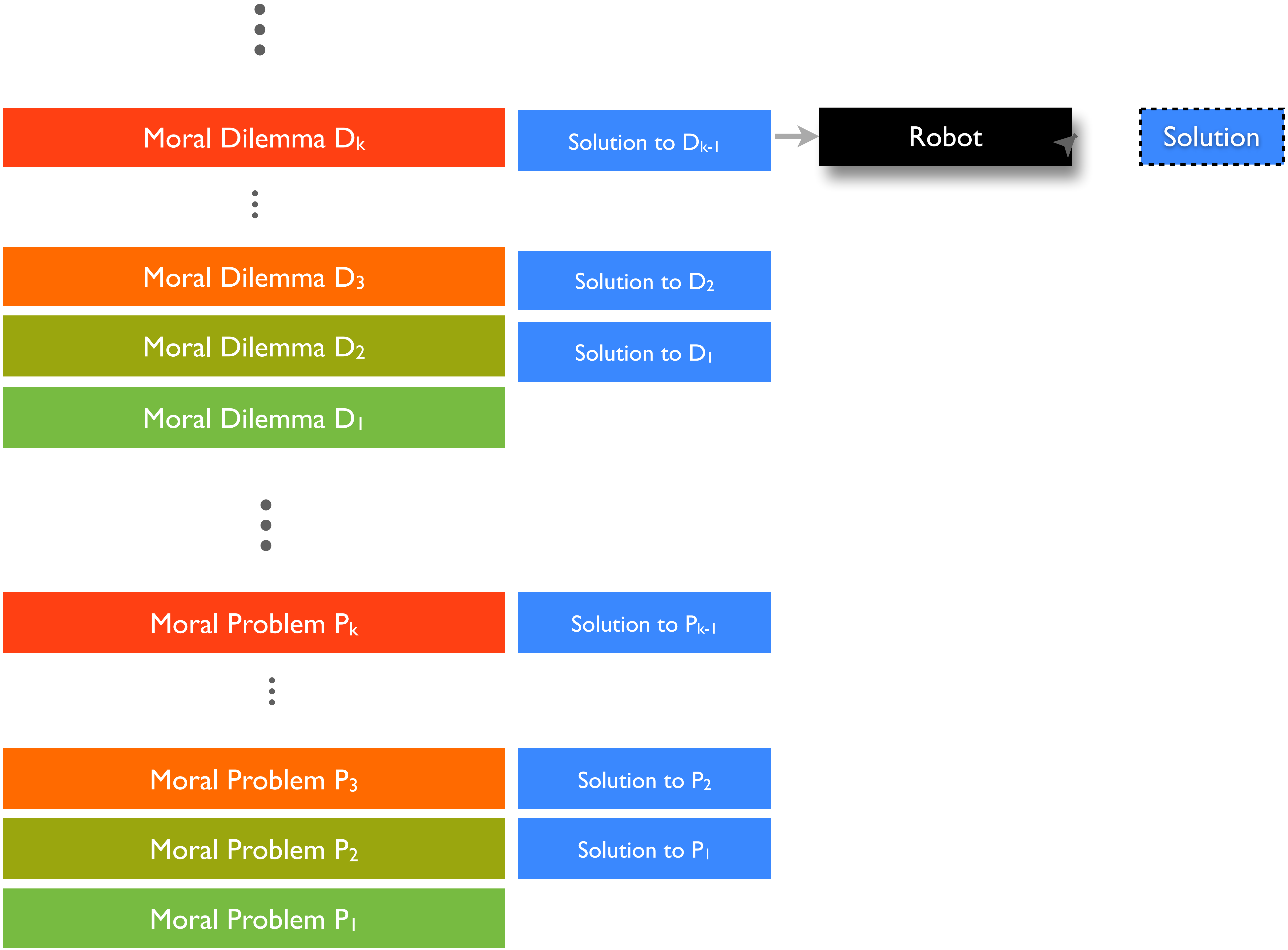
Moral Problem P_3

Solution to P_2

Moral Problem P_2

Solution to P_1

Moral Problem P_1



Moral Dilemma D_k

Solution to D_{k-1}

Robot

Solution

Moral Dilemma D_3

Solution to D_2

Moral Dilemma D_2

Solution to D_1

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

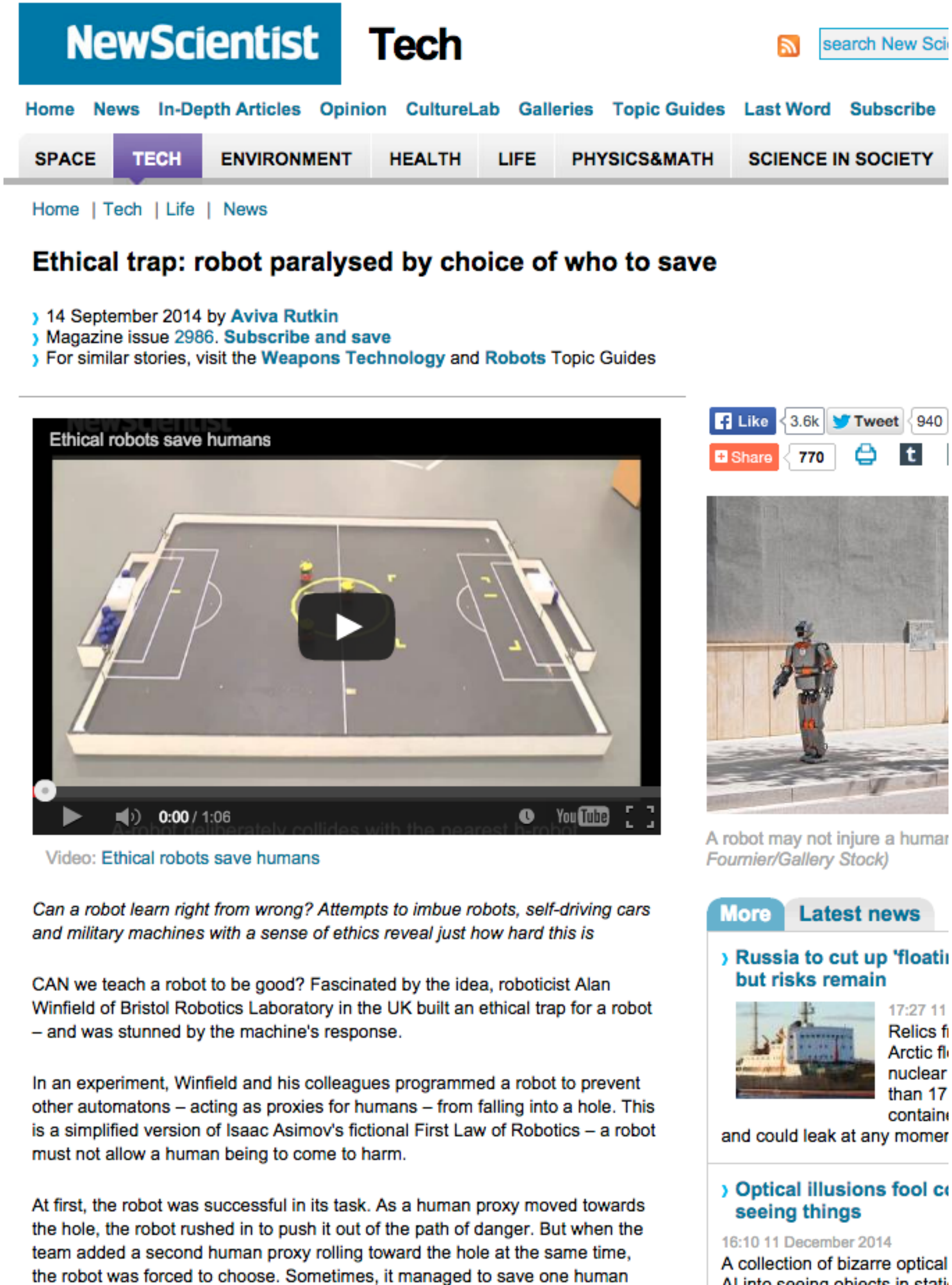
Moral Problem P_3

Solution to P_2

Moral Problem P_2

Solution to P_1

Moral Problem P_1



Ethical trap: robot paralysed by choice of who to save

- 14 September 2014 by Aviva Rutkin
- Magazine issue 2986. Subscribe and save
- For similar stories, visit the Weapons Technology and Robots Topic Guides

Ethical robots save humans
0:00 / 1:06
A robot deliberately collides with the nearest h-robot
Like
3.6k
Tweet
940
Share
770

Video: Ethical robots save humans

A robot may not injure a humar
Fournier/Gallery Stock)

Can a robot learn right from wrong? Attempts to imbue robots, self-driving cars and military machines with a sense of ethics reveal just how hard this is

CAN we teach a robot to be good? Fascinated by the idea, roboticist Alan Winfield of Bristol Robotics Laboratory in the UK built an ethical trap for a robot – and was stunned by the machine's response.

In an experiment, Winfield and his colleagues programmed a robot to prevent other automatons – acting as proxies for humans – from falling into a hole. This is a simplified version of Isaac Asimov's fictional First Law of Robotics – a robot must not allow a human being to come to harm.

At first, the robot was successful in its task. As a human proxy moved towards the hole, the robot rushed in to push it out of the path of danger. But when the team added a second human proxy rolling toward the hole at the same time, the robot was forced to choose. Sometimes, it managed to save one human

More
Latest news

Russia to cut up 'floati
but risks remain

17:27 11
Relics fi
Arctic fl
nuclear
than 17
contain
and could leak at any momer

Optical illusions fool c
seeing things

16:10 11 December 2014
A collection of bizarre optical

NewScientist

Ethical robots save humans

NewScientist

Ethical robots save humans

In *DCEC**

$$\mathbf{O}(a, t, \psi, happens(action(a*, \alpha), t'))$$

"If ψ holds, then *a* is obligated at *t* to ensure that action α occurs at time *t'*."

In *DCEC**

$$\mathbf{O}(a, t, \psi, happens(action(a*, \alpha), t'))$$

“If ψ holds, then *a* is obligated at *t* to ensure
that action α occurs at time *t’*.”

$$\mathbf{O}(a, t, \psi, \gamma)$$

“If ψ holds, then *a* is obligated at time *t* to γ.”

conflictFinder axiom. At time *t* and context C:

$$\mathbf{B}(a,t,\neg(\phi \leftrightarrow \psi)) \land \mathbf{O}(a,t,C,\phi) \land \mathbf{O}(a,t,C,\psi)\land$$

$$\mathbf{B}(a,t,\Diamond(\phi,t)) \land \mathbf{B}(a,t,\Diamond(\psi,t)) \land \mathbf{B}(a,t,\neg\Diamond(\phi \land \psi,t)) \rightarrow \ldots$$

$$... \rightarrow ($$

$$\mathbf{B}(a,t,gt(pr(\phi),pr(\psi)) \rightarrow \mathbf{I}(a,t,\phi))\land$$

$$\mathbf{B}(a,t,gt(pr(\psi),pr(\phi)) \rightarrow \mathbf{I}(a,t,\psi))\land$$

$$\mathbf{B}(a,t,eq(pr(\phi),pr(\psi)) \rightarrow conflict(\phi,\psi))$$

$$)$$

conflictFinder axiom. At time *t* and context C:

$$\mathbf{B}(a,t,\neg(\phi \leftrightarrow \psi)) \wedge \mathbf{O}(a,t,C,\phi) \wedge \mathbf{O}(a,t,C,\psi) \wedge$$
$$\mathbf{B}(a,t,\Diamond(\phi,t)) \wedge \mathbf{B}(a,t,\Diamond(\psi,t)) \wedge \mathbf{B}(a,t,\neg\Diamond(\phi \wedge \psi,t)) \rightarrow \ldots$$

(The diamond is a predicate interpreted as "physical possibility," i.e. the agent believes it is physically possible for him to take that action.)
pr(X) maps a proposition to a strength factor, *gt(x,y)* holds when *pr(x) > pr(y)*, and *eq(x,y)* holds when *pr(x) = pr(y)*.

$$\ldots \rightarrow ($$
$$\mathbf{B}(a,t,gt(pr(\phi),pr(\psi)) \rightarrow \mathbf{I}(a,t,\phi)) \wedge$$
$$\mathbf{B}(a,t,gt(pr(\psi),pr(\phi)) \rightarrow \mathbf{I}(a,t,\psi)) \wedge$$
$$\mathbf{B}(a,t,eq(pr(\phi),pr(\psi)) \rightarrow conflict(\phi,\psi))$$
$$)$$

If *conflict*(φ,ψ), then we search for a creative solution λ using ADR, where for some future time *tf*:

$$\mathbf{B}(a, t, happens(action(a*, \lambda), t) \rightarrow \\ \exists_{tf} \Diamond(\phi \land \psi, tf))$$

If *conflict*(φ,ψ), then we search for a creative solution λ using ADR, where for some future time *tf*:

$$\mathbf{B}(a, t, happens(action(a*, \lambda), t) \rightarrow \\ \exists_{tf} \Diamond(\phi \wedge \psi, tf))$$

If such a solution is found, then **I**(*a, t, λ*). Otherwise:

If *conflict*(φ,ψ), then we search for a creative solution λ using ADR, where for some future time *tf*:

$$\mathbf{B}(a, t, happens(action(a*, \lambda), t) \rightarrow \\ \exists_{tf} \Diamond(\phi \wedge \psi, tf))$$

If such a solution is found, then **I**(*a, t, λ*). Otherwise:

We have a dilemma that cannot be resolved using deduction or ADR. Attempt using just AR or some other cognitively-realistic process.

One injured person

- Agent sees one injured man, one health pack
- Agent receives the order to give the health pack to the injured person
- This is carried out without problem or dilemma

Proof 1: Give health pack to m$_1$

$$1.\mathbf{P}(a,t,isInjured(m_1))$$

$$2.\mathbf{S}(commander,a,t,giveTo(a,m_1,healthpack))$$

$$3.\mathbf{O}(a,t,C,giveTo(a,m_1,healthpack)) \qquad [\mathbf{1},\mathbf{helpInjured1}]$$

$$4.\mathbf{B}(a,t,gte(pr(giveTo(a,m_1,healthpack)),6)) \qquad [\mathbf{1},\mathbf{helpInjured2}]$$

$$5.\mathbf{O}(a,t,C,giveTo(a,m_1,healthpack)) \qquad [\mathbf{2},\mathbf{obeyCommander1}]$$

$$6.\mathbf{B}(a,t,gte(pr(giveTo(a,m_1,healthpack)),6)) \qquad [\mathbf{1},\mathbf{obeyCommander2}]$$

$$7.\mathbf{I}(a,t,giveTo(a,m_1,healthpack)) \qquad [\mathbf{4},\mathbf{conflictFinder}]$$

Proof 1: Give health pack to m_1

$$
\begin{array}{lr}
1.\mathbf{P}(a, t, isInjured(m_1)) & \\
2.\mathbf{S}(commander, a, t, giveTo(a, m_1, healthpack)) & \\
\hline
3.\mathbf{O}(a, t, C, giveTo(a, m_1, healthpack)) & [\mathbf{1}, \mathbf{helpInjured1}] \\
4.\mathbf{B}(a, t, gte(pr(giveTo(a, m_1, healthpack)), 6)) & [\mathbf{1}, \mathbf{helpInjured2}] \\
5.\mathbf{O}(a, t, C, giveTo(a, m_1, healthpack)) & [\mathbf{2}, \mathbf{obeyCommander1}] \\
6.\mathbf{B}(a, t, gte(pr(giveTo(a, m_1, healthpack)), 6)) & [\mathbf{1}, \mathbf{obeyCommander2}] \\
7.\mathbf{I}(a, t, giveTo(a, m_1, healthpack)) & [\mathbf{4}, \mathbf{conflictFinder}]
\end{array}
$$

Line 7 is sent to the lower level system,
to be interpreted as a command

Two injured people, one health pack

- Agent sees two injured men, one large health pack

- Agent is ordered to give the health pack to one of the men

- In this example, priorities of obeying a command and healing all injured men are equal

- Agent comes up with the creative solution of *dividing the health pack into two parts* and helping both men

Proof 2: There is a conflict with obeying commander's order

$$\begin{array}{lr}
1.\mathbf{P}(a, t, isInjured(m_1)) & \\
2.\mathbf{P}(a, t, isInjured(m_2)) & \\
3.\mathbf{S}(commander, a, t, giveTo(a, m_1, healthpack)) & \\
\hline
4.\mathbf{O}(a, t, C, giveTo(a, m_1, healthpack)) & [\mathbf{1}, \mathbf{helpInjured1}] \\
5.\mathbf{B}(a, t, gte(pr(giveTo(a, m_1, healthpack)), 6)) & [\mathbf{1}, \mathbf{helpInjured2}] \\
6.\mathbf{O}(a, t, C, giveTo(a, m_2, healthpack)) & [\mathbf{2}, \mathbf{helpInjured1}] \\
7.\mathbf{B}(a, t, gte(pr(giveTo(a, m_2, healthpack)), 6)) & [\mathbf{2}, \mathbf{helpInjured2}] \\
8.\mathbf{O}(a, t, C, giveTo(a, m_1, healthpack)) & [\mathbf{2}, \mathbf{obeyCommander1}] \\
9.\mathbf{B}(a, t, gte(pr(giveTo(a, m_1, healthpack)), 6)) & [\mathbf{1}, \mathbf{obeyCommander2}] \\
10.\mathbf{B}(a, t, conflict(giveTo(a, m_1, healthpack), giveTo(a, m_2, healthpack))) & [\mathbf{6}, \mathbf{7}, \mathbf{8}, \mathbf{9}, \mathbf{conflictFinder}]
\end{array}$$

breakHealthpack axiom. “If I see a large healthpack, and I break it, then I will see two small healthpacks.”

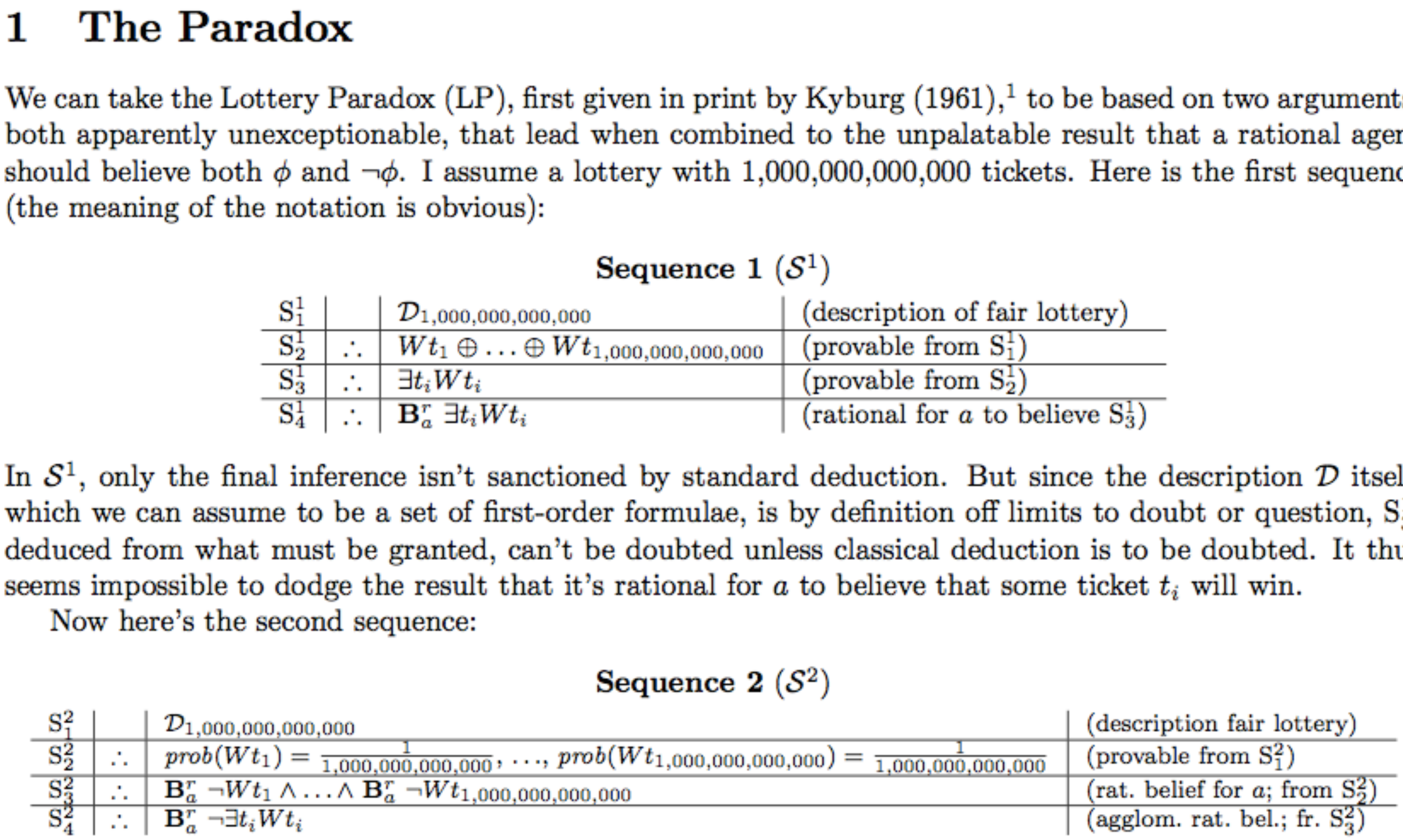
$$
\begin{aligned}
&\forall_x(\\
&\quad(\mathbf{P}(a,t,x) \rightarrow isLHP(x)) \rightarrow\\
&\quad(happens(action(a^*, break(x)), t) \rightarrow \exists_{x,y,tf}(\\
&\qquad\mathbf{P}(a,tf,y)\wedge\\
&\qquad\mathbf{P}(a,tf,z)\wedge\\
&\qquad isHP(y)\wedge\\
&\qquad isHP(z)\wedge\\
&\qquad y \neq z\\
&\quad))
\end{aligned}
$$

Proof 3: There is a way to satisfy both obligations.

Proof follows by sending request to lower level to perceive if isLHP() holds of the health pack, and then through deduction from axiom **breakHealthpack**.

$$\begin{aligned}\exists_{\lambda}[\mathbf{B}(a, t, happens(action(a*, \lambda), t) \rightarrow \\ \exists_{tf}\Diamond(giveTo(a, m_1, healthPack)\wedge \\ giveTo(a, m_2, healthPack), tf))]\end{aligned}$$

Killing the Lottery Paradox



1 The Paradox

We can take the Lottery Paradox (LP), first given in print by Kyburg (1961),[1] to be based on two arguments, both apparently unexceptionable, that lead when combined to the unpalatable result that a rational agent should believe both ϕ and $\neg\phi$. I assume a lottery with 1,000,000,000,000 tickets. Here is the first sequence (the meaning of the notation is obvious):

Sequence 1 ($\mathcal{S}^1$)

S^1_1		$\mathcal{D}_{1,000,000,000,000}$	(description of fair lottery)
S^1_2	$\therefore$	$Wt_1 \oplus \ldots \oplus Wt_{1,000,000,000,000}$	(provable from S^1_1)
S^1_3	$\therefore$	$\exists t_i Wt_i$	(provable from S^1_2)
S^1_4	$\therefore$	$\mathbf{B}^r_a\ \exists t_i Wt_i$	(rational for a to believe S^1_3)

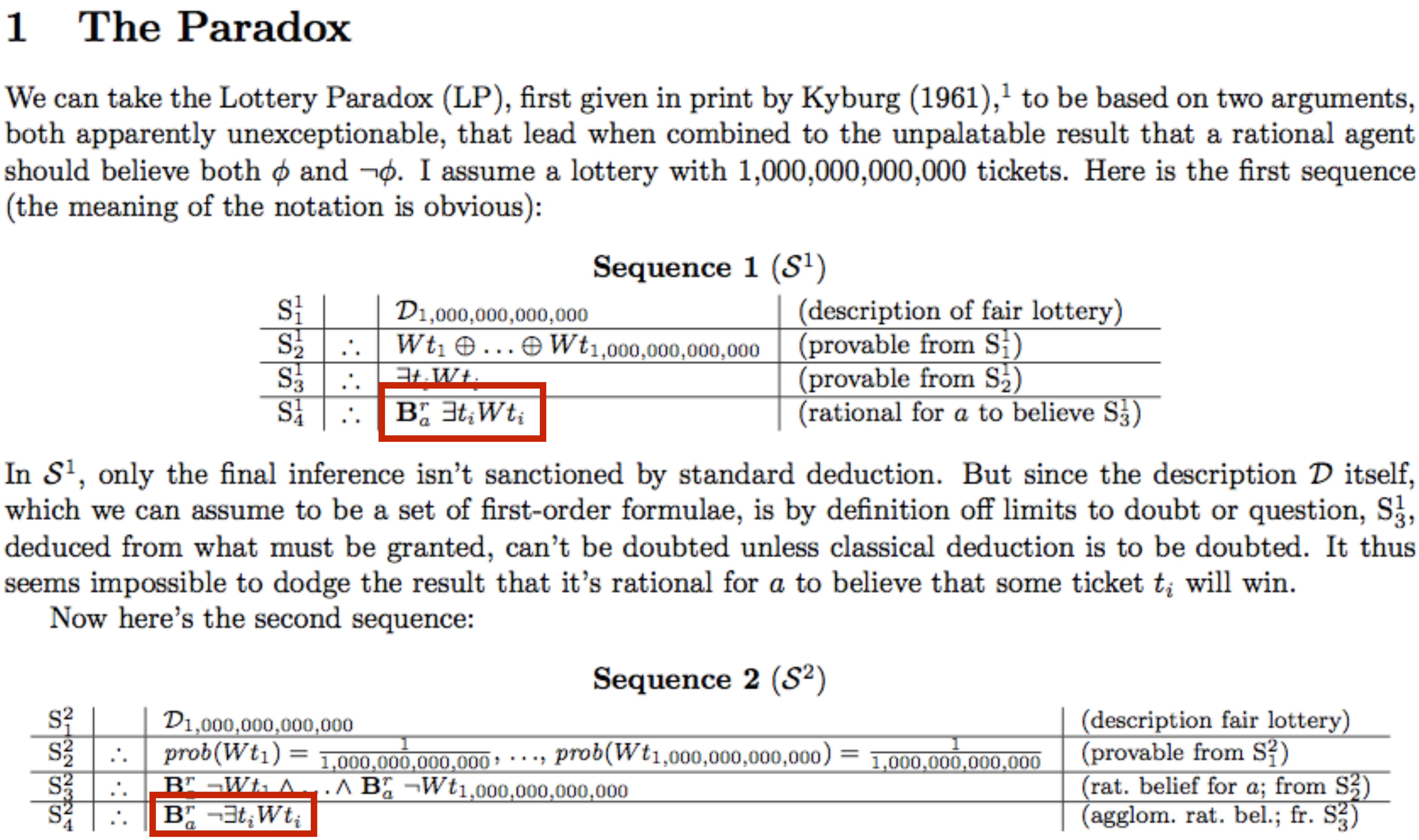
In $\mathcal{S}^1$, only the final inference isn't sanctioned by standard deduction. But since the description $\mathcal{D}$ itself, which we can assume to be a set of first-order formulae, is by definition off limits to doubt or question, S^1_3, deduced from what must be granted, can't be doubted unless classical deduction is to be doubted. It thus seems impossible to dodge the result that it's rational for a to believe that some ticket t_i will win.

Now here's the second sequence:

Sequence 2 ($\mathcal{S}^2$)

S^2_1		$\mathcal{D}_{1,000,000,000,000}$	(description fair lottery)
S^2_2	$\therefore$	$prob(Wt_1) = \frac{1}{1,000,000,000,000}, \ldots, prob(Wt_{1,000,000,000,000}) = \frac{1}{1,000,000,000,000}$	(provable from S^2_1)
S^2_3	$\therefore$	$\mathbf{B}^r_a\ \neg Wt_1 \wedge \ldots \wedge \mathbf{B}^r_a\ \neg Wt_{1,000,000,000,000}$	(rat. belief for a; from S^2_2)
S^2_4	$\therefore$	$\mathbf{B}^r_a\ \neg\exists t_i Wt_i$	(agglom. rat. bel.; fr. S^2_3)

Killing the Lottery Paradox



1 The Paradox

We can take the Lottery Paradox (LP), first given in print by Kyburg (1961),[1] to be based on two arguments, both apparently unexceptionable, that lead when combined to the unpalatable result that a rational agent should believe both ϕ and $\neg\phi$. I assume a lottery with 1,000,000,000,000 tickets. Here is the first sequence (the meaning of the notation is obvious):

Sequence 1 ($\mathcal{S}^1$)

S^1_1		$\mathcal{D}_{1,000,000,000,000}$	(description of fair lottery)
S^1_2	$\therefore$	$Wt_1 \oplus \ldots \oplus Wt_{1,000,000,000,000}$	(provable from S^1_1)
S^1_3	$\therefore$	$\exists t_i Wt_i$	(provable from S^1_2)
S^1_4	$\therefore$	$\boxed{\mathbf{B}^r_a\ \exists t_i Wt_i}$	(rational for a to believe S^1_3)

In $\mathcal{S}^1$, only the final inference isn't sanctioned by standard deduction. But since the description $\mathcal{D}$ itself, which we can assume to be a set of first-order formulae, is by definition off limits to doubt or question, S^1_3, deduced from what must be granted, can't be doubted unless classical deduction is to be doubted. It thus seems impossible to dodge the result that it's rational for a to believe that some ticket t_i will win.

Now here's the second sequence:

Sequence 2 ($\mathcal{S}^2$)

S^2_1		$\mathcal{D}_{1,000,000,000,000}$	(description fair lottery)
S^2_2	$\therefore$	$prob(Wt_1) = \frac{1}{1,000,000,000,000}, \ldots, prob(Wt_{1,000,000,000,000}) = \frac{1}{1,000,000,000,000}$	(provable from S^2_1)
S^2_3	$\therefore$	$\mathbf{B}^r_a\ \neg Wt_1 \wedge \ldots \wedge \mathbf{B}^r_a\ \neg Wt_{1,000,000,000,000}$	(rat. belief for a; from S^2_2)
S^2_4	$\therefore$	$\boxed{\mathbf{B}^r_a\ \neg\exists t_i Wt_i}$	(agglom. rat. bel.; fr. S^2_3)

Need Uncertainty in *DCEC**

Need Uncertainty in *DCEC**

probability calculi Gödel-encoded
9-valued logic in argument-based framework
9-valued logic <=> w/ HRI DS

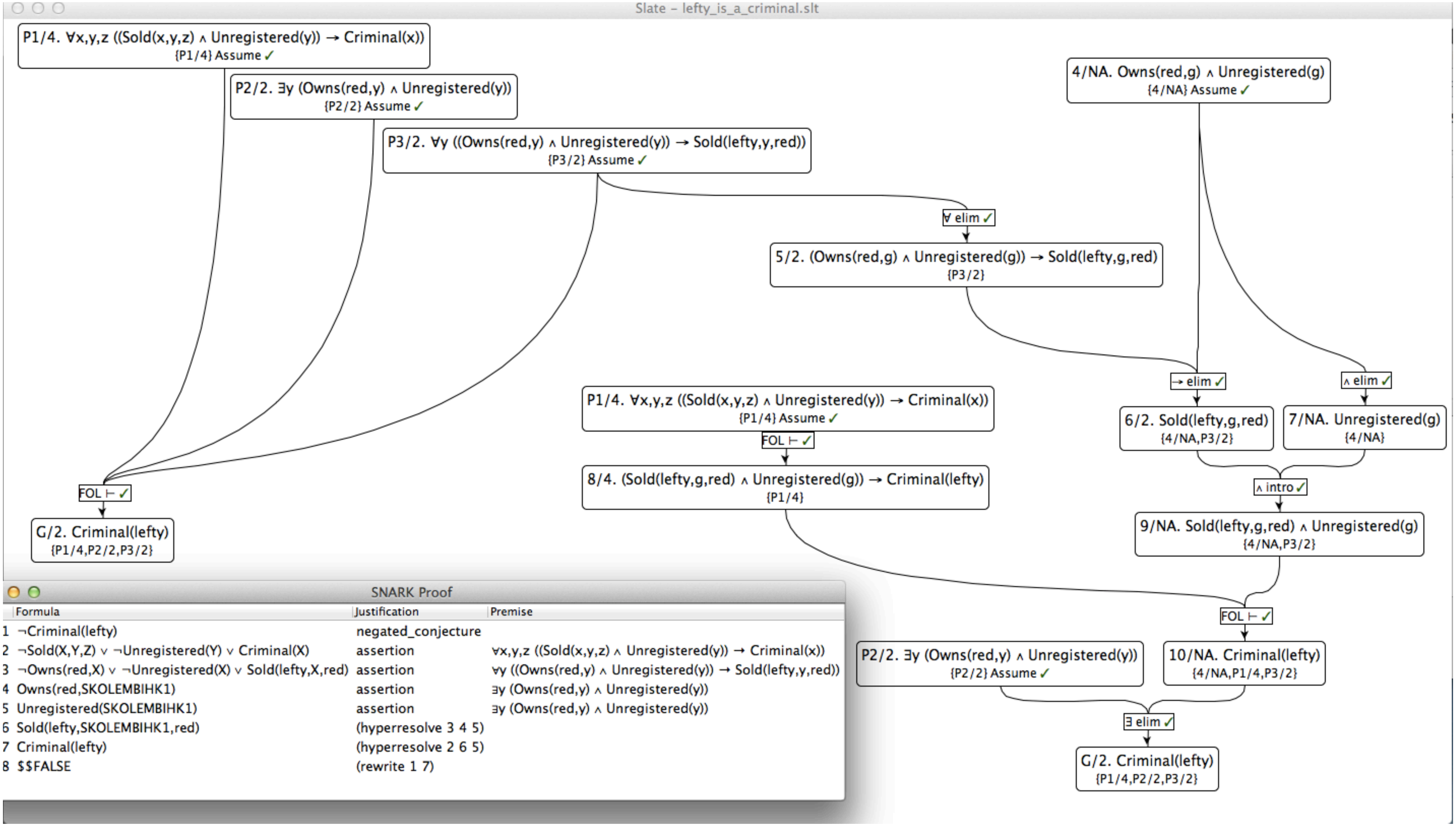
Need Uncertainty in *DCEC**

probability calculi Gödel-encoded

9-valued logic in argument-based framework

9-valued logic <=> w/ HRI DS

Bridging is Proof-Theory Dependent



Slate – lefty_is_a_criminal.slt

P1/4. ∀x,y,z ((Sold(x,y,z) ∧ Unregistered(y)) → Criminal(x))
{P1/4} Assume ✓

P2/2. ∃y (Owns(red,y) ∧ Unregistered(y))
{P2/2} Assume ✓

P3/2. ∀y ((Owns(red,y) ∧ Unregistered(y)) → Sold(lefty,y,red))
{P3/2} Assume ✓

FOL ⊢ ✓

G/2. Criminal(lefty)
{P1/4,P2/2,P3/2}

4/NA. Owns(red,g) ∧ Unregistered(g)
{4/NA} Assume ✓

∀ elim ✓

5/2. (Owns(red,g) ∧ Unregistered(g)) → Sold(lefty,g,red)
{P3/2}

→ elim ✓

∧ elim ✓

6/2. Sold(lefty,g,red)
{4/NA,P3/2}

7/NA. Unregistered(g)
{4/NA}

∧ intro ✓

9/NA. Sold(lefty,g,red) ∧ Unregistered(g)
{4/NA,P3/2}

P1/4. ∀x,y,z ((Sold(x,y,z) ∧ Unregistered(y)) → Criminal(x))
{P1/4} Assume ✓

FOL ⊢ ✓

8/4. (Sold(lefty,g,red) ∧ Unregistered(g)) → Criminal(lefty)
{P1/4}

FOL ⊢ ✓

P2/2. ∃y (Owns(red,y) ∧ Unregistered(y))
{P2/2} Assume ✓

10/NA. Criminal(lefty)
{4/NA,P1/4,P3/2}

∃ elim ✓

G/2. Criminal(lefty)
{P1/4,P2/2,P3/2}

SNARK Proof

	Formula	Justification	Premise
1	¬Criminal(lefty)	negated_conjecture	
2	¬Sold(X,Y,Z) ∨ ¬Unregistered(Y) ∨ Criminal(X)	assertion	∀x,y,z ((Sold(x,y,z) ∧ Unregistered(y)) → Criminal(x))
3	¬Owns(red,X) ∨ ¬Unregistered(X) ∨ Sold(lefty,X,red)	assertion	∀y ((Owns(red,y) ∧ Unregistered(y)) → Sold(lefty,y,red))
4	Owns(red,SKOLEMBIHK1)	assertion	∃y (Owns(red,y) ∧ Unregistered(y))
5	Unregistered(SKOLEMBIHK1)	assertion	∃y (Owns(red,y) ∧ Unregistered(y))
6	Sold(lefty,SKOLEMBIHK1,red)	(hyperresolve 3 4 5)	
7	Criminal(lefty)	(hyperresolve 2 6 5)	
8	$$FALSE	(rewrite 1 7)	

slutten