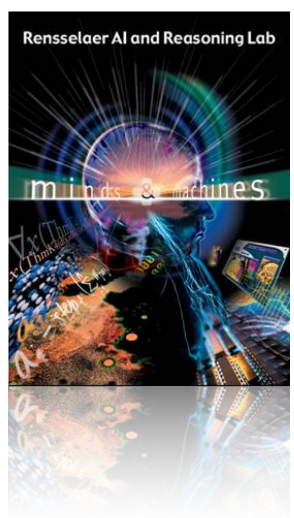
Robot-Ethics Background for OFAI Position Paper (“Engineer at the level of the OS!”)

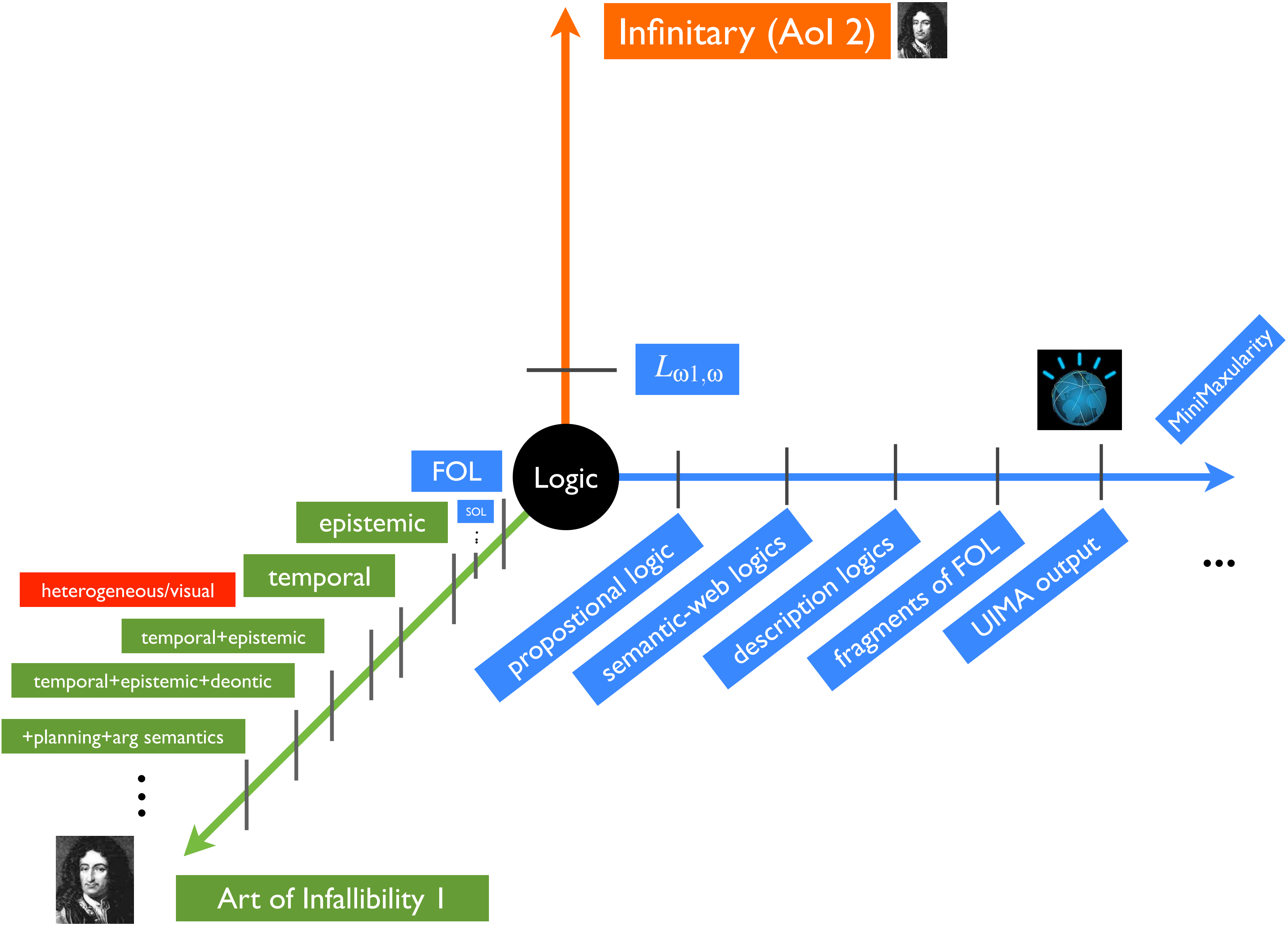
Selmer Bringsjord[1] • Naveen Sundar G.[2]
Rensselaer AI & Reasoning (RAIR) Lab[1,2]
Department of Cognitive Science[1]
Department of Computer Science[1,2]
Lally School of Management & Technology[1]
Rensselaer Polytechnic Institute (RPI)
Troy, New York 12180 USA

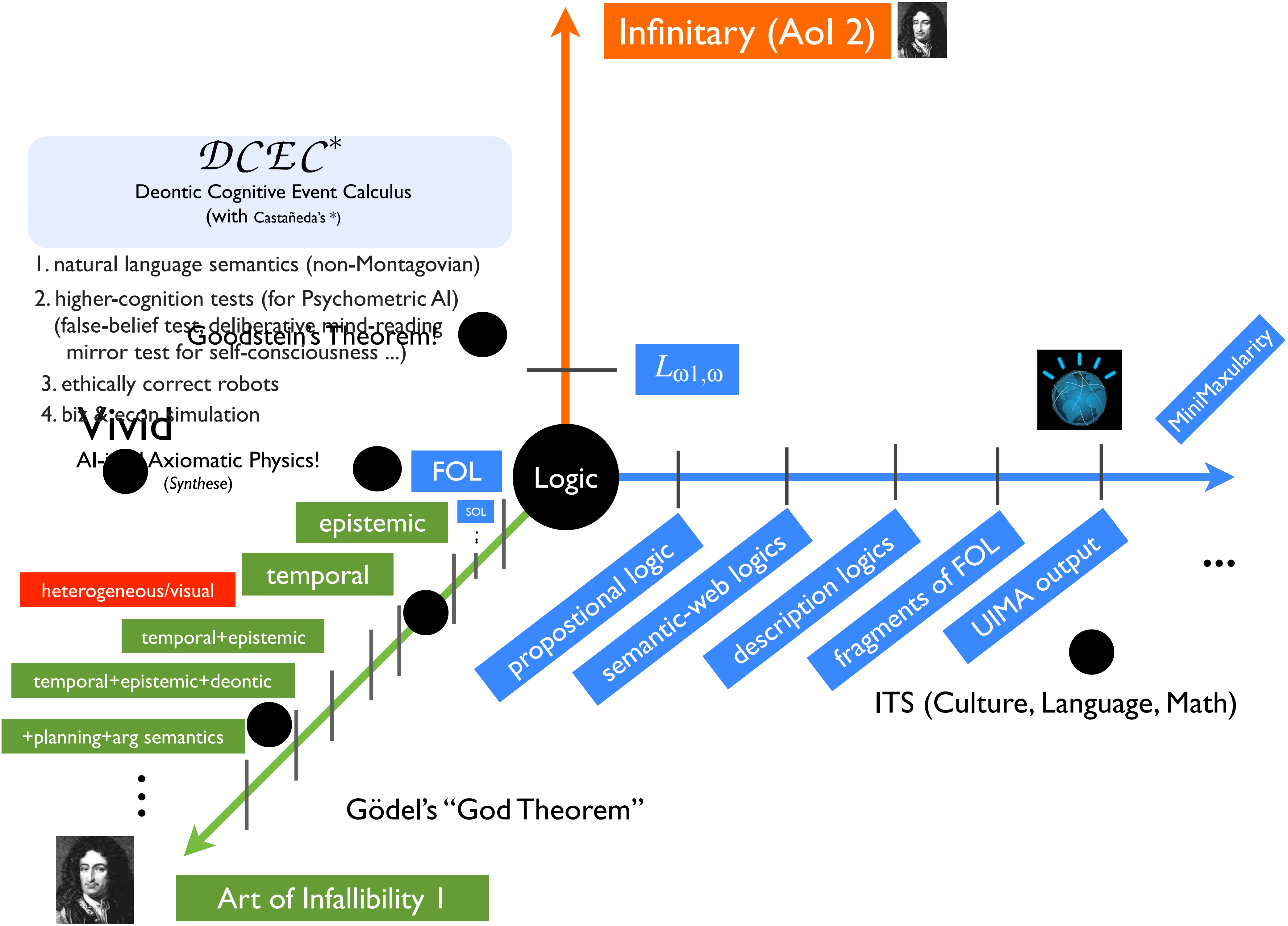
OFAI
Vienna, AT
9/27/2013



Made possible by the generosity of:

IBM
RPI’s pursuit of ITS/IILE
NSF
JTF
AFOSR
ONR x 2
DARPA


Infinitary (AoI 2)
$L_{\omega1,\omega}$
MiniMaxularity
FOL
Logic
SOL
epistemic
temporal
heterogeneous/visual
temporal+epistemic
temporal+epistemic+deontic
+planning+arg semantics
propostional logic
semantic-web logics
description logics
fragments of FOL
UIMA output
...
Art of Infallibility I

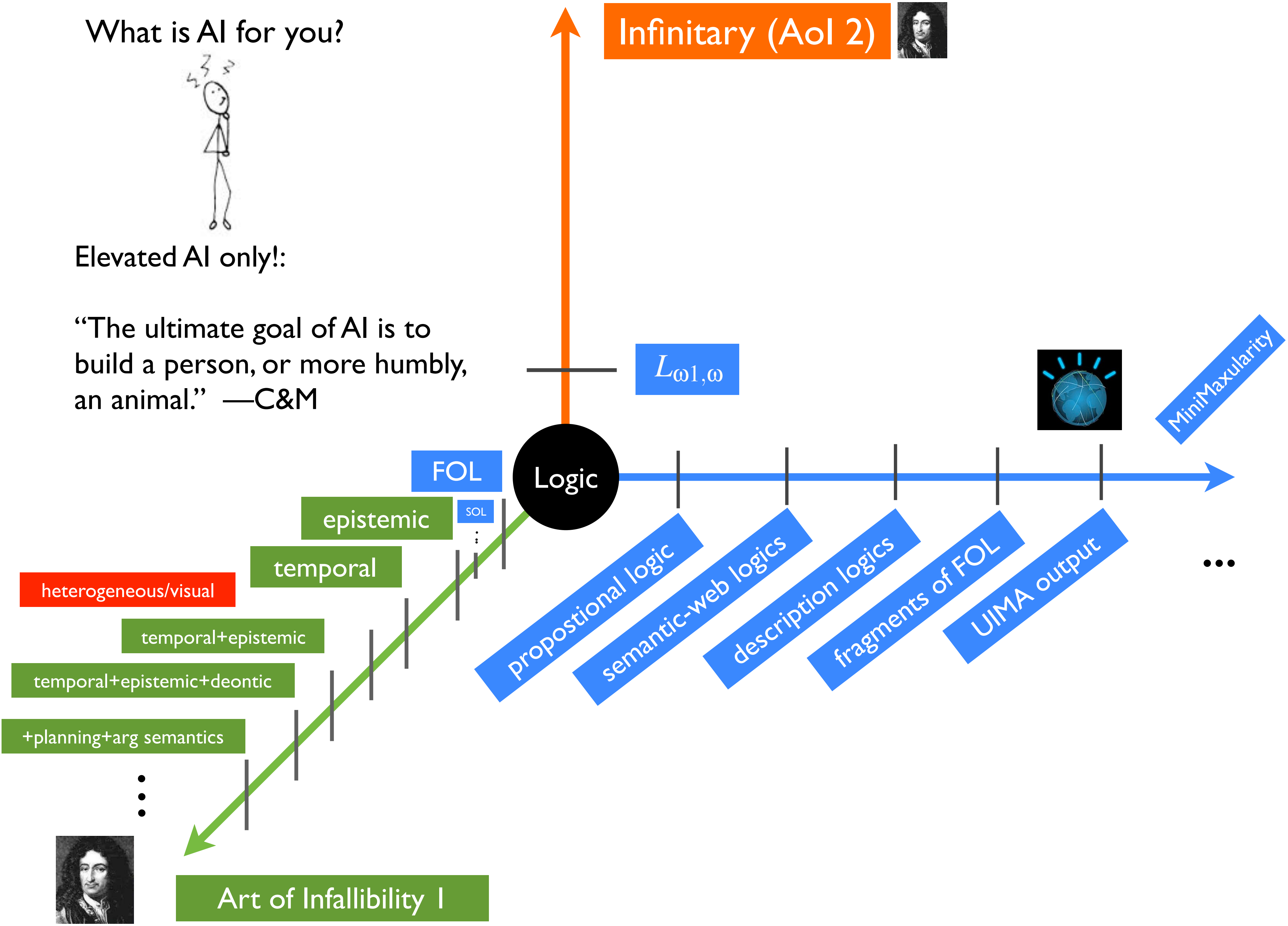


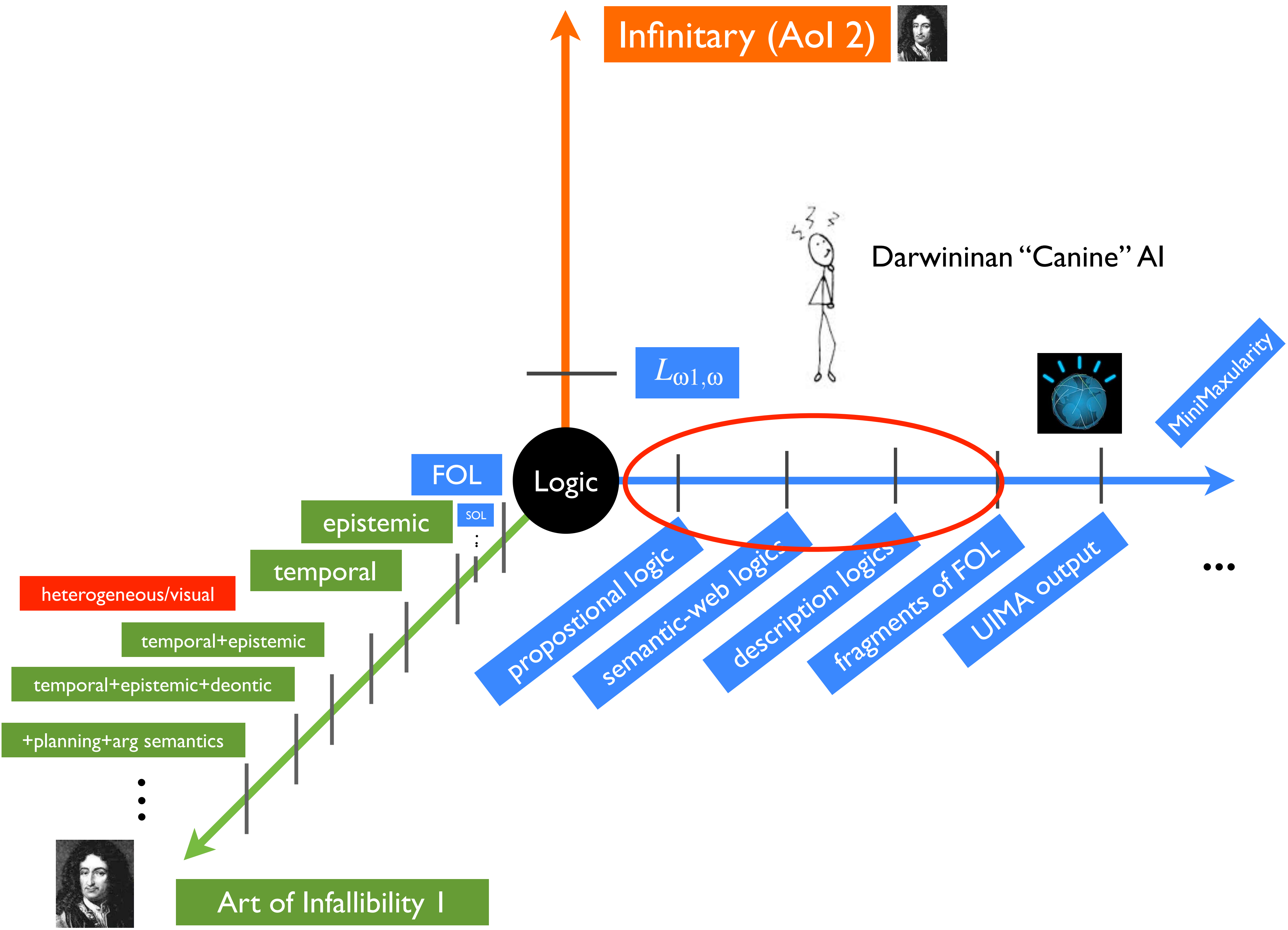
Infinitary (AoI 2)
DCEC*
Deontic Cognitive Event Calculus
(with Castañeda's *)
1. natural language semantics (non-Montagovian)
2. higher-cognition tests (for Psychometric AI)
(false-belief test, deliberative mind-reading
mirror test for self-consciousness ...)
3. ethically correct robots
4. biz & econ simulation
Goodstein's Theorem!
Vivid
Axiomatic Physics!
(Synthese)
$L_{\omega 1,\omega}$
FOL
Logic
SOL
epistemic
temporal
heterogeneous/visual
temporal+epistemic
temporal+epistemic+deontic
+planning+arg semantics
propostional logic
semantic-web logics
description logics
fragments of FOL
UIMA output
MiniMaxularity
...
ITS (Culture, Language, Math)
Gödel's "God Theorem"
Art of Infallibility 1

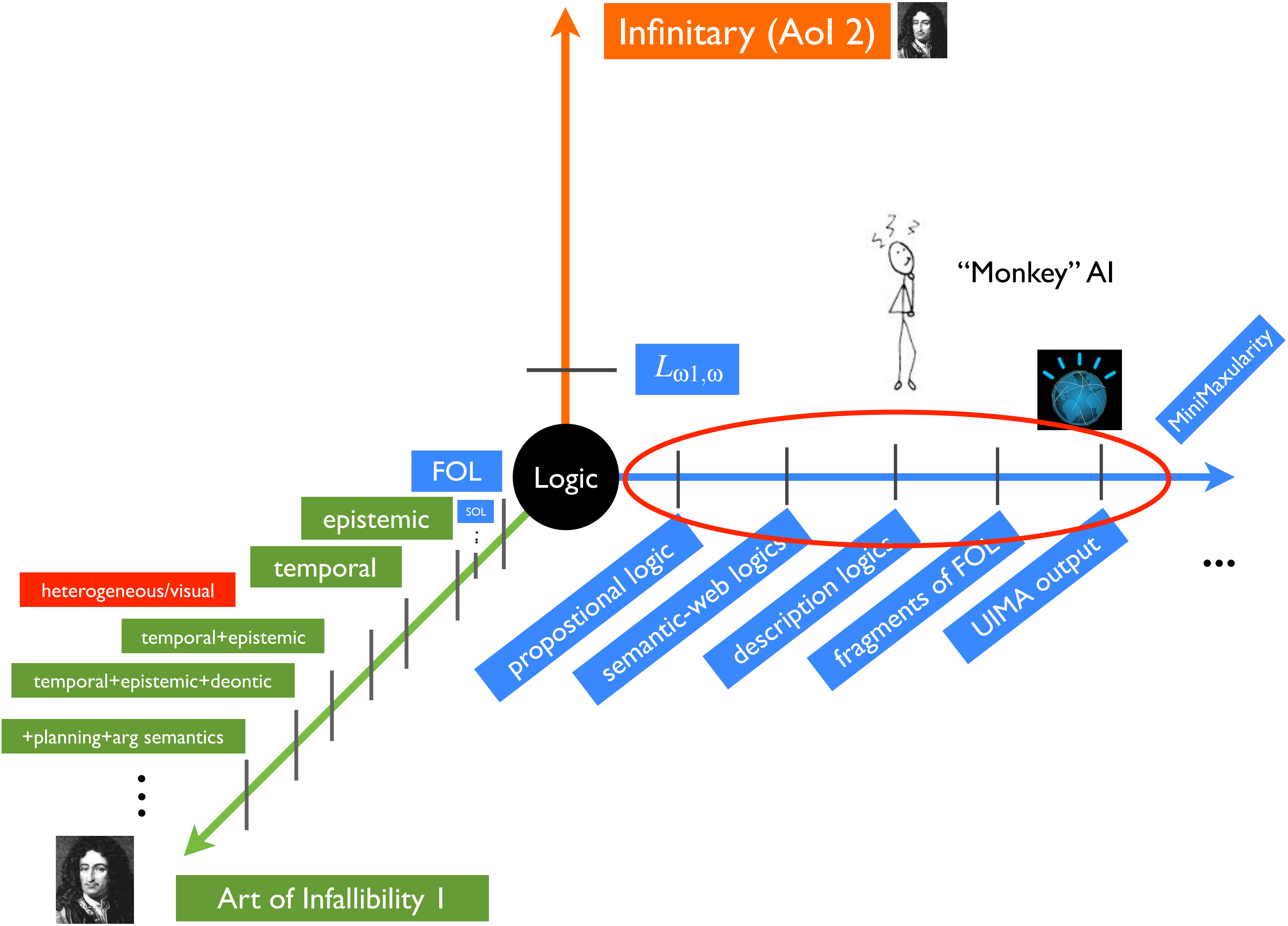
What is AI for you?

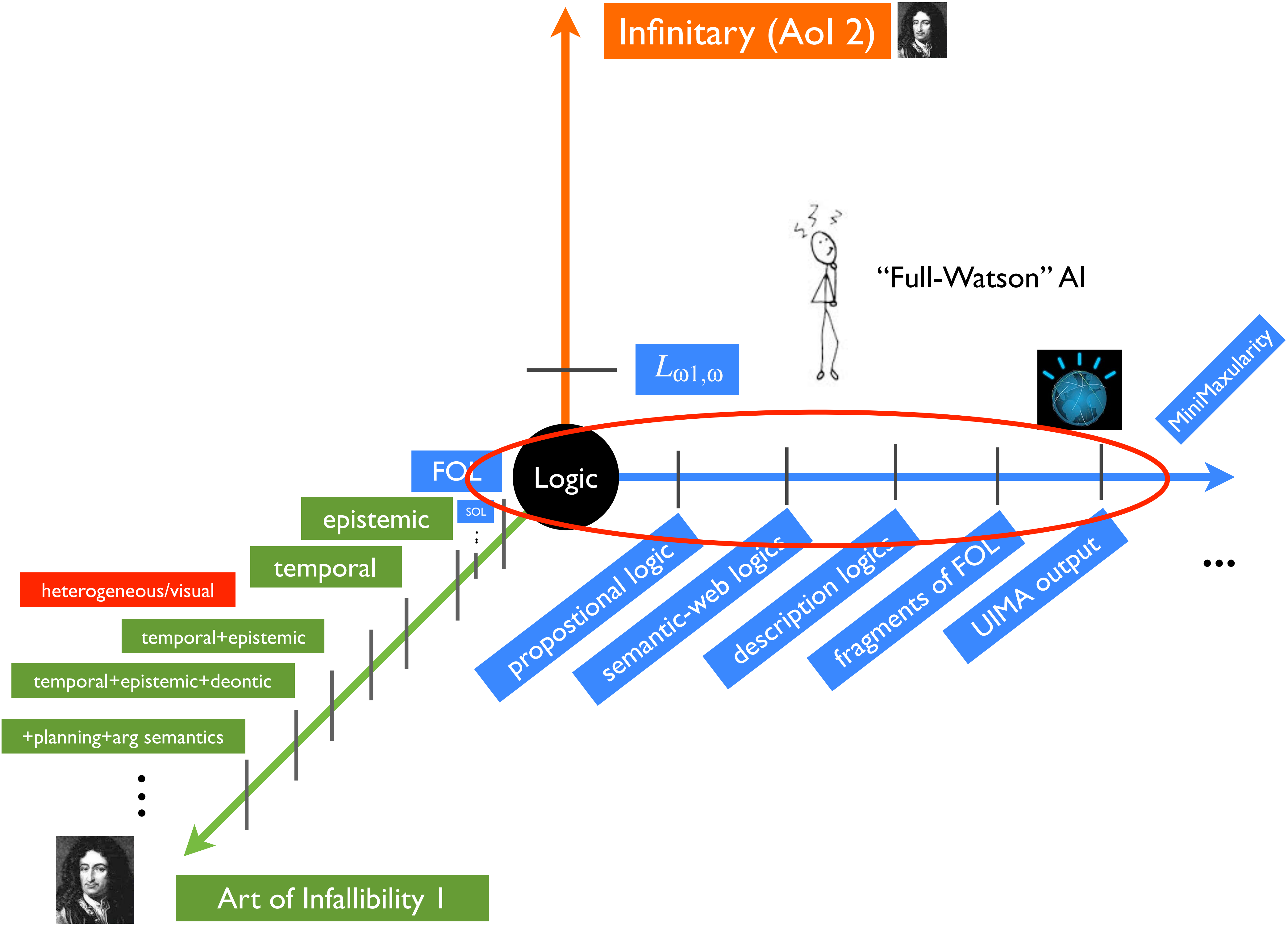
Elevated AI only!:

“The ultimate goal of AI is to build a person, or more humbly, an animal.” —C&M

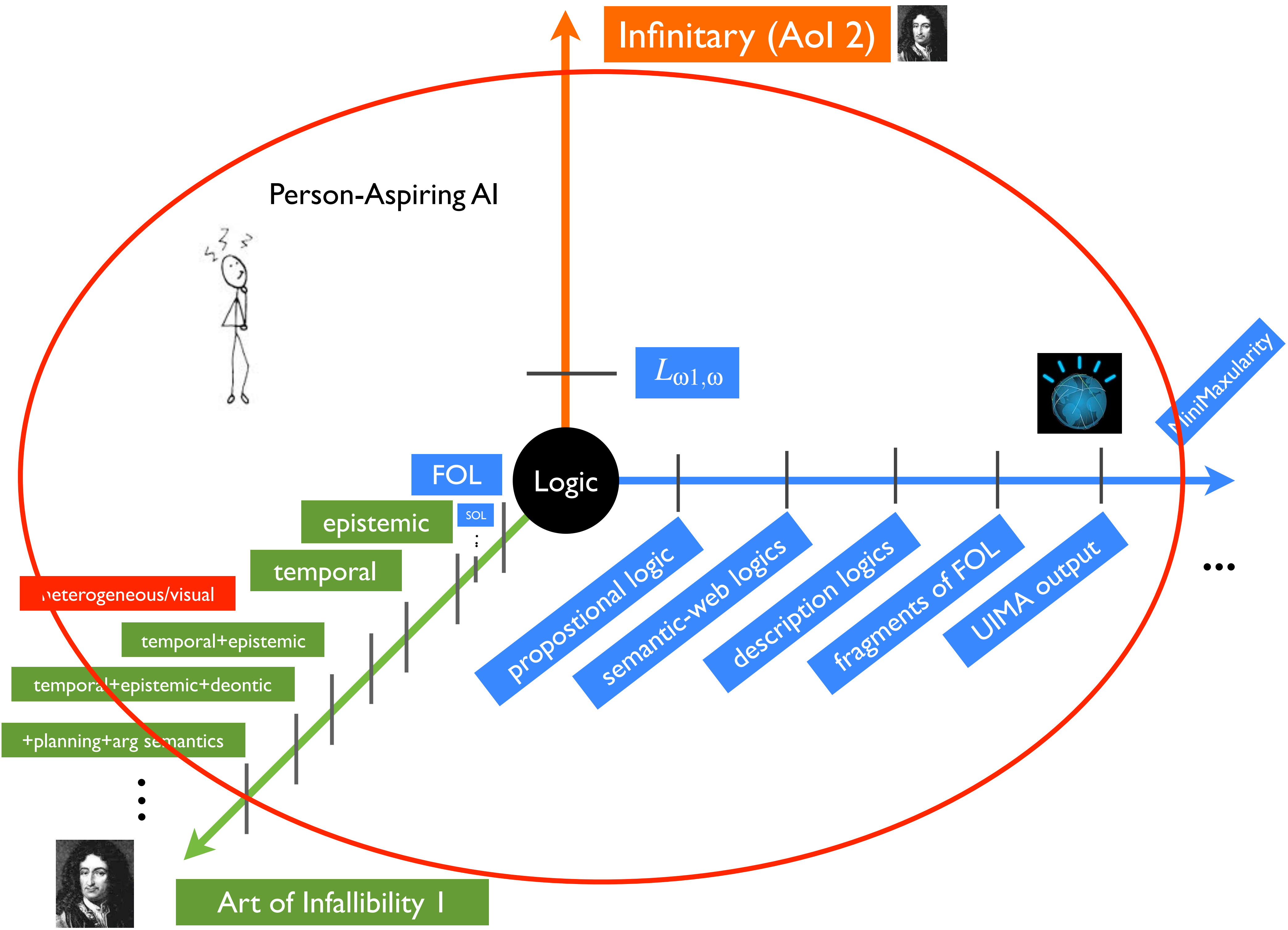



Infinitary (AoI 2)
Darwininan "Canine" AI
$L_{\omega 1,\omega}$
MiniMaxularity
FOL
Logic
epistemic
SOL
temporal
heterogeneous/visual
temporal+epistemic
temporal+epistemic+deontic
+planning+arg semantics
propostional logic
semantic-web logics
description logics
fragments of FOL
UIMA output
Art of Infallibility I


Infinitary (AoI 2)
"Monkey" AI
$L_{\omega1,\omega}$
MiniMaxularity
FOL
Logic
epistemic
SOL
temporal
heterogeneous/visual
temporal+epistemic
temporal+epistemic+deontic
+planning+arg semantics
propostional logic
semantic-web logics
description logics
fragments of FOL
UIMA output
...
Art of Infallibility I



Infinitary (AoI 2)
"Full-Watson" AI
$L_{\omega1,\omega}$
MiniMaxularity
FOL
Logic
SOL
epistemic
temporal
heterogeneous/visual
temporal+epistemic
temporal+epistemic+deontic
+planning+arg semantics
propostional logic
semantic-web logics
description logics
fragments of FOL
UIMA output
...
Art of Infallibility I



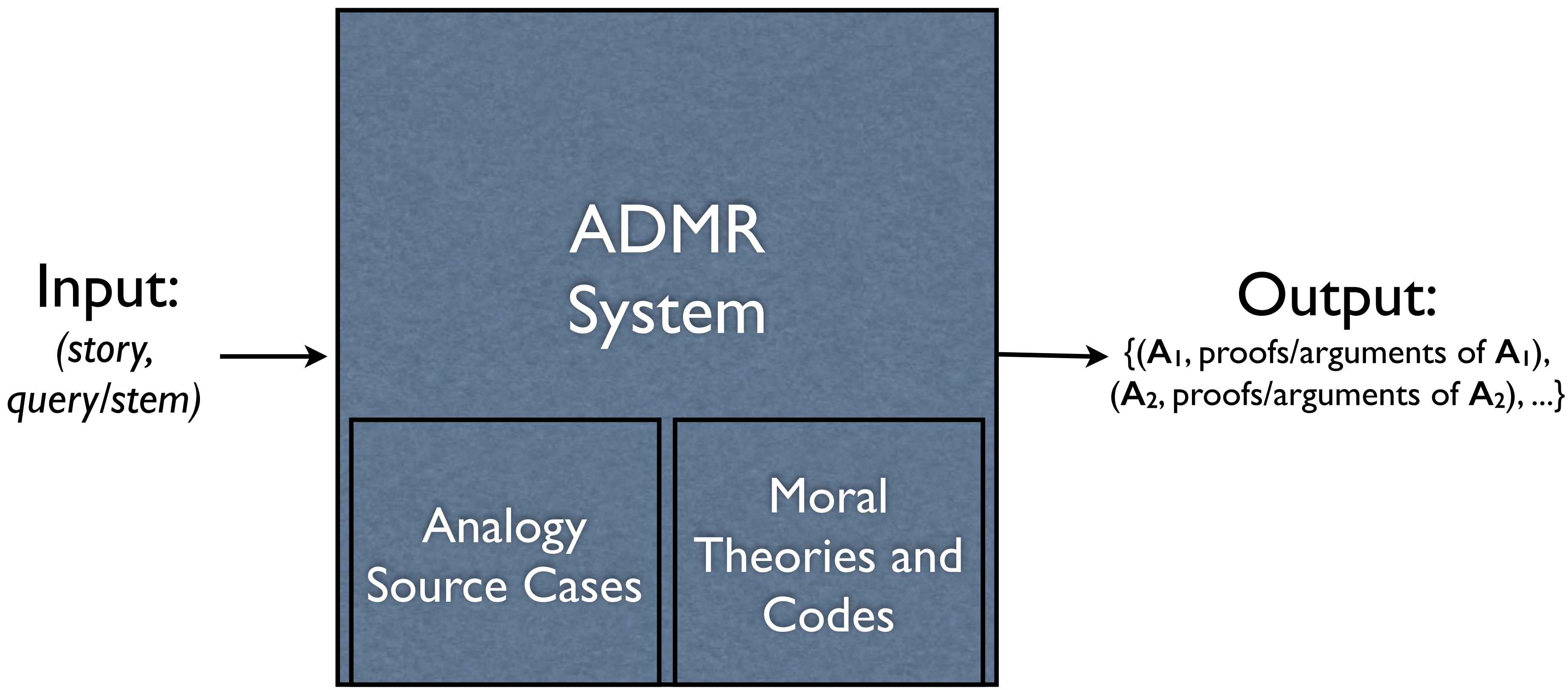
Infinitary (AoI 2)
Person-Aspiring AI
$L_{\omega 1,\omega}$
MiniMaxularity
FOL
Logic
epistemic
SOL
temporal
heterogeneous/visual
temporal+epistemic
temporal+epistemic+deontic
+planning+arg semantics
propostional logic
semantic-web logics
description logics
fragments of FOL
UIMA output
...
Art of Infallibility I

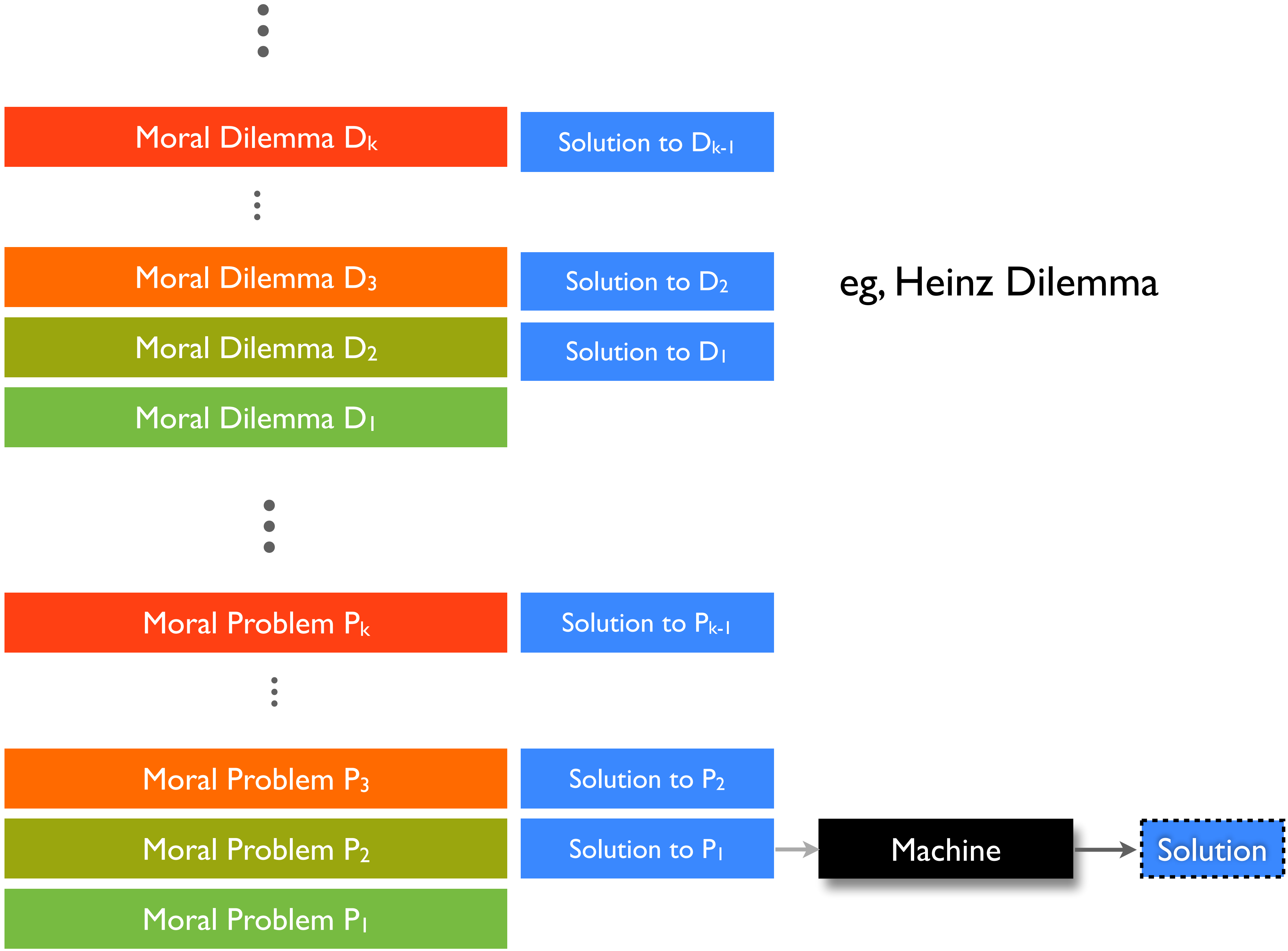
Analogico-Deductive Moral Reasoning (ADMR)

- Moral problem presented as *story* (in psychometric sense) and a *stem*, or *query*.
- A *stem* has correct answer **A** and a set P_i of correct proofs or arguments establishing **A**, relative to:
  - An associated implicit moral theory, and
  - A corresponding moral code

But moral *dilemmas* often have multiple theory codes, and competing answers!

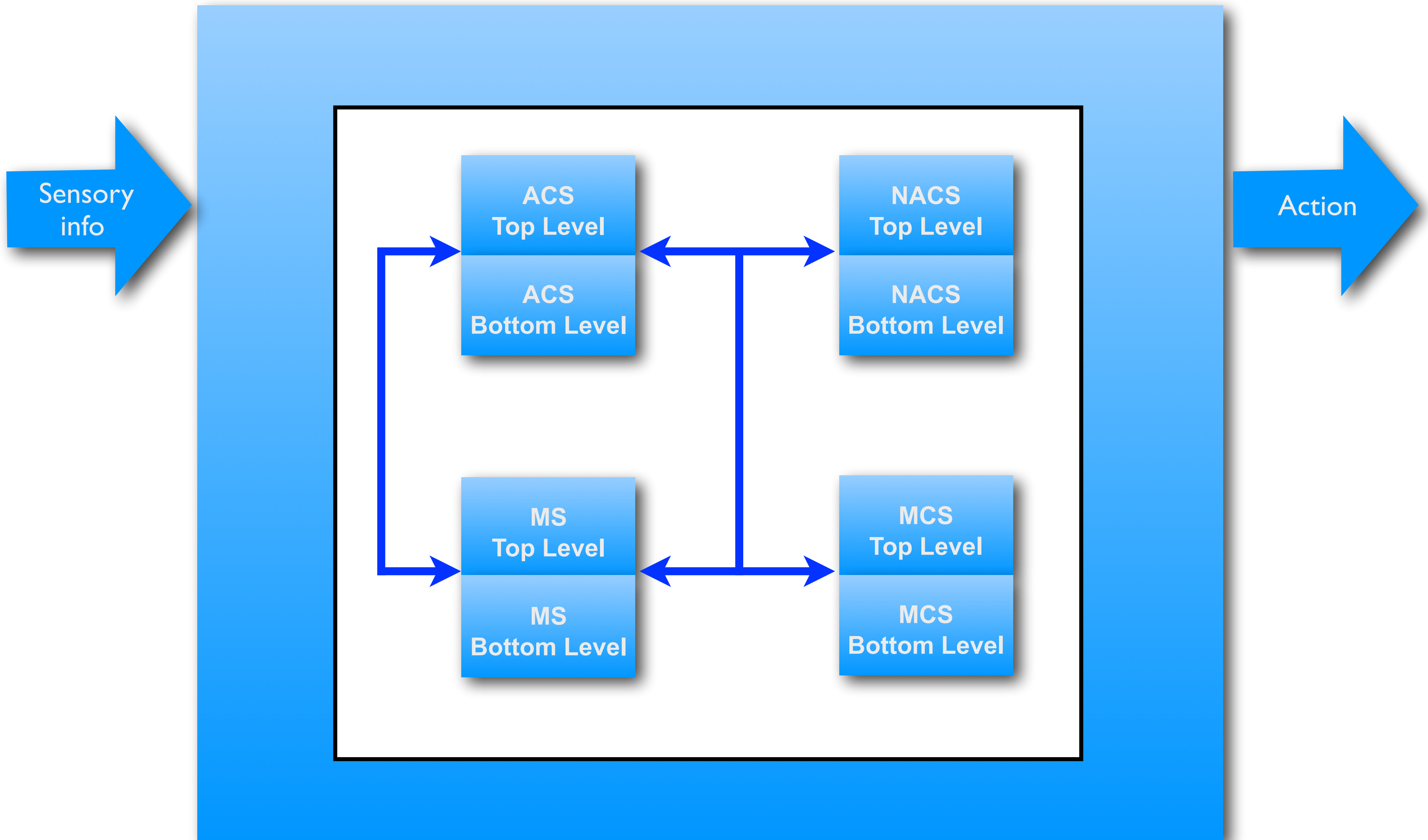
Analogico-Deductive Moral Reasoning (ADMR)




Moral Dilemma D_k
Solution to D_{k-1}
Moral Dilemma D_3
Solution to D_2
eg, Heinz Dilemma
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
Machine
Solution
Moral Problem P_1

But can this be done in a
cognitively-psychologically realistic way?

CLARION Subsystems



The Heinz Dilemma (Kohlberg)

"In Europe, a woman was near death from a special kind of cancer. There was one drug that the doctors thought might save her. It was a form of radium that a druggist in the same town had recently discovered. The drug was expensive to make, but the druggist was charging ten times what the drug cost him to make. He paid $200 for the radium and charged $2,000 for a small dose of the drug.

The sick woman's husband, Heinz, went to everyone he knew to borrow the money, but he could only get together about $1,000, which is half of what it cost. He told the druggist that his wife was dying and asked him to sell it cheaper or let him pay later. But the druggist said: "No, I discovered the drug and I'm going to make money from it." So Heinz got desperate and broke into the man's store to steal the drug for his wife. *Should the husband have done that?*"

A *simple* example in DCEC*

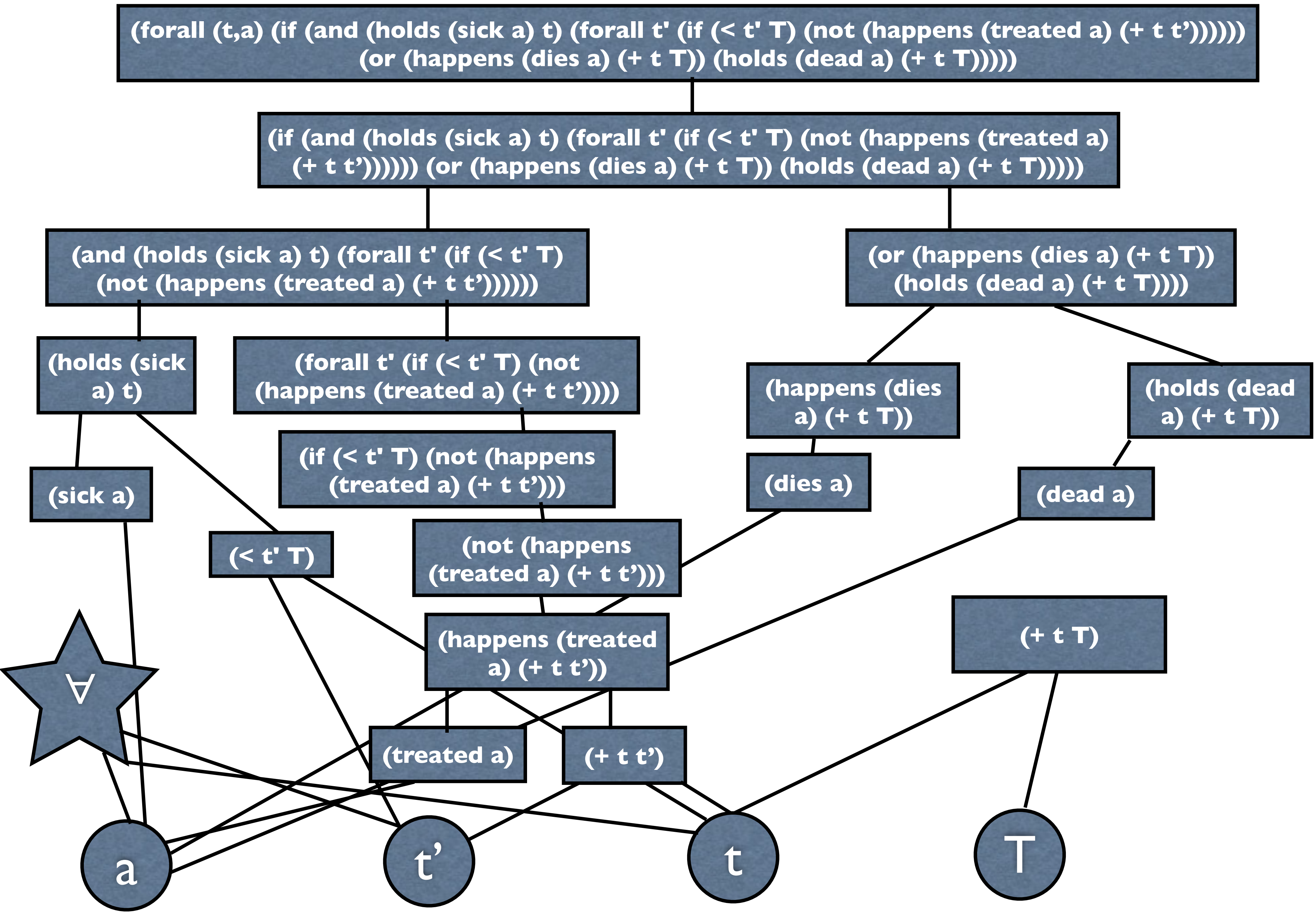
P1 $$\forall t : \mathsf{Moment}, a : \mathsf{Agent}\Bigg(holds(sick(a),t) \land \Big(\forall t' : \mathsf{Moment}\ t' < T \Rightarrow \neg happens(treated(a), t+t')\Big) \Rightarrow \big(happens(dies(a), t+T) \lor holds(dead(a), t+T)\big)\Bigg)$$

P2 $$holds(sick(wife(\mathsf{I}*)), t_0) \land \Big(\forall t' : \mathsf{Moment}\ t' < T \Rightarrow \neg happens(treated(wife(\mathsf{I}*)), t_0+t')$$

Q $$happens(dies(wife(\mathsf{I}*)), t_0+T) \lor holds(dead(wife(\mathsf{I}*)), t_0+T)$$

Note: This adheres strictly to the syntax of DCEC*

P1 in CLARION's NACS (simplified version)



We may need the DCEC*: Far beyond the reach of all cognitive architectures (at the moment)

Syntax

$$S ::= \text{Object} \mid \text{Agent} \mid \text{Self} \sqsubset \text{Agent} \mid \text{ActionType} \mid \text{Action} \sqsubseteq \text{Event} \mid \text{Moment} \mid \text{Boolean} \mid \text{Fluent} \mid \text{Numeric}$$

$$t ::= x : S \mid c : S \mid f(t_1, \ldots, t_n)$$

$$\phi ::= p : \text{Boolean} \mid \neg\phi \mid \phi \land \psi \mid \phi \lor \psi \mid \phi \rightarrow \psi \mid \phi \leftrightarrow \psi \mid \forall x : S.\ \phi \mid \exists x : S.\ \phi$$
$$\mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \mid \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,holds(f,t')) \mid \mathbf{I}(a,t,happens(action(a^*,\alpha),t')) \mid \mathbf{O}(a,t,\phi,happens(action(a^*,\alpha),t'))$$

$f ::=$

$action$: Agent $\times$ ActionType $\rightarrow$ Action

$initially$: Fluent $\rightarrow$ Boolean

$holds$: Fluent $\times$ Moment $\rightarrow$ Boolean

$happens$: Event $\times$ Moment $\rightarrow$ Boolean

$clipped$: Moment $\times$ Fluent $\times$ Moment $\rightarrow$ $Boolean$

$initiates$: Event $\times$ Fluent $\times$ Moment $\rightarrow$ Boolean

$terminates$: Event $\times$ Fluent $\times$ Moment $\rightarrow$ Boolean

$prior$: Moment $\times$ Moment $\rightarrow$ Boolean

$interval$: Moment $\times$ Boolean

$*$: Agent $\rightarrow$ Self

$payoff$: Agent $\times$ ActionType $\times$ Moment $\rightarrow$ Numeric

Rules of Inference

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

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

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

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

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

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

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

$$\frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\phi \rightarrow \psi)}{\mathbf{B}(a,t,\psi)}\ [R_{11a}] \qquad \frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi \land \phi)}\ [R_{11b}]$$

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

$$\frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\mathbf{O}(a^*,t,\phi,happens(action(a^*,\alpha),t')))\quad \mathbf{O}(a,t,\phi,happens(action(a^*,\alpha),t'))}{\mathbf{K}(a,t,\mathbf{I}(a^*,t,happens(action(a^*,\alpha),t')))}\ [R_{14}]$$

$$\frac{\phi \leftrightarrow \psi}{\mathbf{O}(a,t,\phi,\gamma) \leftrightarrow \mathbf{O}(a,t,\psi,\gamma)}\ [R_{15}]$$

More Complex DCEC* Specimen from Heinz Dilemma

Given $\mathbf{B}\Big(\mathsf{I}, \mathsf{now}, \forall t : \mathsf{Moment}, a : \mathsf{Agent}\Big(holds(sick(a), t) \wedge \big(\forall t' : \mathsf{Moment}\ t' < T \Rightarrow \neg happens(treated(a), t + t')\big) \Rightarrow \big(happens(dies(a), t + T) \vee holds(dead(a), t + T)\big)\Big)\Big)$

Given $\mathbf{K}\Big(\mathsf{I}, \mathsf{now}, holds(sick(wife(\mathsf{I}*)), t_0) \wedge \big(\forall t' : \mathsf{Moment}\ t' < T \Rightarrow \neg happens(treated(wife(\mathsf{I}*)), t + t')\big)\Big)$

Inferred $\mathbf{B}\big(\mathsf{I}, \mathsf{now}, happens(dies(wife(\mathsf{I}*)), t_0 + T) \vee holds(dead(wife(\mathsf{I}*)), t_0 + T)\big)$

Given $\mathbf{K}\Big(\mathsf{I}, \mathsf{now}, \mathsf{EventCalculus} \Rightarrow \big(happens(dies(wife(\mathsf{I}*)), t_0 + T) \vee holds(dead(wife(\mathsf{I}*)), t_0 + T) \Rightarrow \neg holds(alive(wife(\mathsf{I}*)), t_0 + T)\big)\Big)$

Inferred $\mathbf{B}\big(\mathsf{I}, \mathsf{now}, \neg holds(alive(wife(\mathsf{I}*)), t_0 + T)\big)$

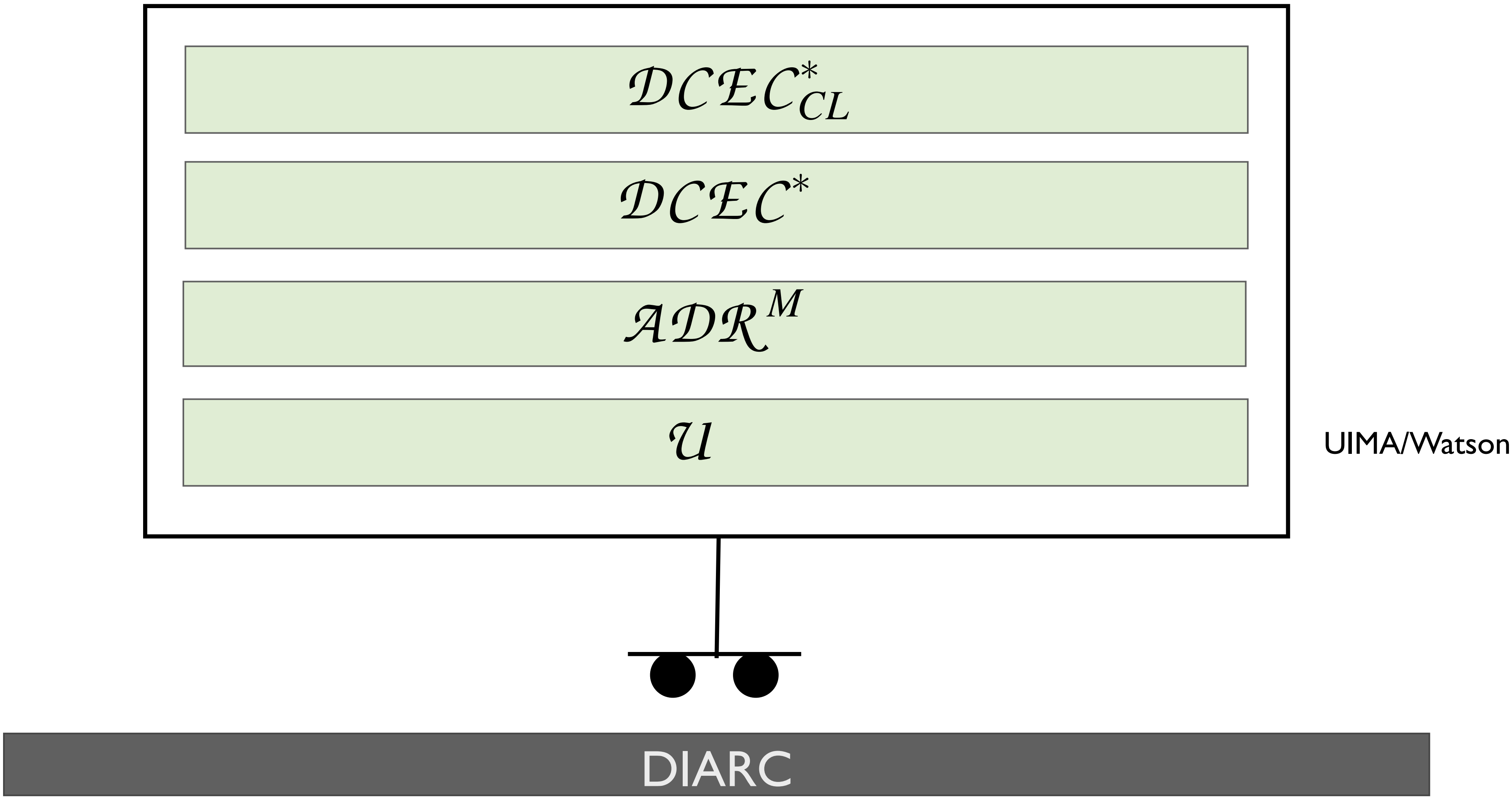
Given $\mathbf{D}\big(\mathsf{I}, \mathsf{now}, holds(alive(wife(\mathsf{I}*)), t_0 + T)\big)$

Given $\big(\mathbf{B}(\mathsf{I}, \mathsf{now}, \neg holds(f, t)) \wedge \mathbf{D}(\mathsf{I}, \mathsf{now}, holds(f, t)) \wedge \mathbf{K}(\mathsf{I}, \mathsf{now}, happens(action(\mathsf{I}*, \alpha), \mathsf{now}) \Rightarrow holds(f, t))\big) \Rightarrow \mathbf{I}(\mathsf{I}, \mathsf{now}, happens(action(\mathsf{I}*, \alpha), \mathsf{now}))$

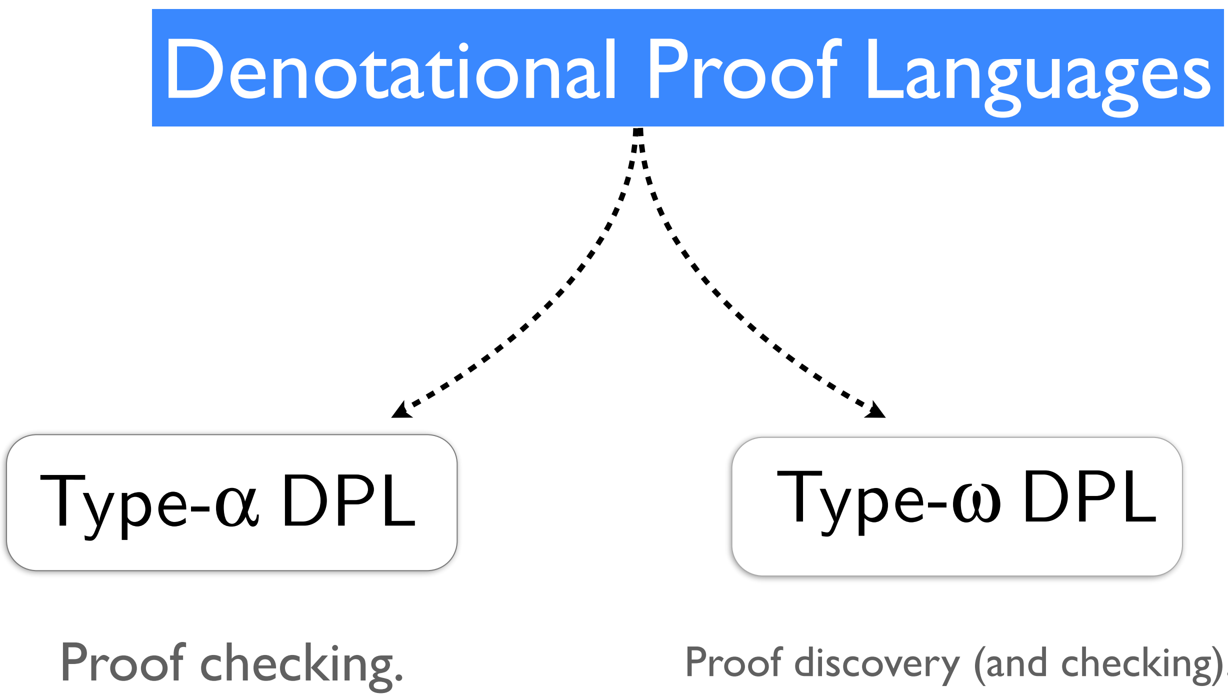
Given $\mathbf{K}\big(\mathsf{I}, \mathsf{now}, happens(action(\mathsf{I}*, treat), \mathsf{now}) \Rightarrow holds(alive(wife(\mathsf{I}*)), t_0 + T)\big)$

Inferred $\mathbf{I}\big(\mathsf{I}, \mathsf{now}, happens(action(\mathsf{I}*, treat), \mathsf{now})\big)$

The Overall Approach



Automation of Reasoning



DPLs for $\mathcal{DCEC}^*$ under construction ...

K. Arkoudas. *Denotational Proof Languages*. PhD thesis, MIT, 2000.

K. Arkoudas and S. Bringsjord. Propositional Attitudes and Causation. *International Journal of Software and Informatics*, 3(1):47–65, 2009.

Logicist NLP

Two Major Approaches

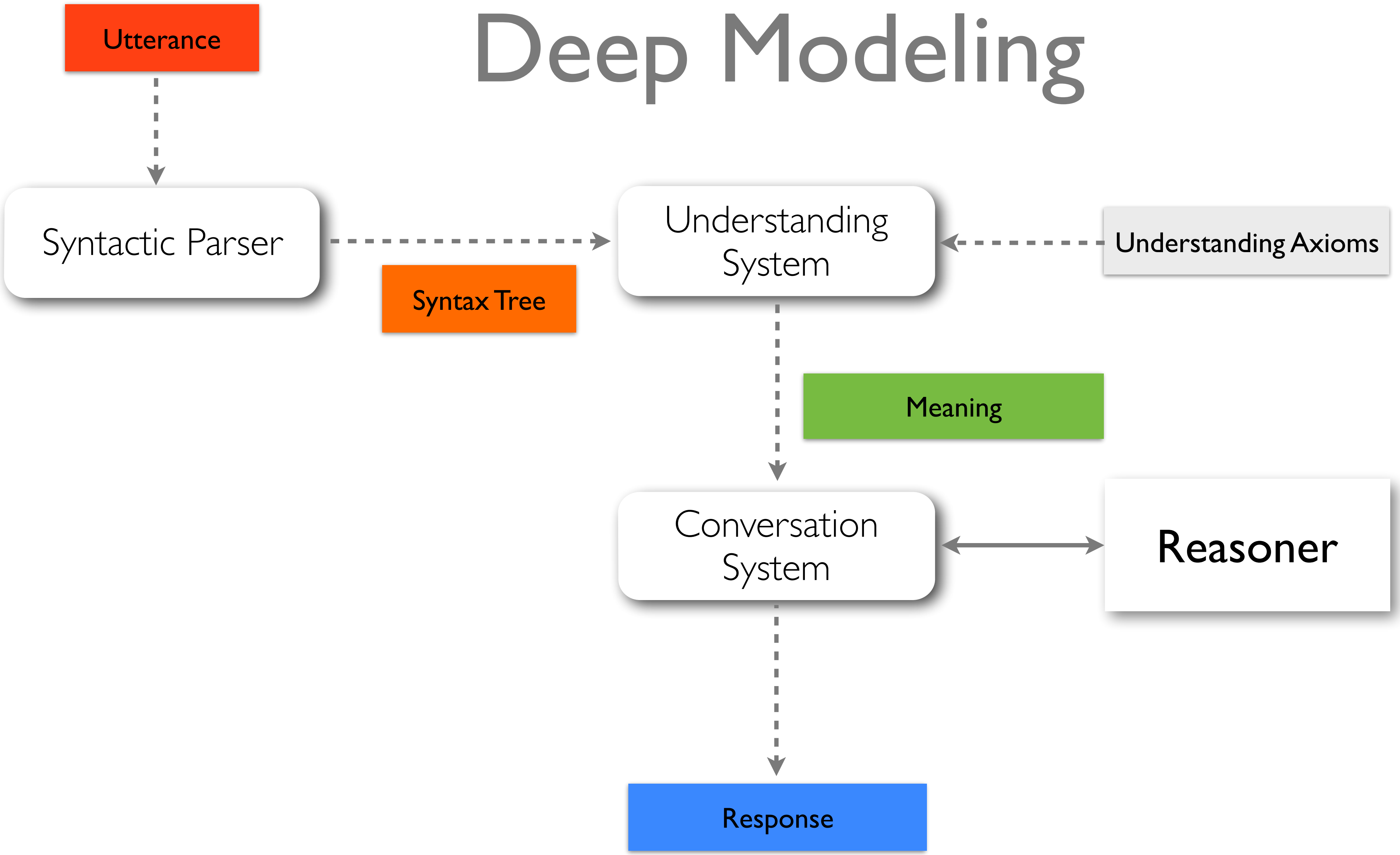
- Deep Modeling
- Controlled English

On Deep Computational Formalization of Natural Language

Naveen Sundar Govindarajulu, John Licato and Selmer Bringsjord

Workshop on Formalizing Mechanisms for Artificial General Intelligence, 2013, AGI 2013

Deep Modeling



Controlled English

$\mathcal{DCEC}^*_{CL}$ corresponds to a subset of English!

RLCNL: RAIR Lab Controlled Natural Language

$$\mathbf{K}(\text{ugv}, \text{now}, \mathit{holds}(\mathit{carrying}(\text{ugv}, \text{soldier}), \text{now}))$$

The ugv now knows that the fluent, 'the ugv is carrying the soldier,' holds now.

$$\mathbf{B}(\text{ugv}, \text{now}, \mathbf{B}(\text{commander}, t_1, \neg\mathbf{P}(\text{ugv}, \text{anytime}, \mathit{happens}(\mathit{firefight}, \text{anytime}))))$$

The ugv now believes that the commander at moment t1 believes that it is not the case that the ugv at any time perceives that a firefight happens at any time.

$$\mathbf{K}(\text{I}, \text{now}, \mathbf{O}(\text{I}^*, \text{now}, \mathit{mission}(\mathit{main}), \mathit{happens}(\mathit{action}(\text{I}^*, \text{silence}), \text{alltime})))$$

I now know that it is obligatory for myself under the condition that the main mission being carried out, that I myself should see to it that silence is maintained at all times.

Partial Implementation: http://naveensundarg.github.io/RLCNL/