Moral Competence in Computational Architectures for Robots: Foundations, Implementations, and Demonstrations

Logico-Mathematical Foundations

Selmer Bringsjord • Naveen Sundar G. • Mei Si

Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer AI & Reasoning (RAIR) Lab
selmer@rpi.edu • govinn@rpi.edu
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Moral Competence in Computational Architectures for Robots: Foundations, Implementations, and Demonstrations

Logico-Mathematical Foundations

Selmer Bringsjord • Naveen Sundar G. • Mei Si

Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer AI & Reasoning (RAIR) Lab
selmer@rpi.edu • govinn@rpi.edu
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Moral Competence in Computational Architectures for Robots: Foundations, Implementations, and Demonstrations

Logico-Mathematical Foundations

Selmer Bringsjord • Naveen Sundar G. • Mei Si

Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer AI & Reasoning (RAIR) Lab
selmer@rpi.edu • govinn@rpi.edu
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Moral Competence in Computational Architectures for Robots: Foundations, Implementations, and Demonstrations

Logico-Mathematical Foundations

Selmer Bringsjord • Naveen Sundar G. • Mei Si

Department of Computer Science
Department of Cognitive Science
Lally School of Management & Technology
Rensselaer AI & Reasoning (RAIR) Lab
selmer@rpi.edu • govinn@rpi.edu
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA

Hierarchy of Ethical Reasoning

Hierarchy of Ethical Reasoning

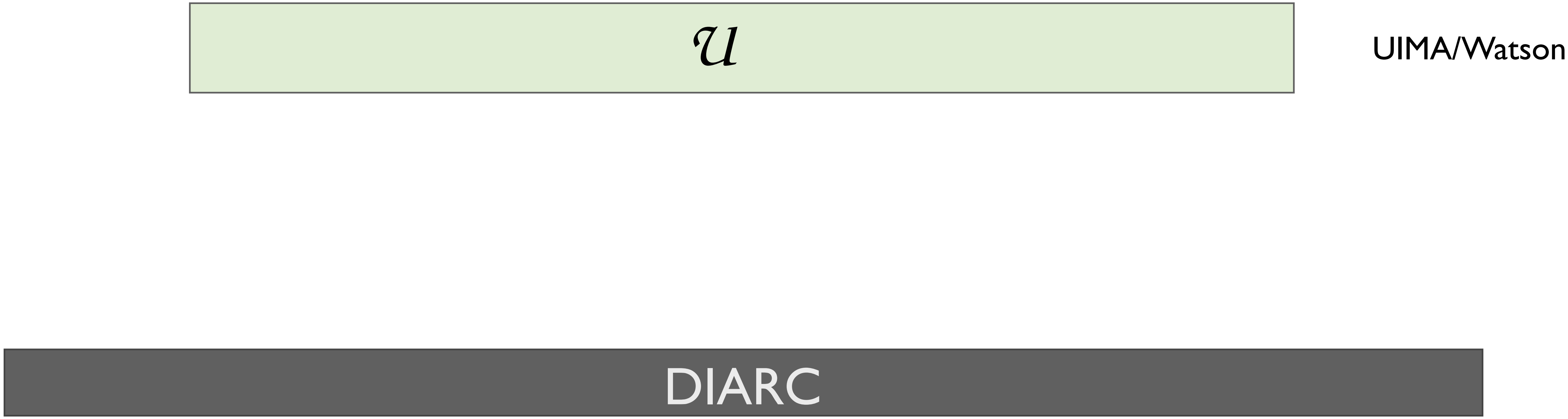
DIARC

Hierarchy of Ethical Reasoning

$\mathcal{U}$

DIARC

Hierarchy of Ethical Reasoning

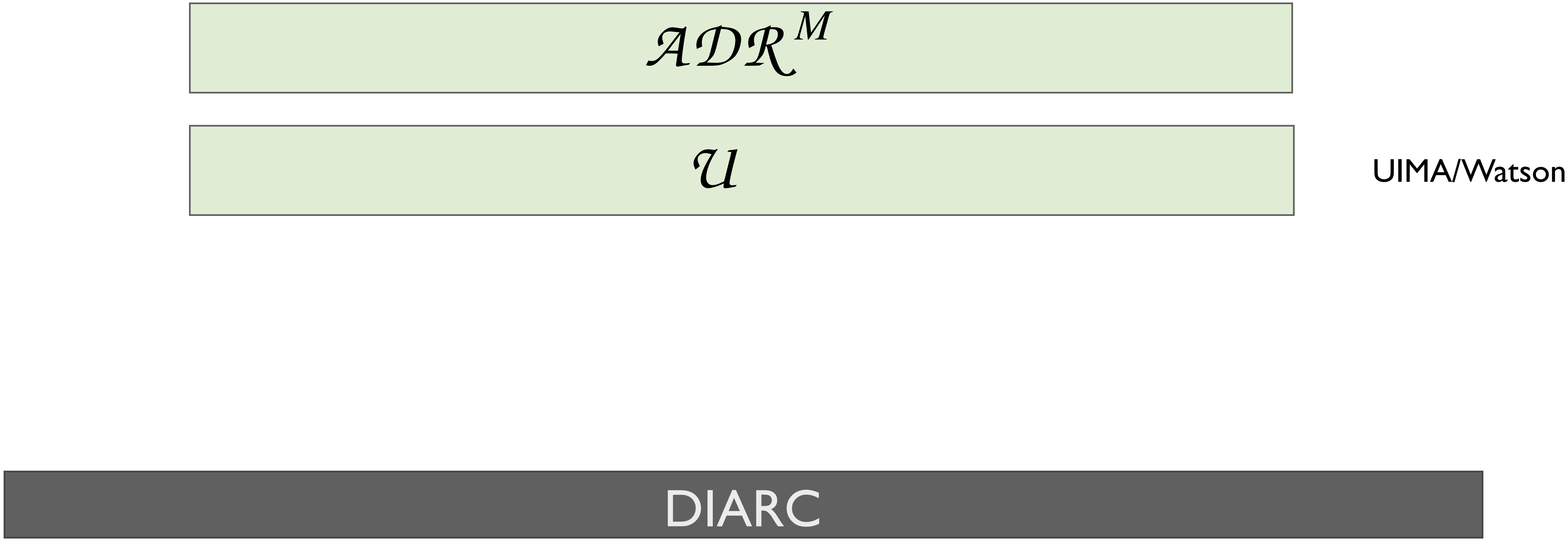


$\mathcal{U}$

UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



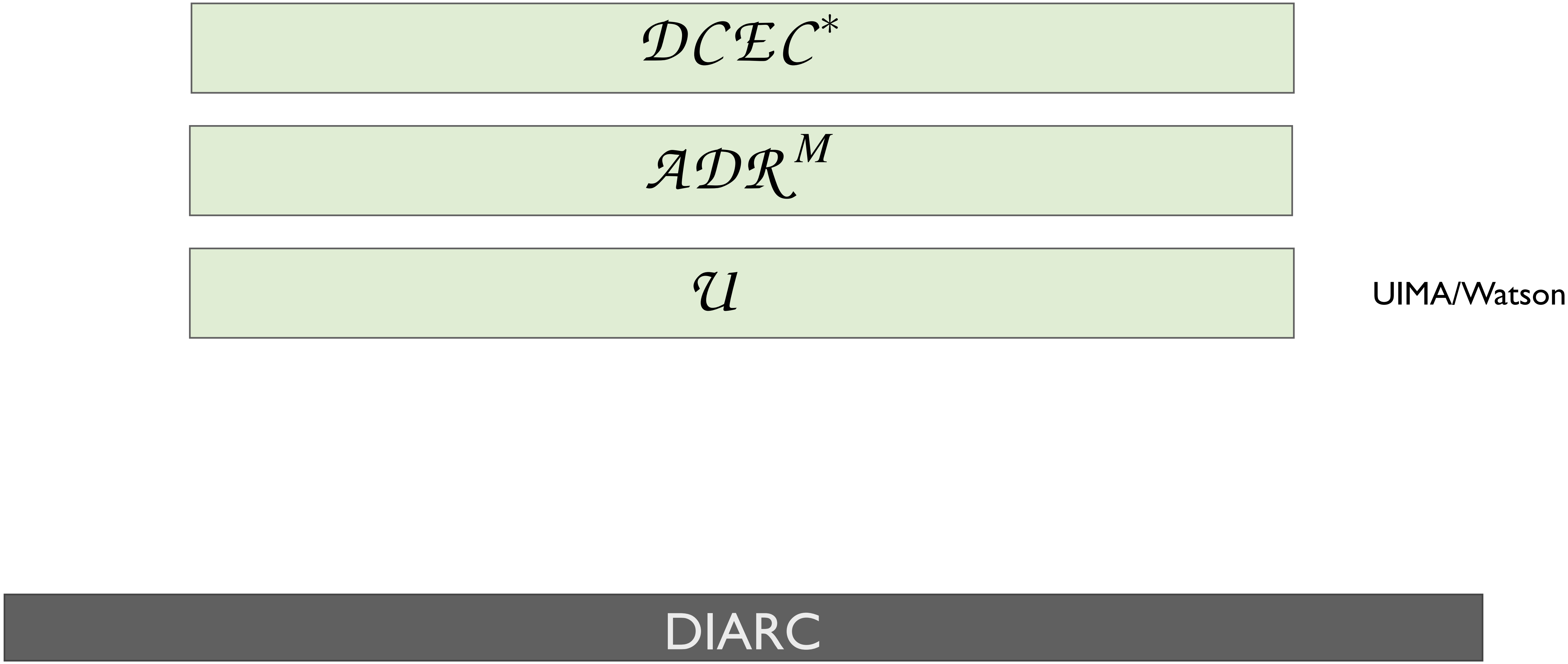
$\mathcal{ADR}^{M}$

$\mathcal{U}$

UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



$\mathcal{DCEC}^*$

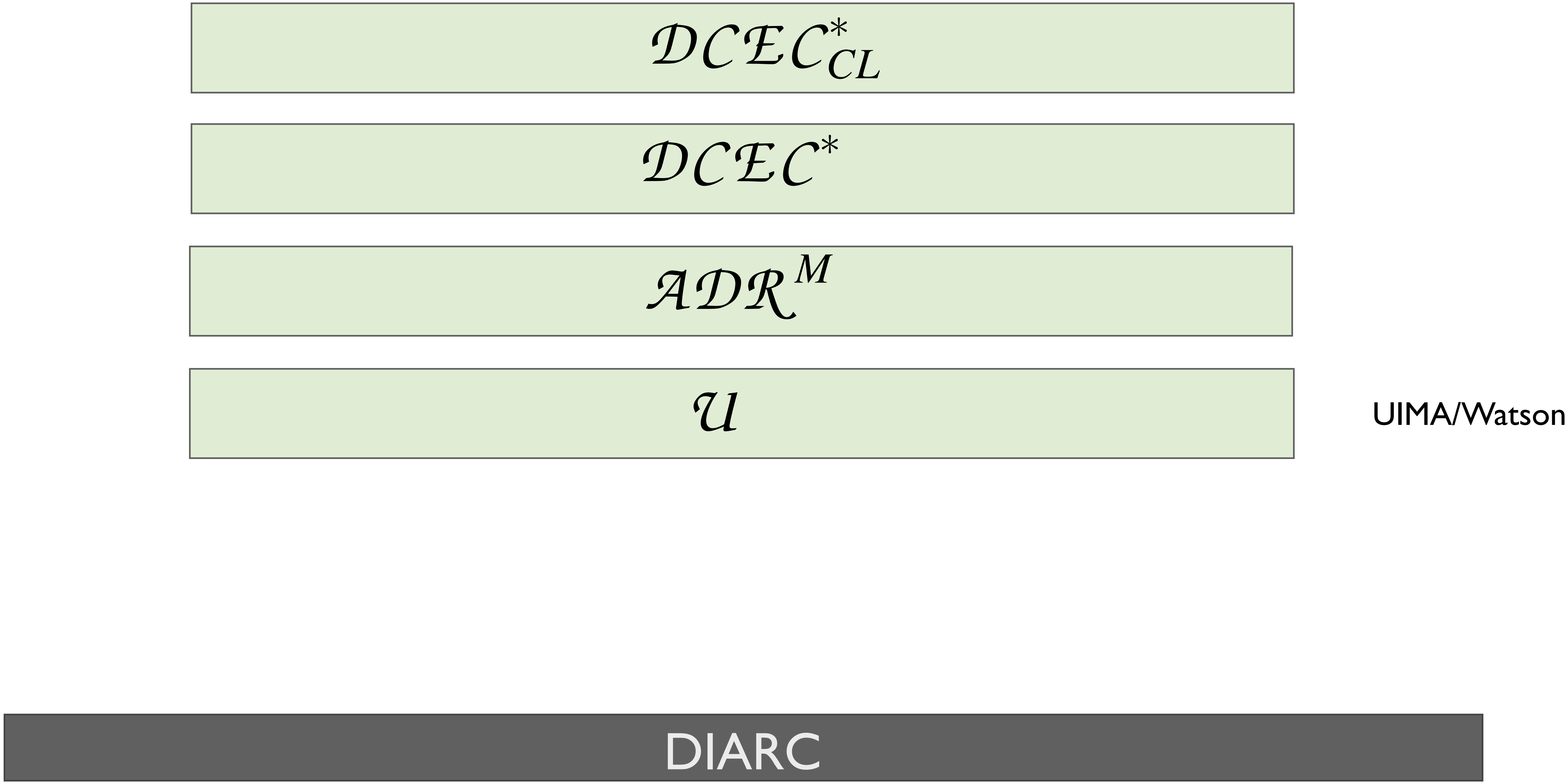
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



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

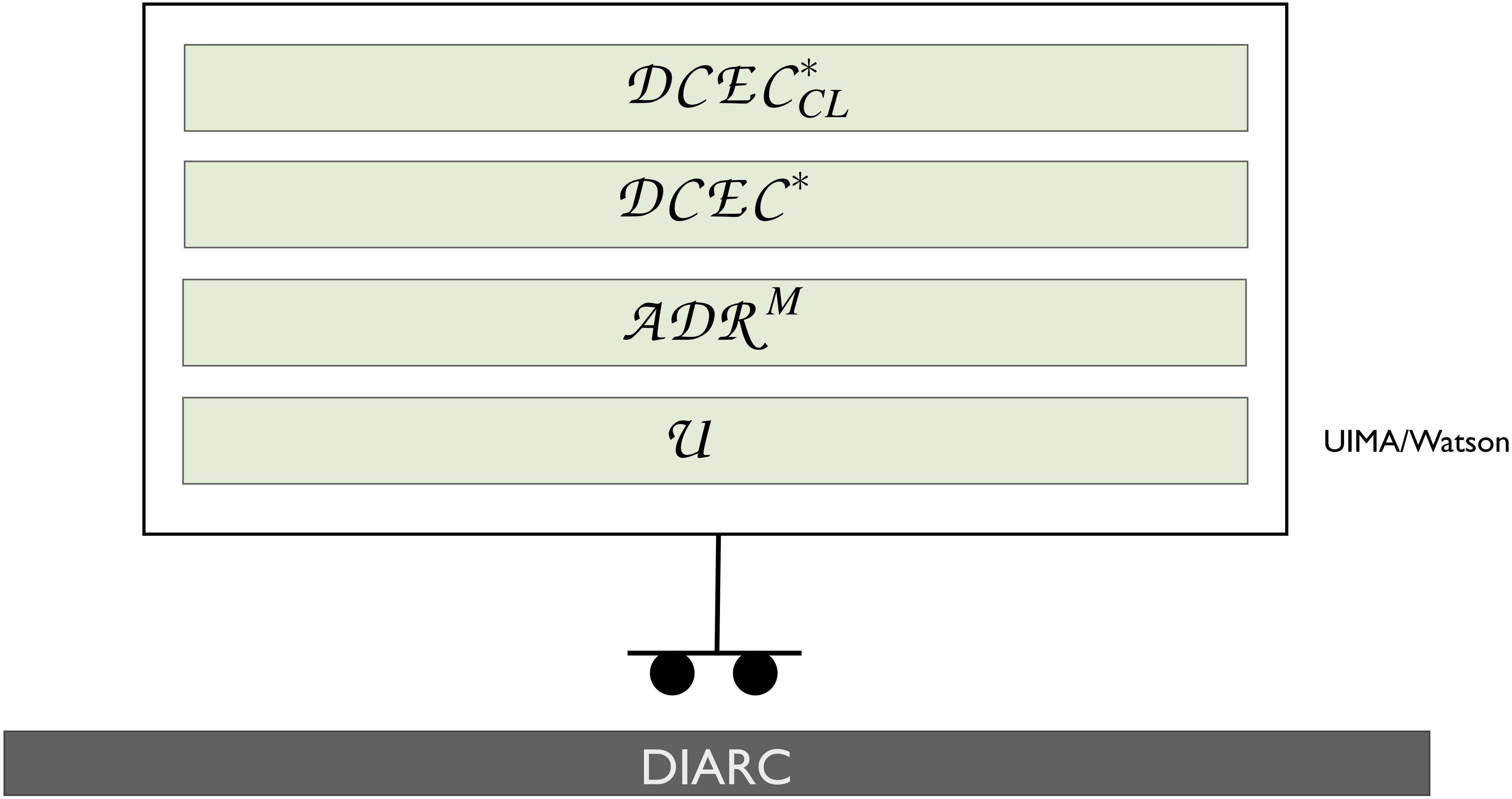
$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$ UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

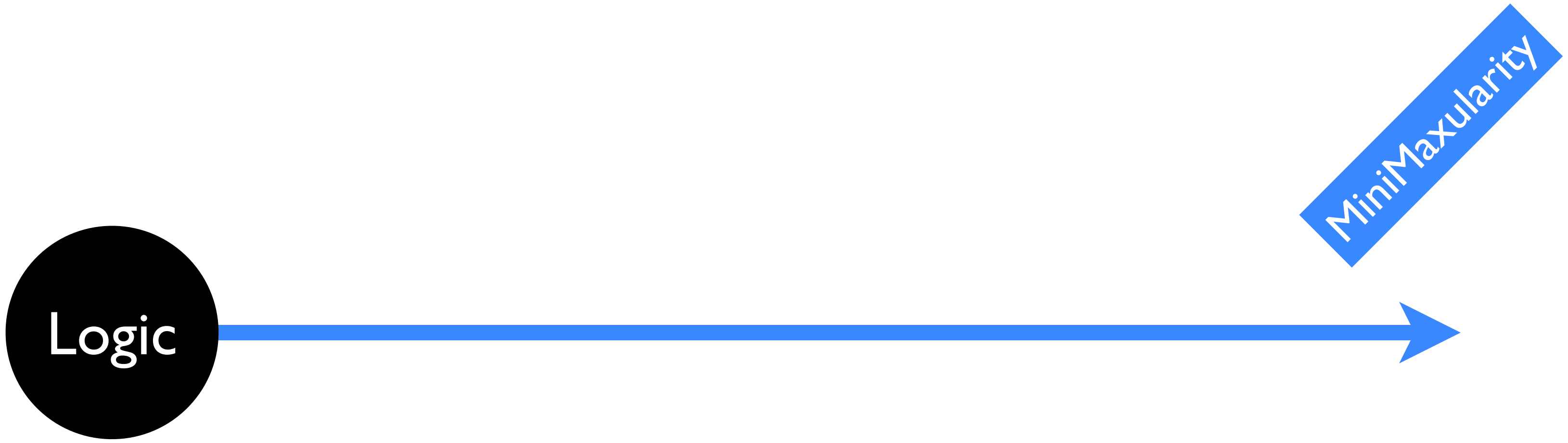
$\mathcal{U}$

UIMA/Watson

DIARC

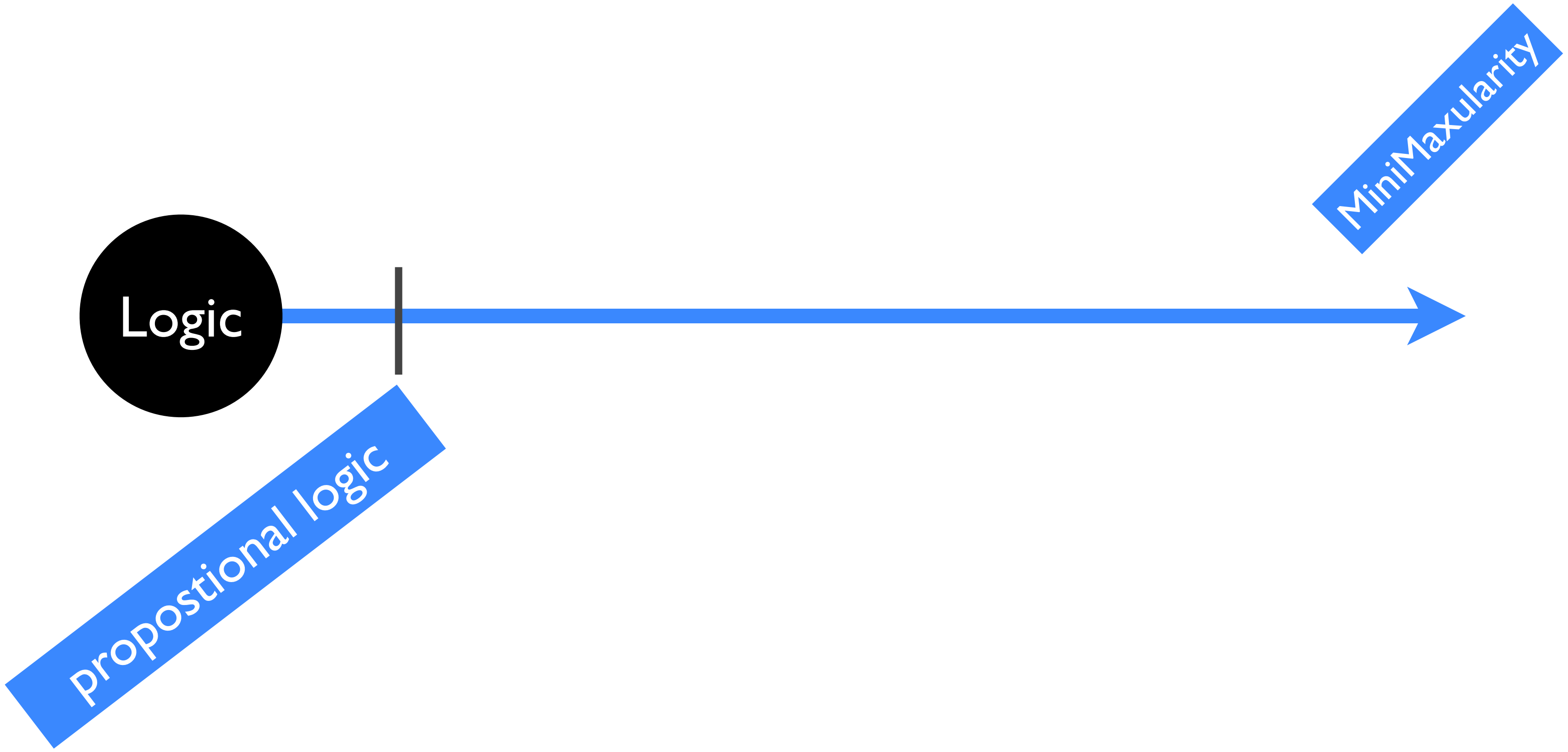
Logic

Logic



Logic

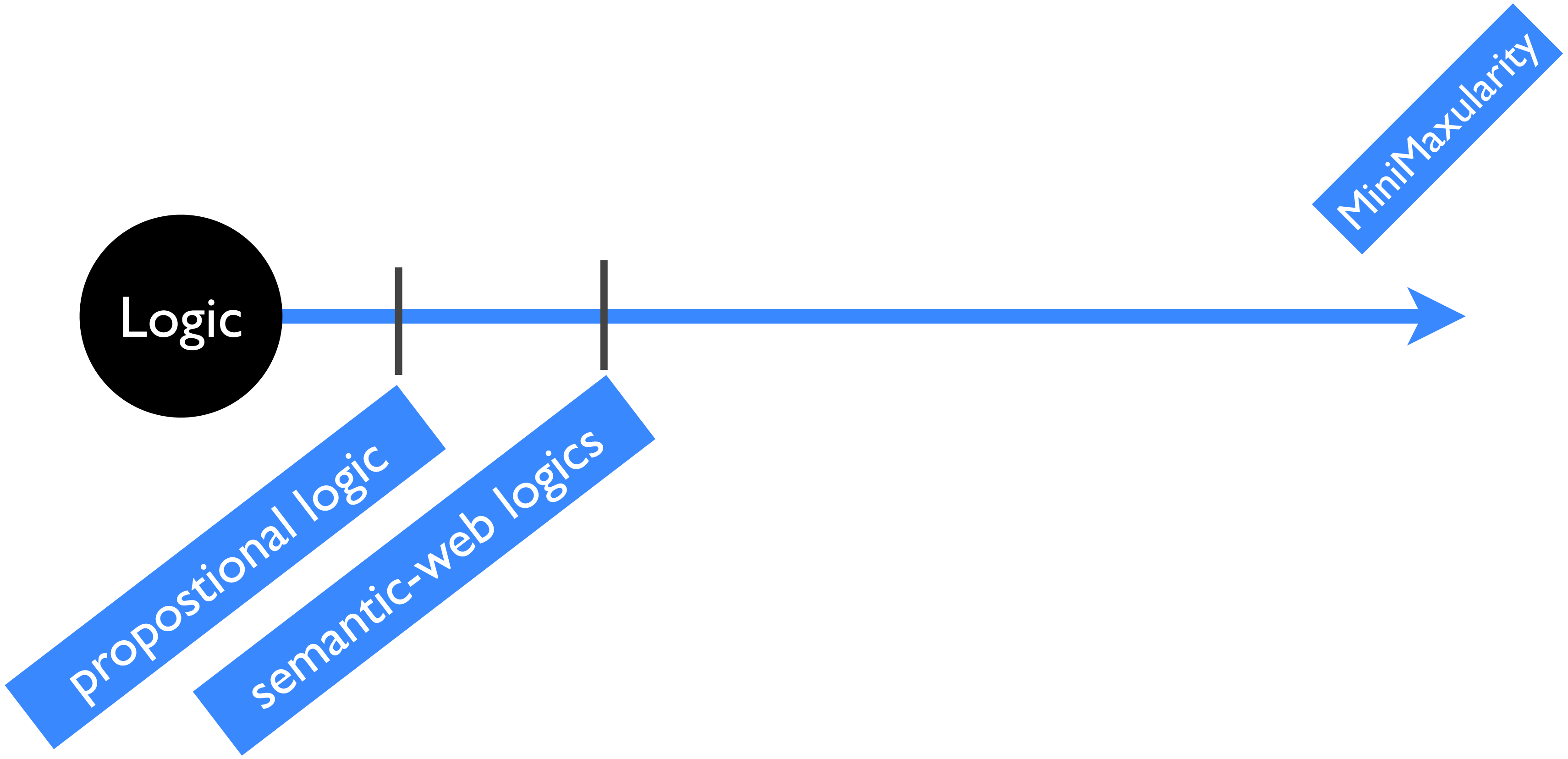
MiniMaxularity



Logic

propostional logic

MiniMaxularity

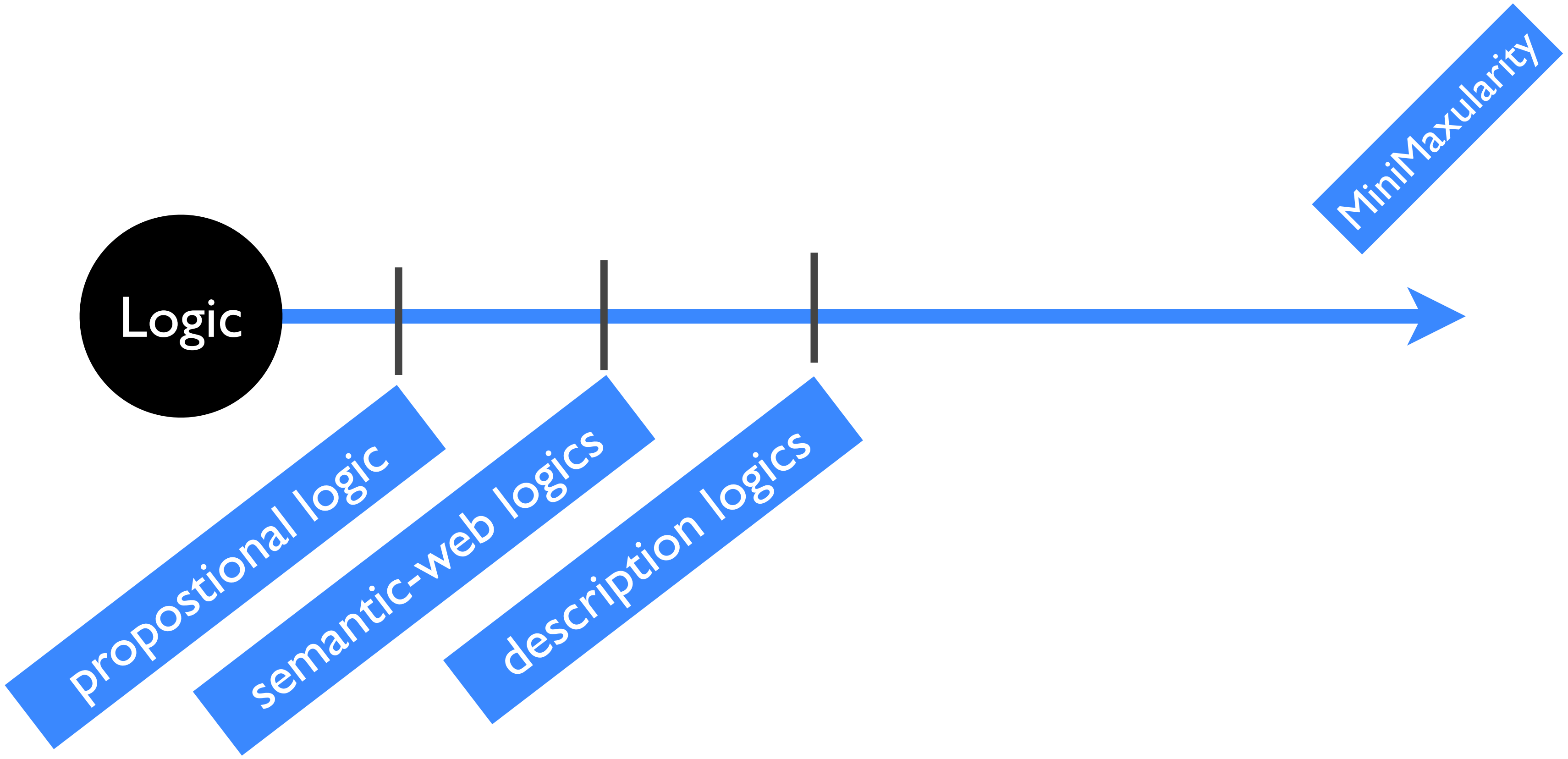


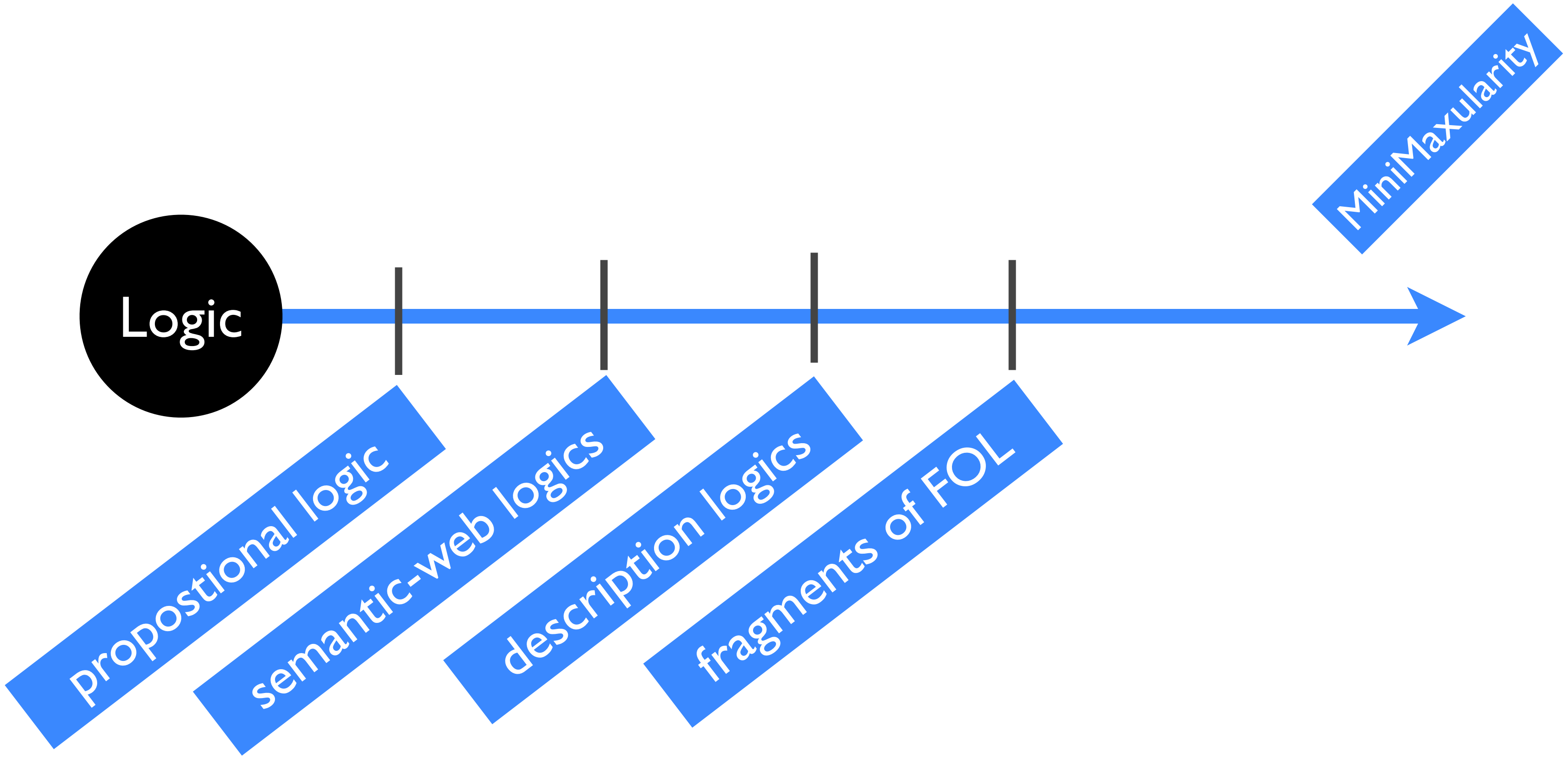
Logic

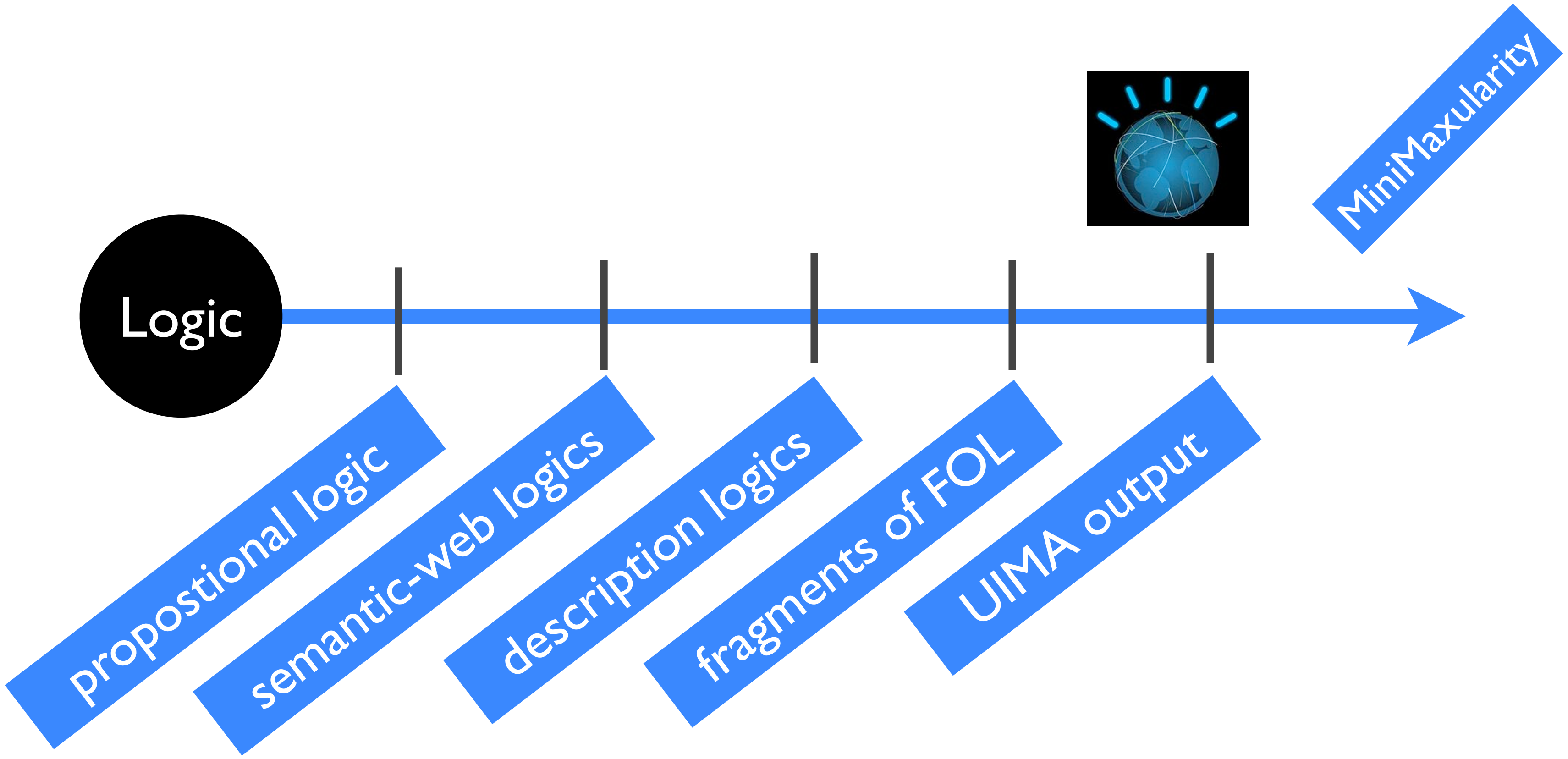
propostional logic

semantic-web logics

MiniMaxularity


Logic
propostional logic
semantic-web logics
description logics
MiniMaxularity


Logic
propostional logic
semantic-web logics
description logics
fragments of FOL
MiniMaxularity



Logic

propostional logic

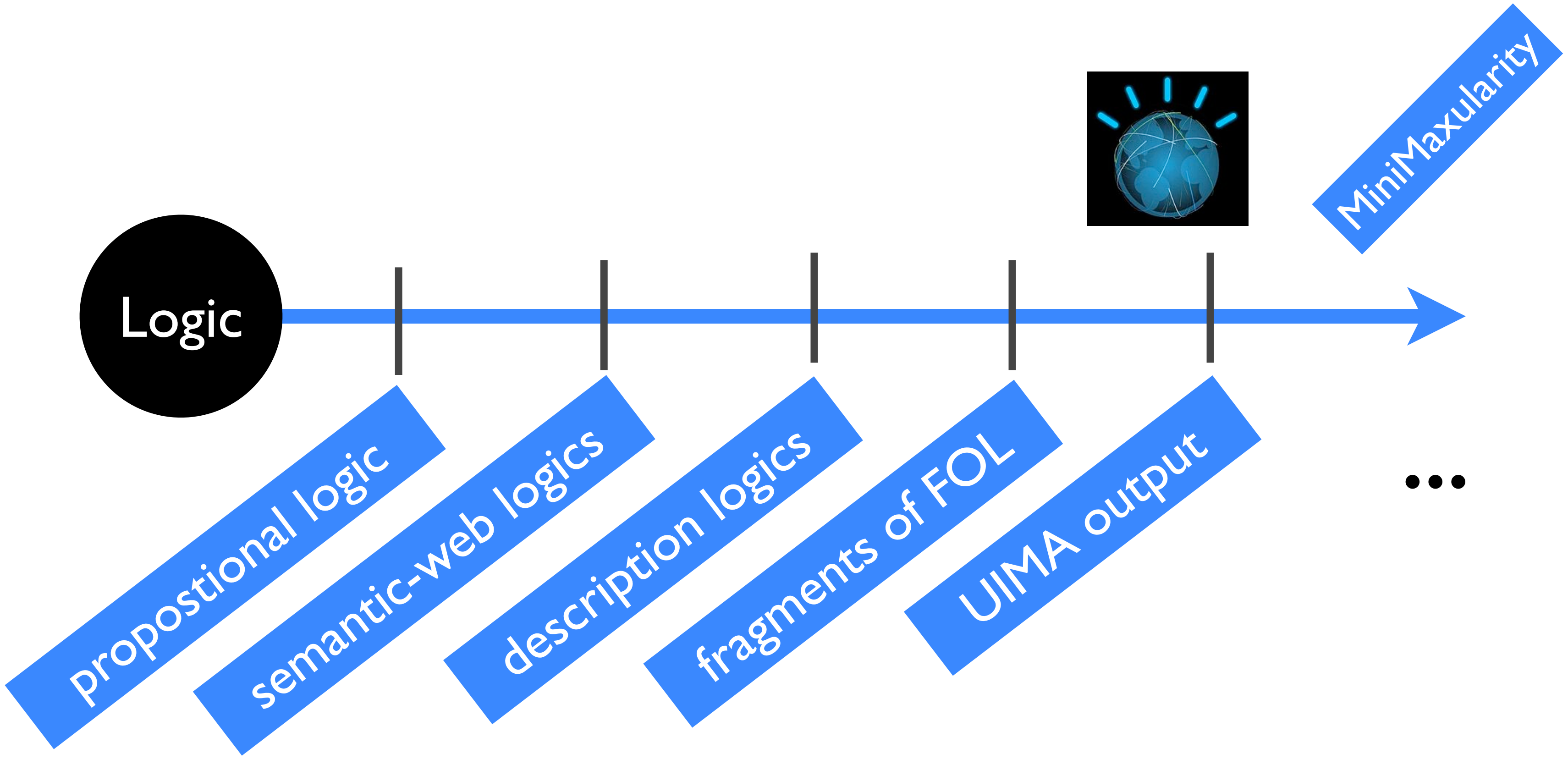
semantic-web logics

description logics

fragments of FOL

UIMA output

MiniMaxularity



Logic

propostional logic

semantic-web logics

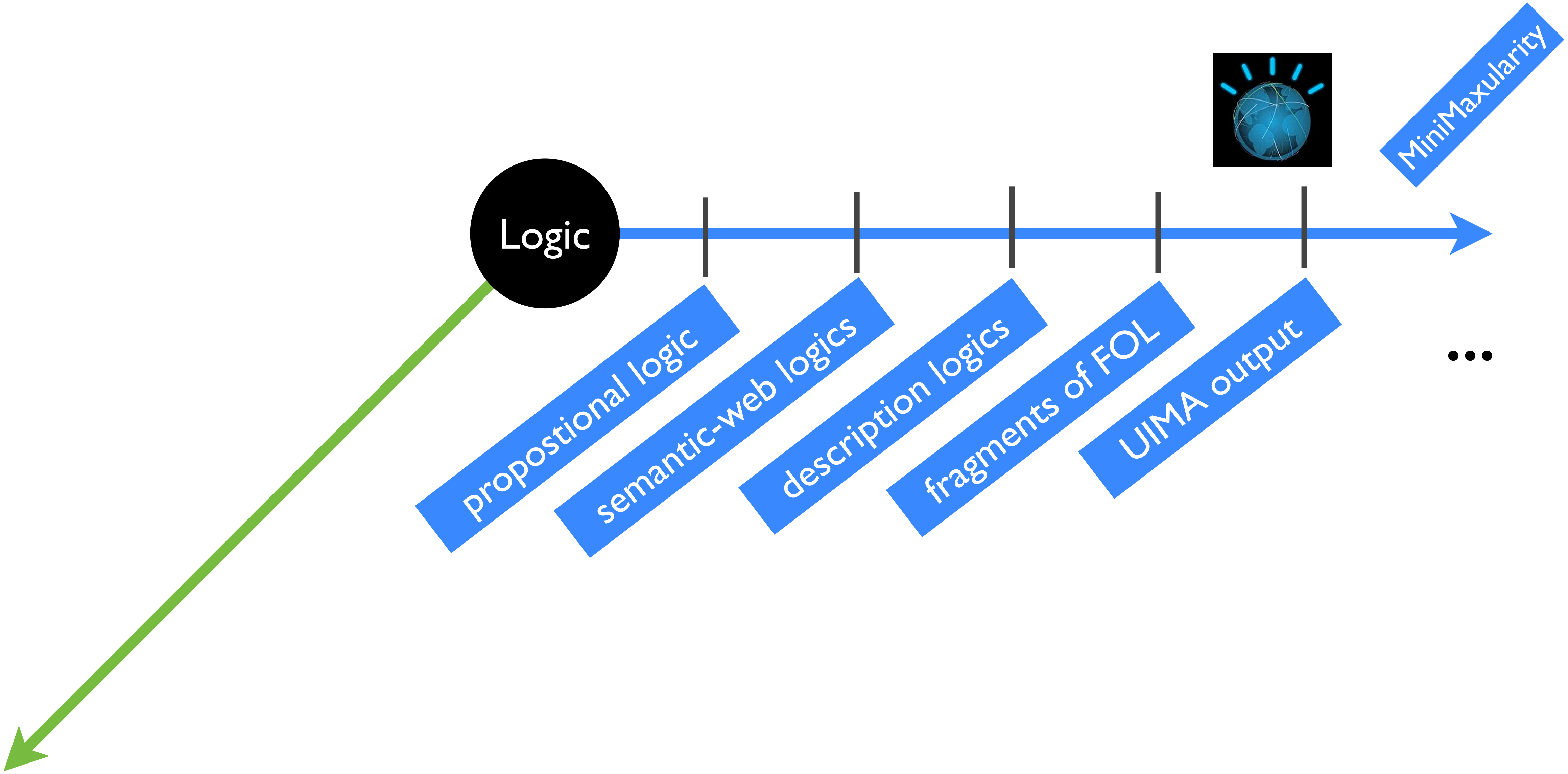
description logics

fragments of FOL

UIMA output

MiniMaxularity

...



Logic

propostional logic

semantic-web logics

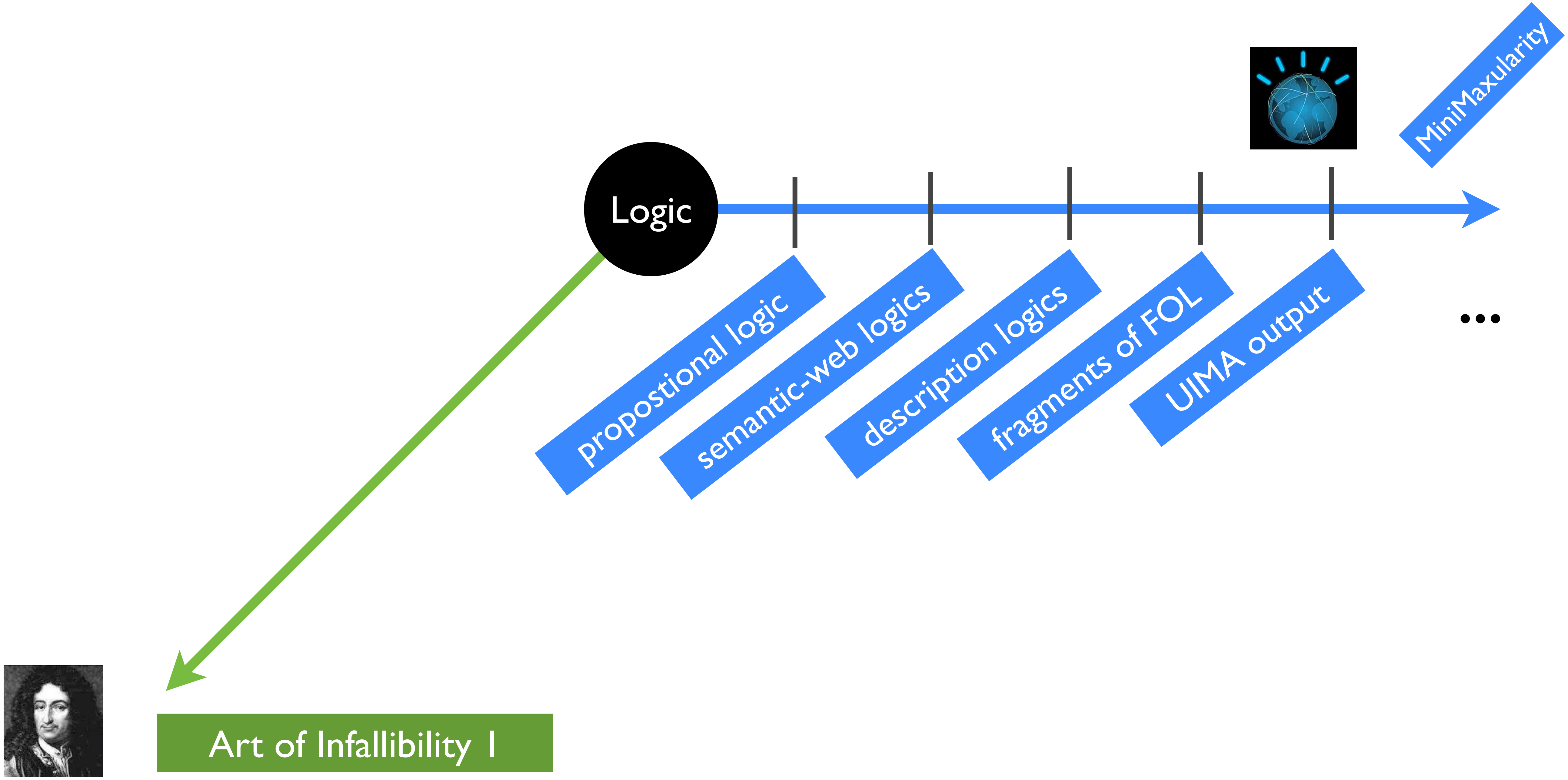
description logics

fragments of FOL

UIMA output

MiniMaxularity

...



Logic

propostional logic

semantic-web logics

description logics

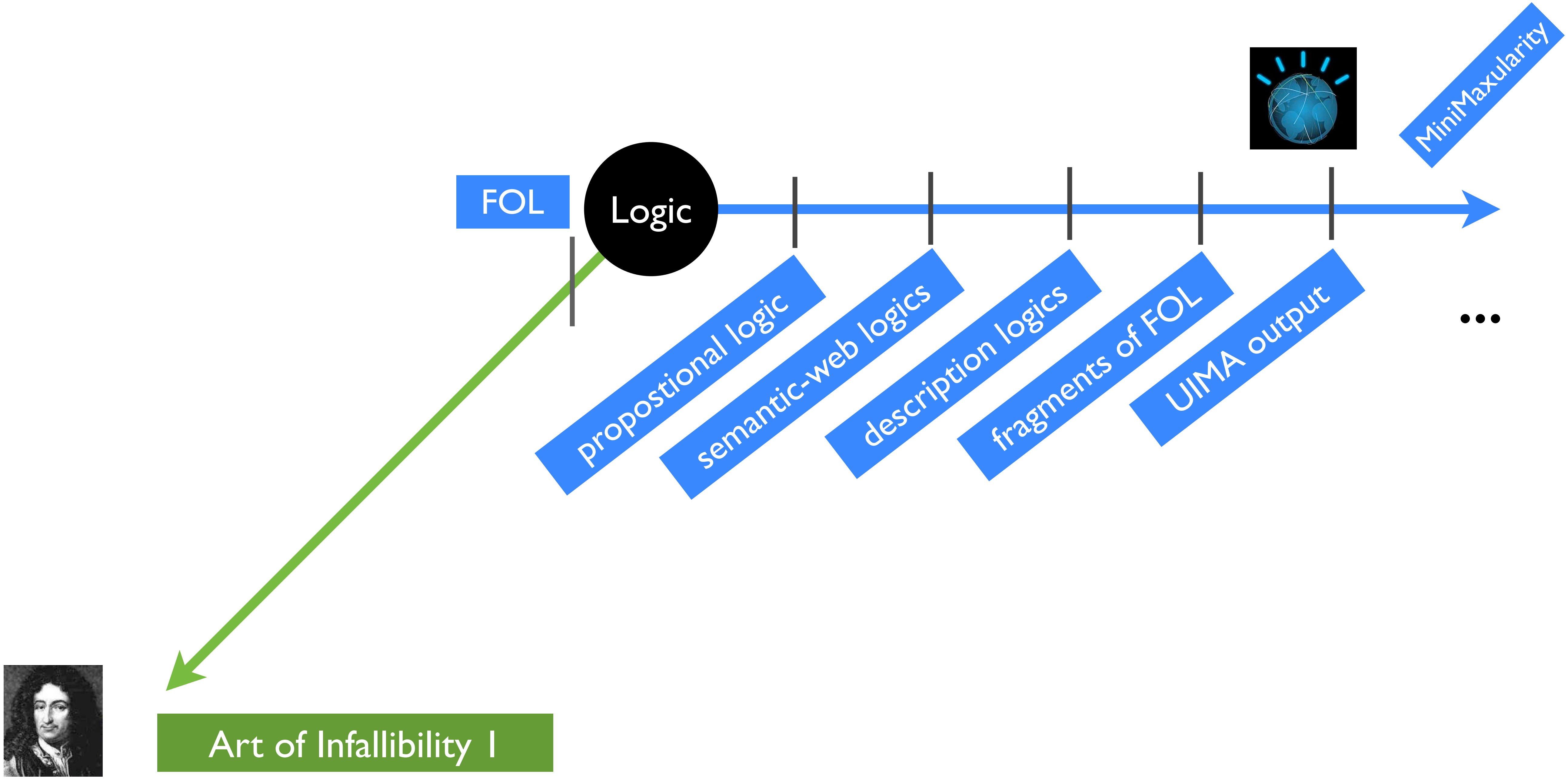
fragments of FOL

UIMA output

MiniMaxularity

...

Art of Infallibility I



FOL

Logic

propostional logic

semantic-web logics

description logics

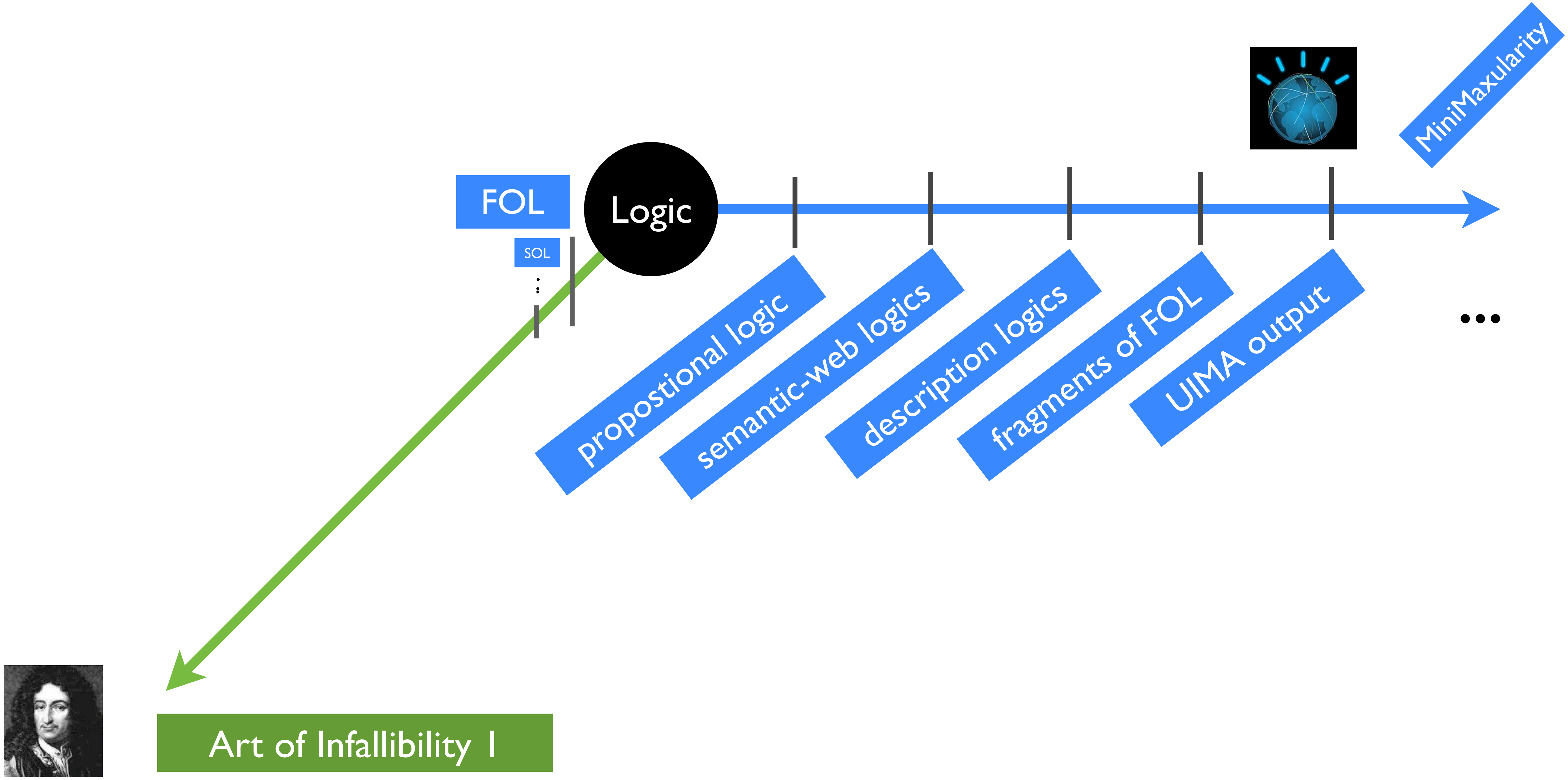
fragments of FOL

UIMA output

MiniMaxularity

...

Art of Infallibility I



FOL

SOL

⋮

Logic

propostional logic

semantic-web logics

description logics

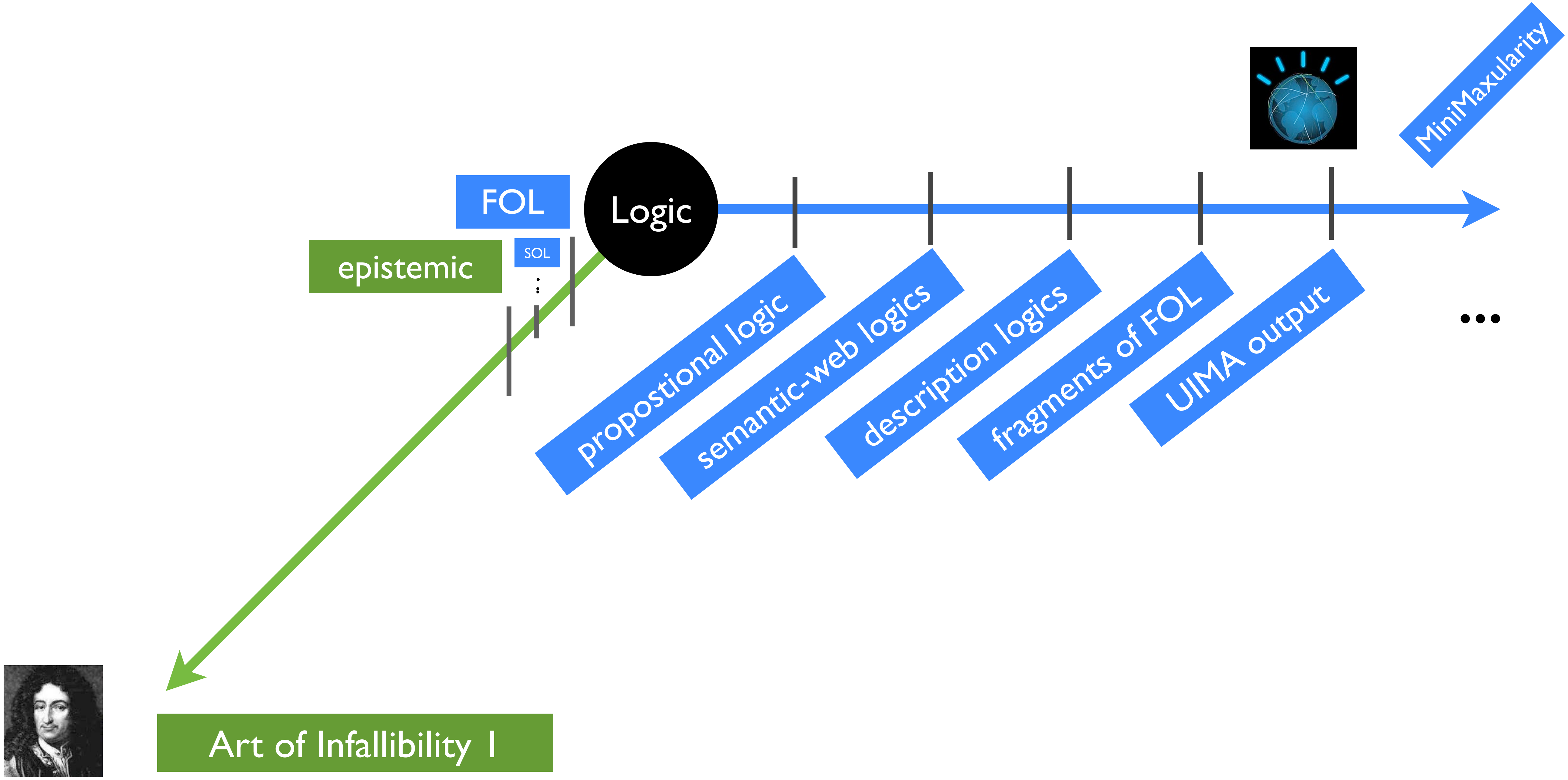
fragments of FOL

UIMA output

MiniMaxularity

...

Art of Infallibility I



FOL

Logic

epistemic

SOL

⋮

propostional logic

semantic-web logics

description logics

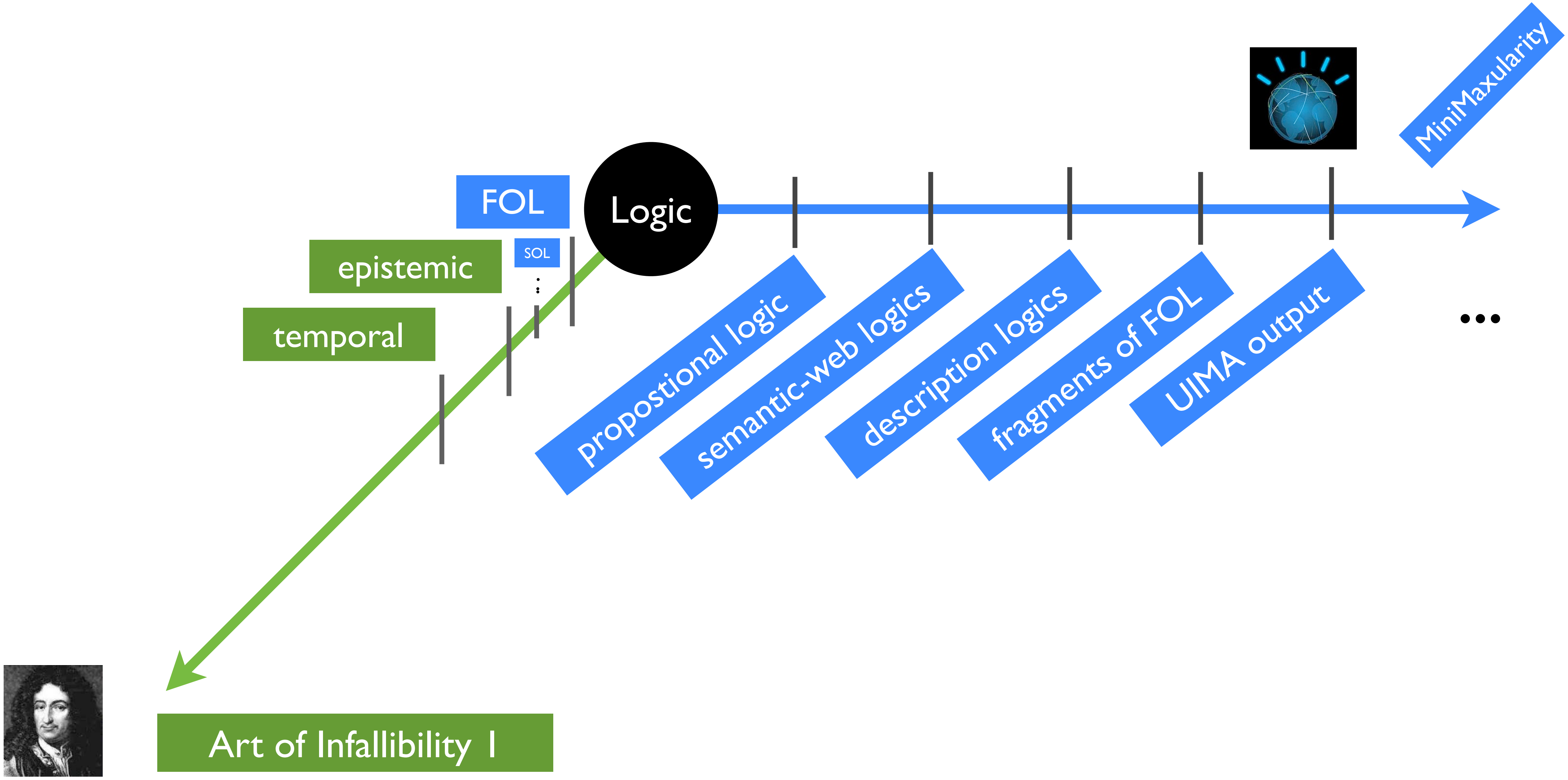
fragments of FOL

UIMA output

MiniMaxularity

...

Art of Infallibility I



FOL

Logic

epistemic

SOL

⋮

temporal

propostional logic

semantic-web logics

description logics

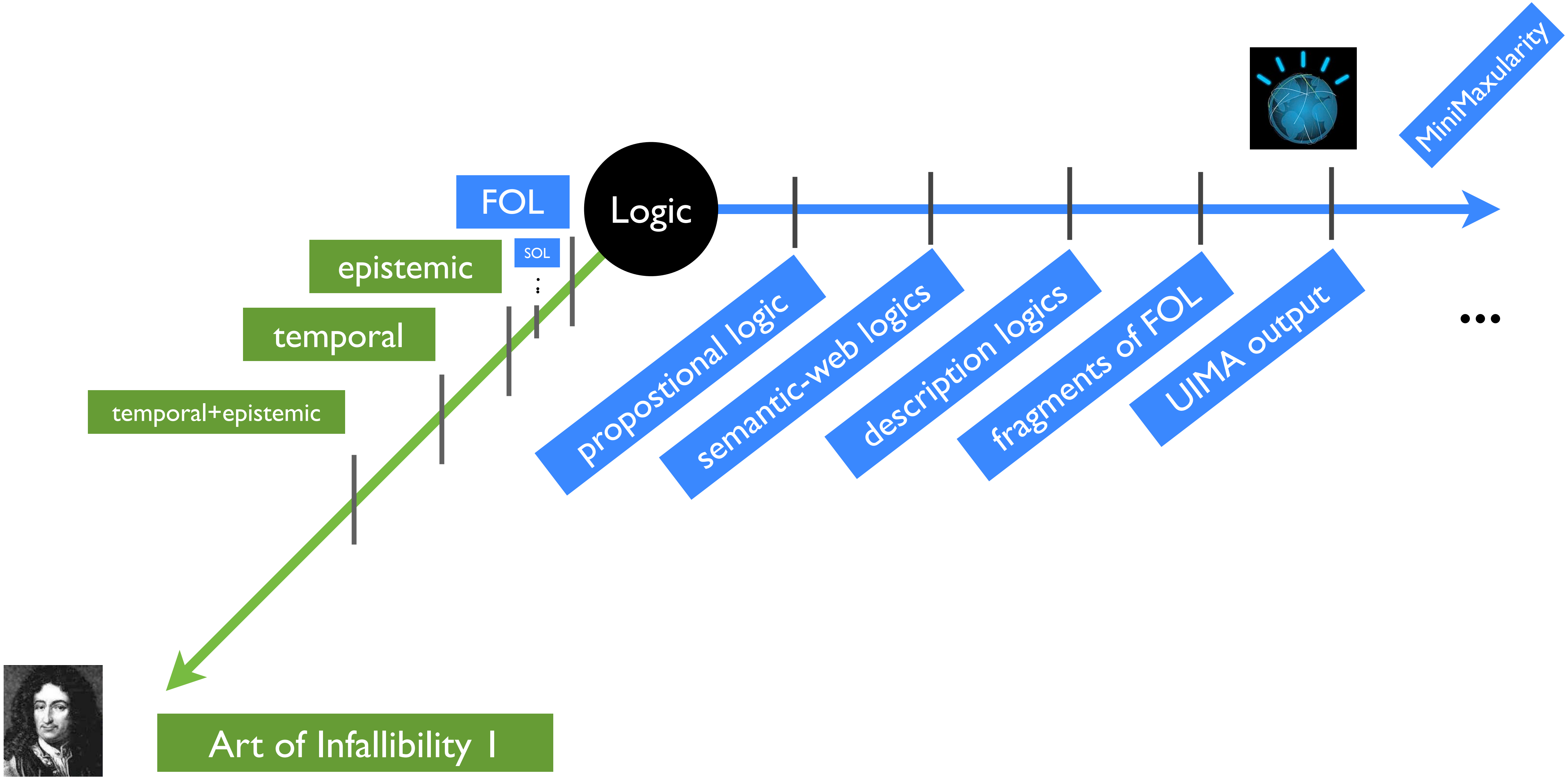
fragments of FOL

UIMA output

MiniMaxularity

…

Art of Infallibility I



FOL

Logic

SOL

⋮

epistemic

temporal

temporal+epistemic

propostional logic

semantic-web logics

description logics

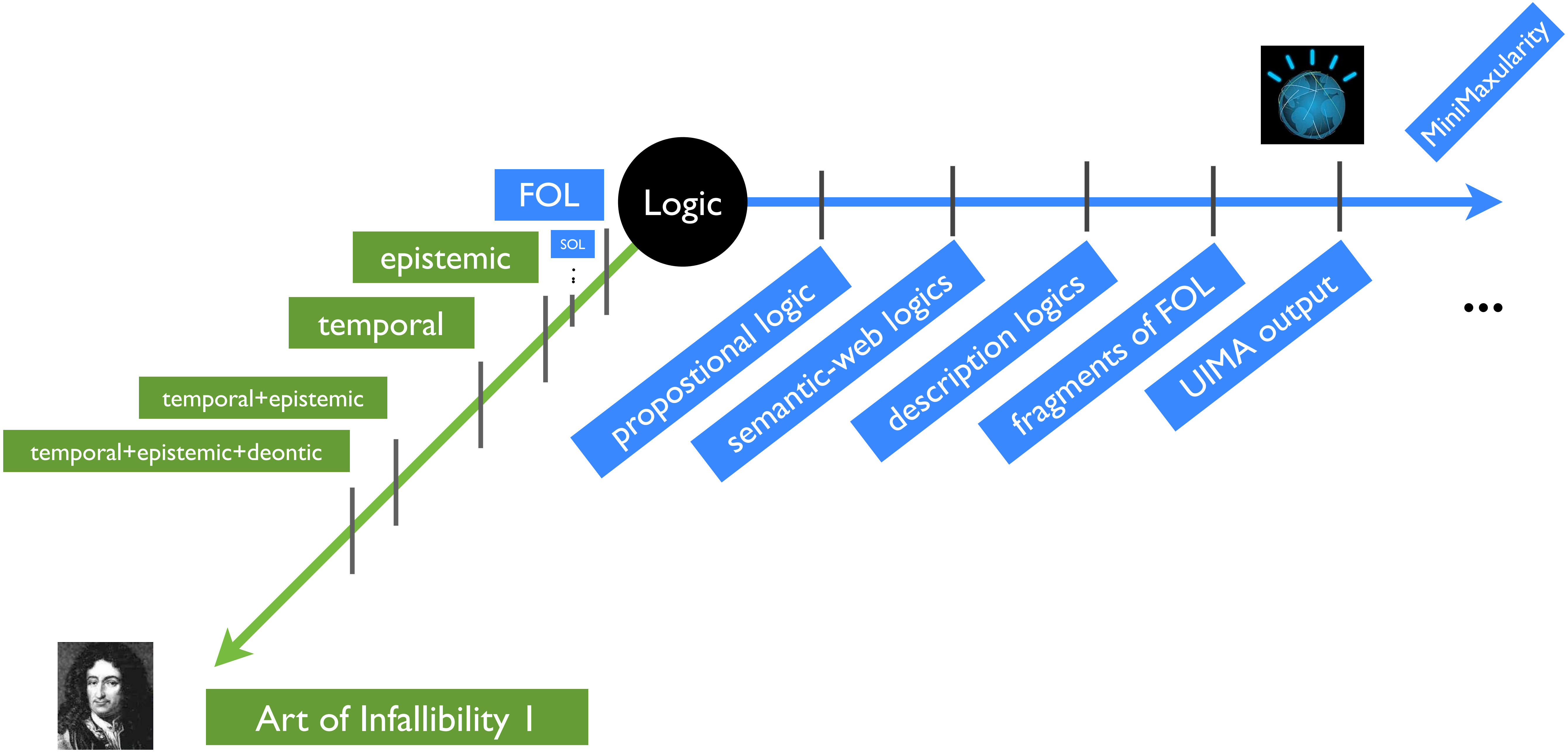
fragments of FOL

UIMA output

MiniMaxularity

...

Art of Infallibility I



Logic

FOL

SOL

epistemic

temporal

temporal+epistemic

temporal+epistemic+deontic

Art of Infallibility I

propostional logic

semantic-web logics

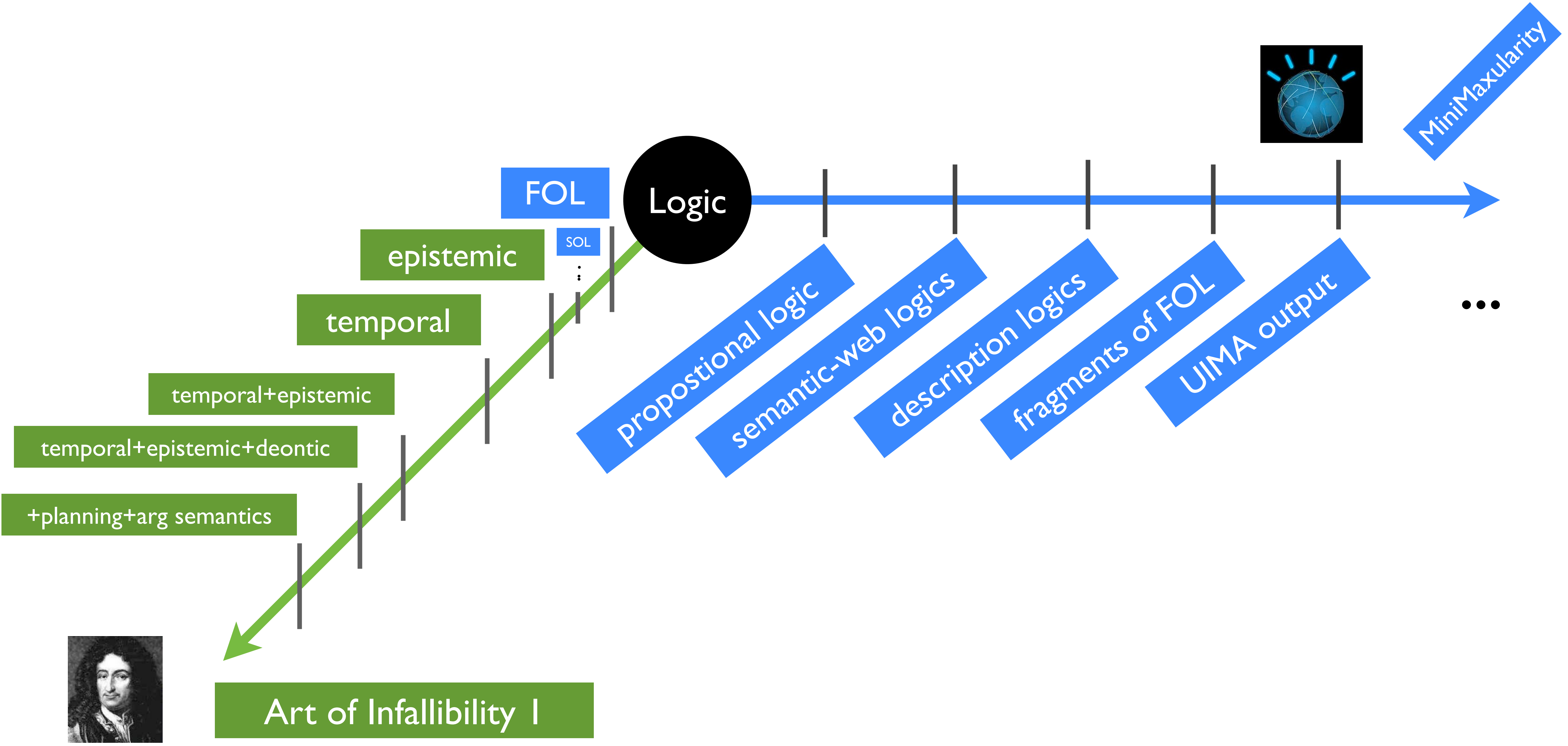
description logics

fragments of FOL

UIMA output

MiniMaxularity

…



MiniMaxularity

FOL

Logic

SOL

⋮

epistemic

temporal

temporal+epistemic

temporal+epistemic+deontic

+planning+arg semantics

Art of Infallibility I

propostional logic

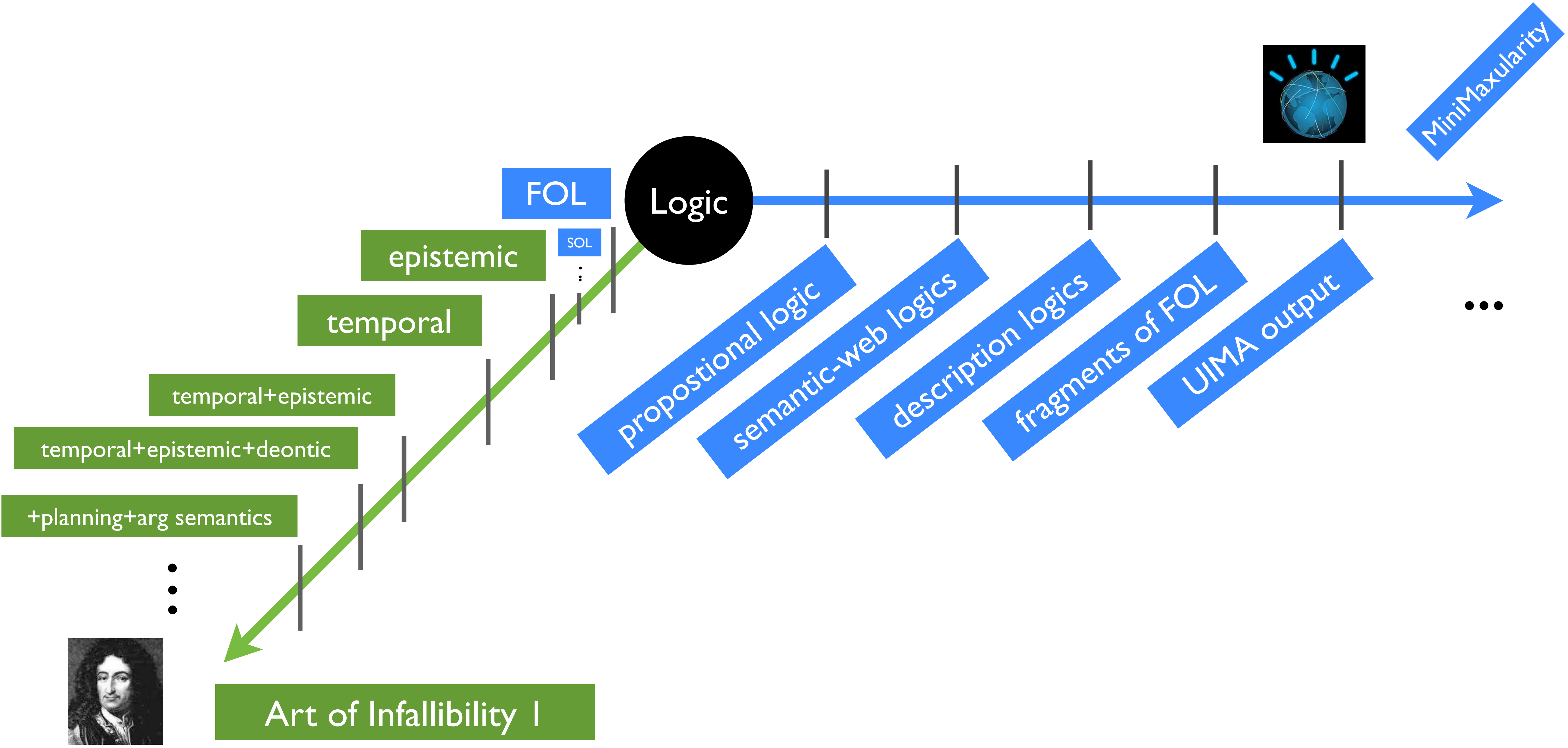
semantic-web logics

description logics

fragments of FOL

UIMA output

…



Logic

FOL

SOL

epistemic

temporal

temporal+epistemic

temporal+epistemic+deontic

+planning+arg semantics

Art of Infallibility I

propostional logic

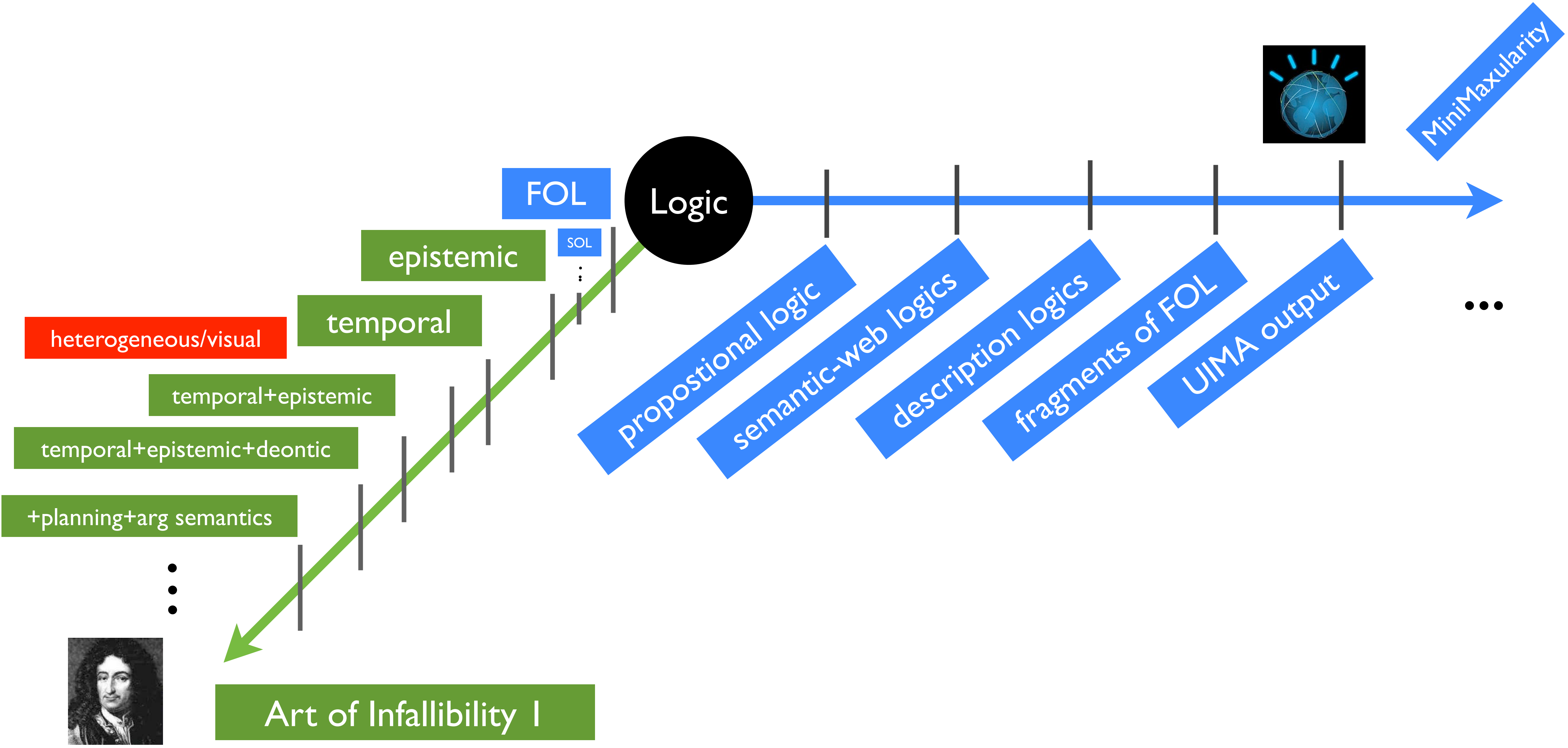
semantic-web logics

description logics

fragments of FOL

UIMA output

MiniMaxularity



Logic

FOL

SOL

⋮

epistemic

temporal

heterogeneous/visual

temporal+epistemic

temporal+epistemic+deontic

+planning+arg semantics

⋮

Art of Infallibility I

propostional logic

semantic-web logics

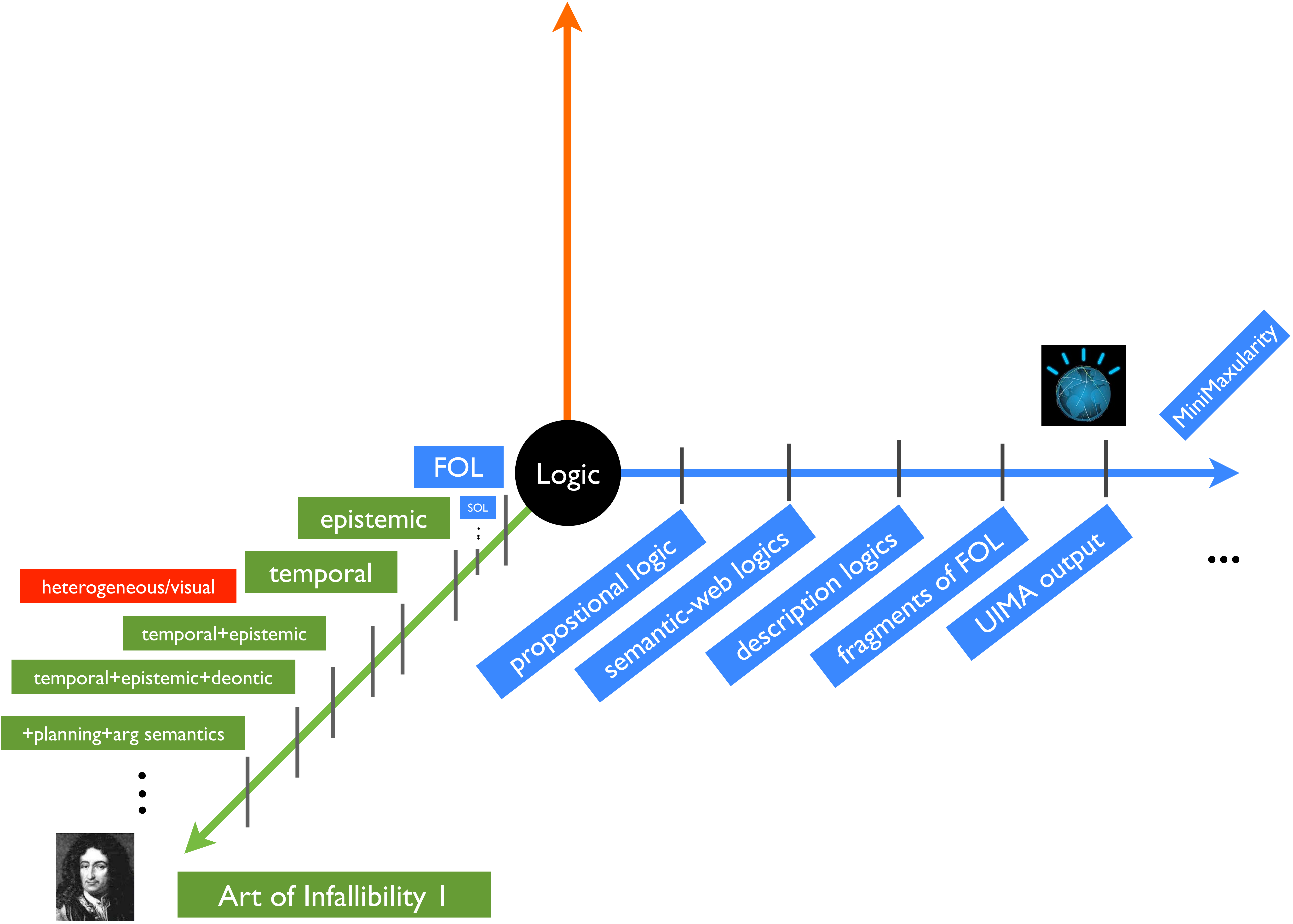
description logics

fragments of FOL

UIMA output

MiniMaxularity

…



Logic

FOL

SOL

epistemic

temporal

heterogeneous/visual

temporal+epistemic

temporal+epistemic+deontic

+planning+arg semantics

Art of Infallibility I

propostional logic

semantic-web logics

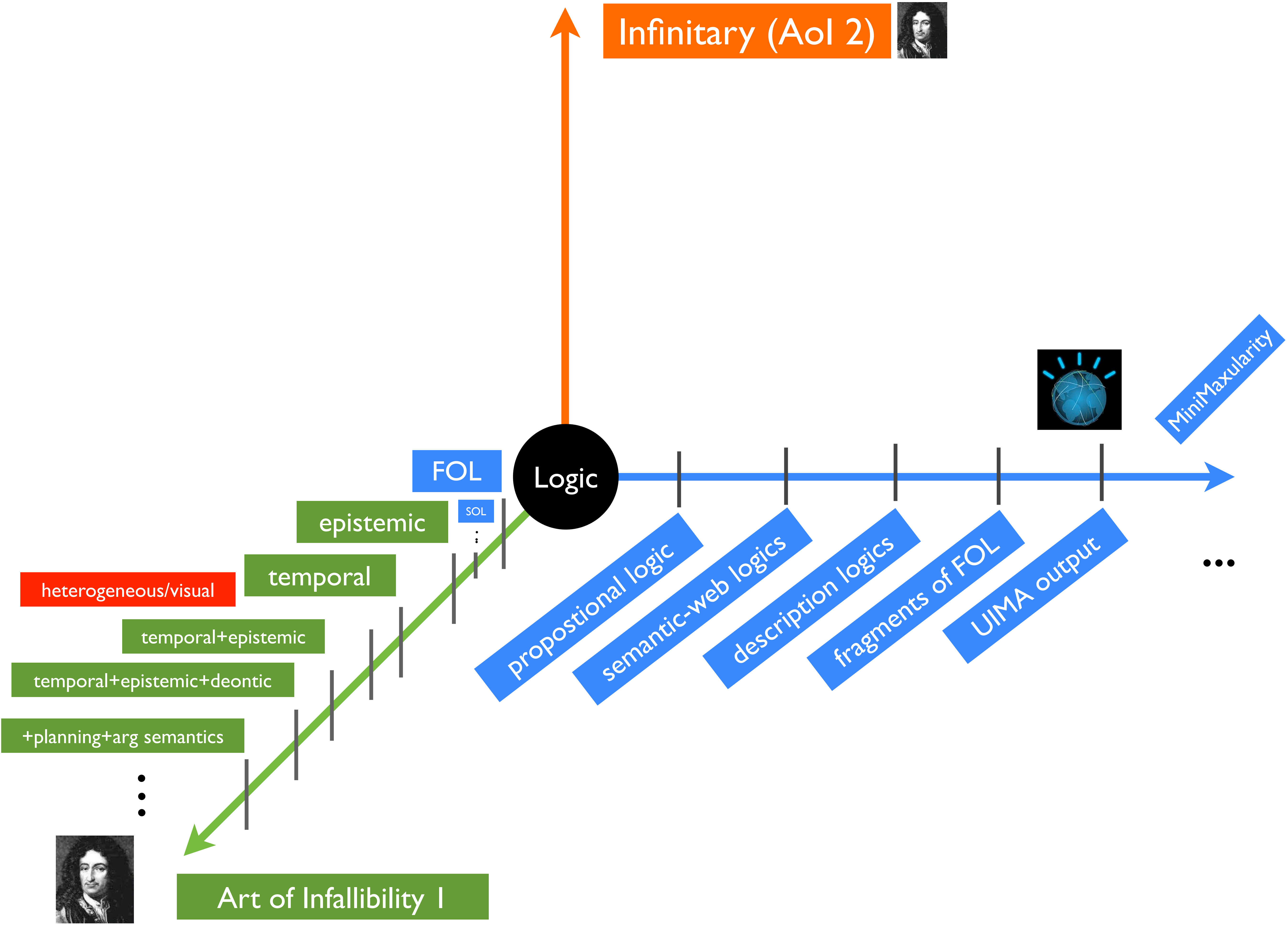
description logics

fragments of FOL

UIMA output

MiniMaxularity

...



Infinitary (AoI 2)

Logic

FOL

SOL

epistemic

temporal

heterogeneous/visual

temporal+epistemic

temporal+epistemic+deontic

+planning+arg semantics

Art of Infallibility I

propostional logic

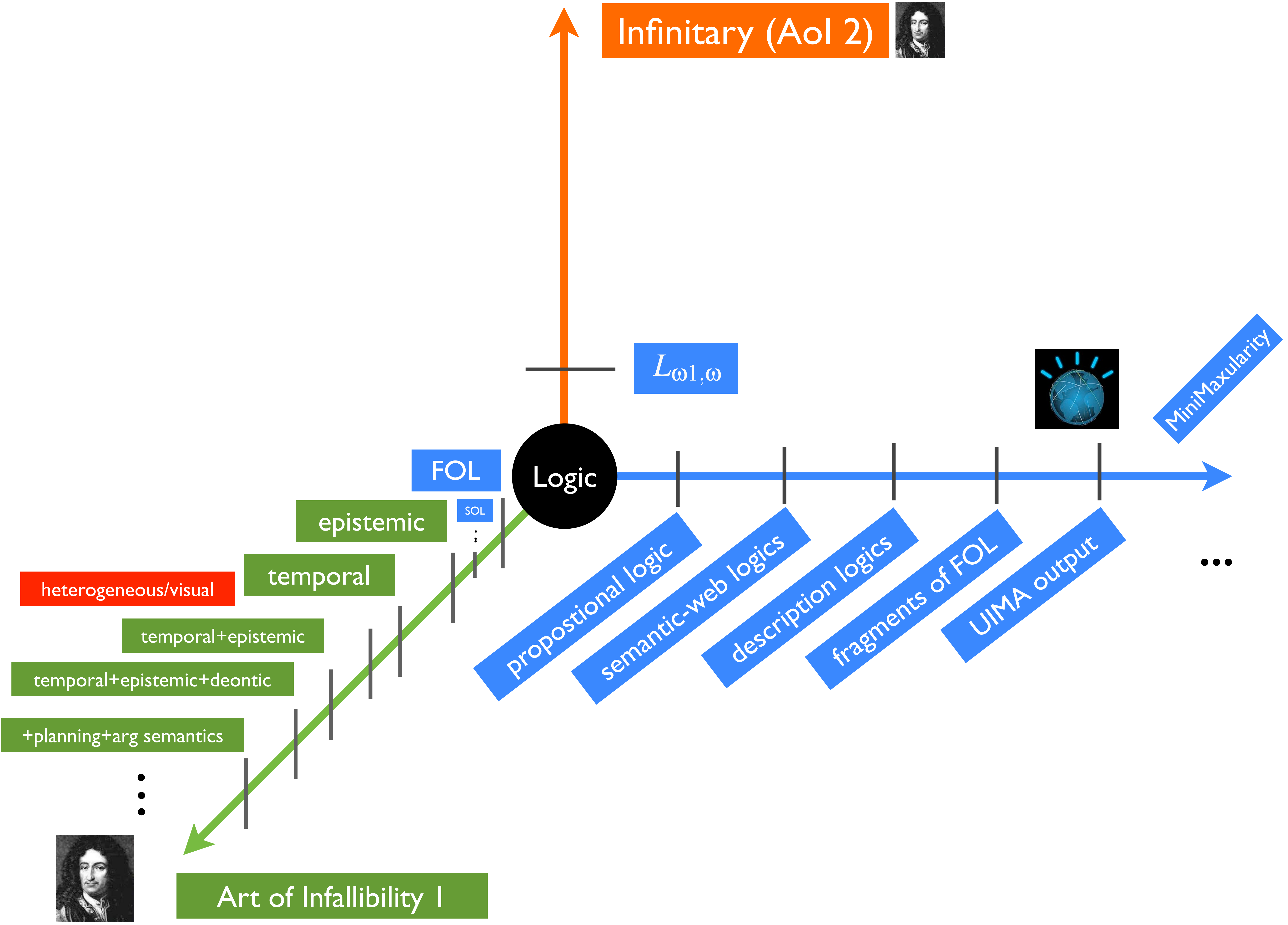
semantic-web logics

description logics

fragments of FOL

UIMA output

MiniMaxularity



Infinitary (AoI 2)

$L_{\omega1,\omega}$

MiniMaxularity

FOL

Logic

SOL

epistemic

temporal

heterogeneous/visual

temporal+epistemic

temporal+epistemic+deontic

+planning+arg semantics

Art of Infallibility I

propostional logic

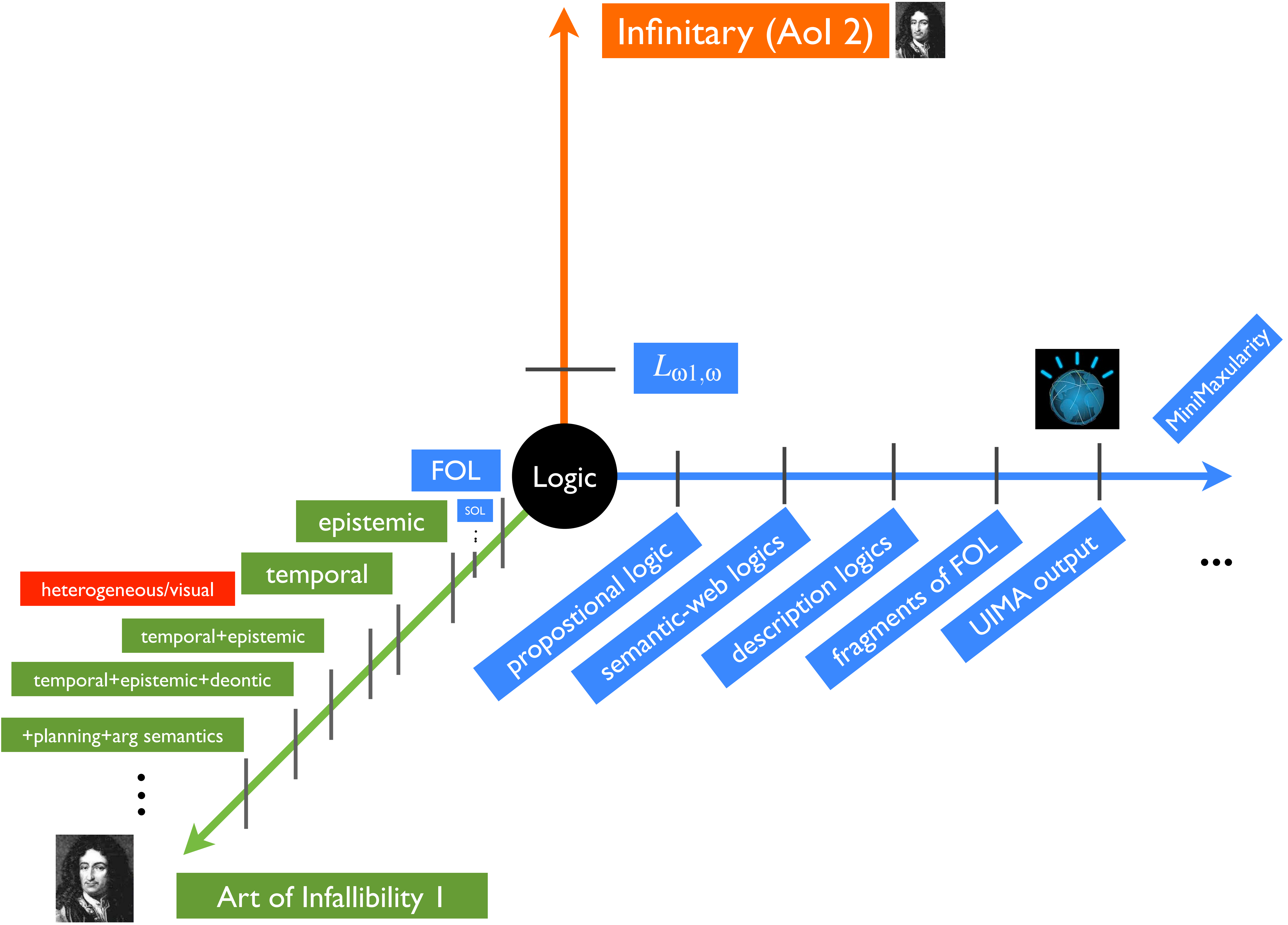
semantic-web logics

description logics

fragments of FOL

UIMA output

...



Infinitary (AoI 2)

$L_{\omega 1,\omega}$

MiniMaxularity

FOL

Logic

SOL

propostional logic

semantic-web logics

description logics

fragments of FOL

UIMA output

...

epistemic

temporal

heterogeneous/visual

temporal+epistemic

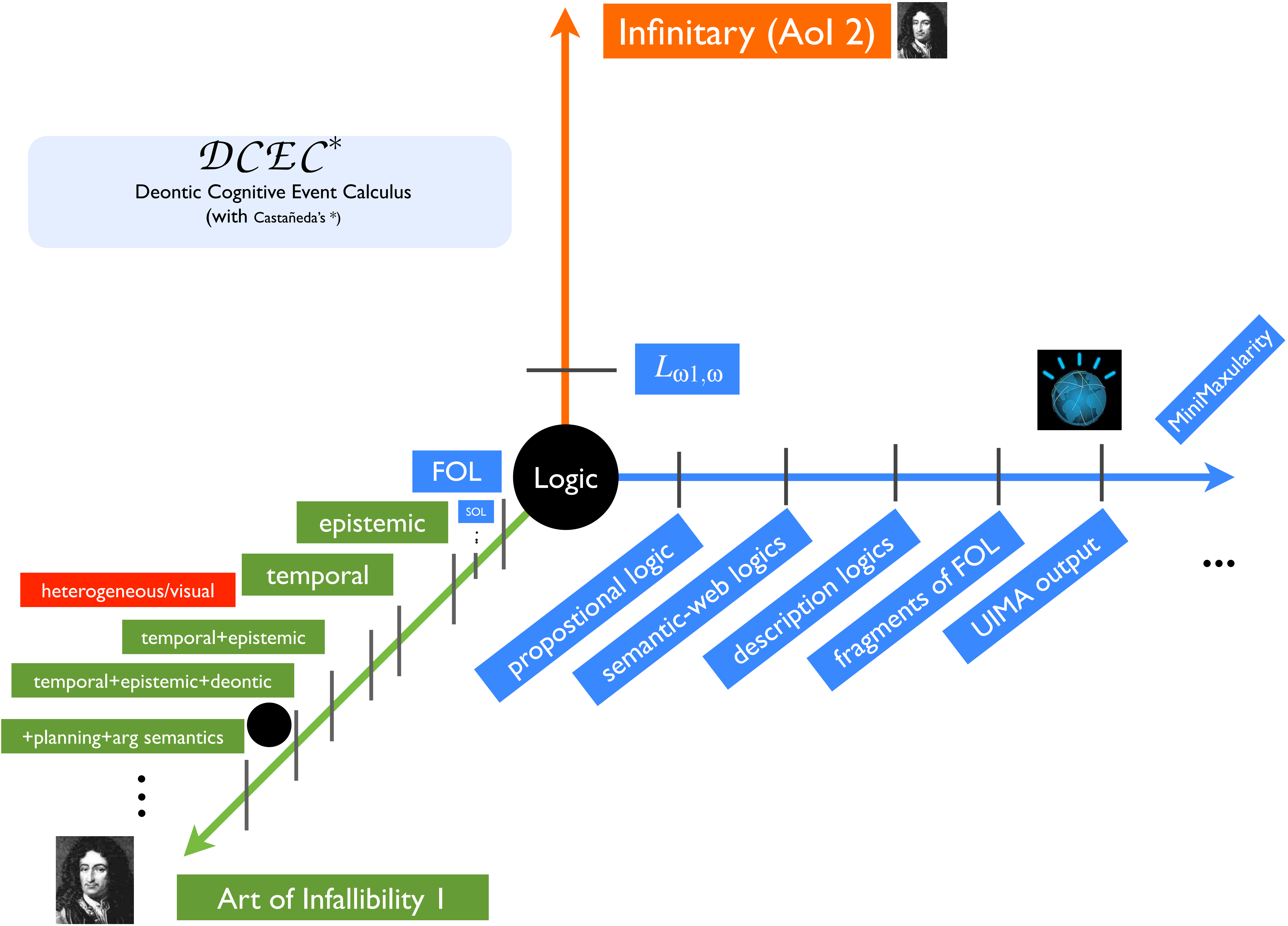
temporal+epistemic+deontic

+planning+arg semantics

Art of Infallibility I

$\mathcal{DCEC}^*$

Deontic Cognitive Event Calculus
(with Castañeda's *)



Infinitary (AoI 2)

$L_{\omega 1,\omega}$

Logic

FOL

SOL

⋮

epistemic

temporal

heterogeneous/visual

temporal+epistemic

temporal+epistemic+deontic

+planning+arg semantics

⋮

Art of Infallibility I

propostional logic

semantic-web logics

description logics

fragments of FOL

UIMA output

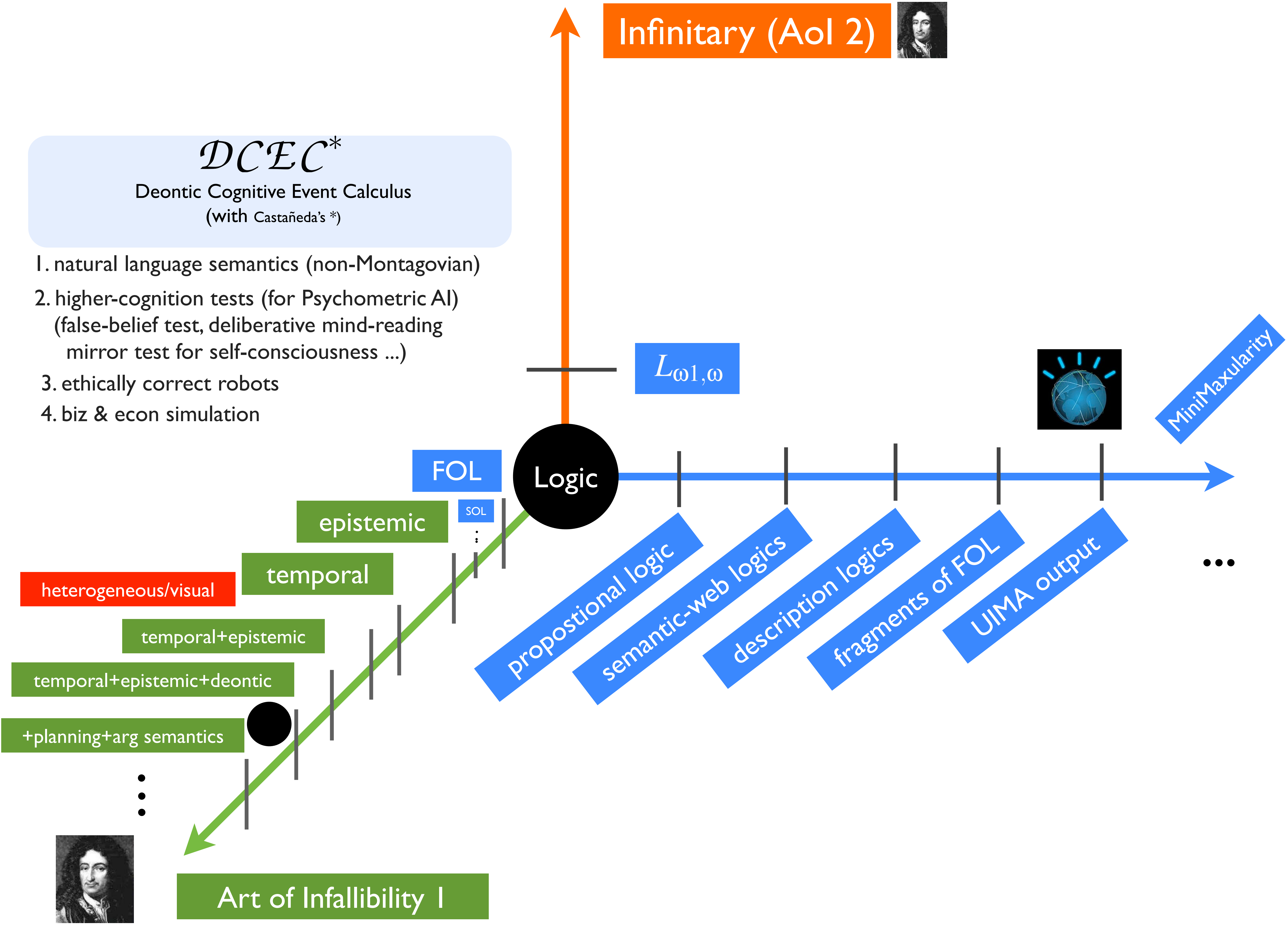
MiniMaxularity

...

$\mathcal{DCEC}^*$

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



Infinitary (AoI 2)

$L_{\omega1,\omega}$

Logic

FOL

SOL

propostional logic

semantic-web logics

description logics

fragments of FOL

UIMA output

MiniMaxularity

...

epistemic

temporal

heterogeneous/visual

temporal+epistemic

temporal+epistemic+deontic

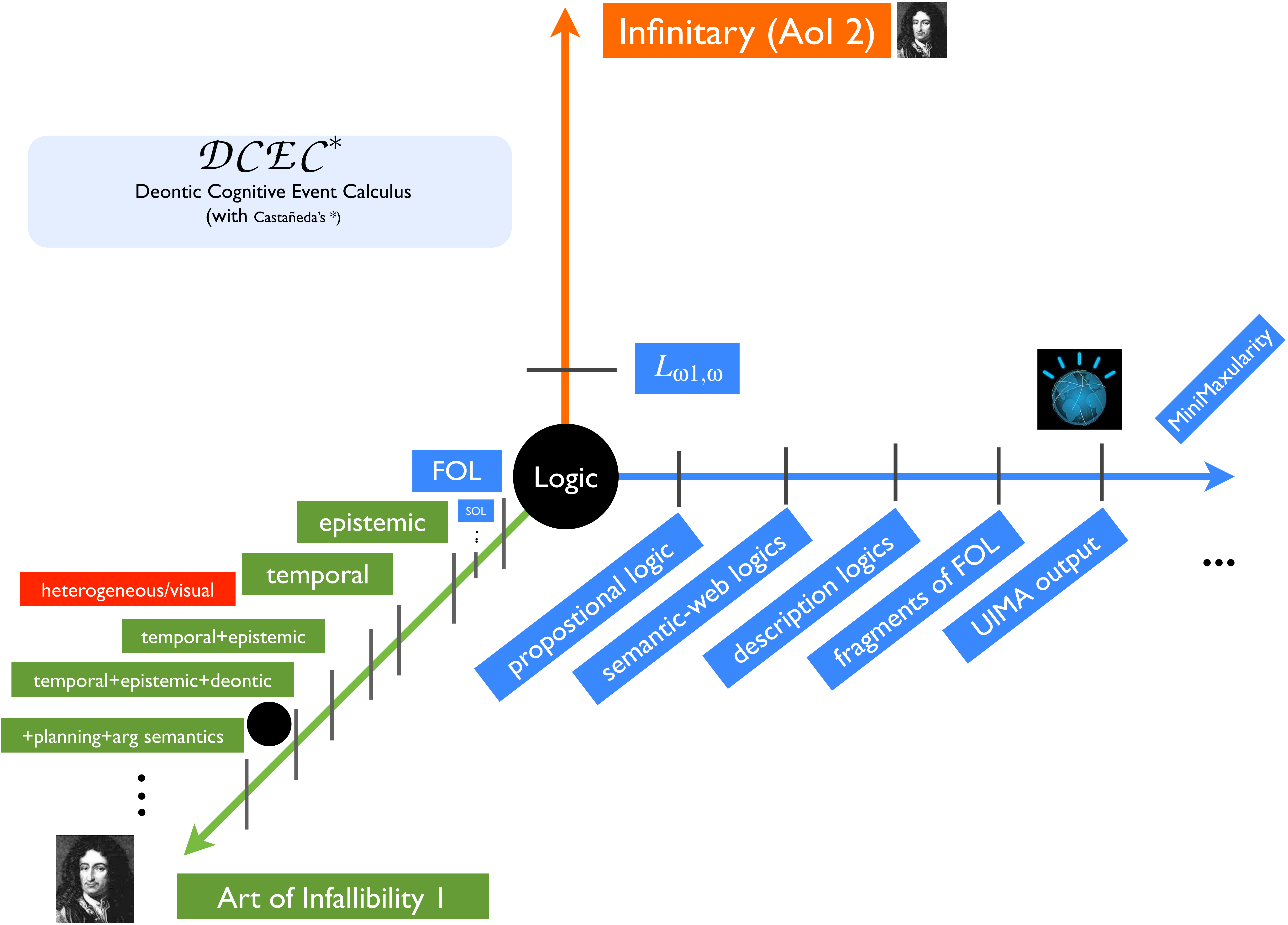
+planning+arg semantics

⋮

Art of Infallibility I

$\mathcal{DCEC}^*$

Deontic Cognitive Event Calculus
(with Castañeda's *)



Infinitary (AoI 2)

$L_{\omega 1,\omega}$

Logic

FOL

SOL

epistemic

temporal

heterogeneous/visual

temporal+epistemic

temporal+epistemic+deontic

+planning+arg semantics

⋮

Art of Infallibility I

propostional logic

semantic-web logics

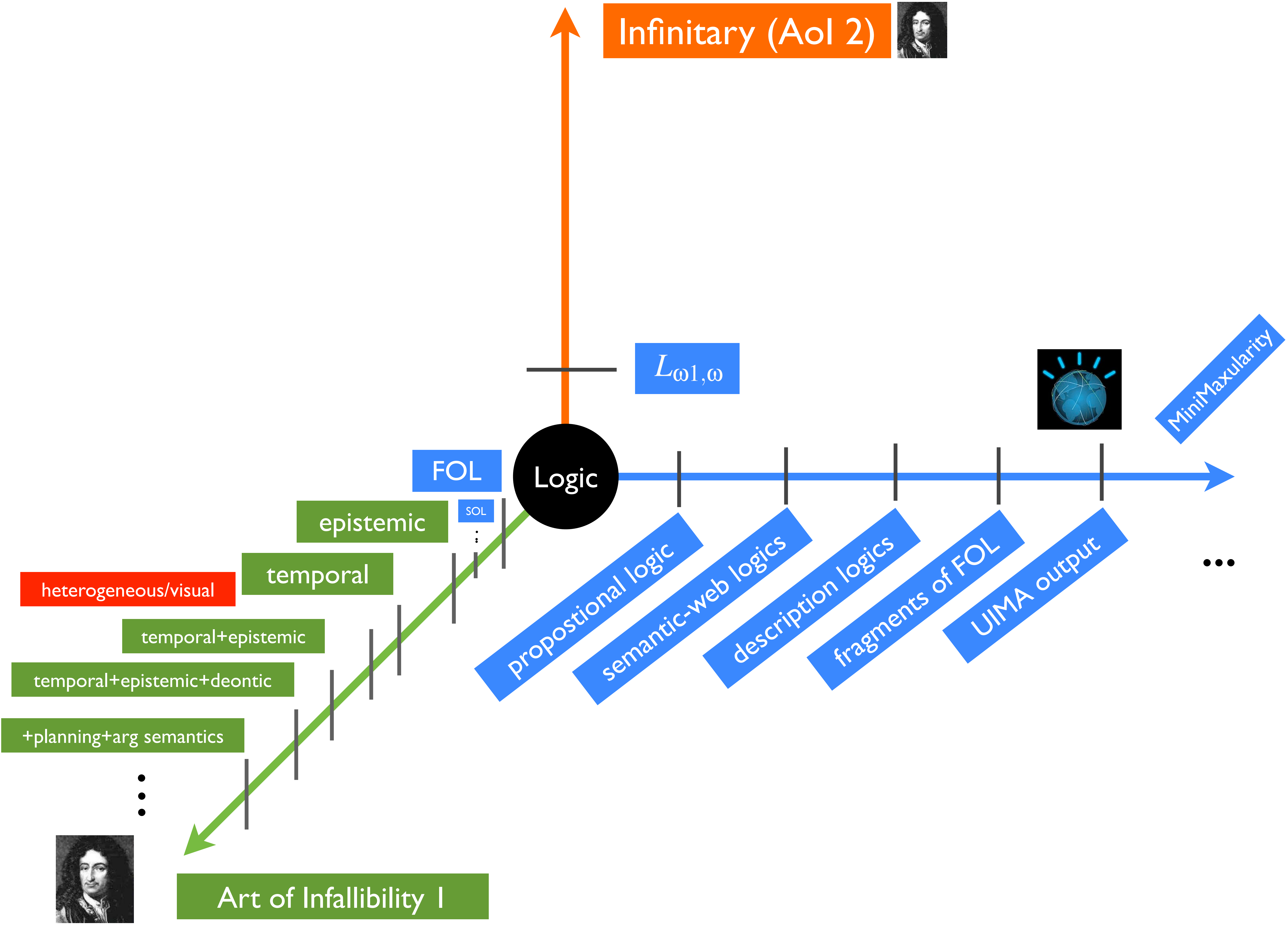
description logics

fragments of FOL

UIMA output

MiniMaxularity

...



Infinitary (AoI 2)

$L_{\omega 1,\omega}$

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

Hierarchy of Ethical Reasoning

Hierarchy of Ethical Reasoning

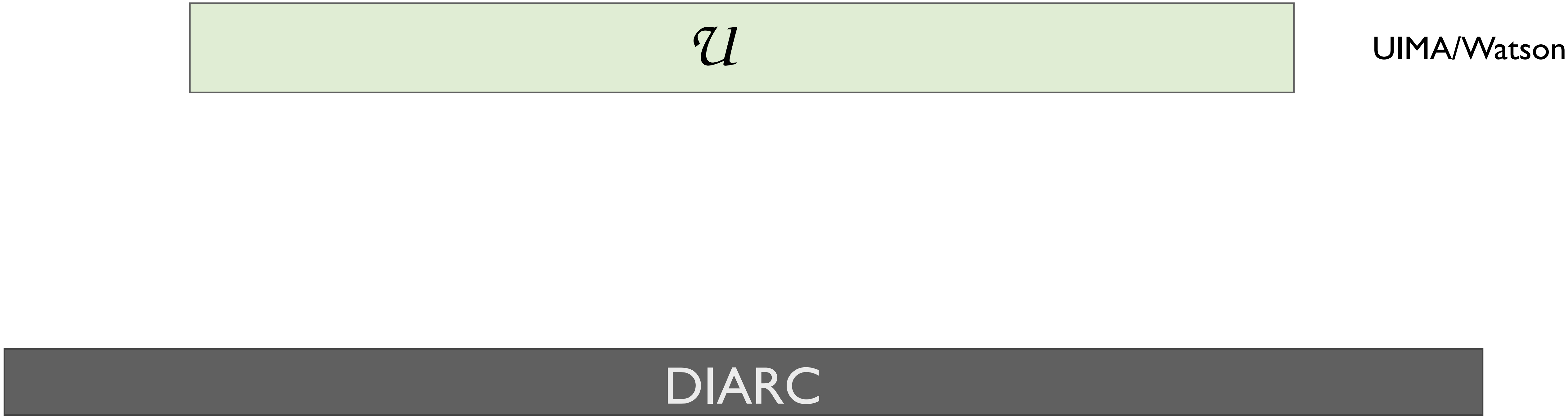
DIARC

Hierarchy of Ethical Reasoning

$\mathcal{U}$

DIARC

Hierarchy of Ethical Reasoning

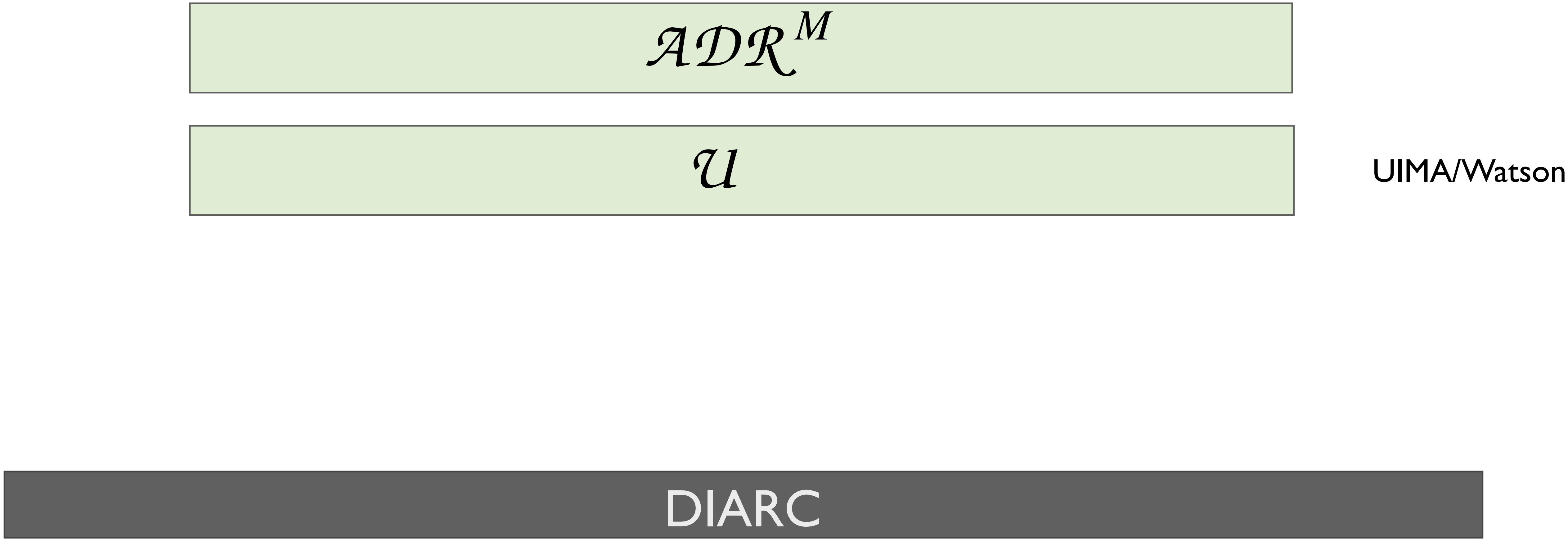


$\mathcal{U}$

UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



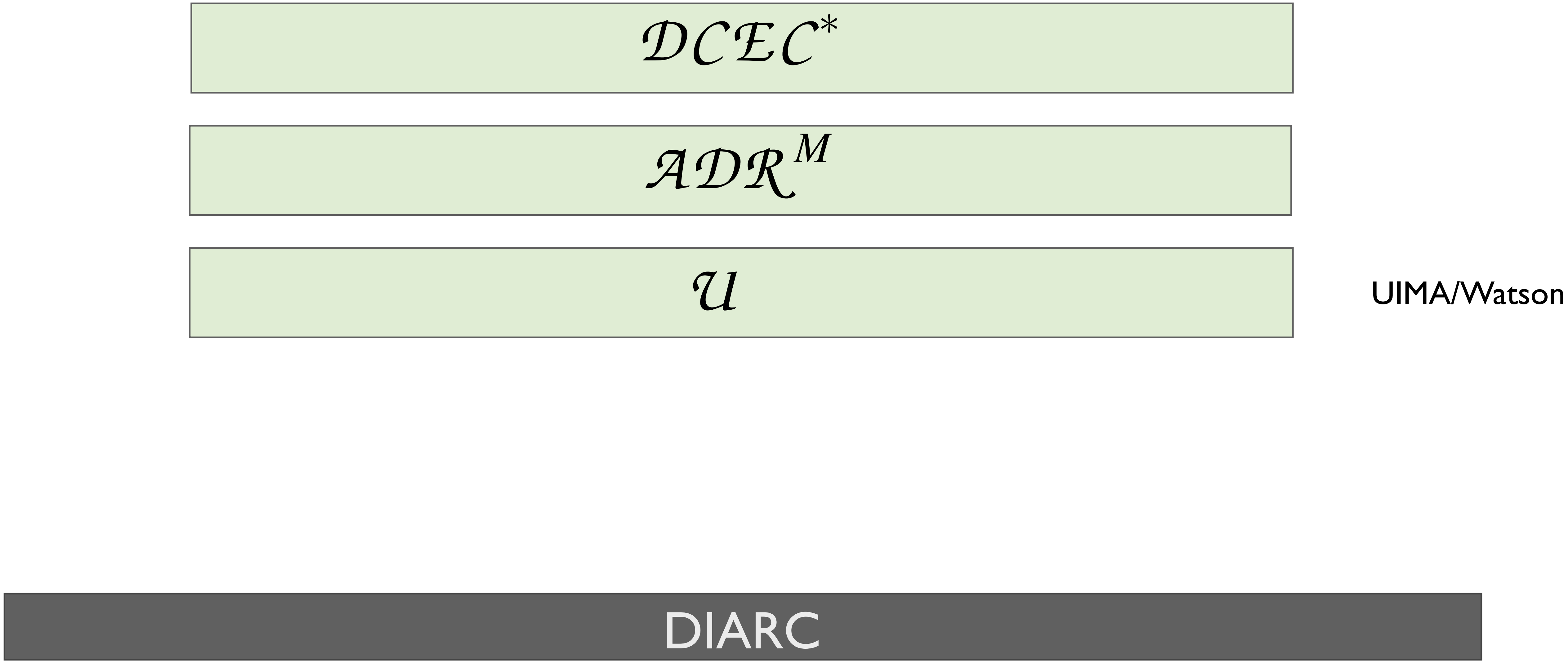
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



$\mathcal{DCEC}^*$

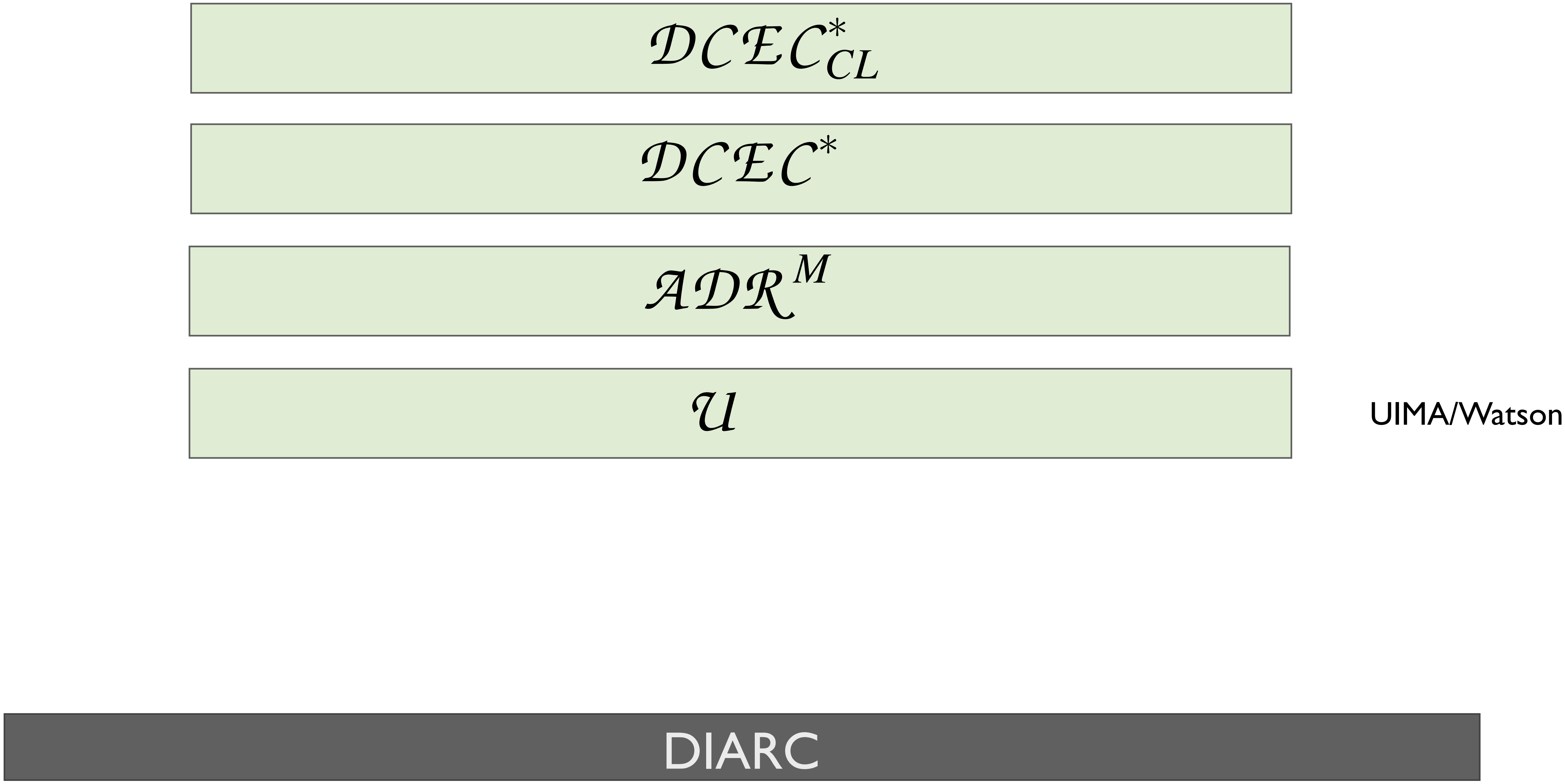
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

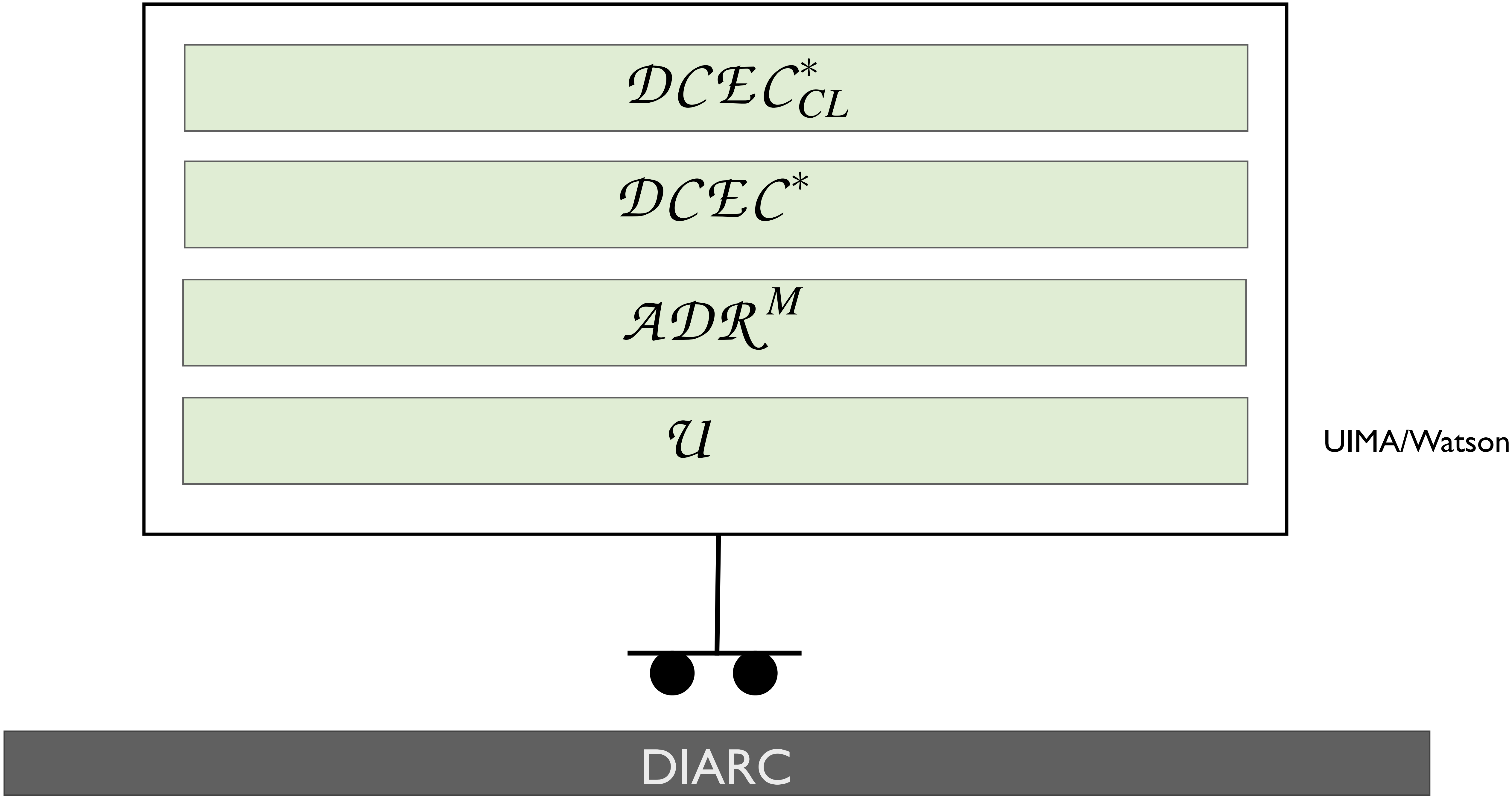
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

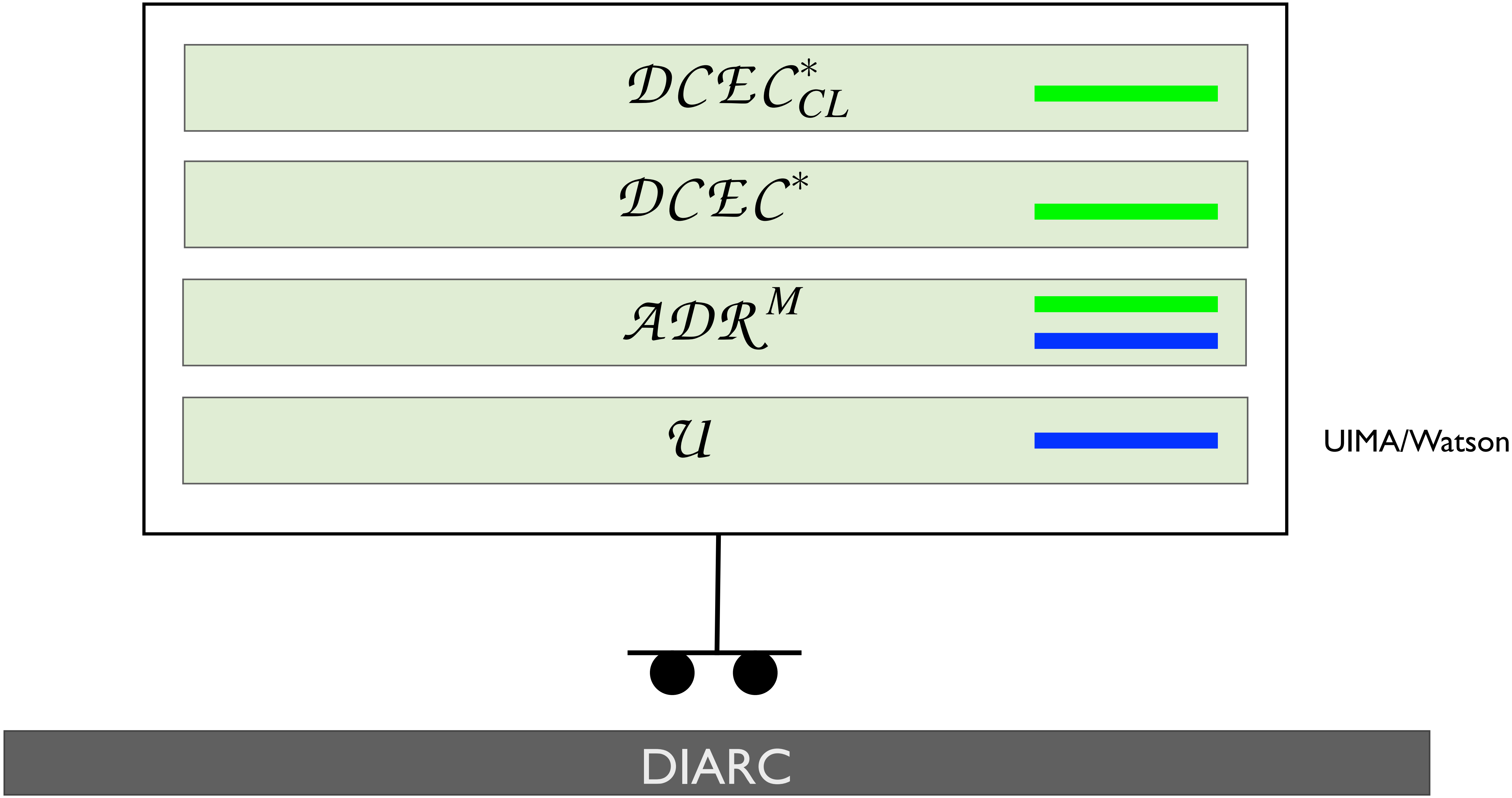
$\mathcal{ADR}^M$

$\mathcal{U}$

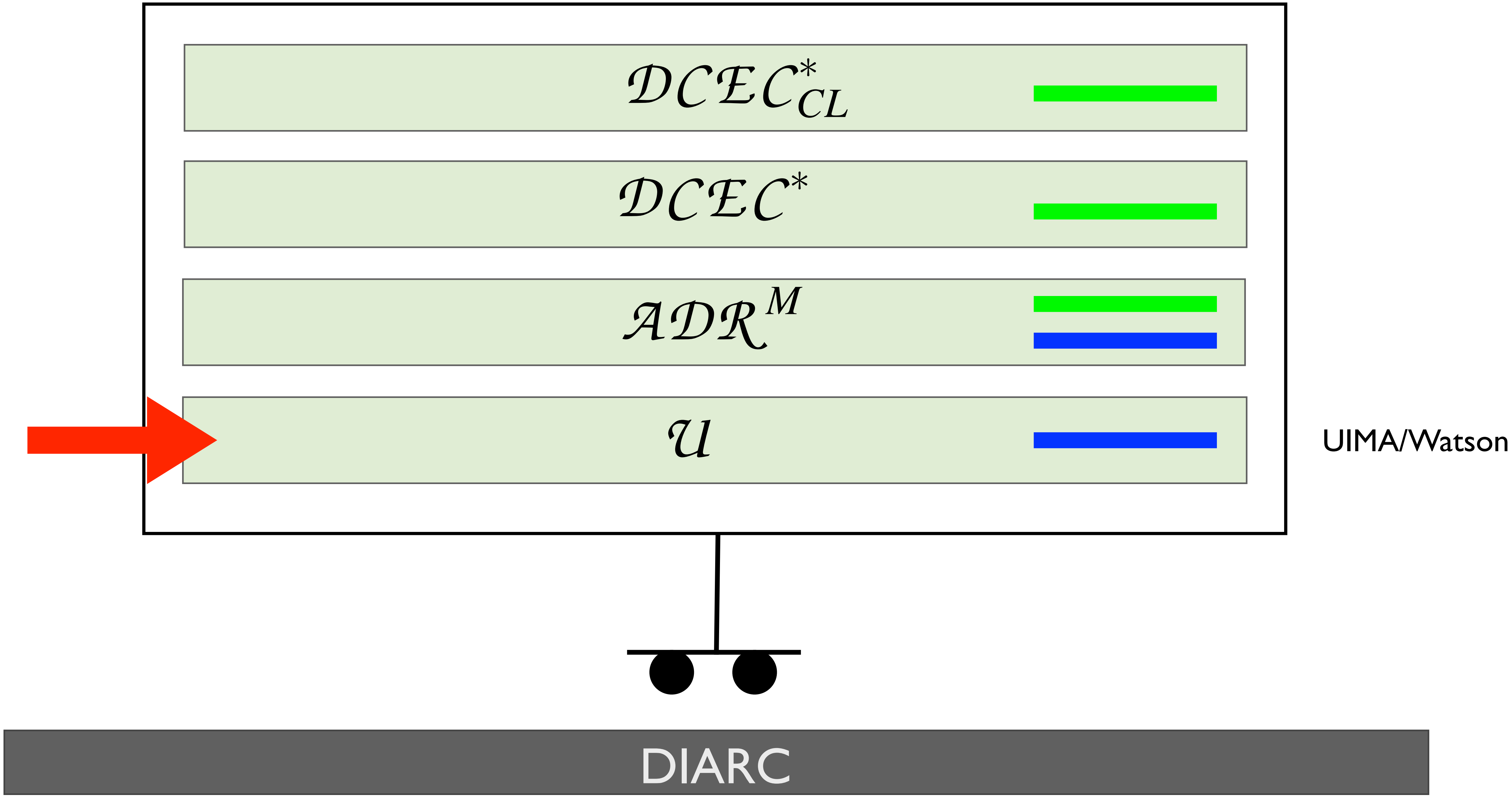
UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson

DIARC

JEOPARDY!
Ian
Sarah
Nigel
CLICK HERE TO JOIN NOW

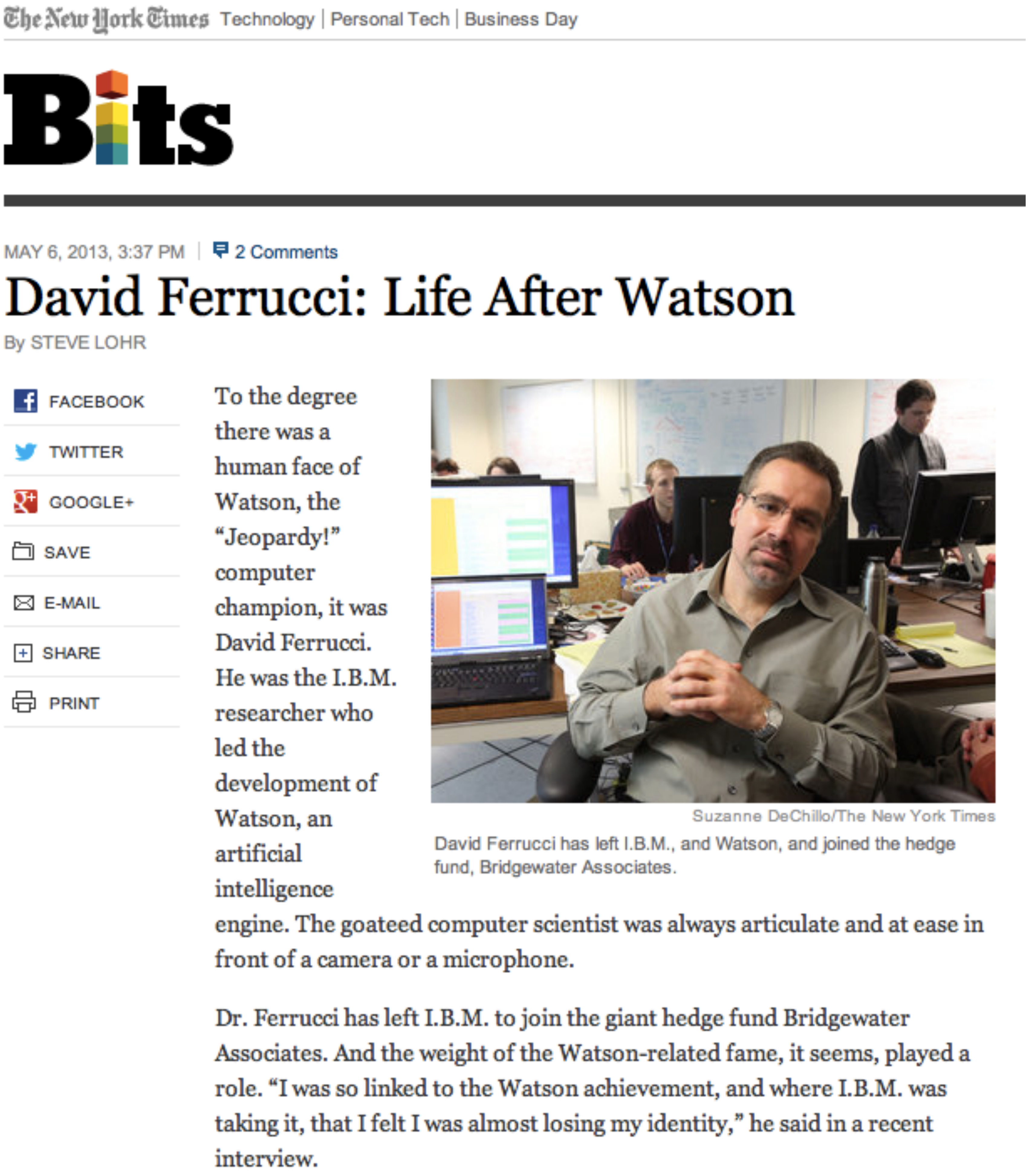
JEOPARDY!
Ian
Sarah
Nigel
CLICK HERE TO JOIN NOW

Many experts to IBM: “Can’t be done!”

Many experts to IBM: “Can’t be done!”

No one asked me.

From computational logic for configuration and design to ...



The New York Times Technology | Personal Tech | Business Day

Bits

MAY 6, 2013, 3:37 PM | 2 Comments

David Ferrucci: Life After Watson

By STEVE LOHR

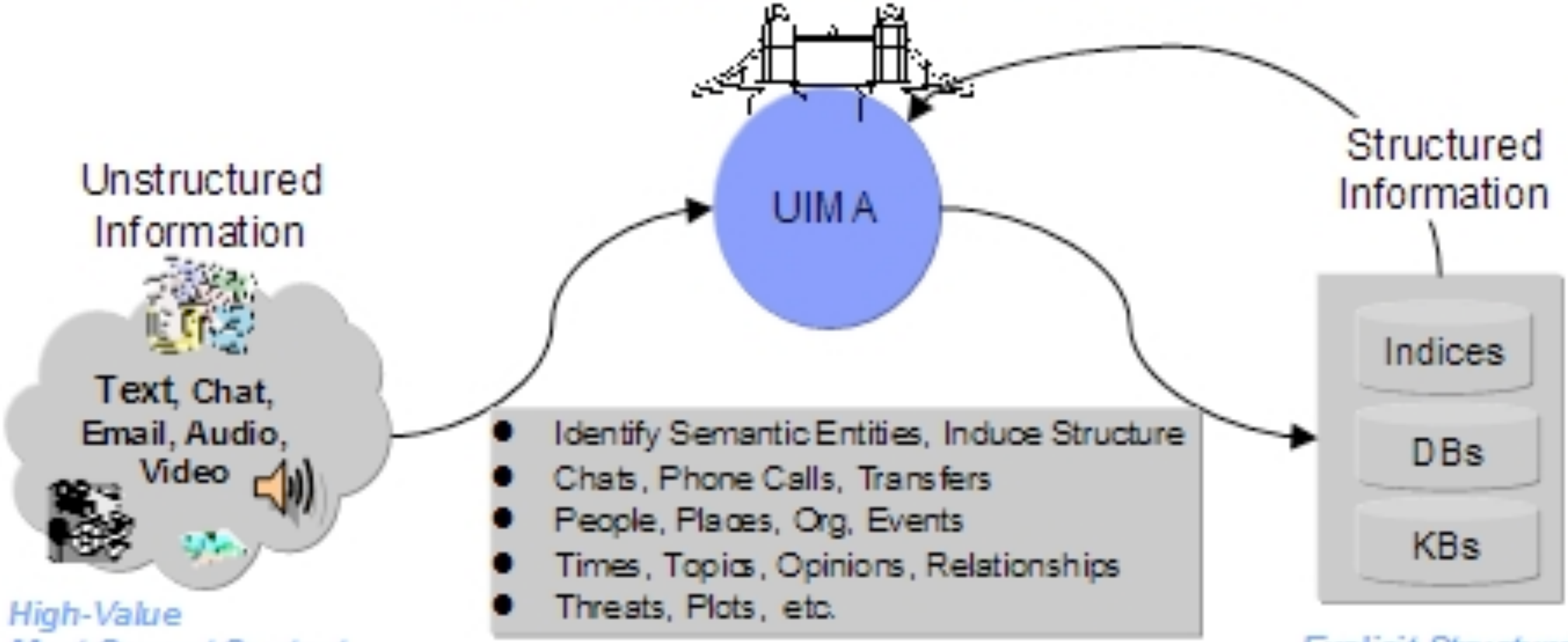
Suzanne DeChillo/The New York Times

David Ferrucci has left I.B.M., and Watson, and joined the hedge fund, Bridgewater Associates.

To the degree there was a human face of Watson, the "Jeopardy!" computer champion, it was David Ferrucci. He was the I.B.M. researcher who led the development of Watson, an artificial intelligence engine. The goateed computer scientist was always articulate and at ease in front of a camera or a microphone.

Dr. Ferrucci has left I.B.M. to join the giant hedge fund Bridgewater Associates. And the weight of the Watson-related fame, it seems, played a role. "I was so linked to the Watson achievement, and where I.B.M. was taking it, that I felt I was almost losing my identity," he said in a recent interview.

Analytics bridge the Unstructured & Structured worlds

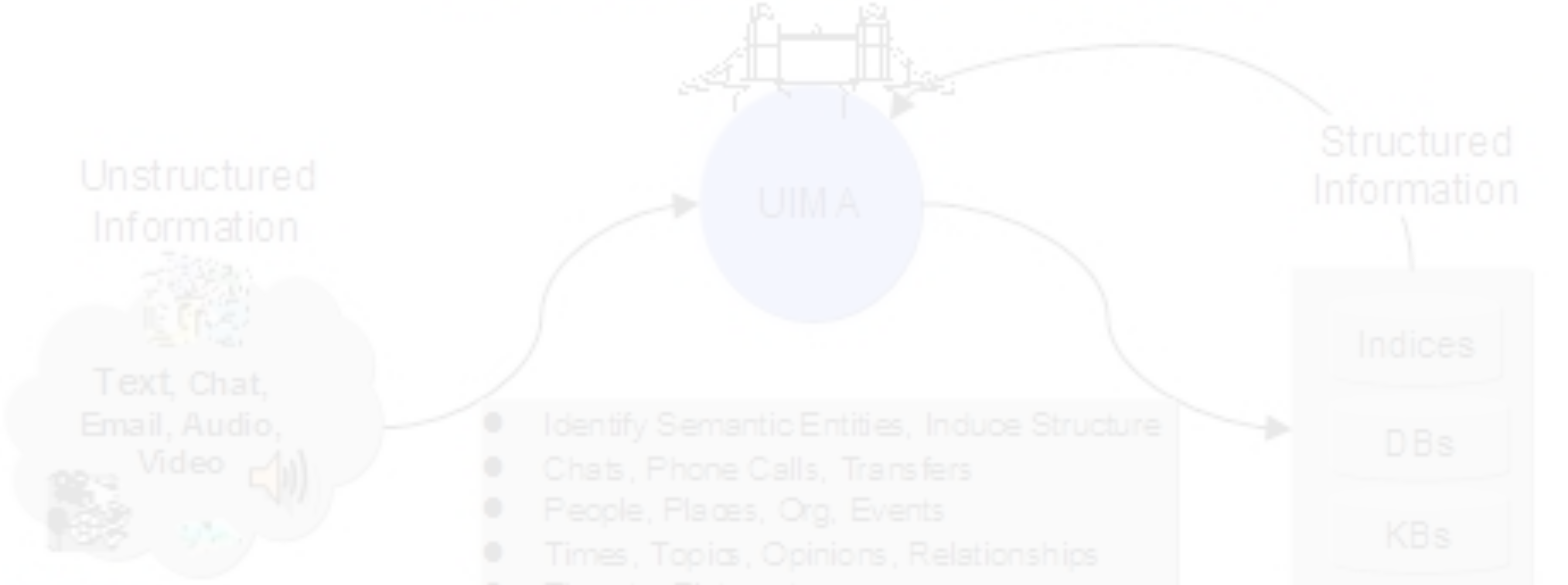


- Identify Semantic Entities, Induce Structure
- Chats, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

High-Value
Most Current Content
BUT ...
Buried in Huge Volumes
Lots of Noise, Implicit Semantics
Inefficient Search

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content

Analytics bridge the Unstructured & Structured worlds

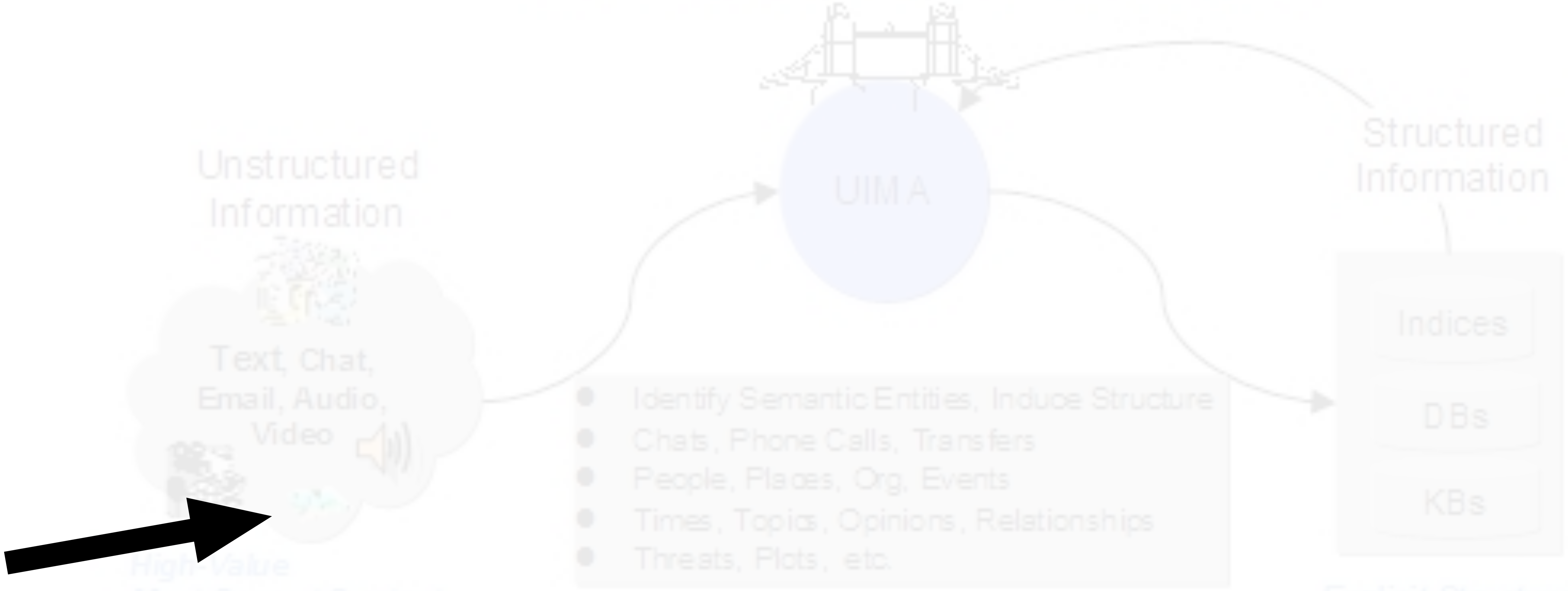


- Identify Semantic Entities, Induce Structure
- Chats, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

High-Value
Most Current Content
BUT…
  Buried in Huge Volumes
  Lots of Noise, Implicit Semantics
  Inefficient Search

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content

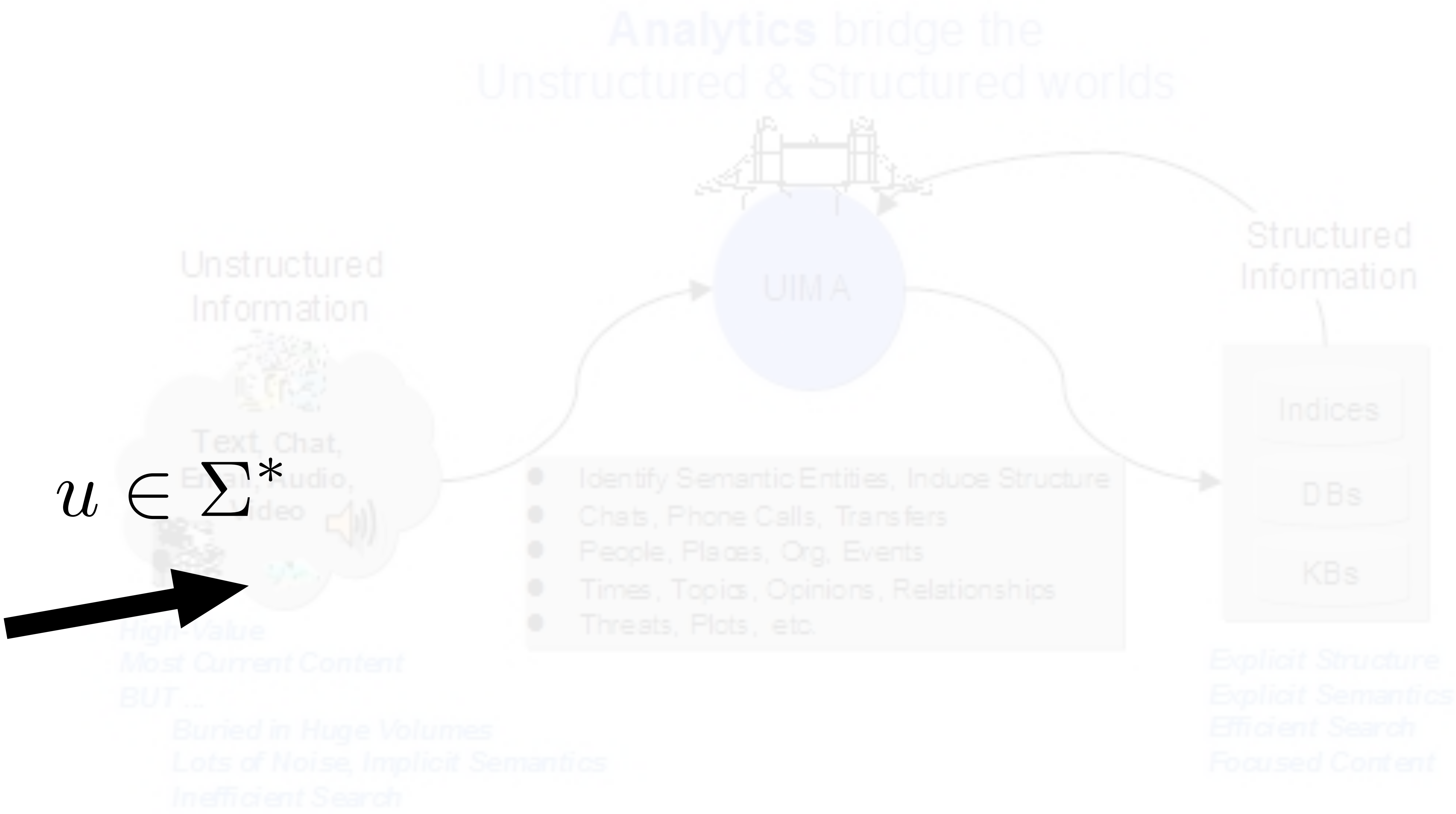
Analytics bridge the Unstructured & Structured worlds



- Identify Semantic Entities, Induce Structure
- Chats, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

High-Value
Most Current Content
BUT ...
Buried in Huge Volumes
Lots of Noise, Implicit Semantics
Inefficient Search

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content



$u \in \Sigma^*$

Analytics bridge the Unstructured & Structured worlds

Unstructured Information

Text, Chat, Email, Audio, Video

UIMA

- Identify Semantic Entities, Induce Structure
- Chats, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

Structured Information

Indices

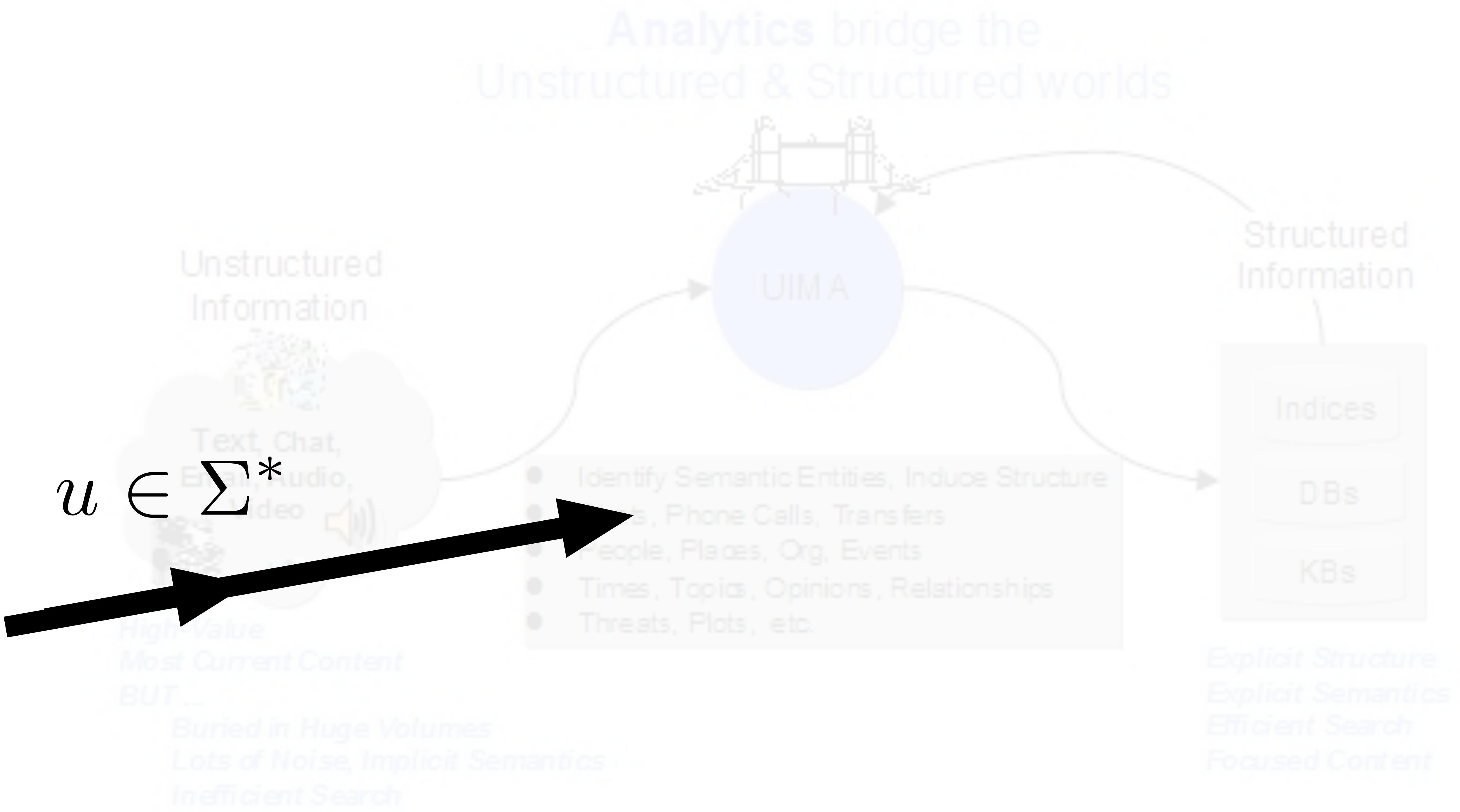
DBs

KBs

High Value
Most Current Content
BUT…
Buried in Huge Volumes
Lots of Noise, Implicit Semantics
Inefficient Search

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content

$u \in \Sigma^*$



Analytics bridge the
Unstructured & Structured worlds

Unstructured
Information

Text, Chat,
Email, Audio,
Video

UIMA

- Identify Semantic Entities, Induce Structure
- s, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

Structured
Information

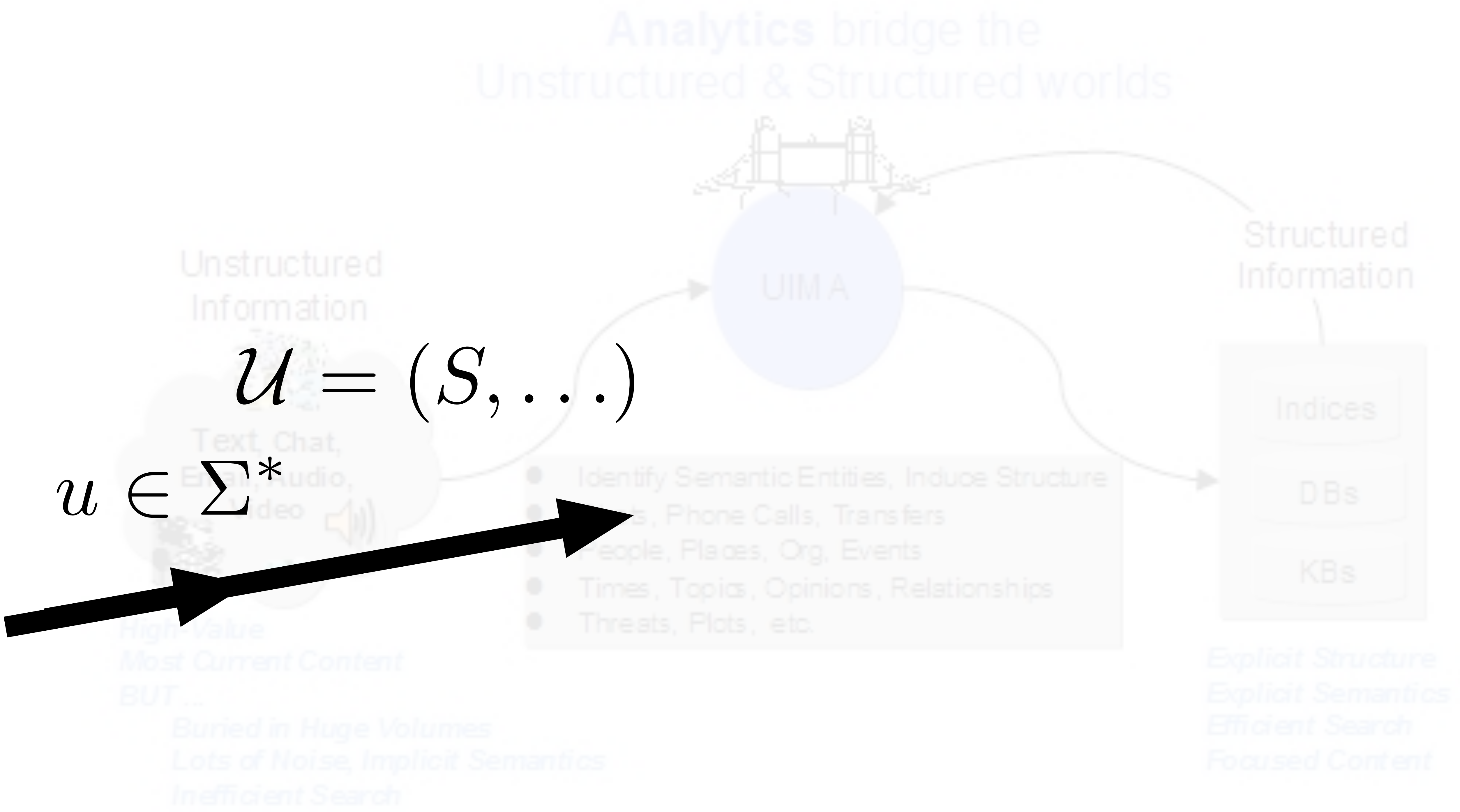
Indices

DBs

KBs

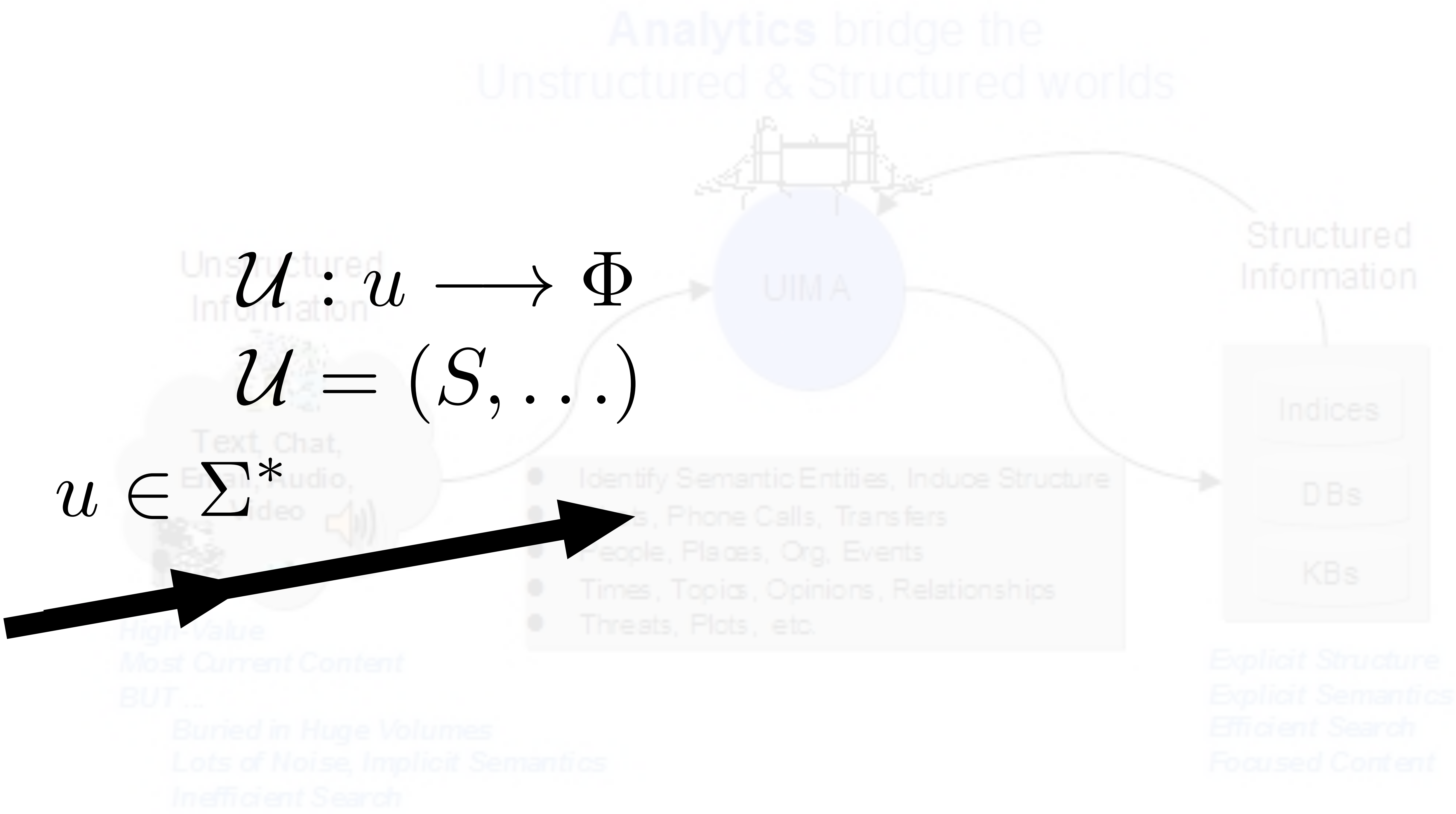
High Value
Most Current Content
BUT ...
Buried in Huge Volumes
Lots of Noise, Implicit Semantics
Inefficient Search

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content



$$\mathcal{U} = (S, \ldots)$$

$$u \in \Sigma^*$$



Analytics bridge the
Unstructured & Structured worlds

$$\mathcal{U} : u \longrightarrow \Phi$$

$$\mathcal{U} = (S, \ldots)$$

$$u \in \Sigma^*$$

Unstructured Information

Text, Chat, Email, Audio, Video

UIMA

- Identify Semantic Entities, Induce Structure
- ts, Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

Structured Information

Indices

DBs

KBs

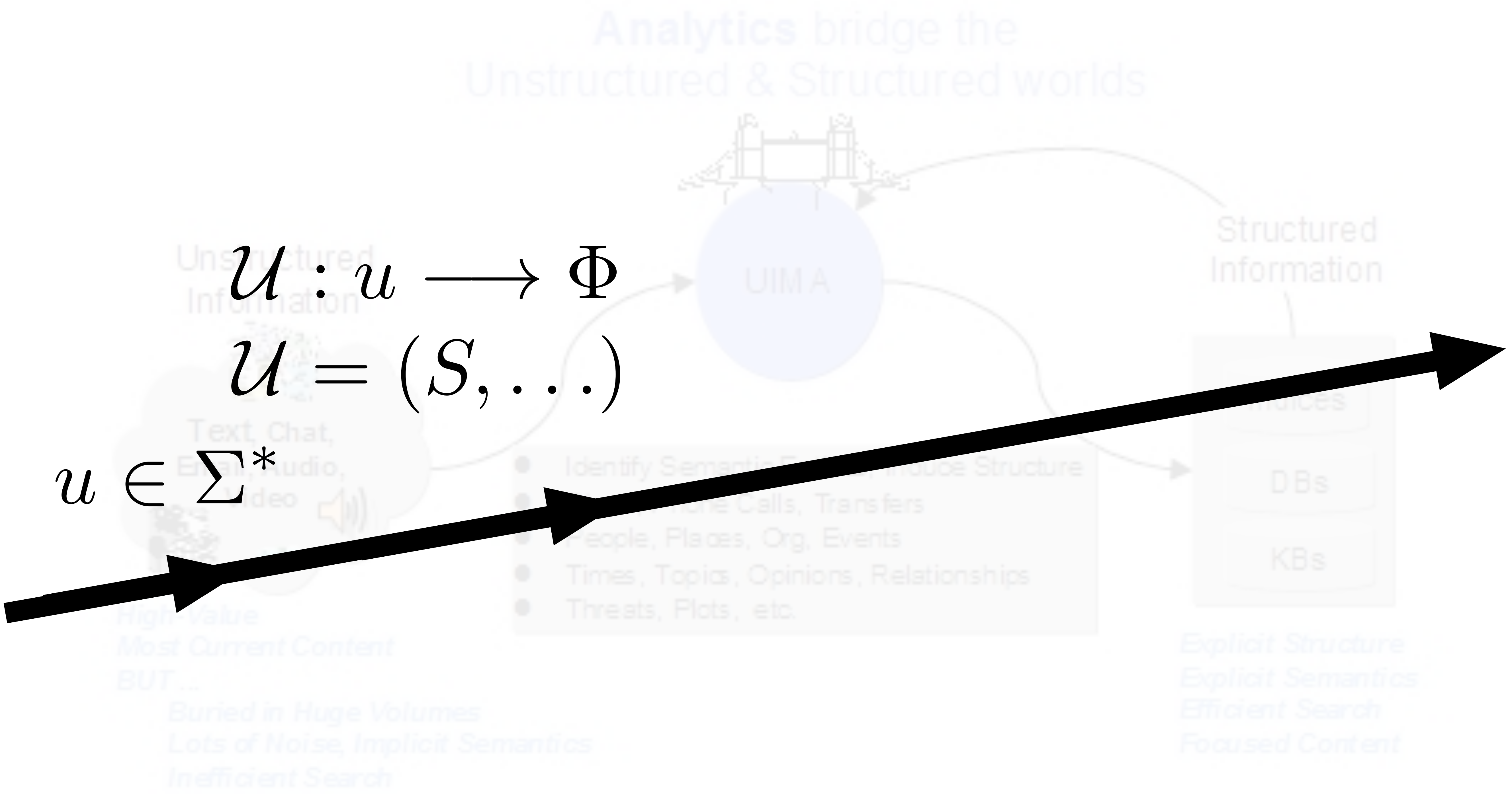
High Value
Most Current Content
BUT…
Buried in Huge Volumes
Lots of Noise, Implicit Semantics
Inefficient Search

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content

$$\mathcal{U} : u \longrightarrow \Phi$$

$$\mathcal{U} = (S, \ldots)$$

$$u \in \Sigma^*$$



Analytics bridge the Unstructured & Structured worlds

- Identify Semantic [illegible], Induce Structure
- Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

High Value
Most Current Content
BUT ...
Buried in Huge Volumes
Lots of Noise, Implicit Semantics
Inefficient Search

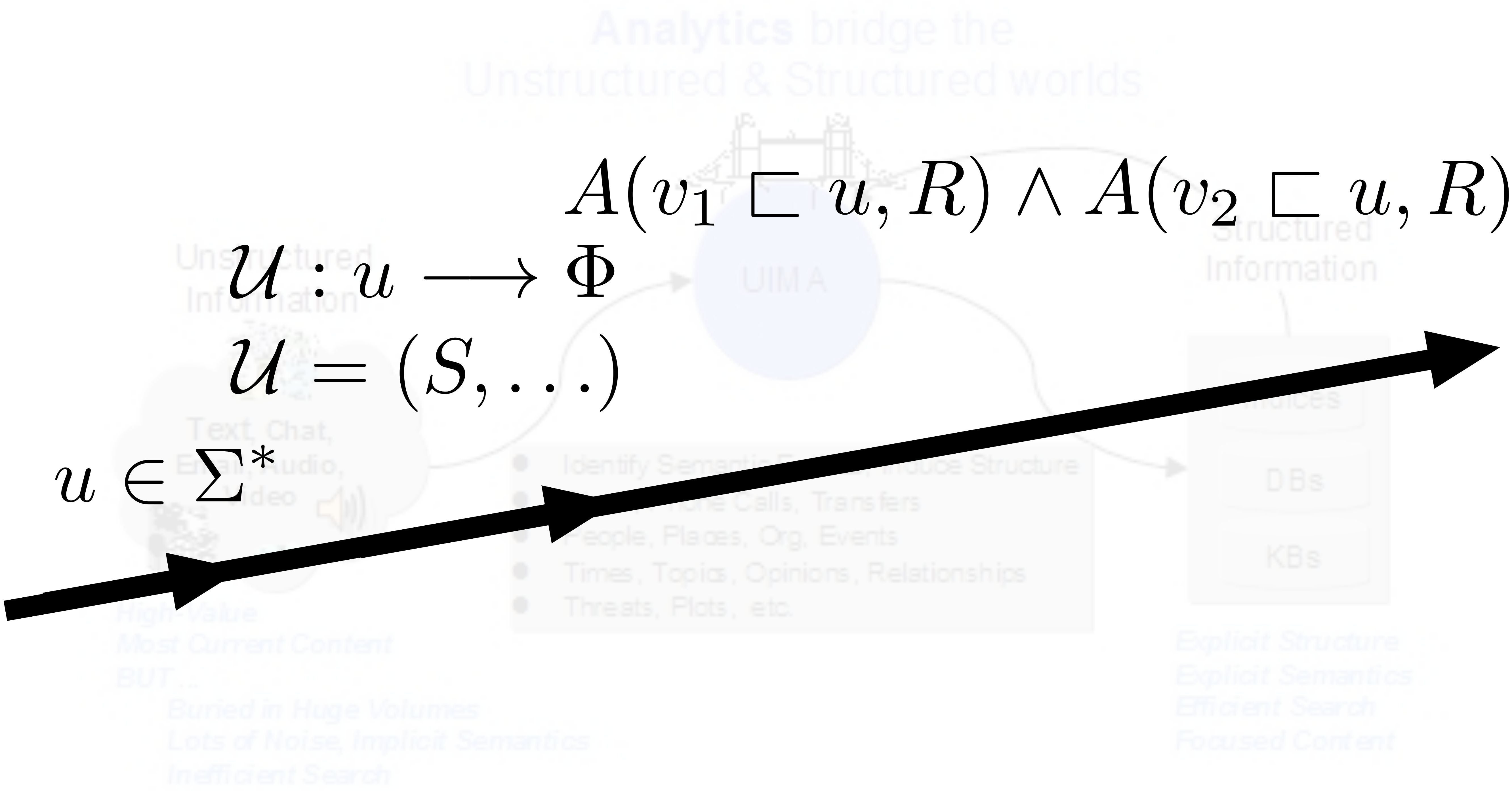
Explicit Structure
Explicit Semantics
Efficient Search
Focused Content

$$A(v_1 \sqsubset u, R) \wedge A(v_2 \sqsubset u, R)$$

$$\mathcal{U} : u \longrightarrow \Phi$$

$$\mathcal{U} = (S, \ldots)$$

$$u \in \Sigma^*$$



Analytics bridge the Unstructured & Structured worlds

- Identify Semantic Entities, Induce Structure
- Telephone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

High Value
Most Current Content
BUT ...
Buried in Huge Volumes
Lots of Noise, Implicit Semantics
Inefficient Search

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content

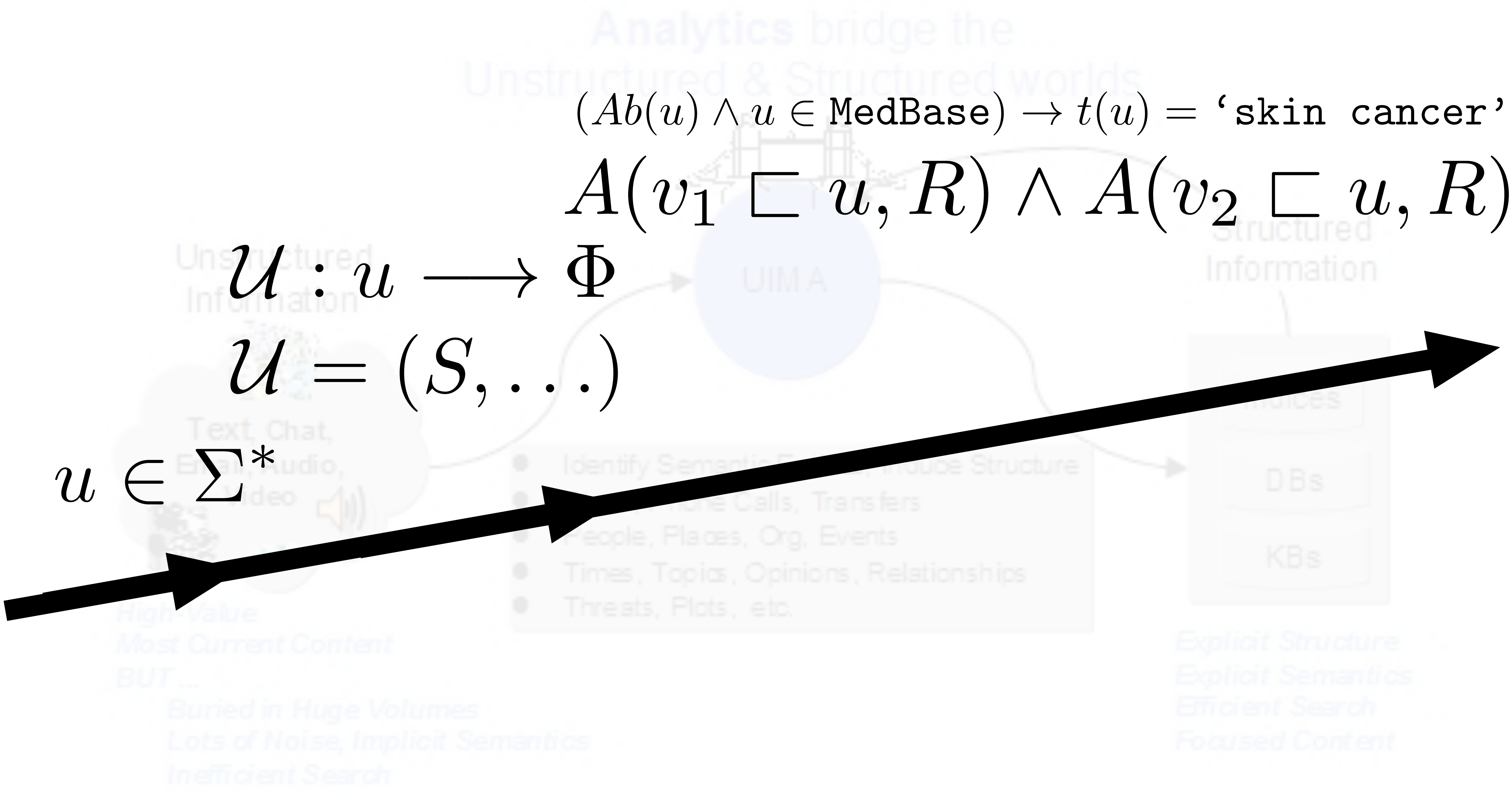
$(Ab(u) \wedge u \in \texttt{MedBase}) \rightarrow t(u) = \text{`}\texttt{skin cancer}\text{'}$

$$A(v_1 \sqsubset u, R) \wedge A(v_2 \sqsubset u, R)$$

$$\mathcal{U} : u \longrightarrow \Phi$$

$$\mathcal{U} = (S, \ldots)$$

$$u \in \Sigma^*$$



Analytics bridge the Unstructured & Structured worlds

Unstructured Information

Text, Chat, Email, Audio, Video

UIMA

- Identify Semantic Entities, Induce Structure
- Phone Calls, Transfers
- People, Places, Org, Events
- Times, Topics, Opinions, Relationships
- Threats, Plots, etc.

Structured Information

Indices

DBs

KBs

High-Value
Most Current Content
BUT…
Buried in Huge Volumes
Lots of Noise, Implicit Semantics
Inefficient Search

Explicit Structure
Explicit Semantics
Efficient Search
Focused Content

$

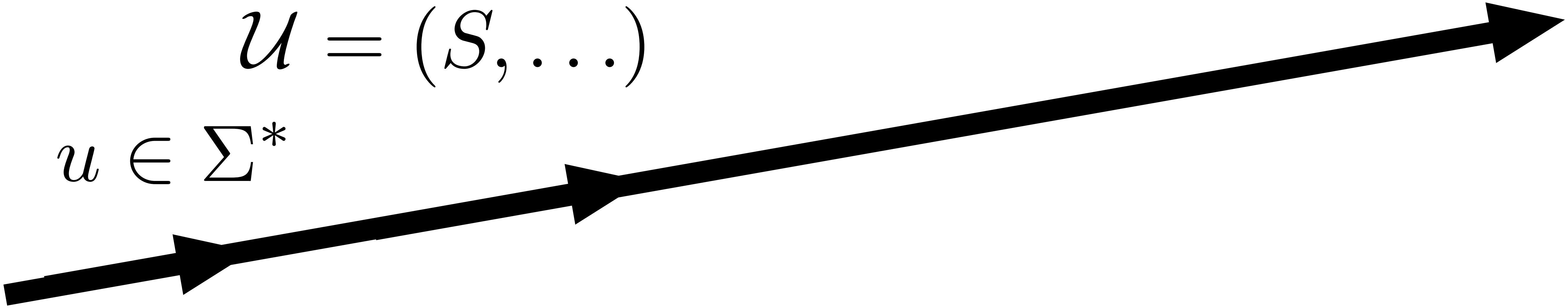
$$(Ab(u) \land u \in \texttt{MedBase}) \to t(u) = \text{`}\texttt{skin cancer}\text{'}$$

$$A(v_1 \sqsubset u, R) \land A(v_2 \sqsubset u, R)$$

$$\mathcal{U} : u \longrightarrow \Phi$$

$$\mathcal{U} = (S, \ldots)$$

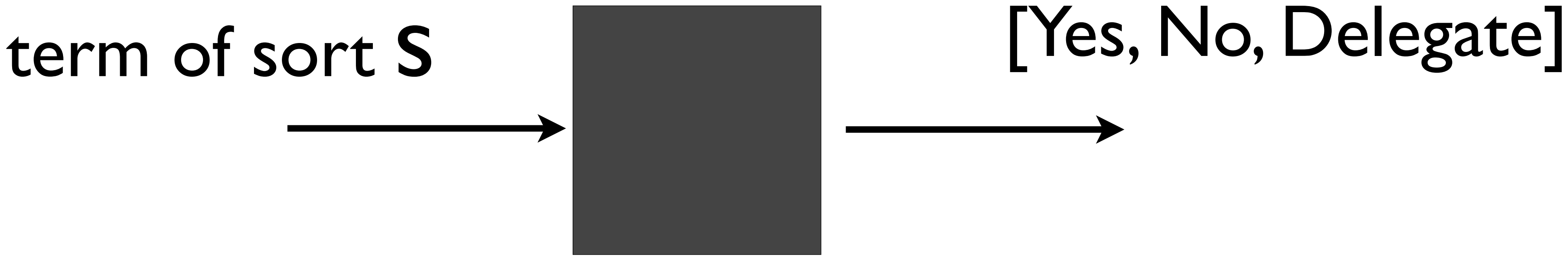
$$u \in \Sigma^*$$



What *is* the “carry over” here?

Hierarchical Ethical Classifier (initial design)

- Preprocessing system for deciding whether a situation warrants deliberate ethical reasoning.
- Made up of atomic ethical classifiers (UIMA's Analysis Engines)



term of sort **S**

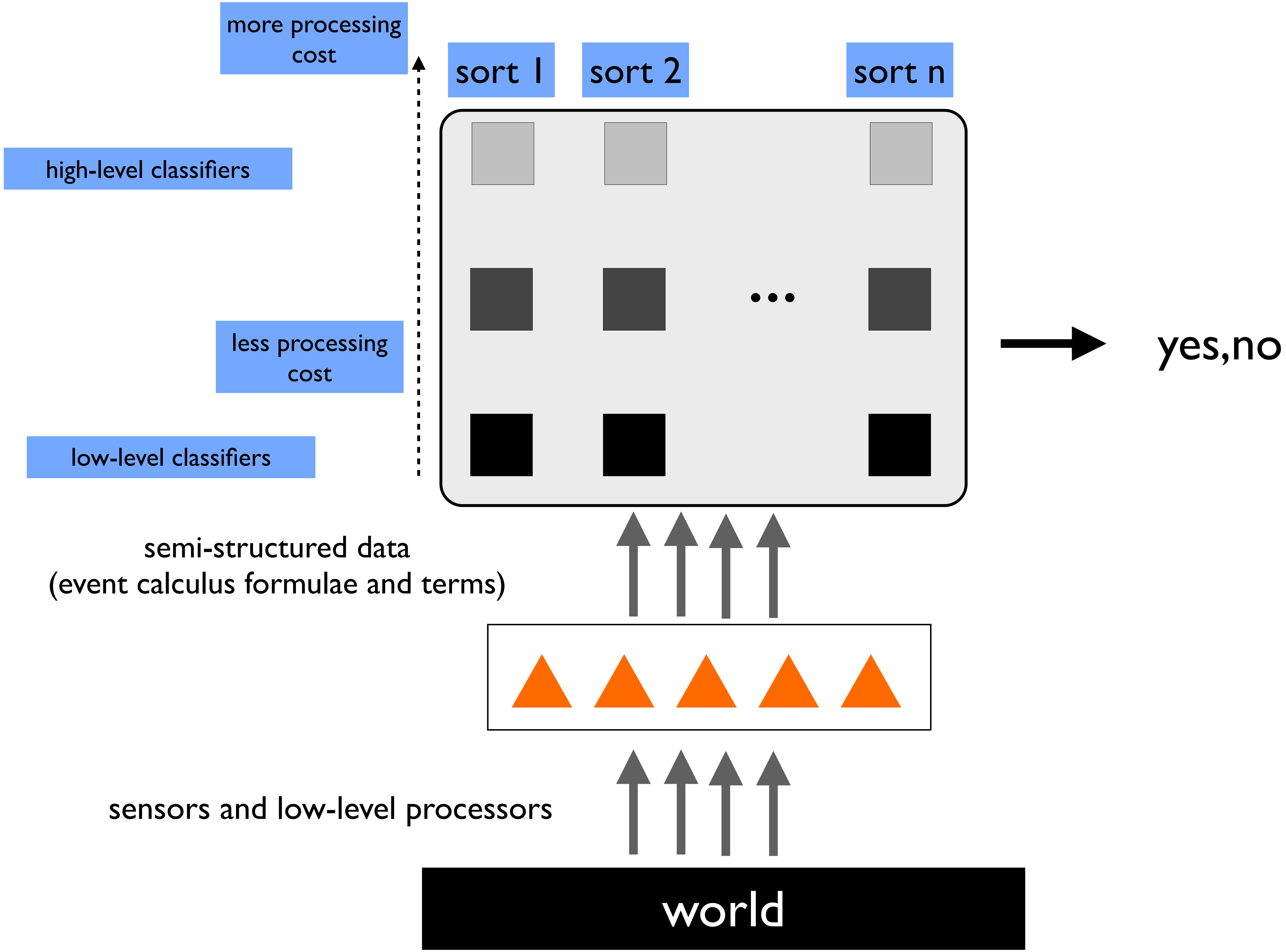
[Yes, No, Delegate]

atomic ethical classifier

Why?

- Not all situations need deliberate deontic reasoning.
- Need to quickly decide at every time instant whether the current situation requires deliberate, deontic reasoning.
- Need many heuristics to do so.
- The design provides a disciplined approach to organize and add new heuristics.

Hierarchical Ethical Classifier (UIMA-Style)



more processing cost

sort 1

sort 2

sort n

high-level classifiers

less processing cost

low-level classifiers

yes,no

semi-structured data
(event calculus formulae and terms)

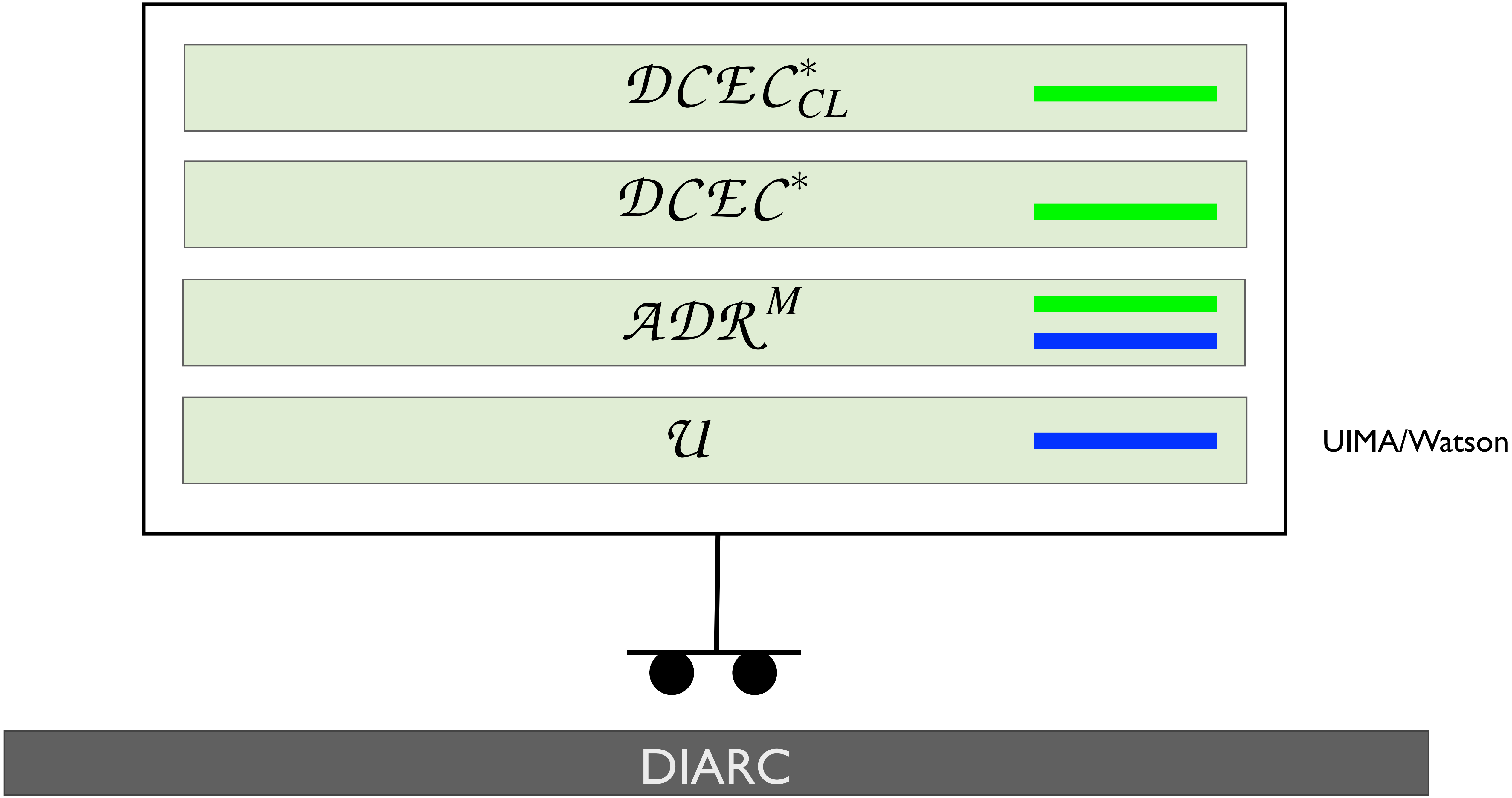
sensors and low-level processors

world

Specification

- Processing goes to a higher-level classifier only if the corresponding lower classifier answers **Delegate**.
- Notion of *top-fired classifiers*.
- Systems answers:
    - **Yes**: If and only if any one of the top-fired classifiers answers **Yes,** or all the top-level atomic classifiers answer **Delegate**.
    - **No**: If and only if all the top-fired classifiers answer **No**.

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

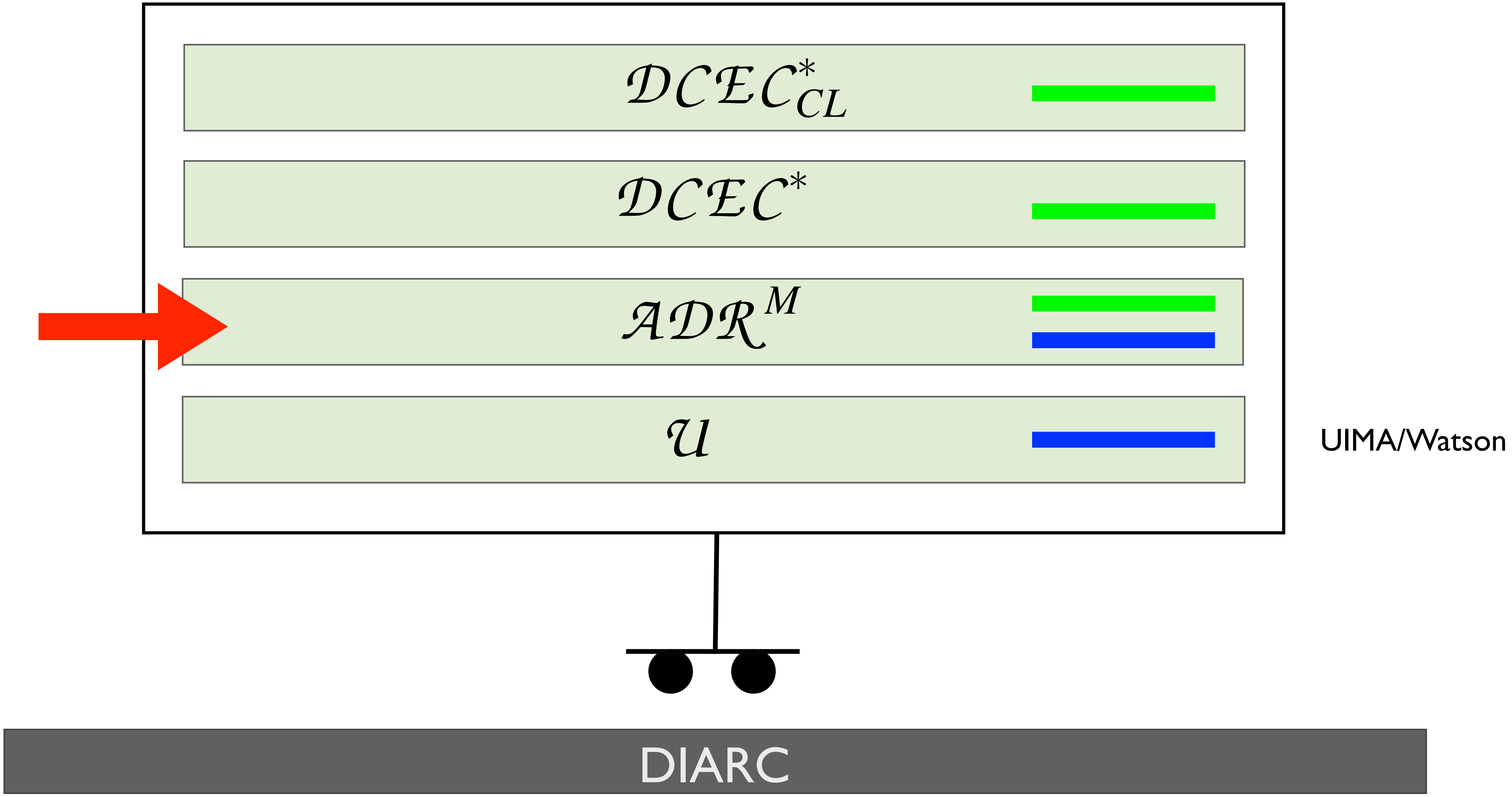
$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson

DIARC

Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson

DIARC

Analogico-Deductive Moral Reasoning (ADMR)

Analogico-Deductive Moral Reasoning (ADMR)

- Moral problem presented as *story* (in psychometric sense) and a *stem*, or *query*.

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:

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

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

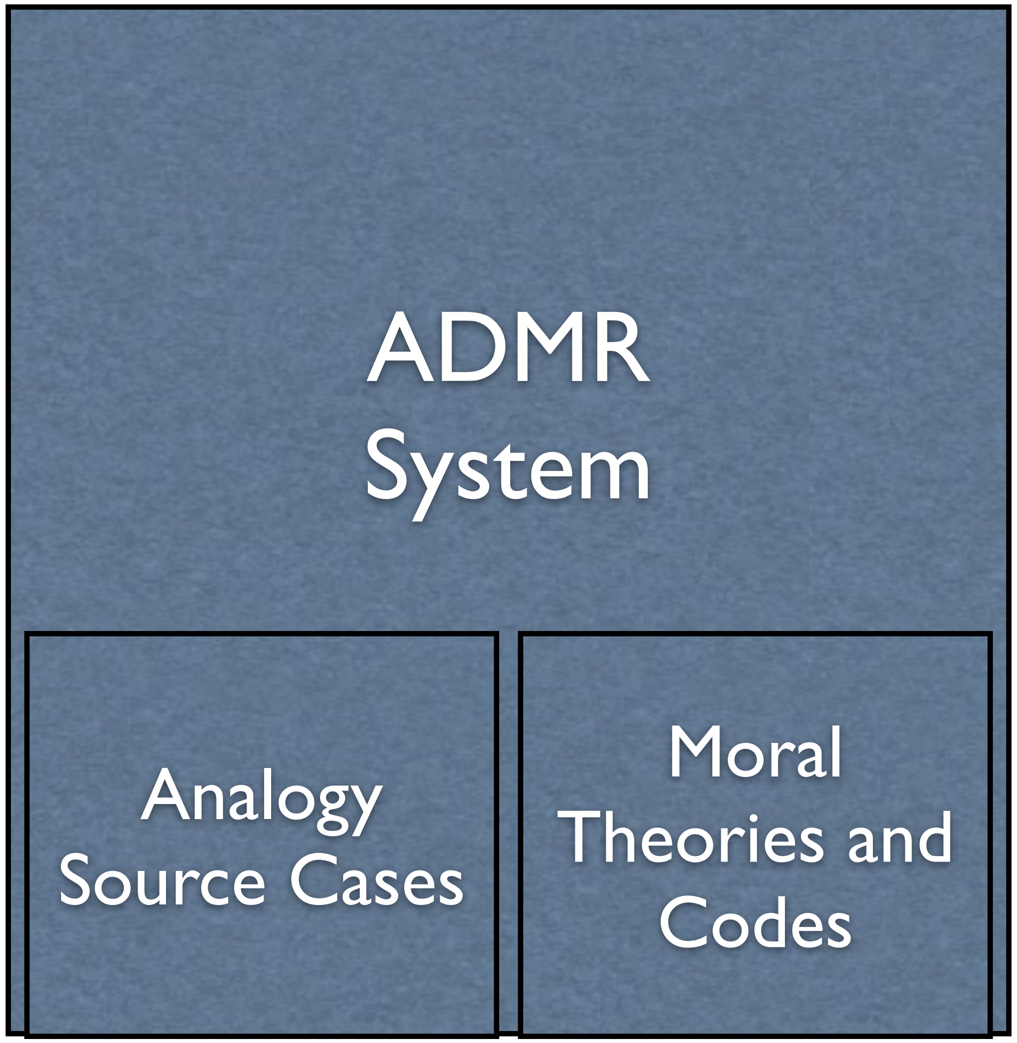
Analogico-Deductive Moral Reasoning (ADMR)

Input:
(story, query/stem)

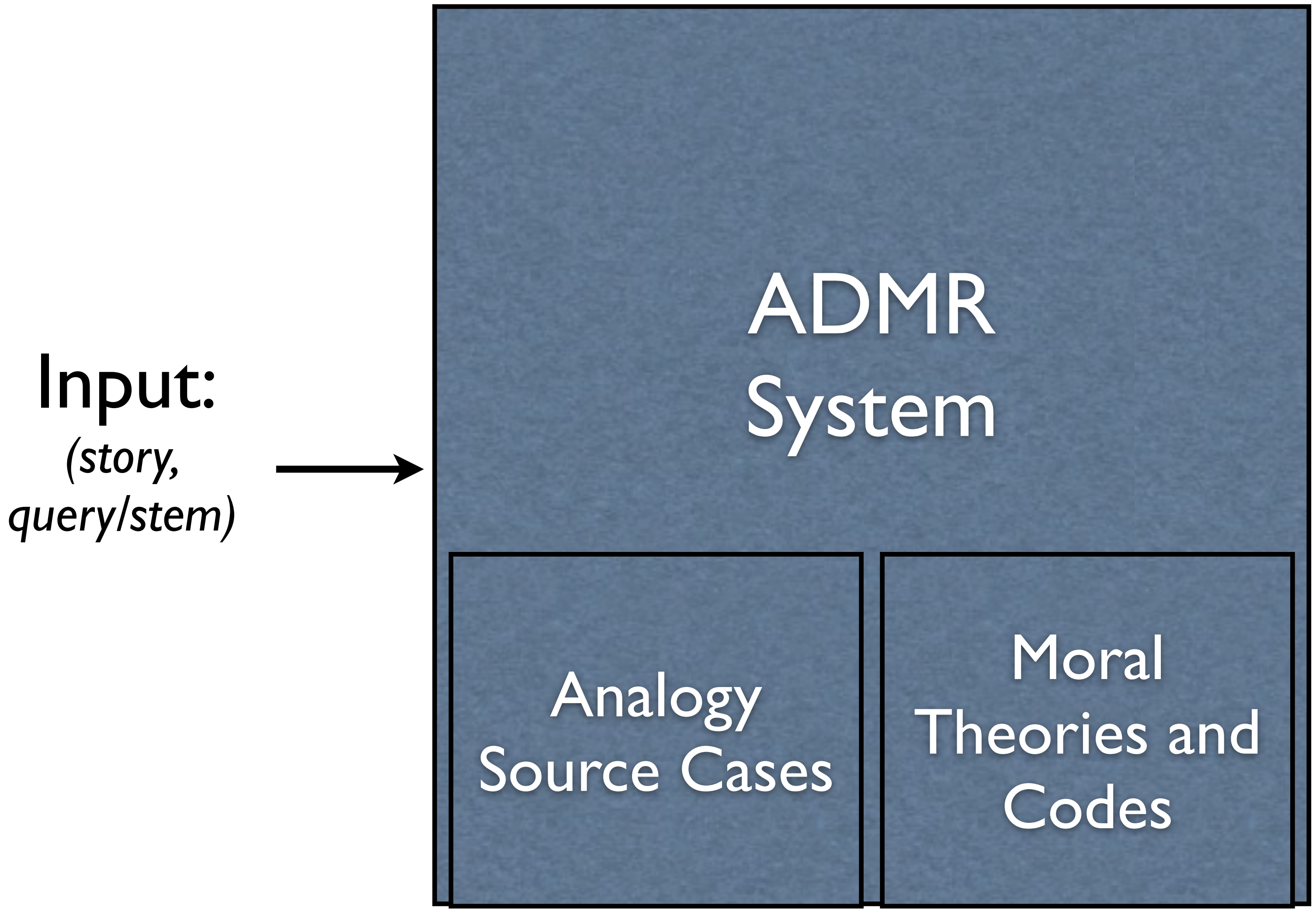
Analogico-Deductive Moral Reasoning (ADMR)

Input:

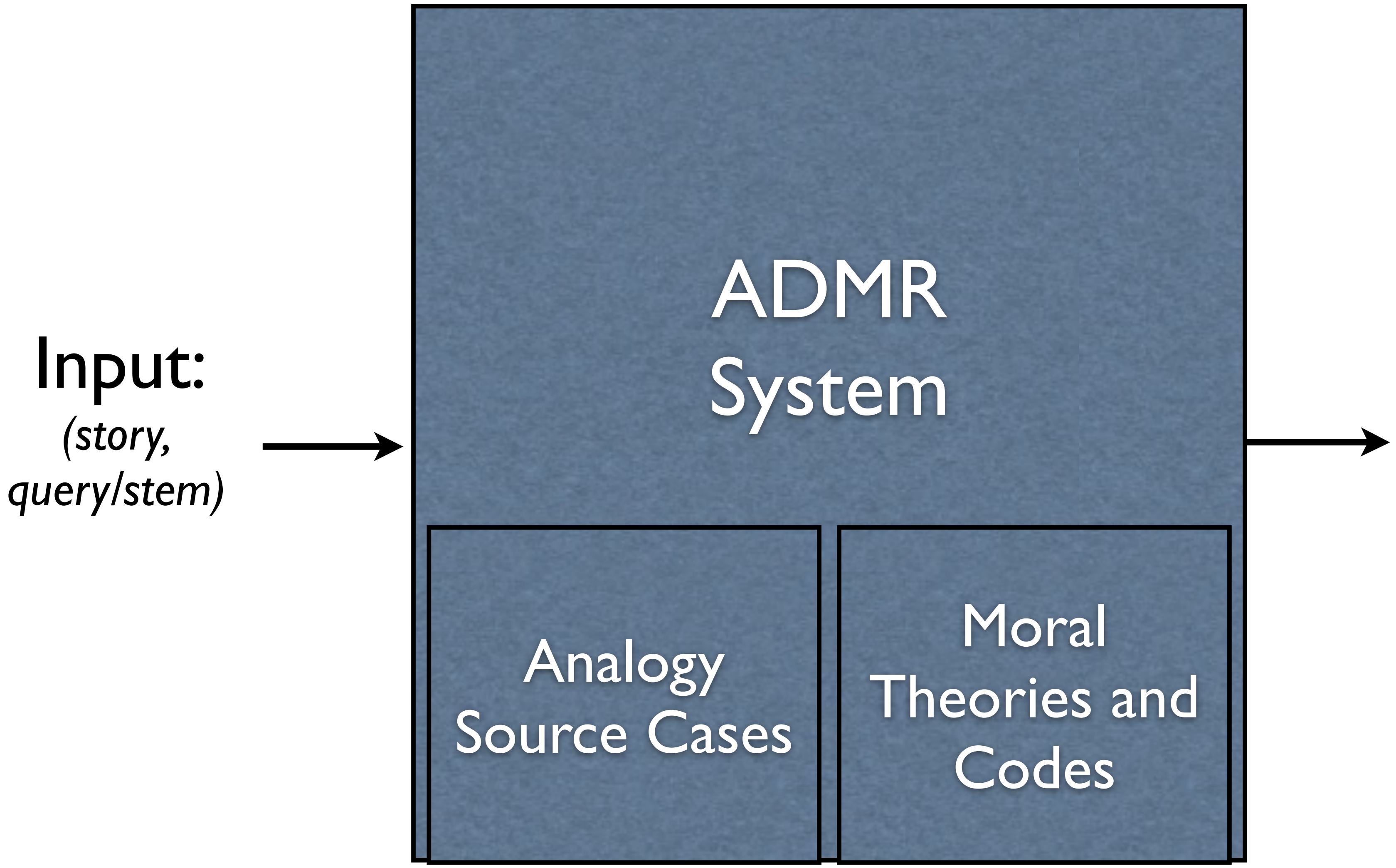
(story, query/stem)



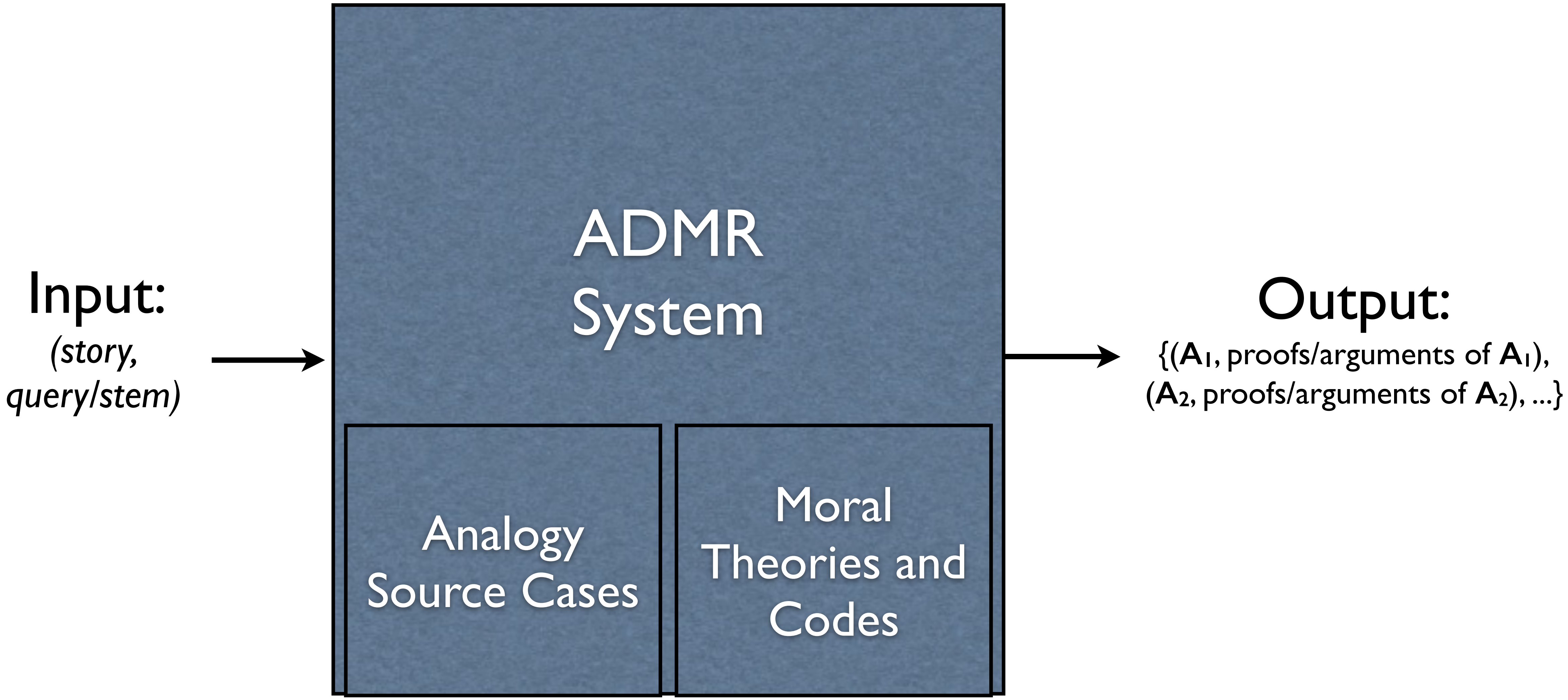
Analogico-Deductive Moral Reasoning (ADMR)



Analogico-Deductive Moral Reasoning (ADMR)



Analogico-Deductive Moral Reasoning (ADMR)



Input:

(story, query/stem)

ADMR System

Analogy Source Cases

Moral Theories and Codes

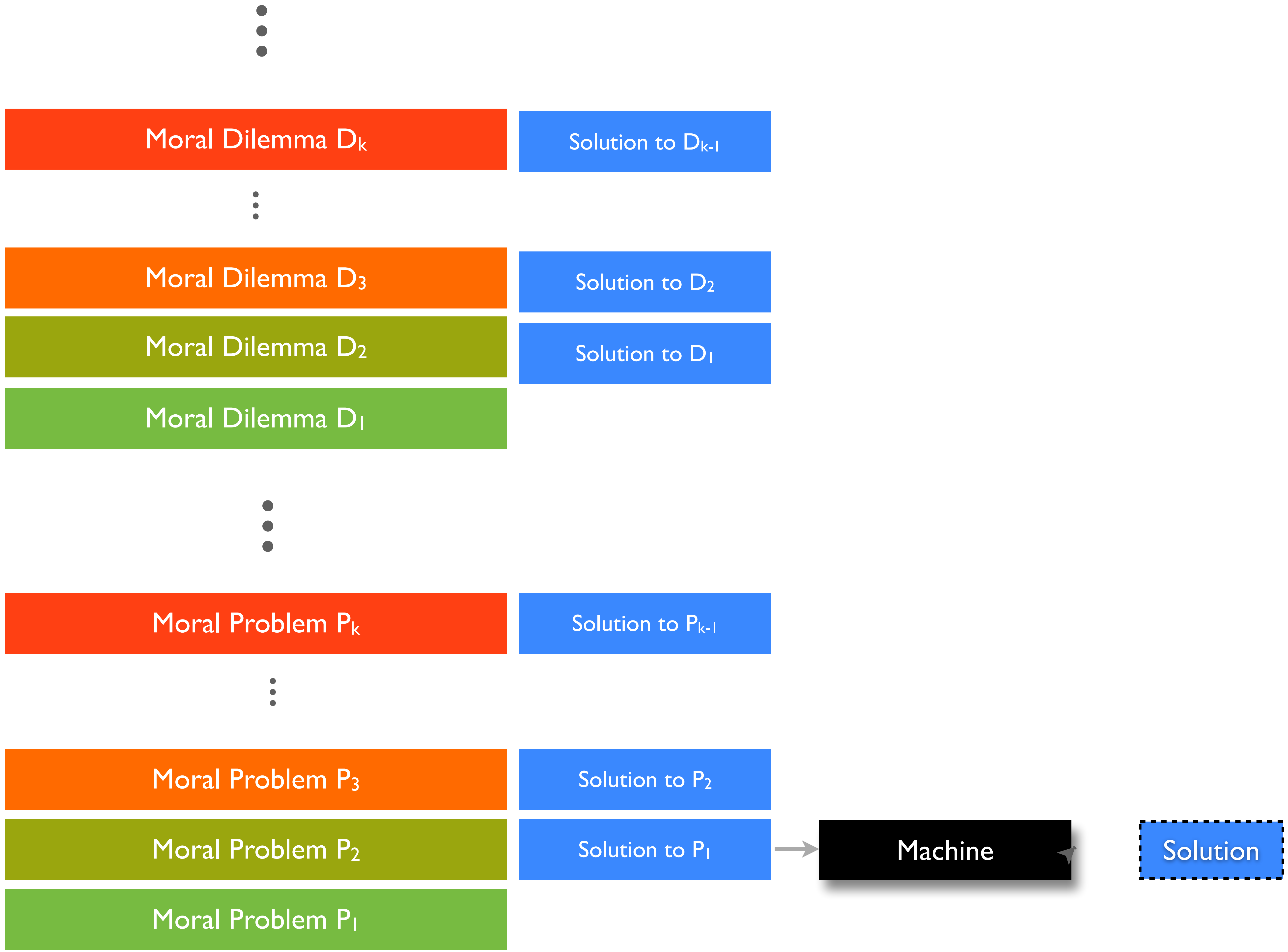
Output:

{($\mathbf{A}_1$, proofs/arguments of $\mathbf{A}_1$), ($\mathbf{A}_2$, proofs/arguments of $\mathbf{A}_2$), ...}

Sample ("Tough") Input:
The Heinz Dilemma (Kolhberg)

"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?*"



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

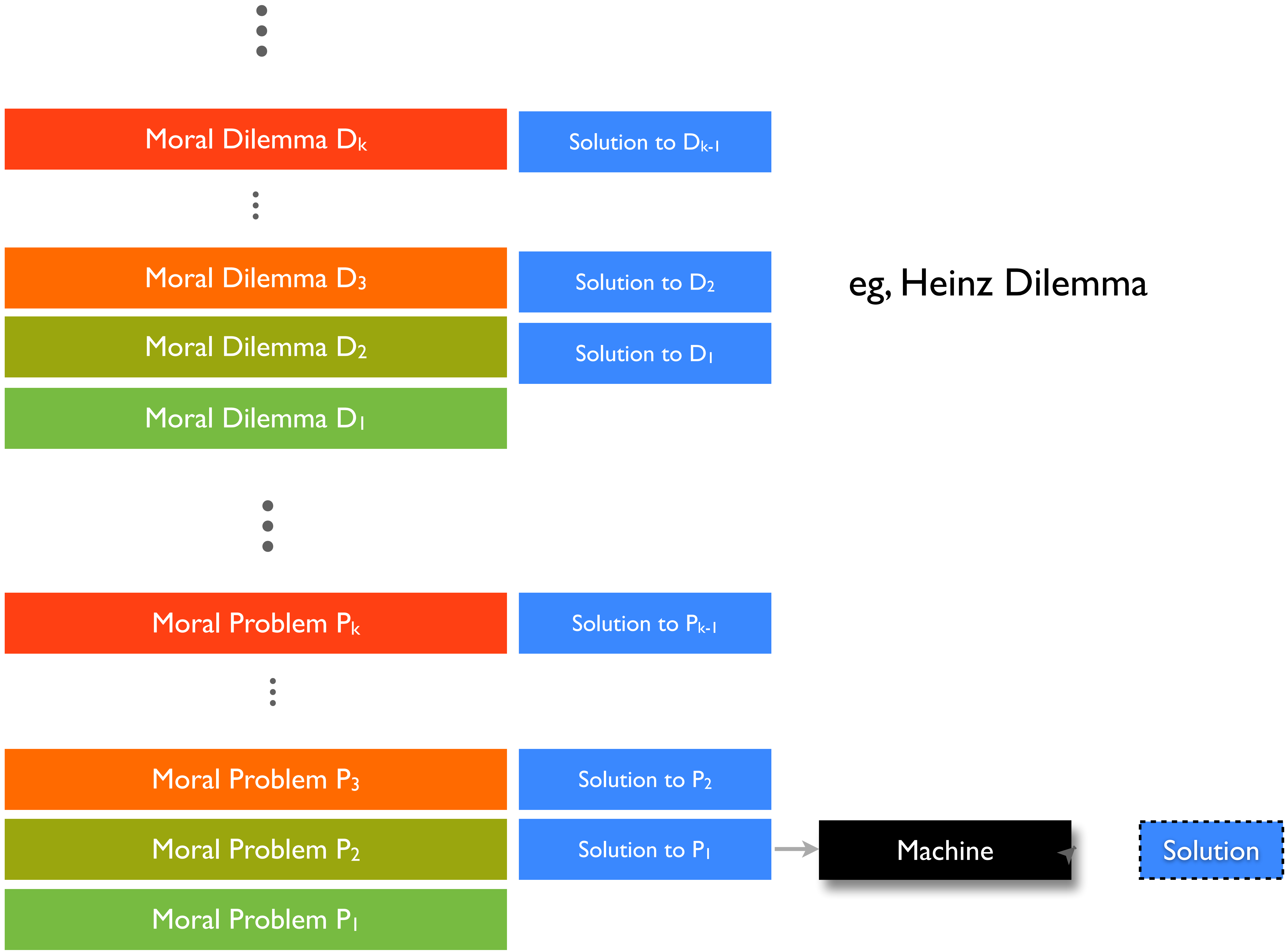
Moral Problem P_2

Solution to P_1

Machine

Solution

Moral Problem P_1



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

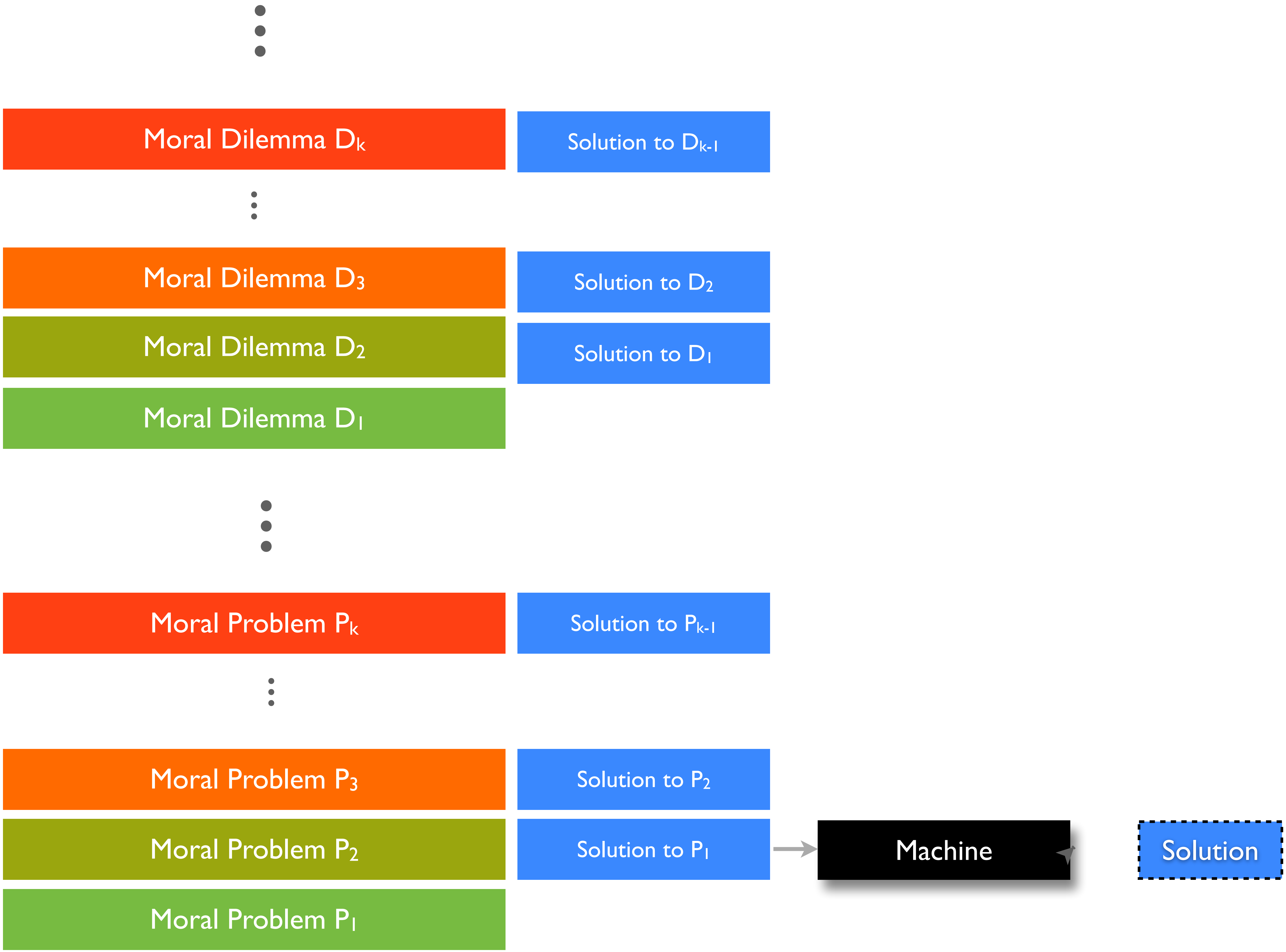
Moral Problem P_2

Solution to P_1

Machine

Solution

Moral Problem P_1



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

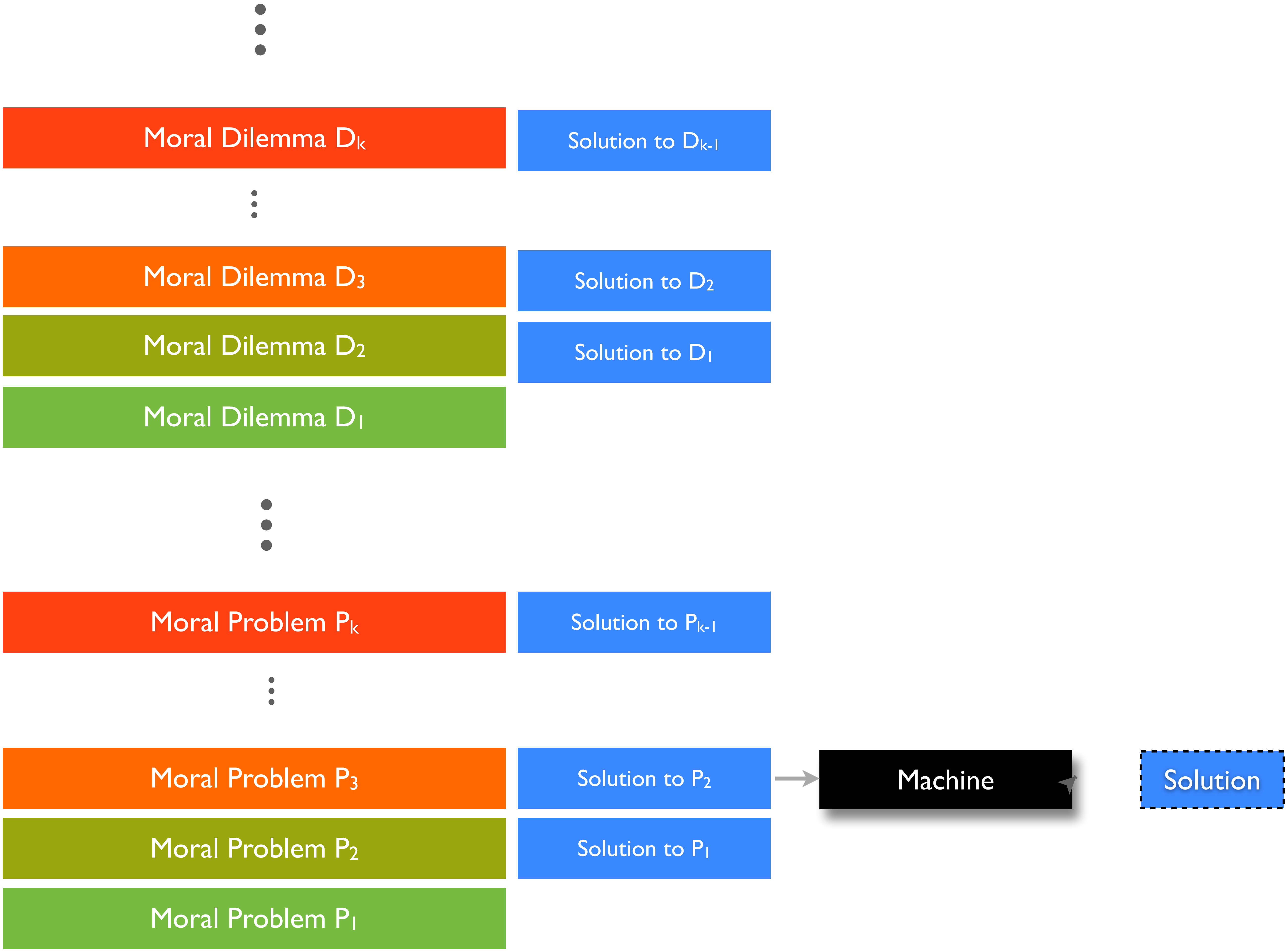
Moral Problem P_2

Solution to P_1

Machine

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

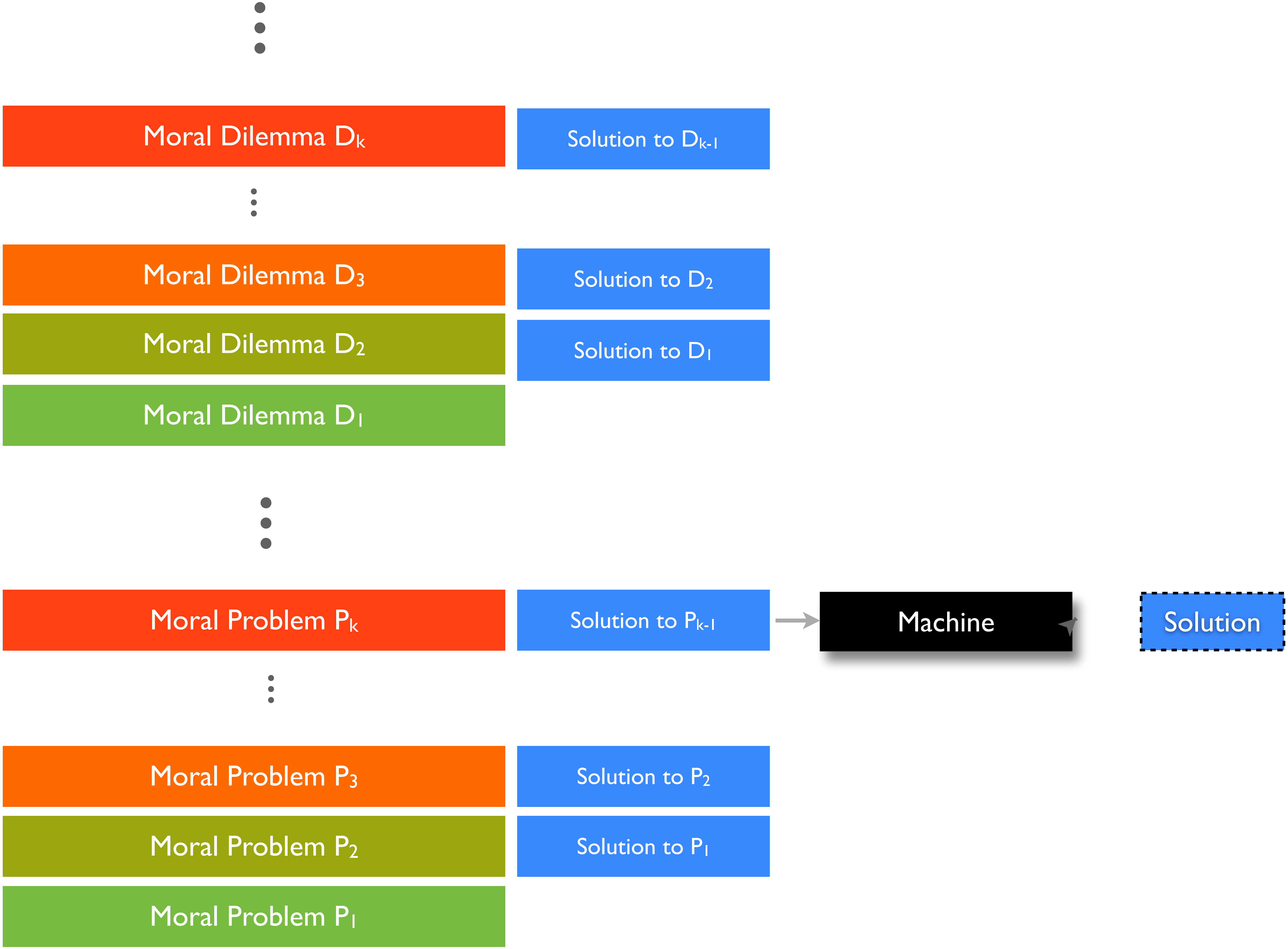
Machine

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}

Machine

Solution

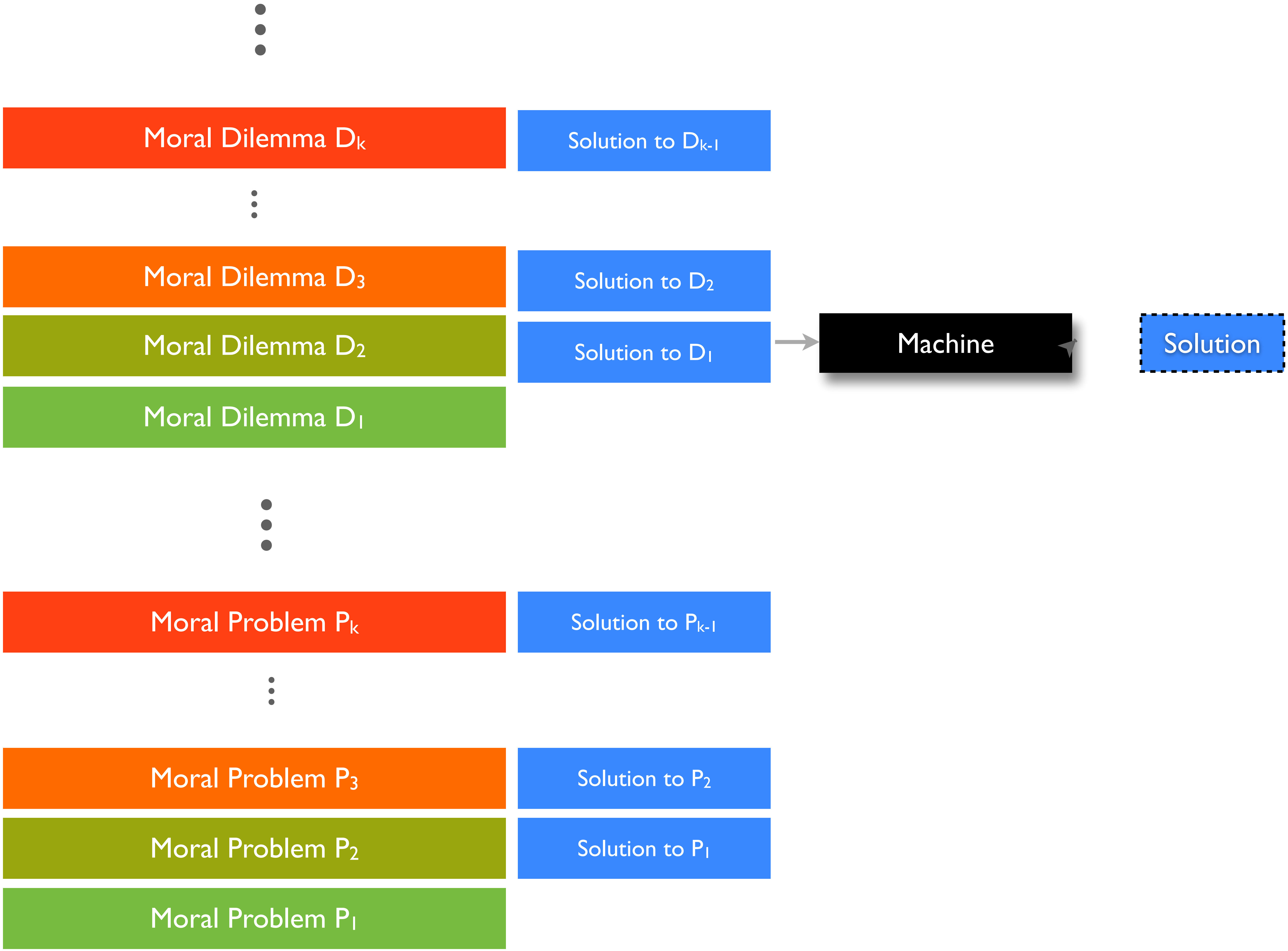
Moral Problem P_3

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

Machine

Solution

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

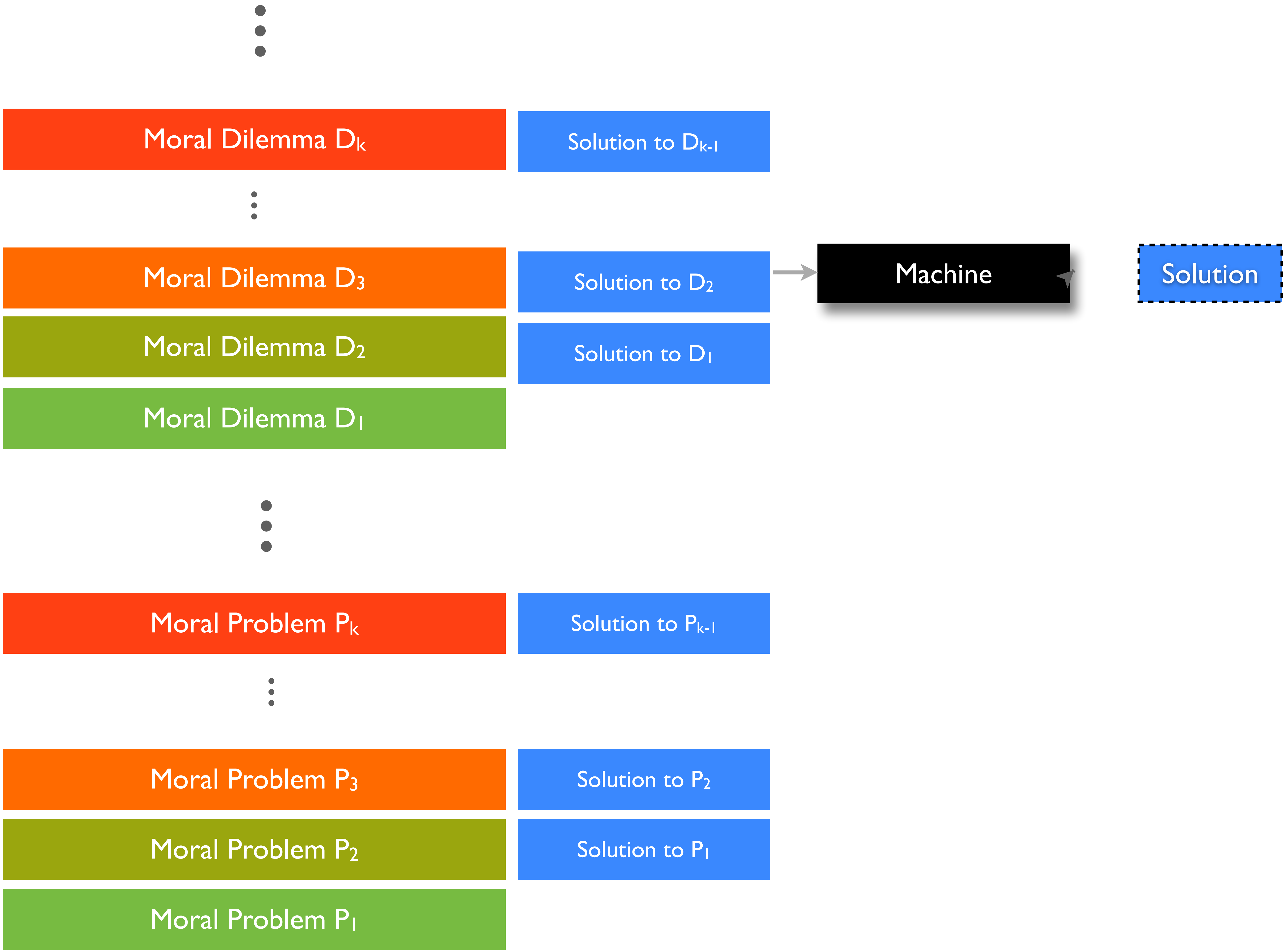
Moral Problem P_3

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

Machine

Solution

Moral Dilemma D_2

Solution to D_1

Moral Dilemma D_1

Moral Problem P_k

Solution to P_{k-1}

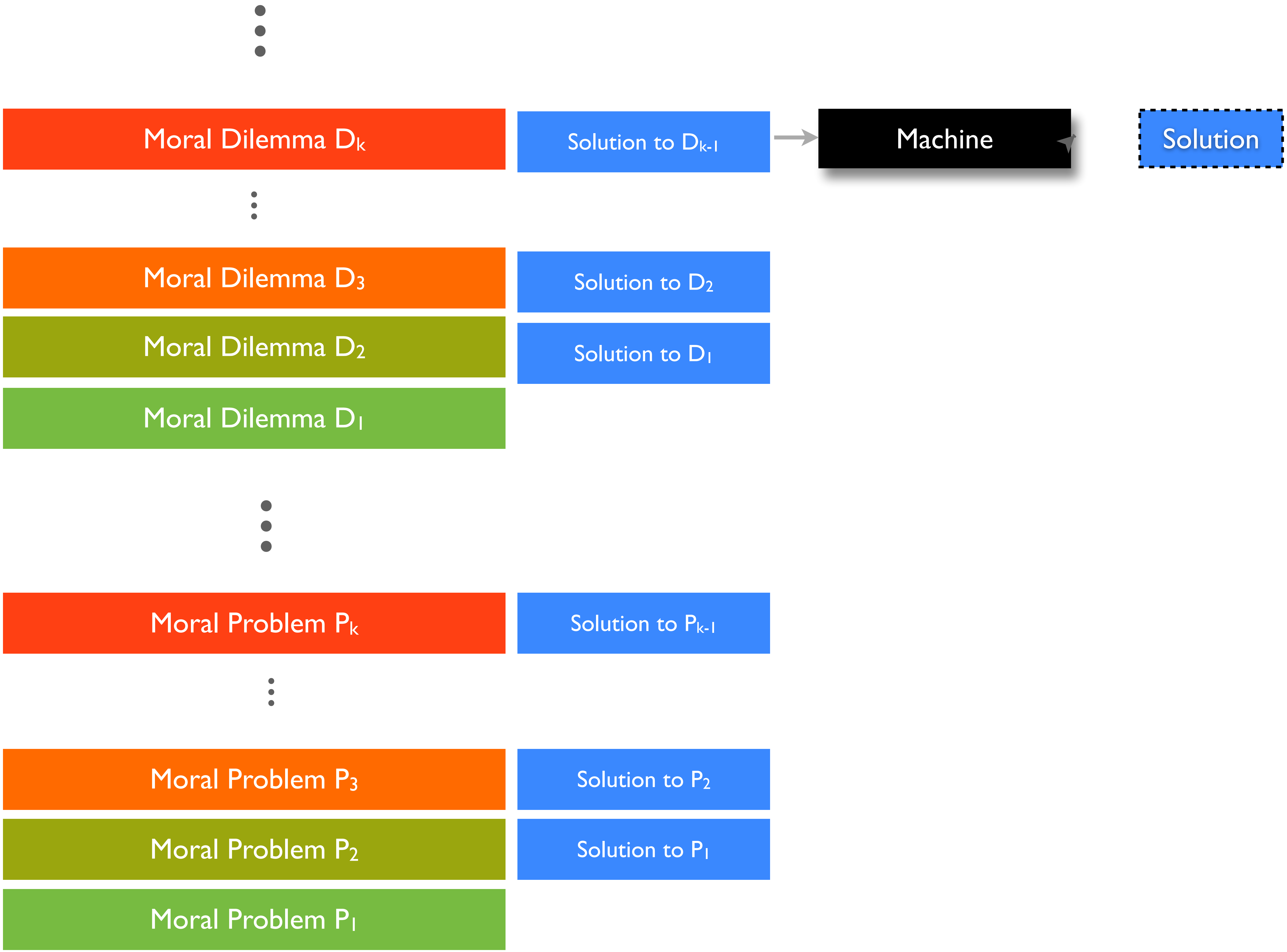
Moral Problem P_3

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}

Machine

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

Fragment of Heinz in DCEC*

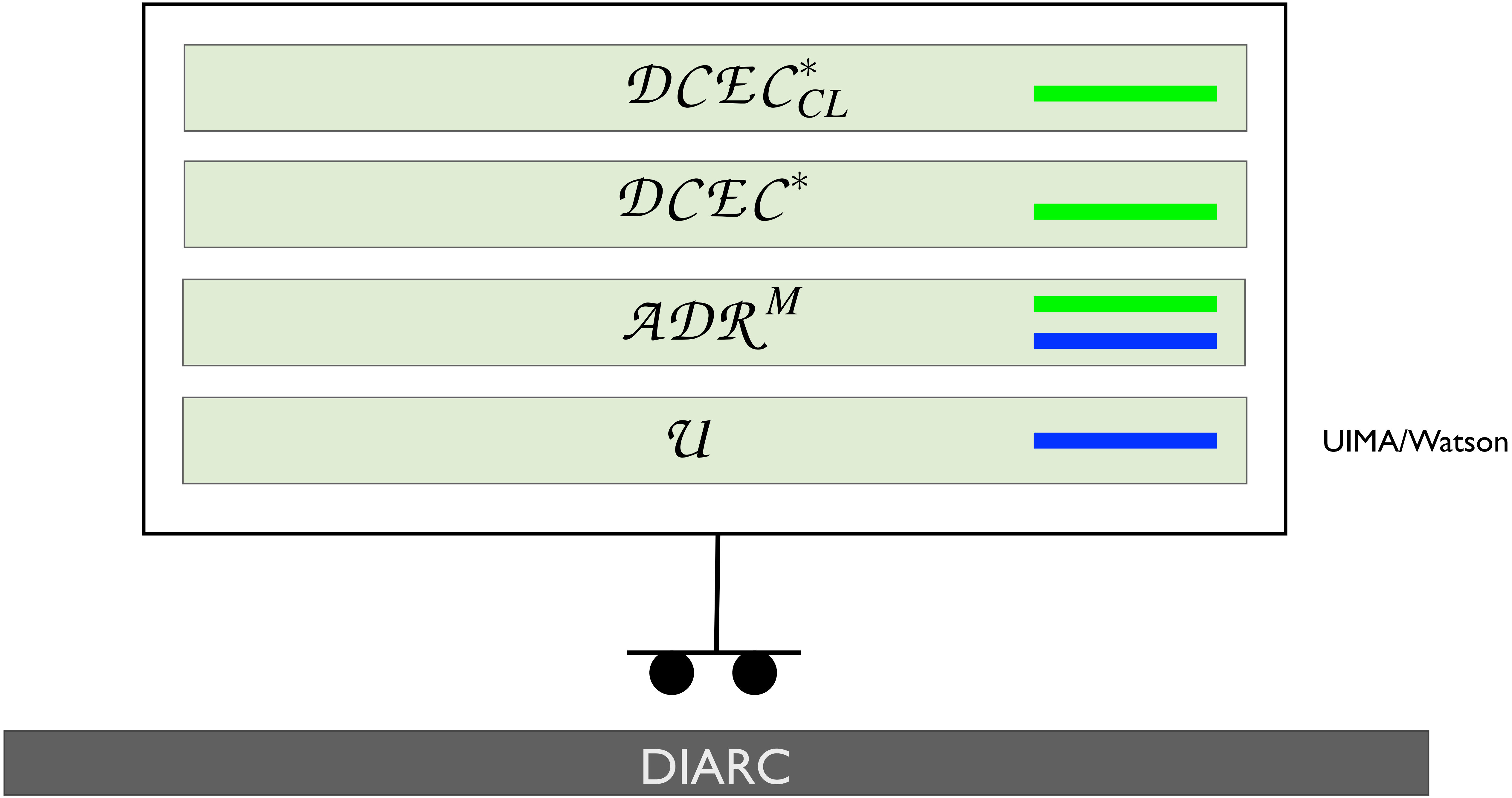
P1 $$\forall t : \text{Moment}, a : \text{Agent}\left(holds(sick(a), t) \land \left(\forall t' : \text{Moment } t' < T \Rightarrow \neg happens(treated(a), t + t')\right) \Rightarrow (happens(dies(a), t + T) \lor holds(dead(a), t + T)\right)$$

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

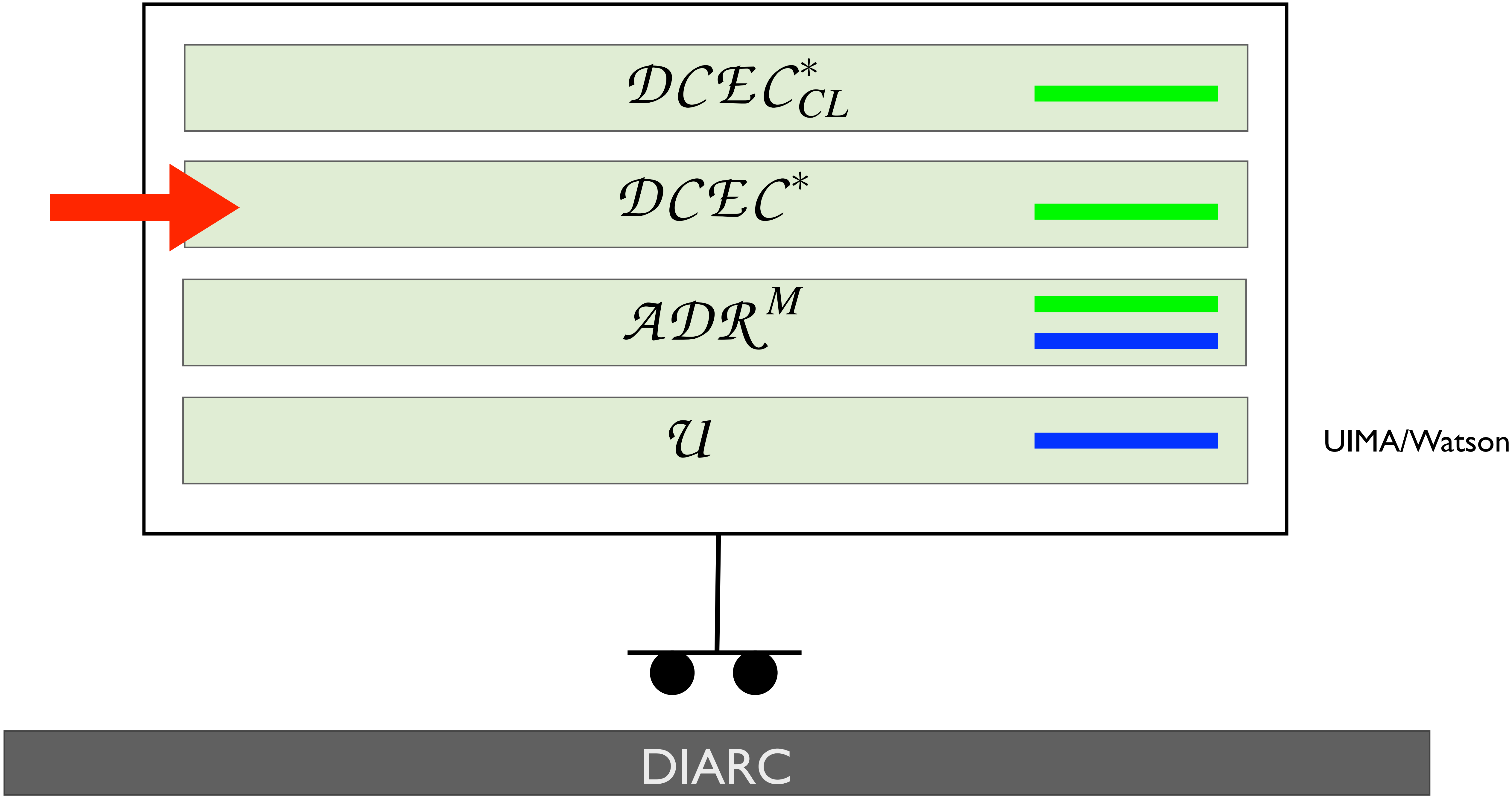
Q $$happens(dies(wife(\text{I}*)), t_0 + T) \lor holds(dead(wife(\text{I}*)), t_0 + T)$$

Note: This adheres strictly to the syntax of DCEC*

Hierarchy of Ethical Reasoning



Hierarchy of Ethical Reasoning



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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$

UIMA/Watson

DIARC

$\mathcal{DCEC}^*$

$\mathcal{DCEC}^*$

Syntax

$$S ::= \begin{array}{l}\textsf{Object} \mid \textsf{Agent} \mid \textsf{Self} \sqsubset \textsf{Agent} \mid \textsf{ActionType} \mid \textsf{Action} \sqsubseteq \textsf{Event} \mid \\ \textsf{Moment} \mid \textsf{Boolean} \mid \textsf{Fluent} \mid \textsf{Numeric}\end{array}$$

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

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

$$f ::= \begin{array}{l} action : \textsf{Agent} \times \textsf{ActionType} \rightarrow \textsf{Action} \\ initially : \textsf{Fluent} \rightarrow \textsf{Boolean} \\ holds : \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ happens : \textsf{Event} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ clipped : \textsf{Moment} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow Boolean \\ initiates : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ terminates : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ prior : \textsf{Moment} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ interval : \textsf{Moment} \times \textsf{Boolean} \\ * : \textsf{Agent} \rightarrow \textsf{Self} \\ payoff : \textsf{Agent} \times \textsf{ActionType} \times \textsf{Moment} \rightarrow \textsf{Numeric} \end{array}$$

Rules of Inference

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

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

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

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

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

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

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

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

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

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

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

$\mathcal{DCEC}^*$

Where are the emotions?

Syntax

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

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

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

action : Agent × ActionType → Action

initially : Fluent → Boolean

holds : Fluent × Moment → Boolean

happens : Event × Moment → Boolean

clipped : Moment × Fluent × Moment → *Boolean*

initiates : Event × Fluent × Moment → Boolean

terminates : Event × Fluent × Moment → Boolean

prior : Moment × Moment → Boolean

interval : Moment × Boolean

$*$: Agent → Self

payoff : Agent × ActionType × Moment → Numeric

Rules of Inference

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

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

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

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

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

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

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

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

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

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

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

$\mathcal{DCEC}^*$

Syntax

$$S ::= \begin{array}{l}\textsf{Object} \mid \textsf{Agent} \mid \textsf{Self} \sqsubset \textsf{Agent} \mid \textsf{ActionType} \mid \textsf{Action} \sqsubseteq \textsf{Event} \mid \\ \textsf{Moment} \mid \textsf{Boolean} \mid \textsf{Fluent} \mid \textsf{Numeric}\end{array}$$

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

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

$$f ::= \begin{array}{l} action : \textsf{Agent} \times \textsf{ActionType} \rightarrow \textsf{Action} \\ initially : \textsf{Fluent} \rightarrow \textsf{Boolean} \\ holds : \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ happens : \textsf{Event} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ clipped : \textsf{Moment} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow Boolean \\ initiates : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ terminates : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ prior : \textsf{Moment} \times \textsf{Moment} \rightarrow \textsf{Boolean} \\ interval : \textsf{Moment} \times \textsf{Boolean} \\ * : \textsf{Agent} \rightarrow \textsf{Self} \\ payoff : \textsf{Agent} \times \textsf{ActionType} \times \textsf{Moment} \rightarrow \textsf{Numeric}\end{array}$$

Rules of Inference

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

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

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

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

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

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

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

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

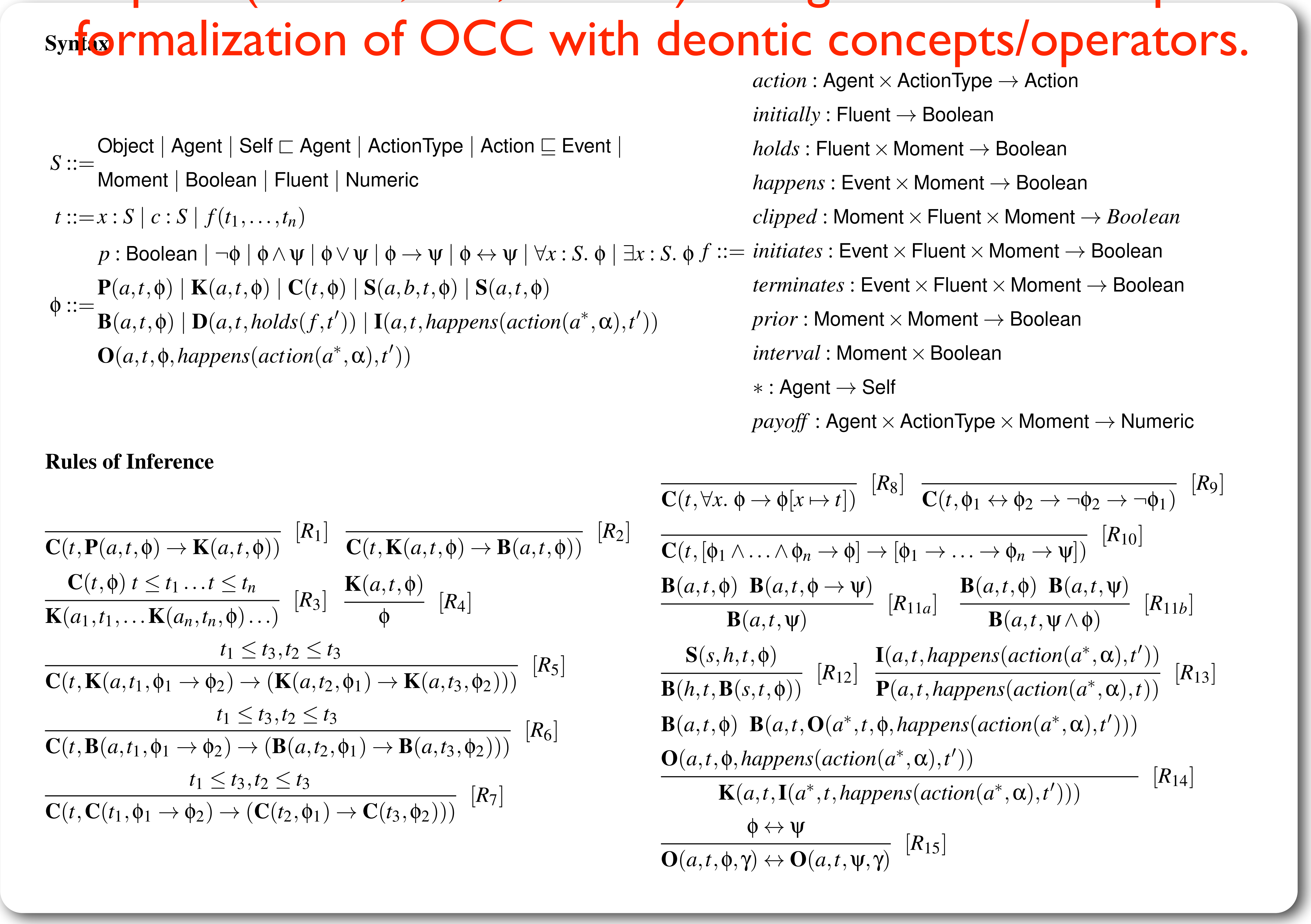
$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}] \qquad \frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\ [R_{13}]$$

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

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

$\mathcal{DCEC}^*$

Step #1 (Selmer, Mei, Naveen): Integrate version of prior formalization of OCC with deontic concepts/operators.



Syntax

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

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

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

action : Agent × ActionType → Action

$\mathit{initially}$: Fluent → Boolean

holds : Fluent × Moment → Boolean

$\mathit{happens}$: Event × Moment → Boolean

$\mathit{clipped}$: Moment × Fluent × Moment → $\mathit{Boolean}$

$\mathit{initiates}$: Event × Fluent × Moment → Boolean

$\mathit{terminates}$: Event × Fluent × Moment → Boolean

prior : Moment × Moment → Boolean

$\mathit{interval}$: Moment × Boolean

$*$: Agent → Self

payoff : Agent × ActionType × Moment → Numeric

Rules of Inference

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

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

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

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

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

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

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

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

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

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

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

$\mathcal{DCEC}^*$

Syntax

$$S ::= \begin{array}{l}\textsf{Object} \mid \textsf{Agent} \mid \textsf{Self} \sqsubset \textsf{Agent} \mid \textsf{ActionType} \mid \textsf{Action} \sqsubseteq \textsf{Event} \mid \\ \textsf{Moment} \mid \textsf{Boolean} \mid \textsf{Fluent} \mid \textsf{Numeric}\end{array}$$

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

$$p : \textsf{Boolean} \mid \neg\phi \mid \phi \wedge \psi \mid \phi \vee \psi \mid \phi \rightarrow \psi \mid \phi \leftrightarrow \psi \mid \forall x : S.\ \phi \mid \exists x : S.\ \phi \ f ::=$$

$$\phi ::= \begin{array}{l}\mathbf{P}(a,t,\phi) \mid \mathbf{K}(a,t,\phi) \mid \mathbf{C}(t,\phi) \mid \mathbf{S}(a,b,t,\phi) \mid \mathbf{S}(a,t,\phi) \\ \mathbf{B}(a,t,\phi) \mid \mathbf{D}(a,t,\mathit{holds}(f,t')) \mid \mathbf{I}(a,t,\mathit{happens}(\mathit{action}(a^*,\alpha),t')) \\ \mathbf{O}(a,t,\phi,\mathit{happens}(\mathit{action}(a^*,\alpha),t'))\end{array}$$

$\mathit{action} : \textsf{Agent} \times \textsf{ActionType} \rightarrow \textsf{Action}$

$\mathit{initially} : \textsf{Fluent} \rightarrow \textsf{Boolean}$

$\mathit{holds} : \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean}$

$\mathit{happens} : \textsf{Event} \times \textsf{Moment} \rightarrow \textsf{Boolean}$

$\mathit{clipped} : \textsf{Moment} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \mathit{Boolean}$

$\mathit{initiates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean}$

$\mathit{terminates} : \textsf{Event} \times \textsf{Fluent} \times \textsf{Moment} \rightarrow \textsf{Boolean}$

$\mathit{prior} : \textsf{Moment} \times \textsf{Moment} \rightarrow \textsf{Boolean}$

$\mathit{interval} : \textsf{Moment} \times \textsf{Boolean}$

$* : \textsf{Agent} \rightarrow \textsf{Self}$

$\mathit{payoff} : \textsf{Agent} \times \textsf{ActionType} \times \textsf{Moment} \rightarrow \textsf{Numeric}$

Rules of Inference

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

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

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

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

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

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

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

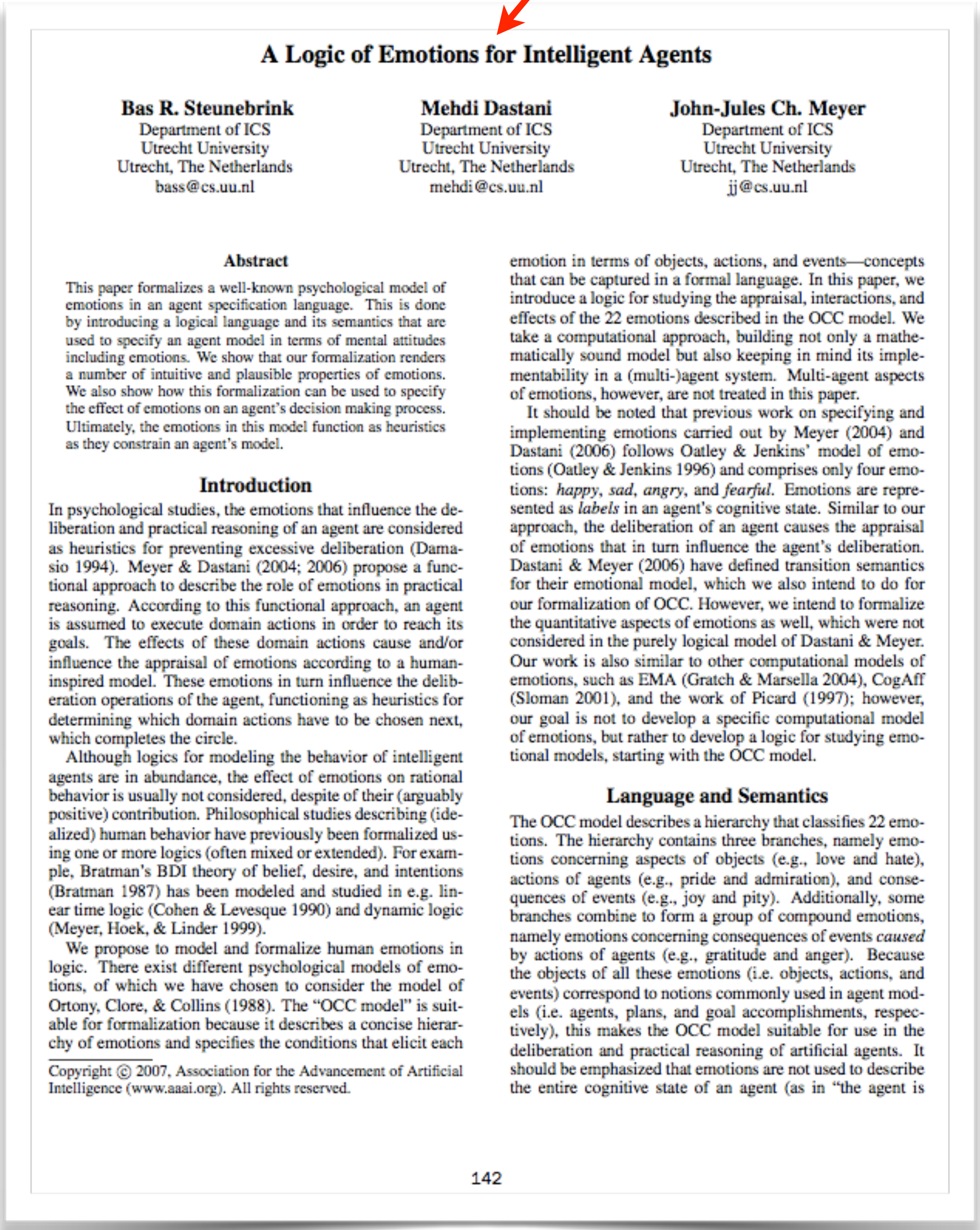
$$\frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\phi \rightarrow \psi)}{\mathbf{B}(a,t,\psi)}\ [R_{11a}] \qquad \frac{\mathbf{B}(a,t,\phi)\quad \mathbf{B}(a,t,\psi)}{\mathbf{B}(a,t,\psi \wedge \phi)}\ [R_{11b}]$$

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

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

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

$\mathcal{DCEC}^*$



A Logic of Emotions for Intelligent Agents

Bas R. Steunebrink
Department of ICS
Utrecht University
Utrecht, The Netherlands
bass@cs.uu.nl

Mehdi Dastani
Department of ICS
Utrecht University
Utrecht, The Netherlands
mehdi@cs.uu.nl

John-Jules Ch. Meyer
Department of ICS
Utrecht University
Utrecht, The Netherlands
jj@cs.uu.nl

Abstract

This paper formalizes a well-known psychological model of emotions in an agent specification language. This is done by introducing a logical language and its semantics that are used to specify an agent model in terms of mental attitudes including emotions. We show that our formalization renders a number of intuitive and plausible properties of emotions. We also show how this formalization can be used to specify the effect of emotions on an agent's decision making process. Ultimately, the emotions in this model function as heuristics as they constrain an agent's model.

Introduction

In psychological studies, the emotions that influence the deliberation and practical reasoning of an agent are considered as heuristics for preventing excessive deliberation (Damasio 1994). Meyer & Dastani (2004; 2006) propose a functional approach to describe the role of emotions in practical reasoning. According to this functional approach, an agent is assumed to execute domain actions in order to reach its goals. The effects of these domain actions cause and/or influence the appraisal of emotions according to a human-inspired model. These emotions in turn influence the deliberation operations of the agent, functioning as heuristics for determining which domain actions have to be chosen next, which completes the circle.

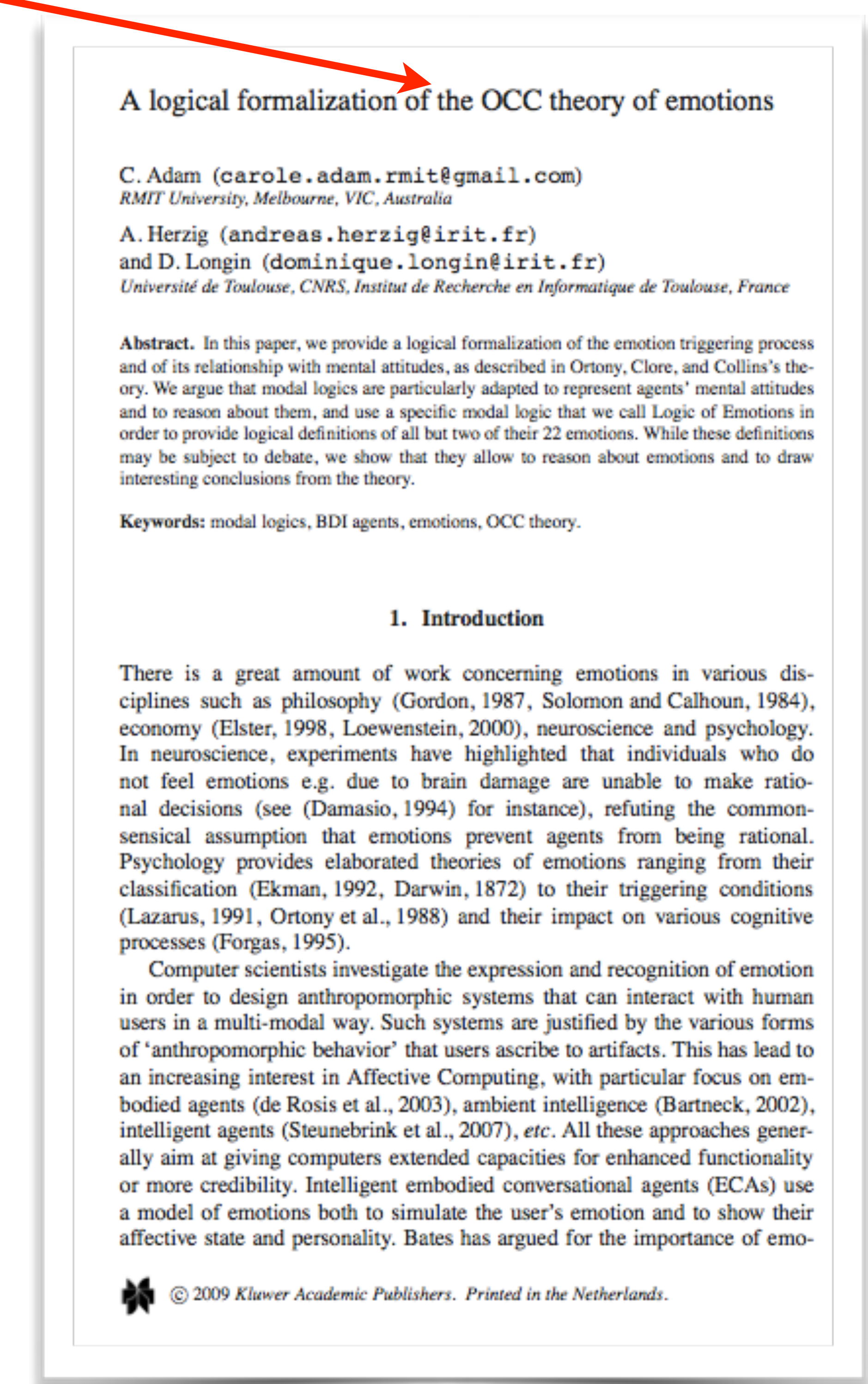
Although logics for modeling the behavior of intelligent agents are in abundance, the effect of emotions on rational behavior is usually not considered, despite of their (arguably positive) contribution. Philosophical studies describing (idealized) human behavior have previously been formalized using one or more logics (often mixed or extended). For example, Bratman's BDI theory of belief, desire, and intentions (Bratman 1987) has been modeled and studied in e.g. linear time logic (Cohen & Levesque 1990) and dynamic logic (Meyer, Hoek, & Linder 1999).

We propose to model and formalize human emotions in logic. There exist different psychological models of emotions, of which we have chosen to consider the model of Ortony, Clore, & Collins (1988). The "OCC model" is suitable for formalization because it describes a concise hierarchy of emotions and specifies the conditions that elicit each emotion in terms of objects, actions, and events—concepts that can be captured in a formal language. In this paper, we introduce a logic for studying the appraisal, interactions, and effects of the 22 emotions described in the OCC model. We take a computational approach, building not only a mathematically sound model but also keeping in mind its implementability in a (multi-)agent system. Multi-agent aspects of emotions, however, are not treated in this paper.

It should be noted that previous work on specifying and implementing emotions carried out by Meyer (2004) and Dastani (2006) follows Oatley & Jenkins' model of emotions (Oatley & Jenkins 1996) and comprises only four emotions: *happy*, *sad*, *angry*, and *fearful*. Emotions are represented as *labels* in an agent's cognitive state. Similar to our approach, the deliberation of an agent causes the appraisal of emotions that in turn influence the agent's deliberation. Dastani & Meyer (2006) have defined transition semantics for their emotional model, which we also intend to do for our formalization of OCC. However, we intend to formalize the quantitative aspects of emotions as well, which were not considered in the purely logical model of Dastani & Meyer. Our work is also similar to other computational models of emotions, such as EMA (Gratch & Marsella 2004), CogAff (Sloman 2001), and the work of Picard (1997); however, our goal is not to develop a specific computational model of emotions, but rather to develop a logic for studying emotional models, starting with the OCC model.

Language and Semantics

The OCC model describes a hierarchy that classifies 22 emotions. The hierarchy contains three branches, namely emotions concerning aspects of objects (e.g., love and hate), actions of agents (e.g., pride and admiration), and consequences of events (e.g., joy and pity). Additionally, some branches combine to form a group of compound emotions, namely emotions concerning consequences of events *caused* by actions of agents (e.g., gratitude and anger). Because the objects of all these emotions (i.e. objects, actions, and events) correspond to notions commonly used in agent models (i.e. agents, plans, and goal accomplishments, respectively), this makes the OCC model suitable for use in the deliberation and practical reasoning of artificial agents. It should be emphasized that emotions are not used to describe the entire cognitive state of an agent (as in "the agent is



A logical formalization of the OCC theory of emotions

C. Adam (carole.adam.rmit@gmail.com)
RMIT University, Melbourne, VIC, Australia

A. Herzig (andreas.herzig@irit.fr)
and D. Longin (dominique.longin@irit.fr)
Université de Toulouse, CNRS, Institut de Recherche en Informatique de Toulouse, France

Abstract. In this paper, we provide a logical formalization of the emotion triggering process and of its relationship with mental attitudes, as described in Ortony, Clore, and Collins's theory. We argue that modal logics are particularly adapted to represent agents' mental attitudes and to reason about them, and use a specific modal logic that we call Logic of Emotions in order to provide logical definitions of all but two of their 22 emotions. While these definitions may be subject to debate, we show that they allow to reason about emotions and to draw interesting conclusions from the theory.

Keywords: modal logics, BDI agents, emotions, OCC theory.

1. Introduction

There is a great amount of work concerning emotions in various disciplines such as philosophy (Gordon, 1987, Solomon and Calhoun, 1984), economy (Elster, 1998, Loewenstein, 2000), neuroscience and psychology. In neuroscience, experiments have highlighted that individuals who do not feel emotions e.g. due to brain damage are unable to make rational decisions (see (Damasio, 1994) for instance), refuting the commonsensical assumption that emotions prevent agents from being rational. Psychology provides elaborated theories of emotions ranging from their classification (Ekman, 1992, Darwin, 1872) to their triggering conditions (Lazarus, 1991, Ortony et al., 1988) and their impact on various cognitive processes (Forgas, 1995).

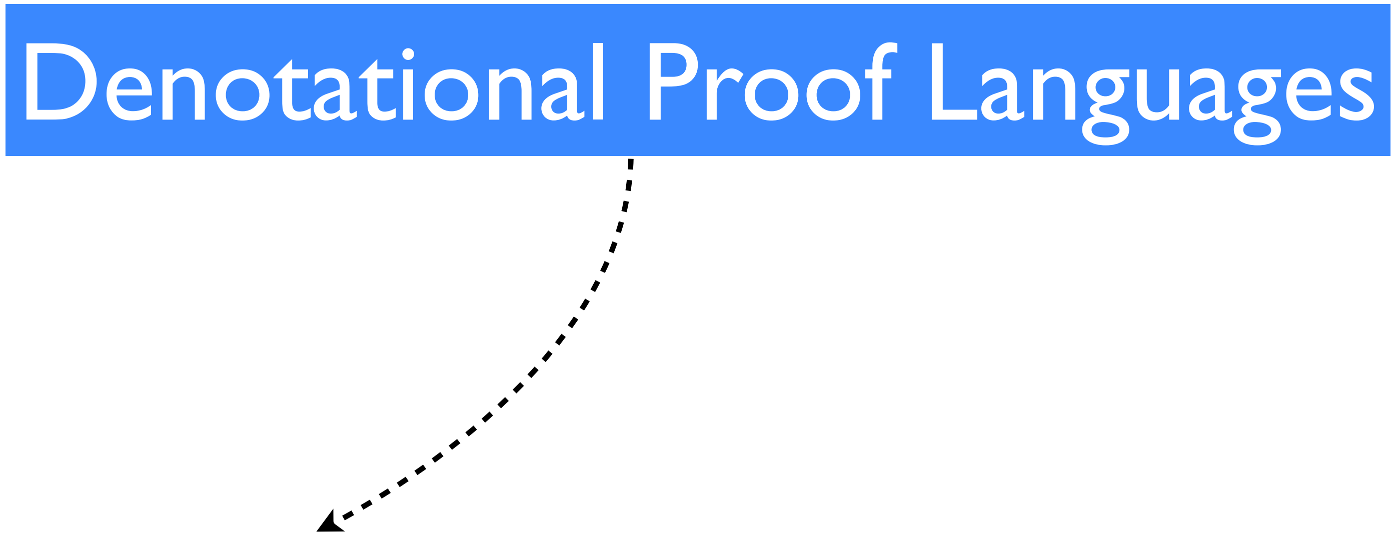
Computer scientists investigate the expression and recognition of emotion in order to design anthropomorphic systems that can interact with human users in a multi-modal way. Such systems are justified by the various forms of 'anthropomorphic behavior' that users ascribe to artifacts. This has lead to an increasing interest in Affective Computing, with particular focus on embodied agents (de Rosis et al., 2003), ambient intelligence (Bartneck, 2002), intelligent agents (Steunebrink et al., 2007), *etc*. All these approaches generally aim at giving computers extended capacities for enhanced functionality or more credibility. Intelligent embodied conversational agents (ECAs) use a model of emotions both to simulate the user's emotion and to show their affective state and personality. Bates has argued for the importance of emo-

Automation of Reasoning

Automation of Reasoning

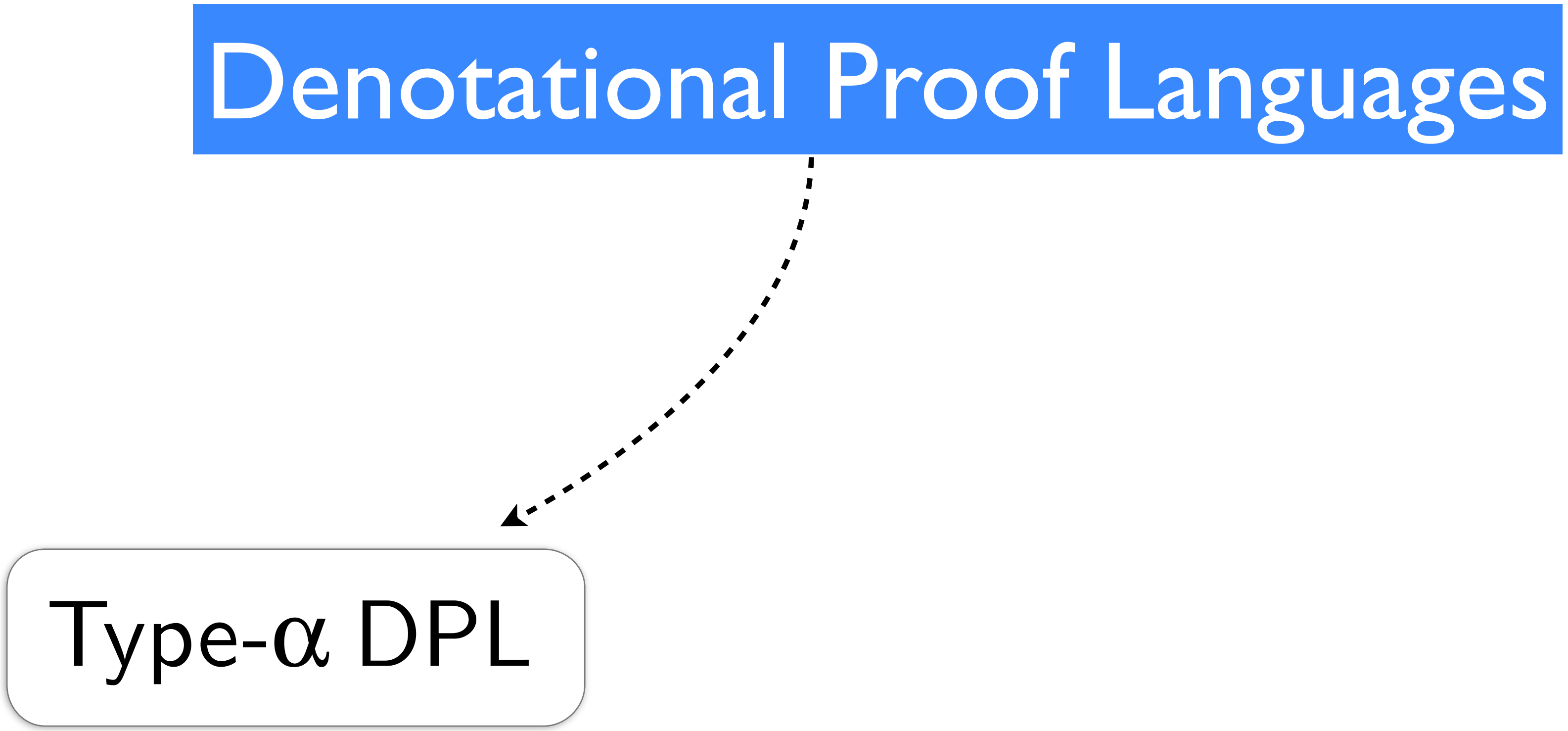
Denotational Proof Languages

Automation of Reasoning



Denotational Proof Languages

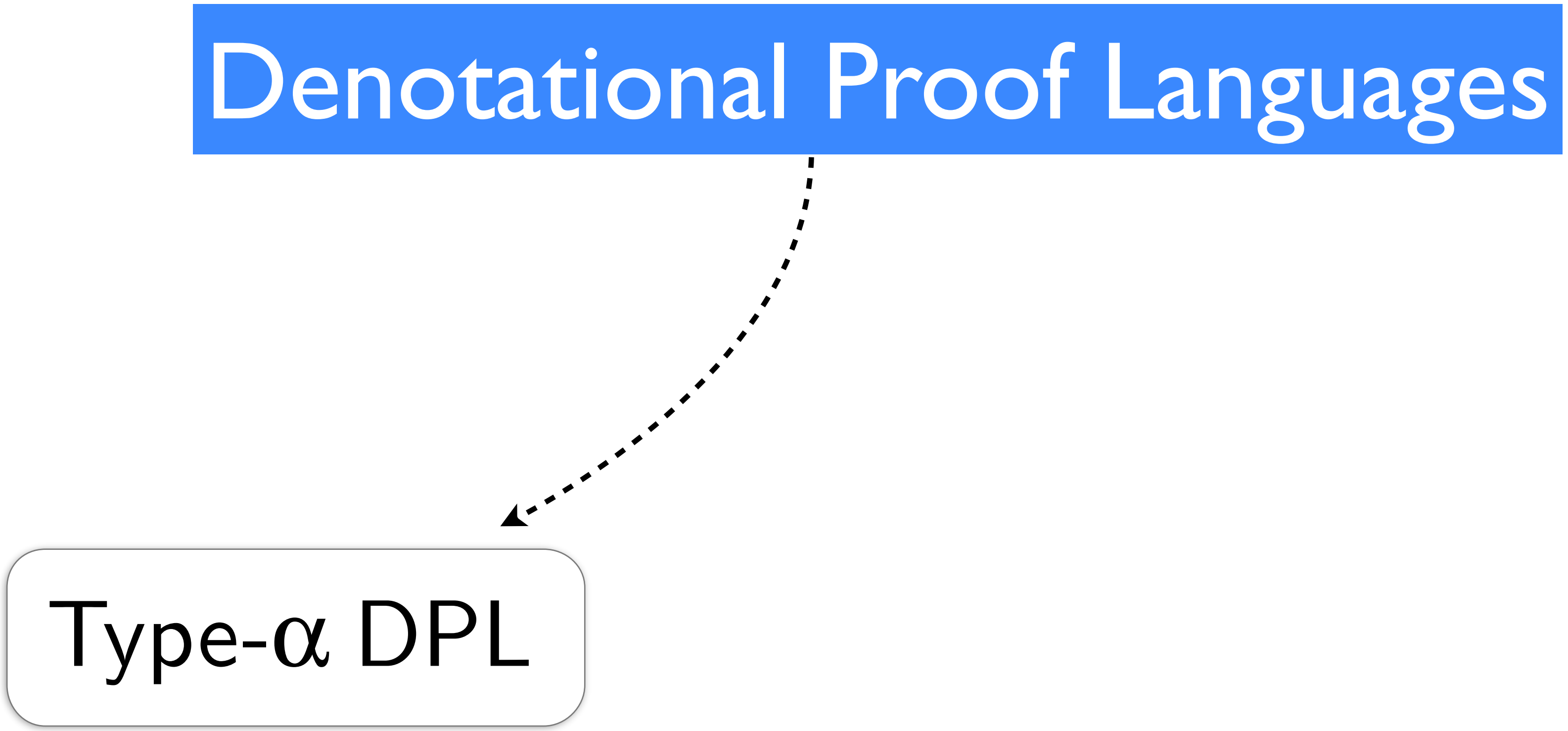
Automation of Reasoning



Denotational Proof Languages

Type-α DPL

Automation of Reasoning

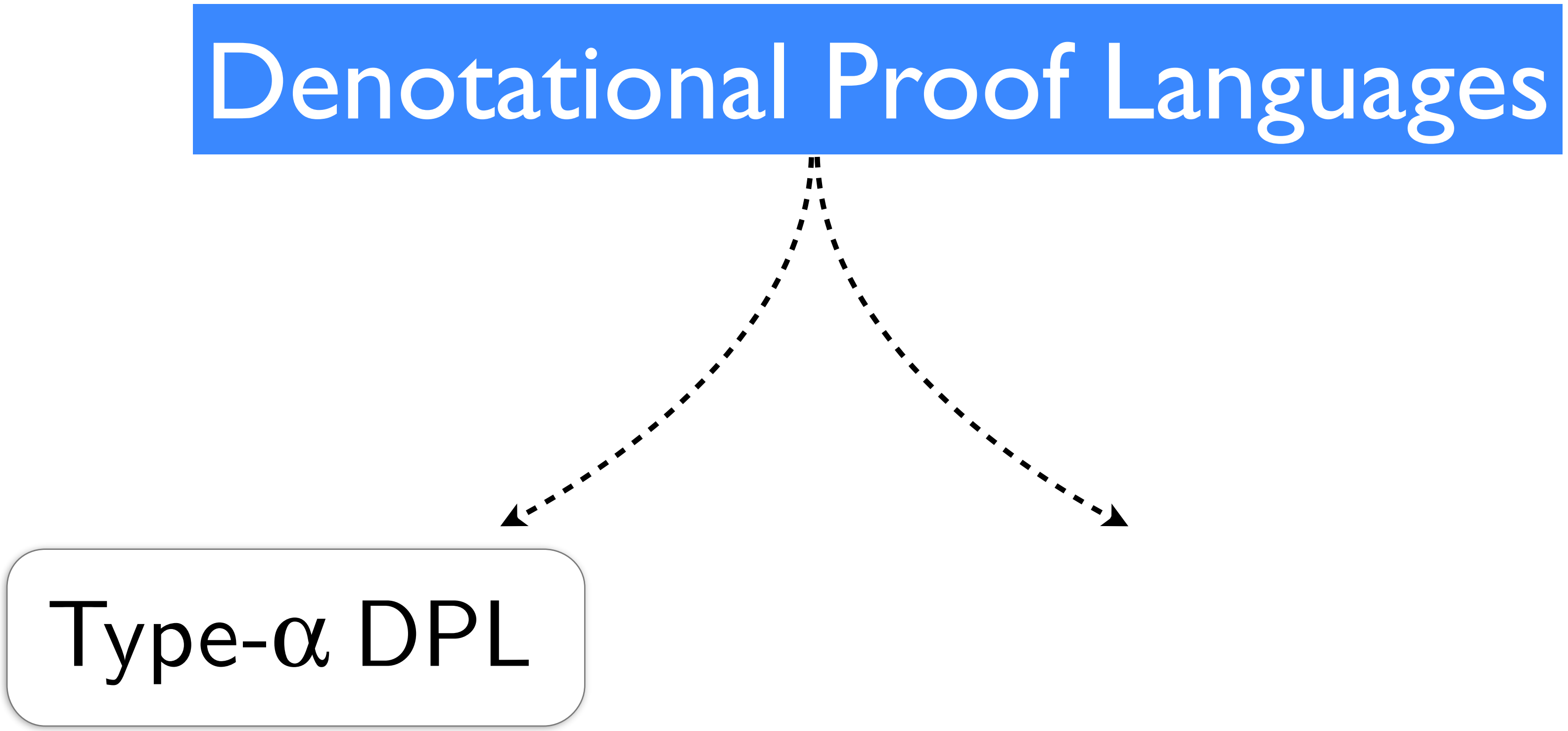


Denotational Proof Languages

Type-α DPL

Proof checking.

Automation of Reasoning

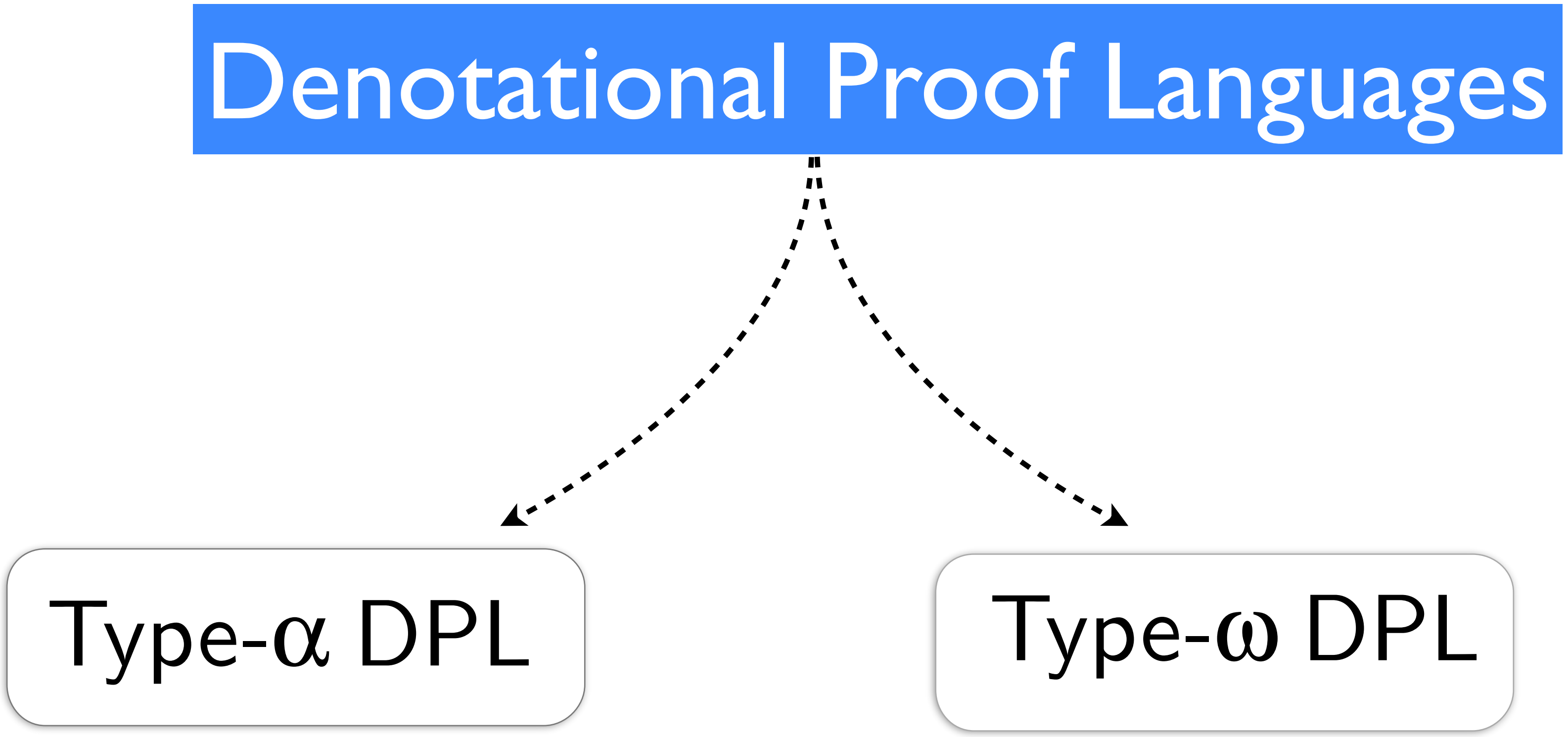


Denotational Proof Languages

Type-α DPL

Proof checking.

Automation of Reasoning



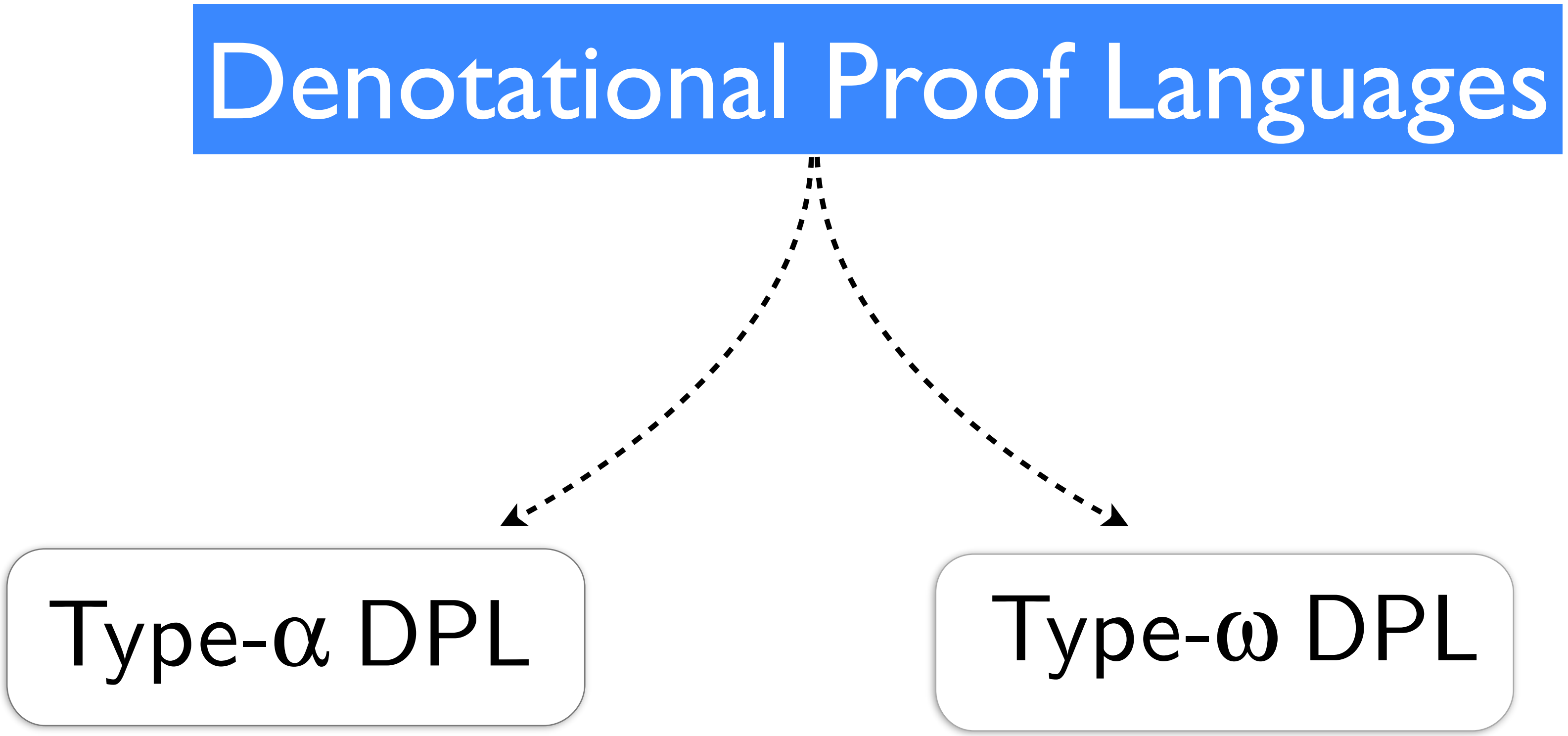
Denotational Proof Languages

Type-α DPL

Type-ω DPL

Proof checking.

Automation of Reasoning



Denotational Proof Languages

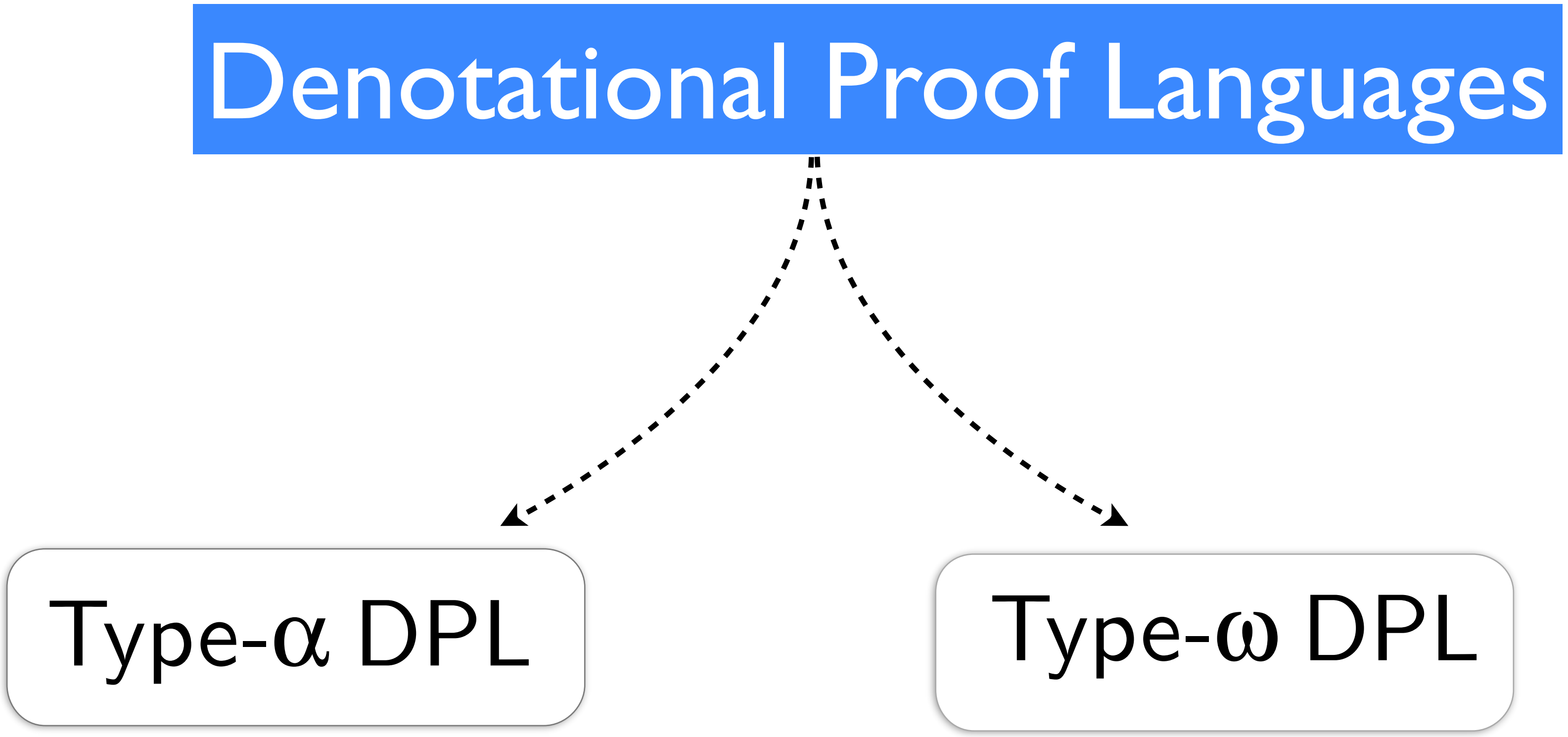
Type-α DPL

Type-ω DPL

Proof checking.

Proof discovery (and checking).

Automation of Reasoning



Denotational Proof Languages

Type-α DPL

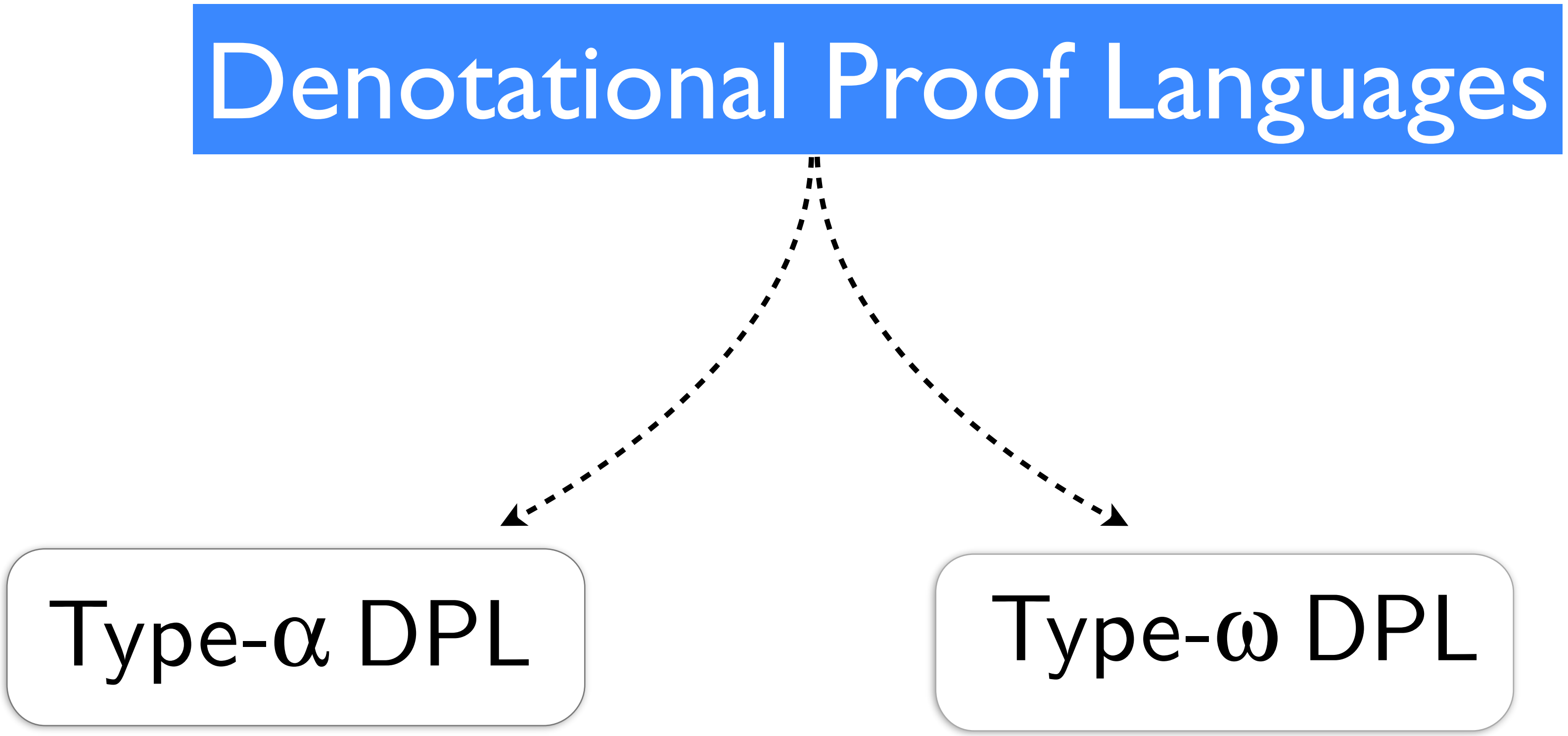
Type-ω DPL

Proof checking.

Proof discovery (and checking).

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

Automation of Reasoning



Denotational Proof Languages

Type-α DPL

Type-ω DPL

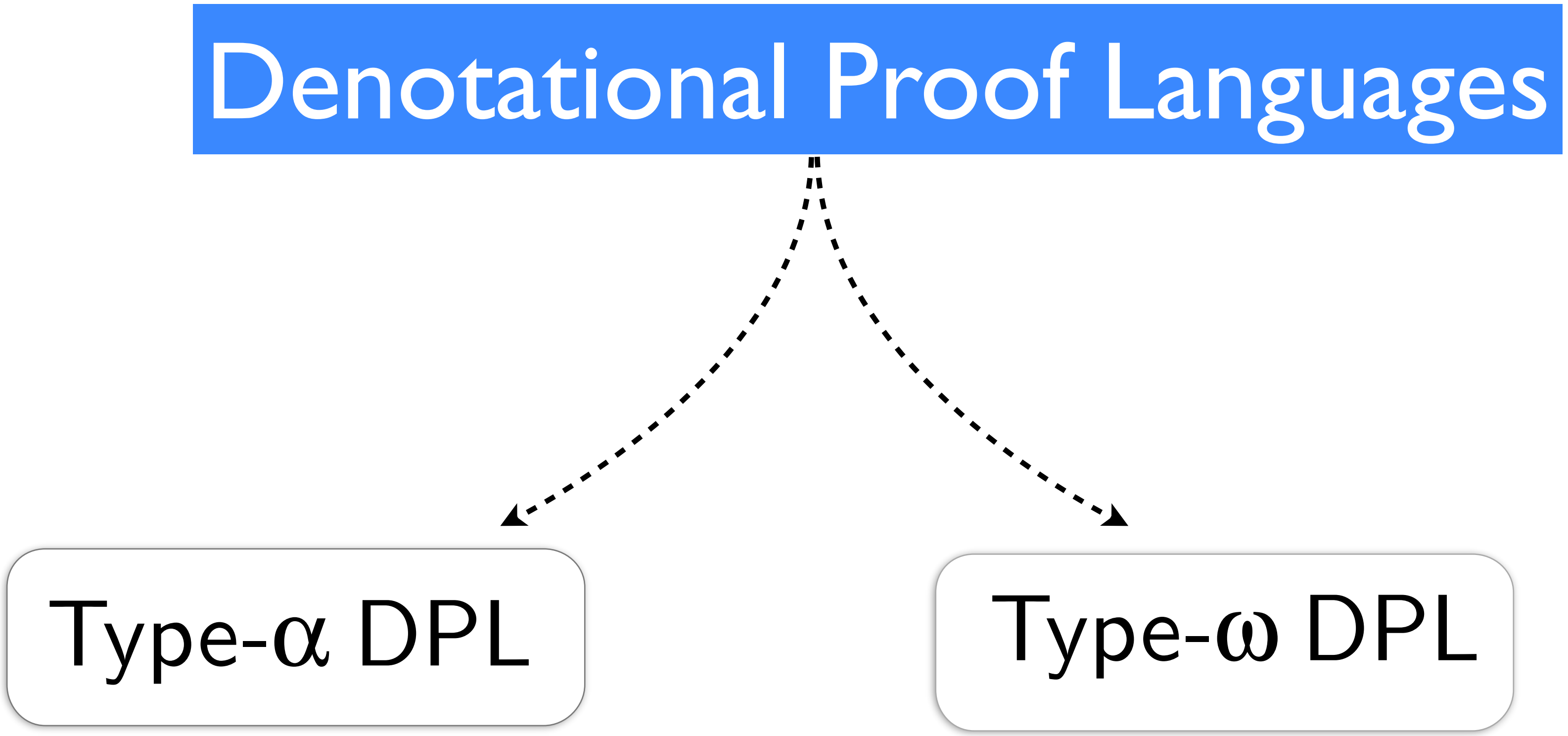
Proof checking.

Proof discovery (and checking).

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.

Automation of Reasoning



Denotational Proof Languages

Type-α DPL

Type-ω DPL

Proof checking.

Proof discovery (and checking).

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

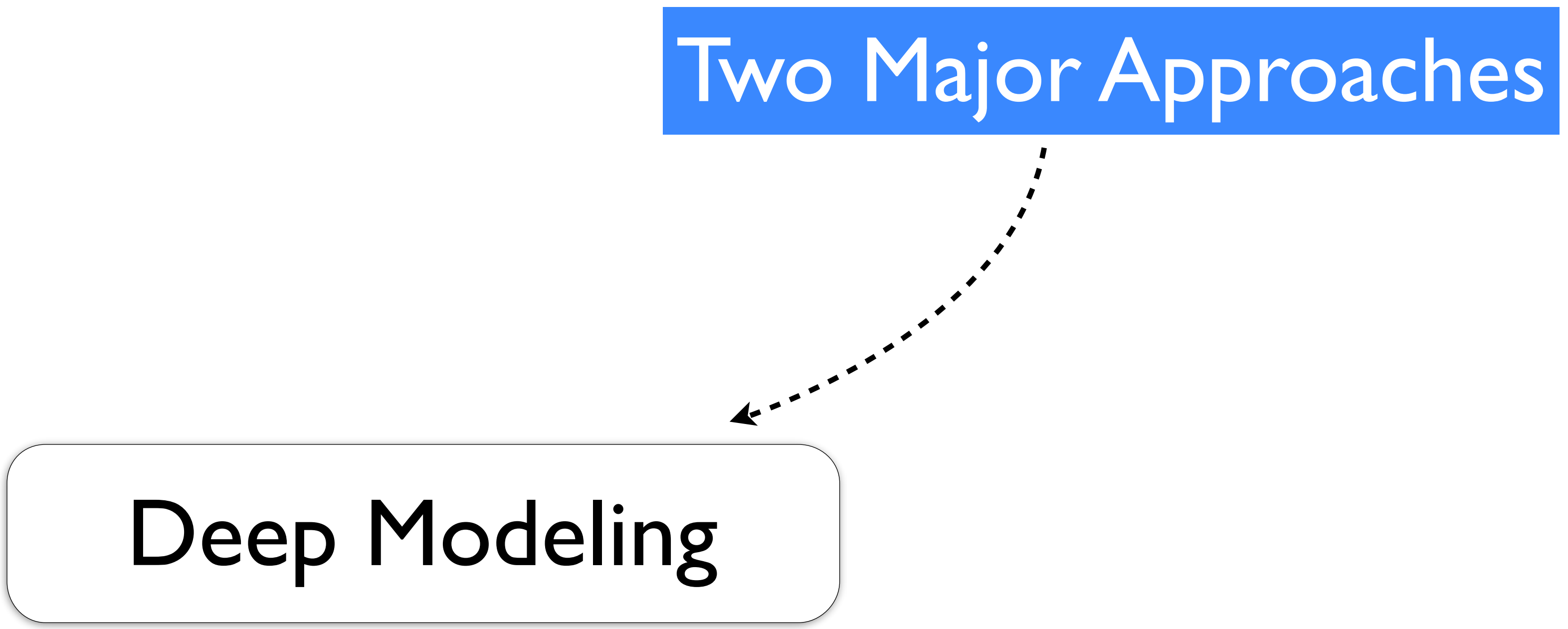
Logicist NLP

Two Major Approaches

Logicist NLP

Two Major Approaches

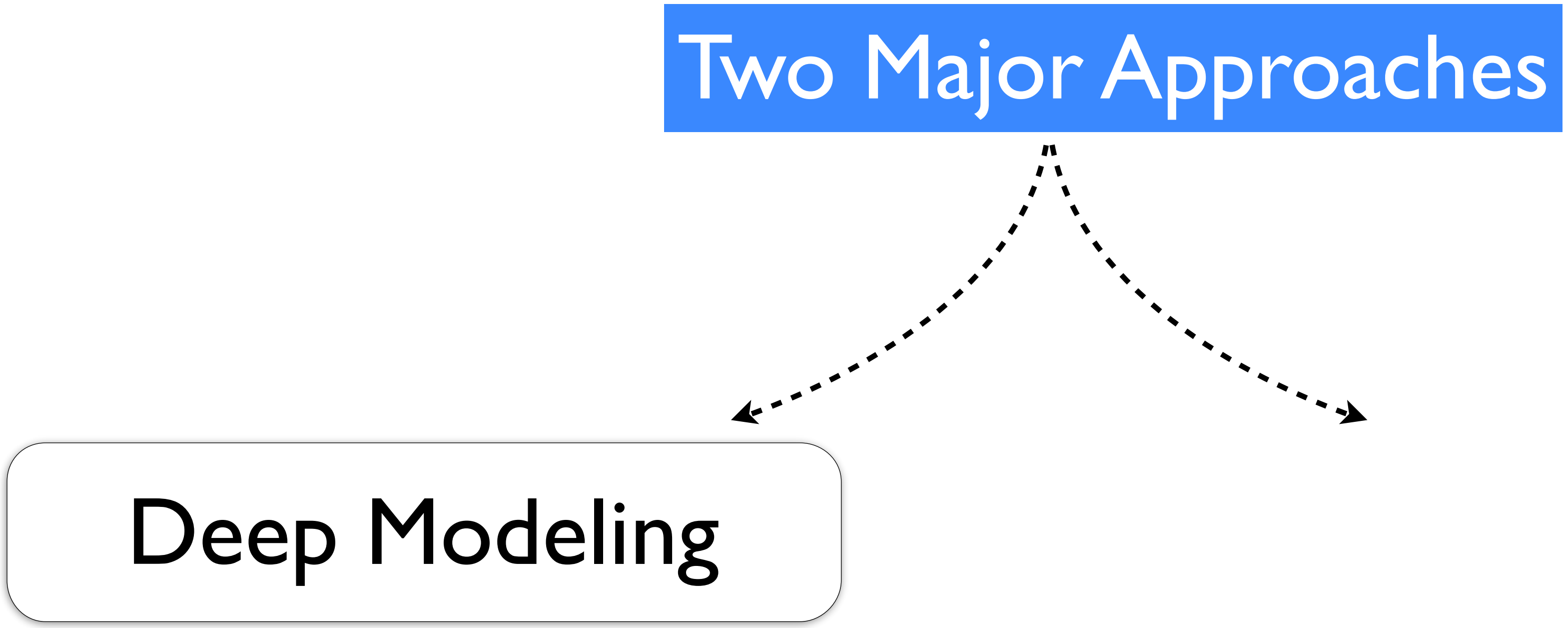
Logicist NLP



Two Major Approaches

Deep Modeling

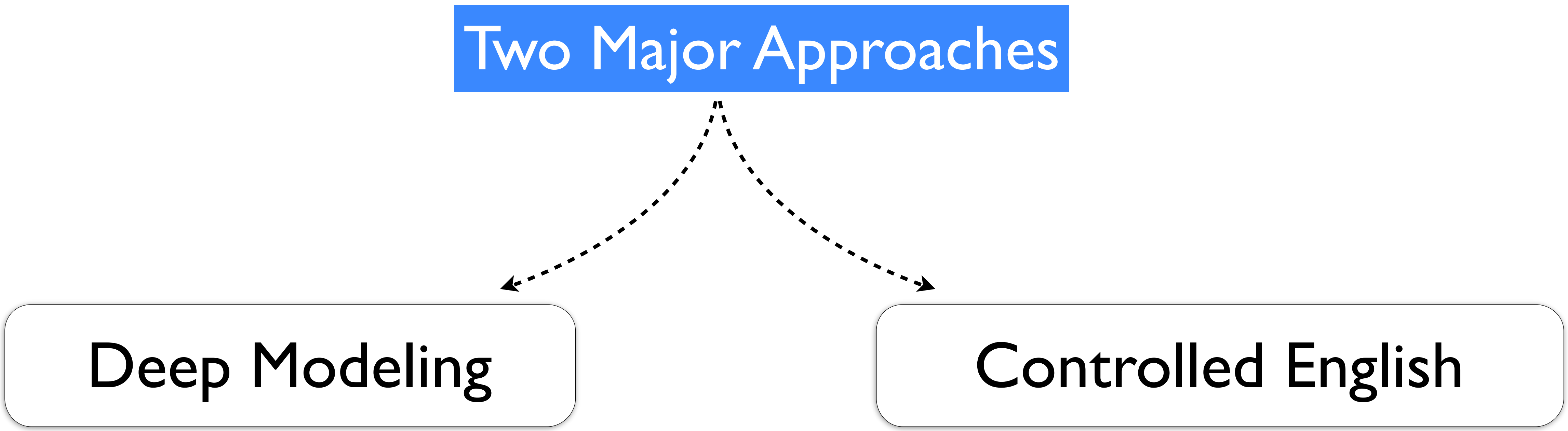
Logicist NLP



Two Major Approaches

Deep Modeling

Logicist NLP

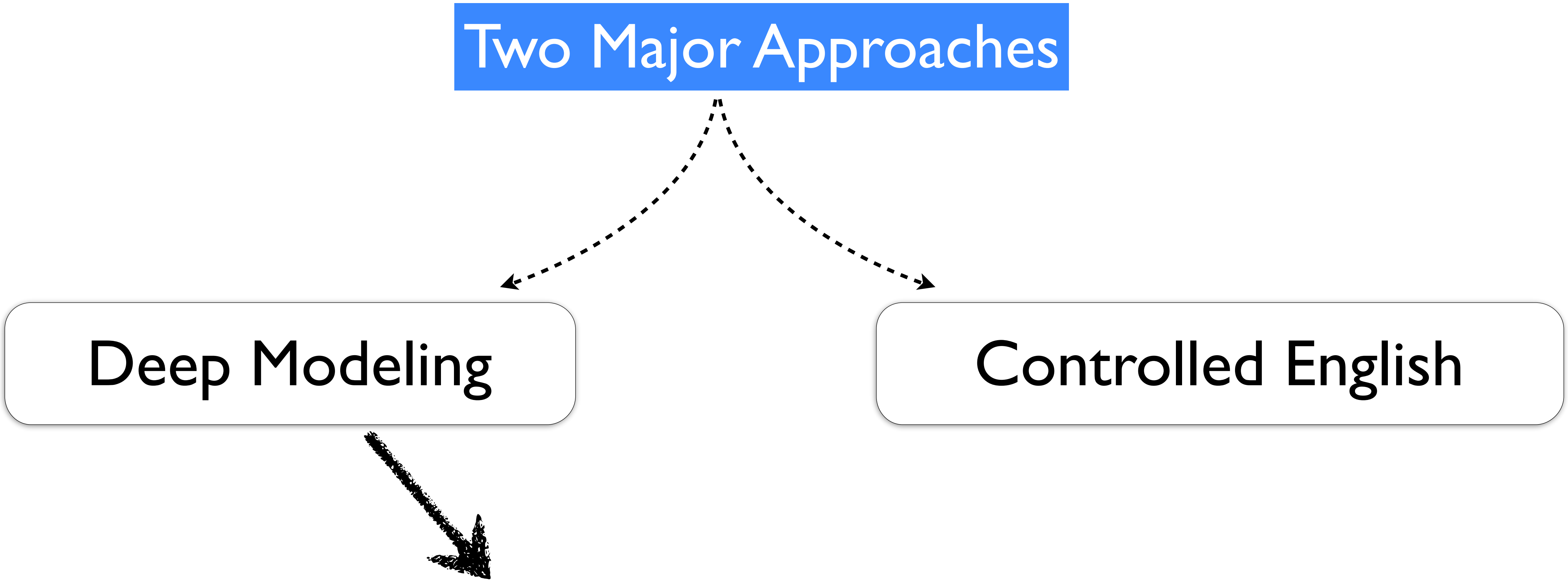


Two Major Approaches

Deep Modeling

Controlled English

Logicist NLP

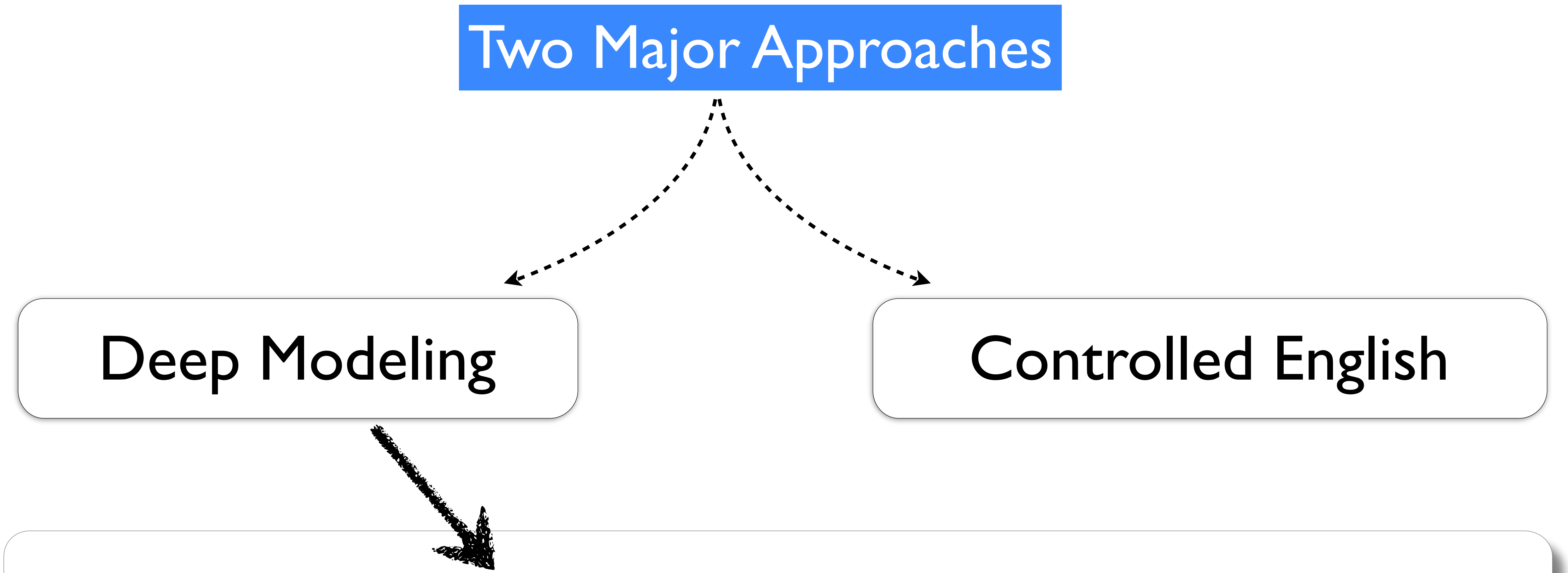


Two Major Approaches

Deep Modeling

Controlled English

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

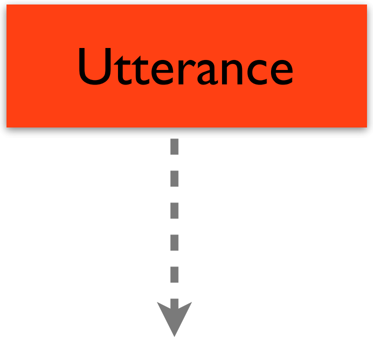
The Sixth Conference on
Artificial General Intelligence
Beijing, July 31 – August 3, 2013

Deep Modeling

Utterance

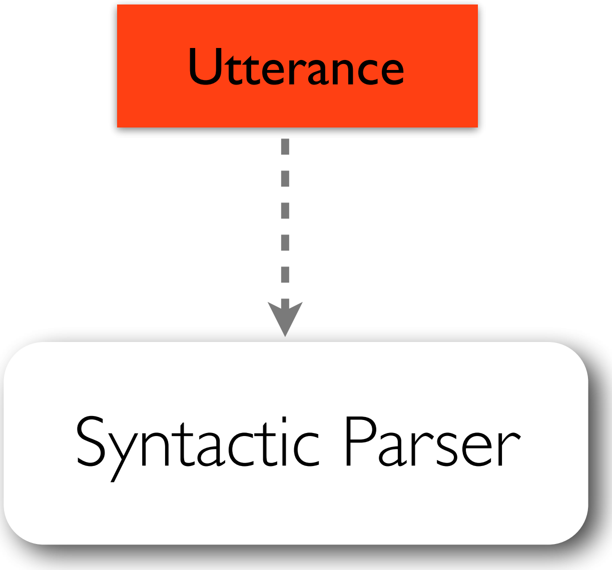
Deep Modeling

Deep Modeling



Utterance

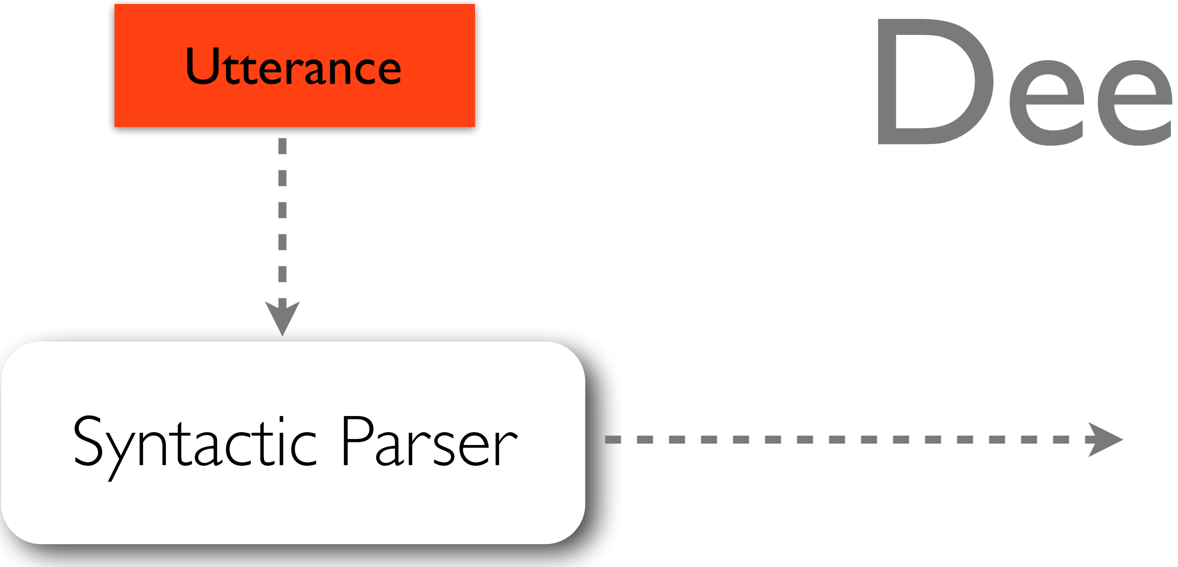
Deep Modeling



Utterance

Syntactic Parser

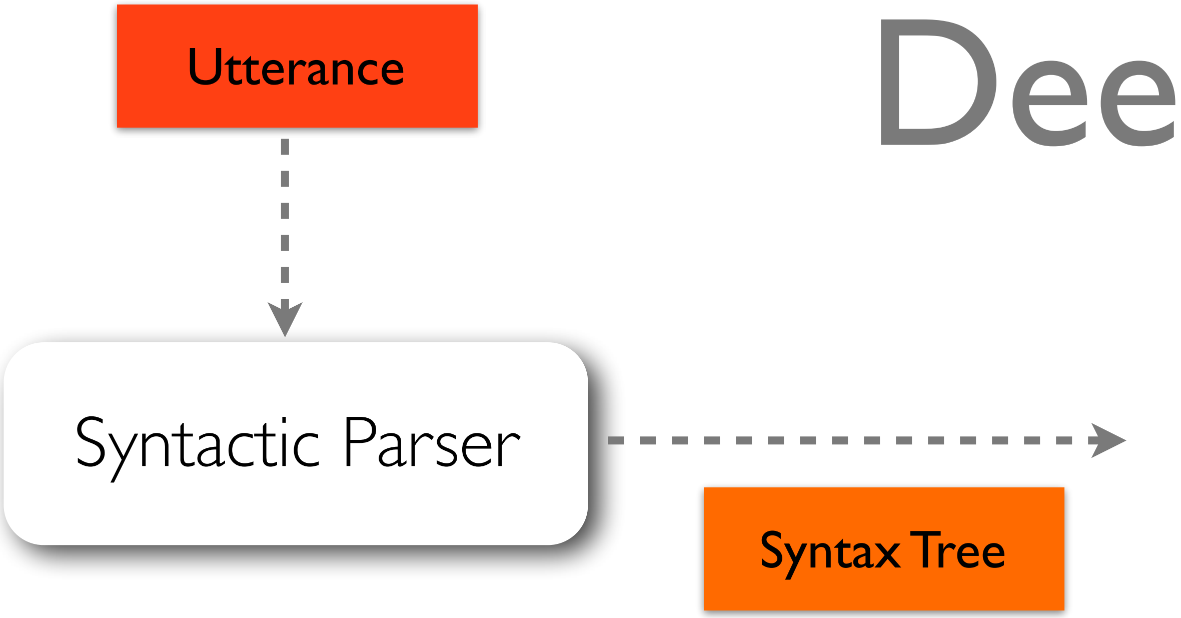
Deep Modeling



Utterance

Syntactic Parser

Deep Modeling

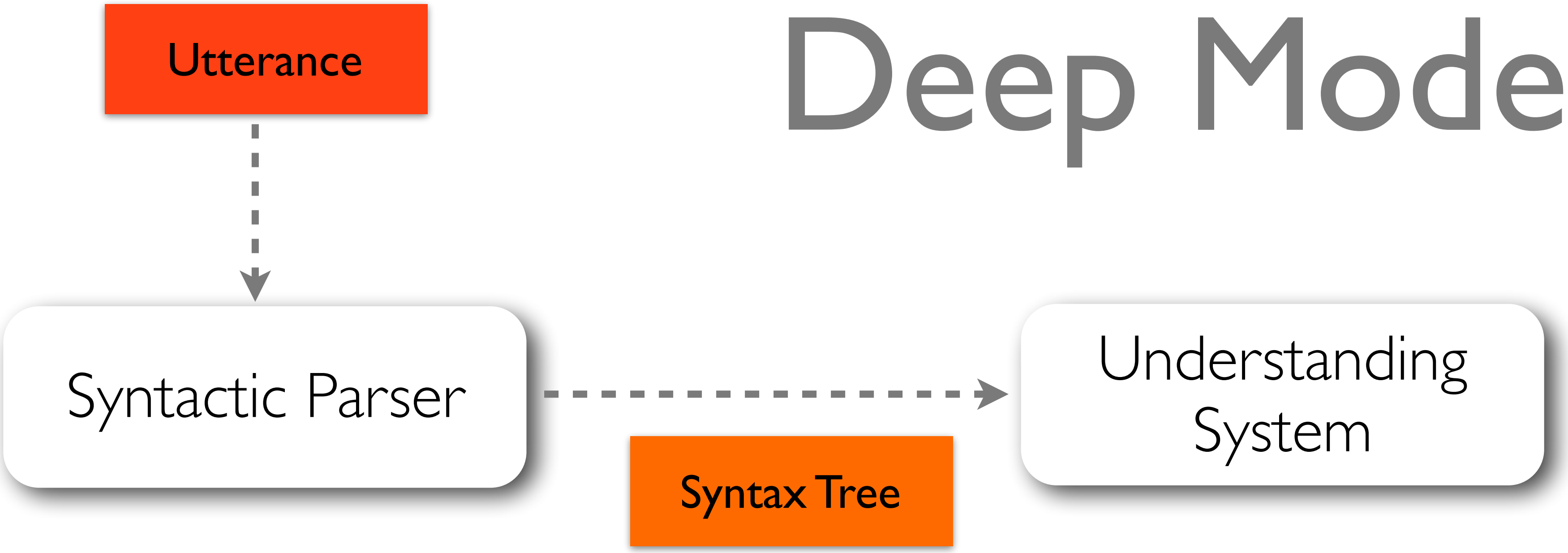


Utterance

Syntactic Parser

Syntax Tree

Deep Modeling



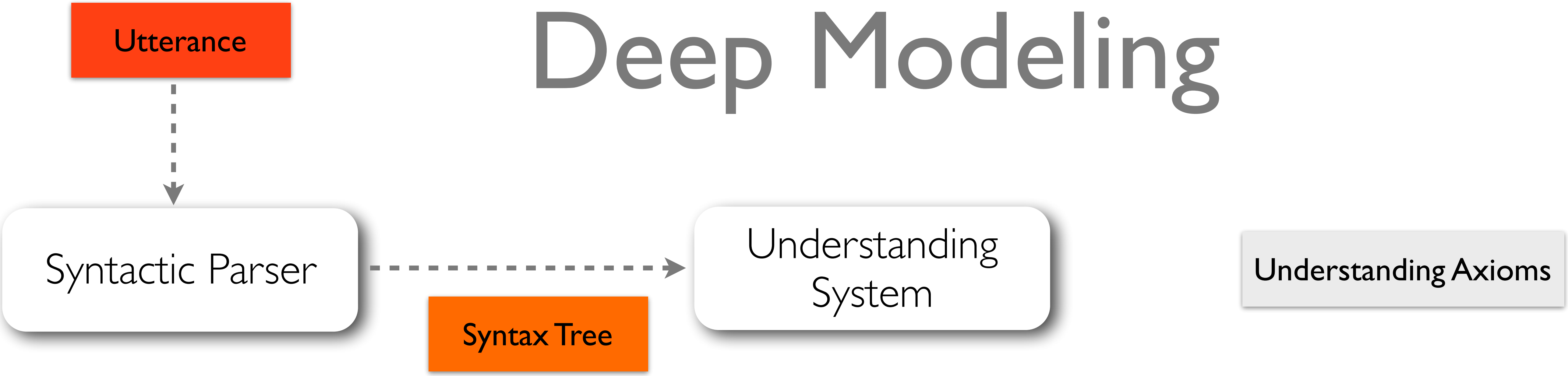
Utterance

Syntactic Parser

Syntax Tree

Understanding System

Deep Modeling


Utterance

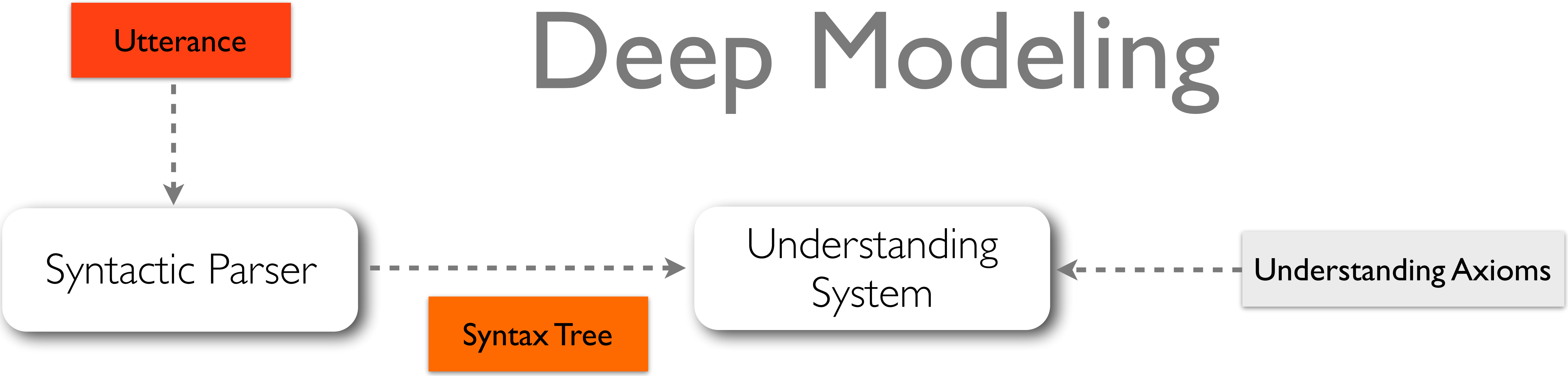
Syntactic Parser

Syntax Tree

Understanding System

Understanding Axioms

Deep Modeling



Utterance

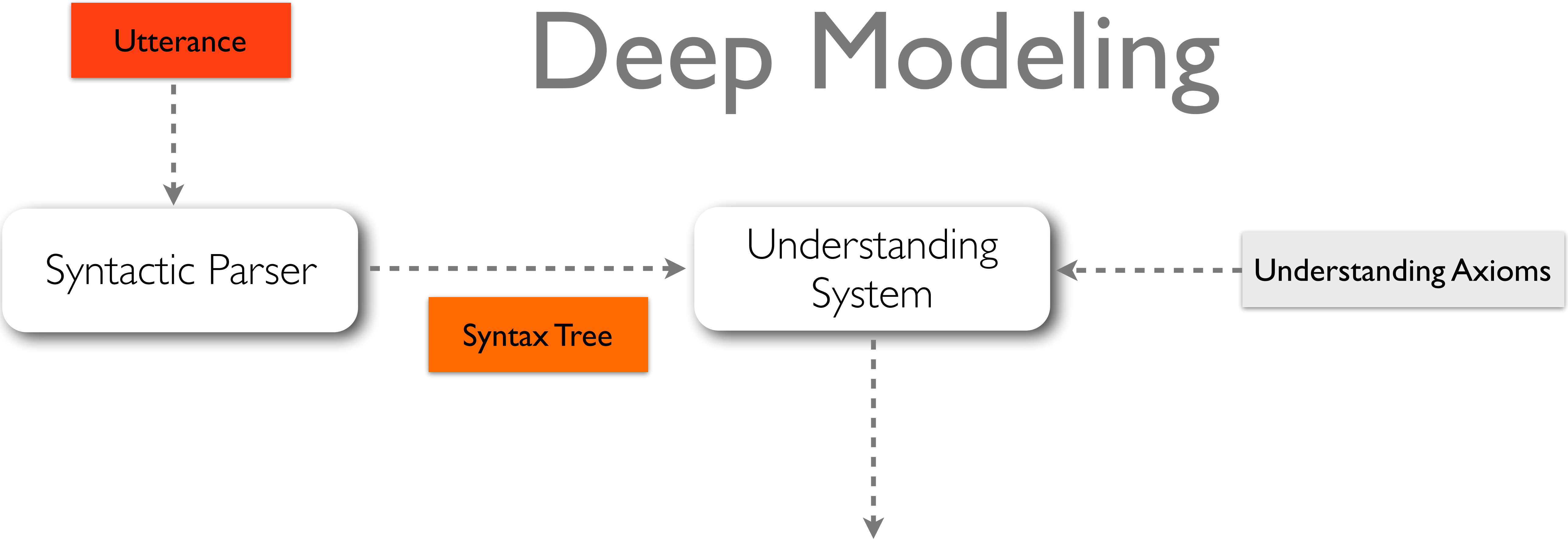
Syntactic Parser

Syntax Tree

Understanding System

Understanding Axioms

Deep Modeling



Utterance

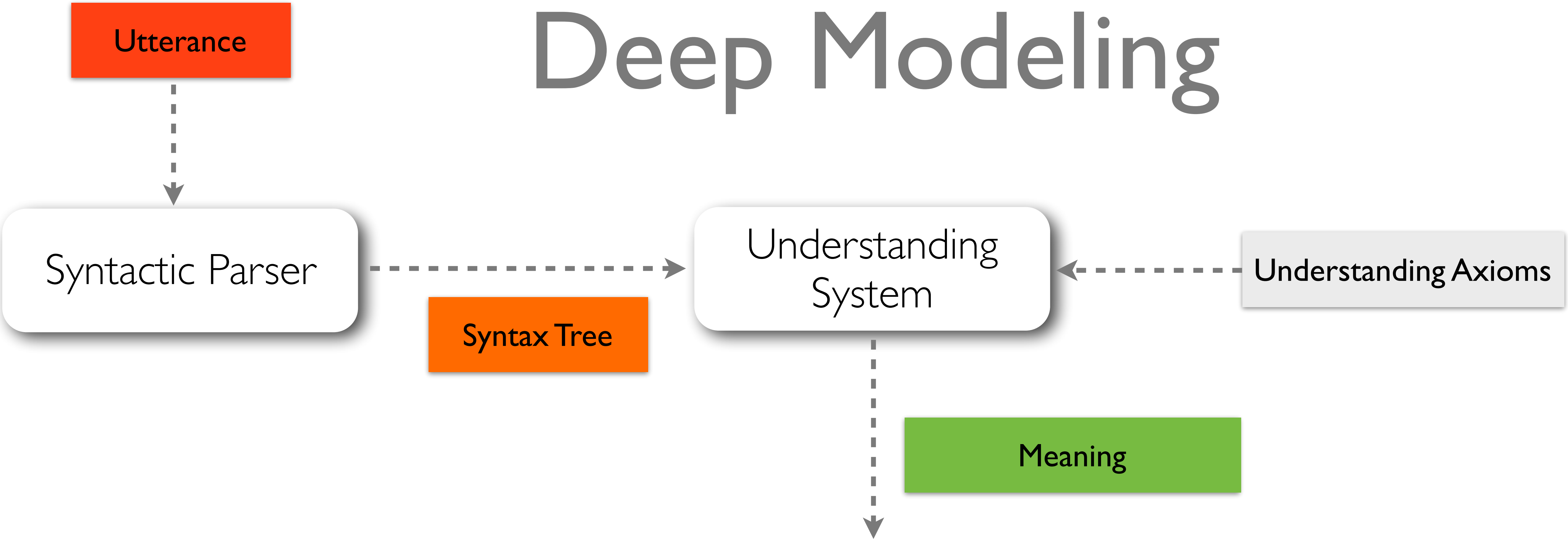
Syntactic Parser

Syntax Tree

Understanding System

Understanding Axioms

Deep Modeling



Utterance

Syntactic Parser

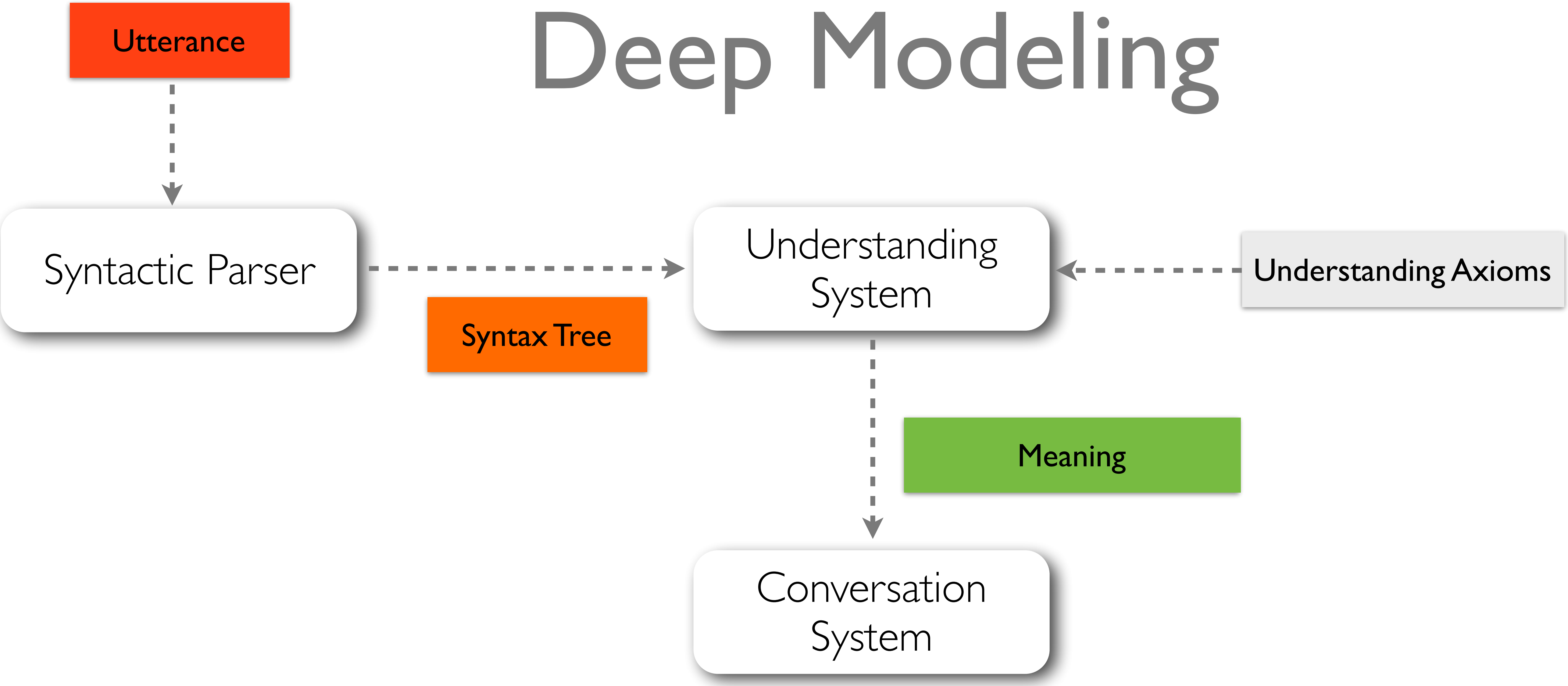
Syntax Tree

Understanding System

Understanding Axioms

Meaning

Deep Modeling



Utterance

Syntactic Parser

Syntax Tree

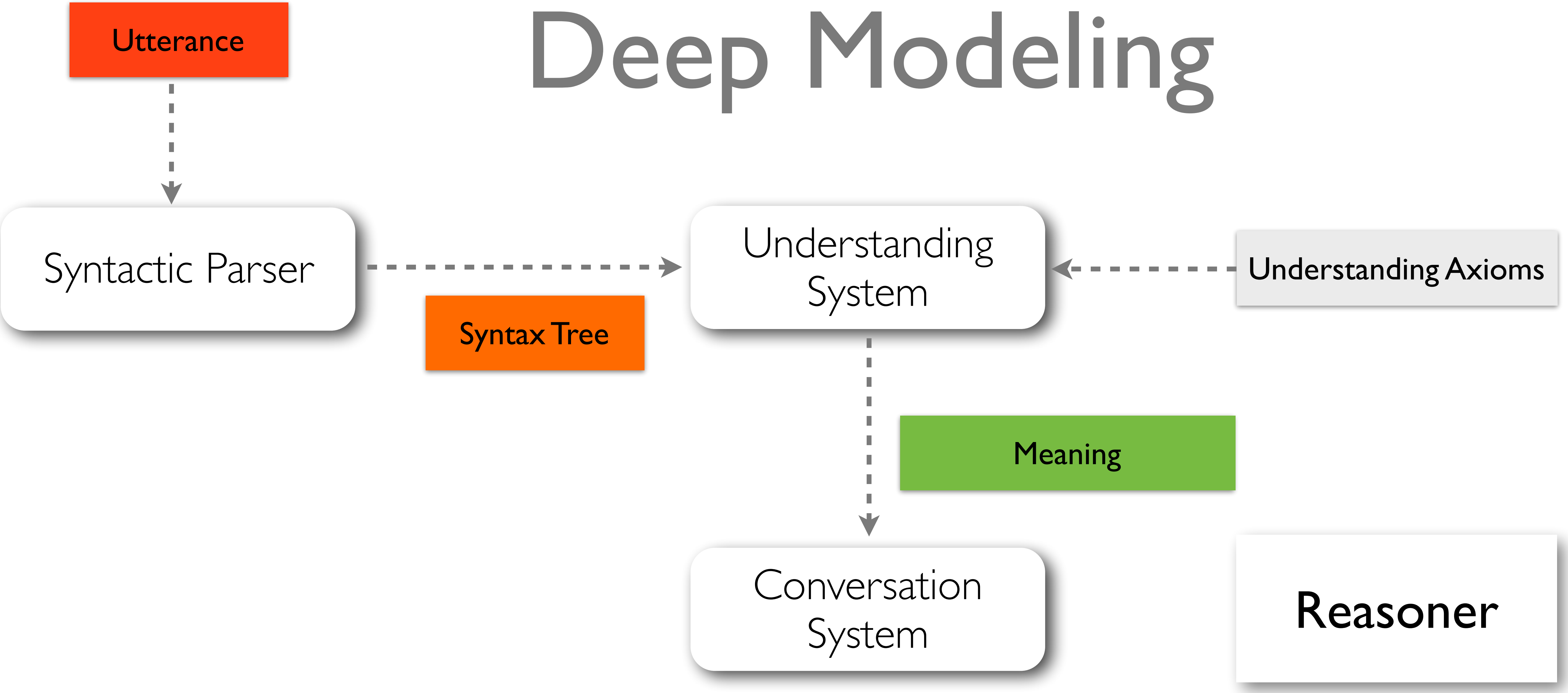
Understanding System

Understanding Axioms

Meaning

Conversation System

Deep Modeling



Utterance

Syntactic Parser

Syntax Tree

Understanding System

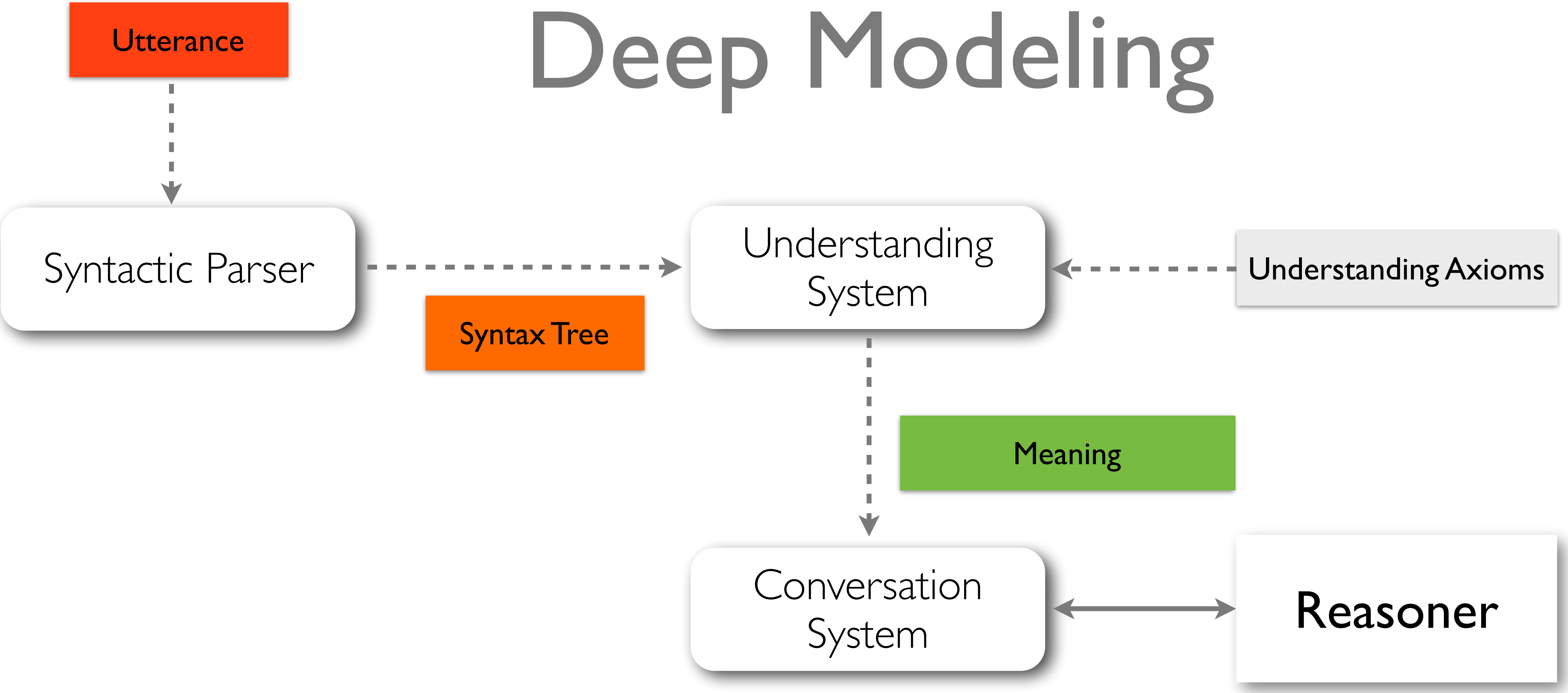
Understanding Axioms

Meaning

Conversation System

Reasoner

Deep Modeling



Utterance

Syntactic Parser

Syntax Tree

Understanding System

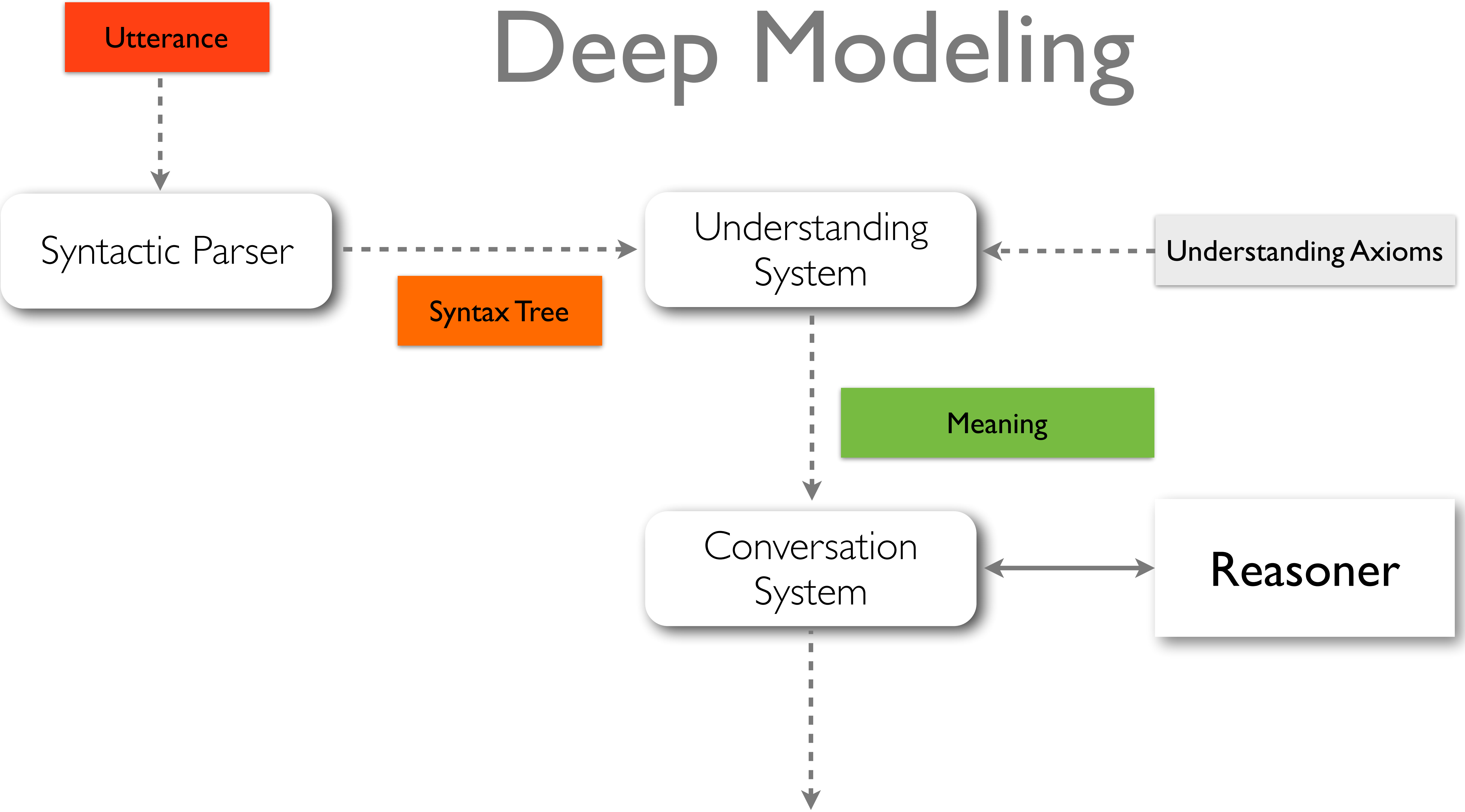
Understanding Axioms

Meaning

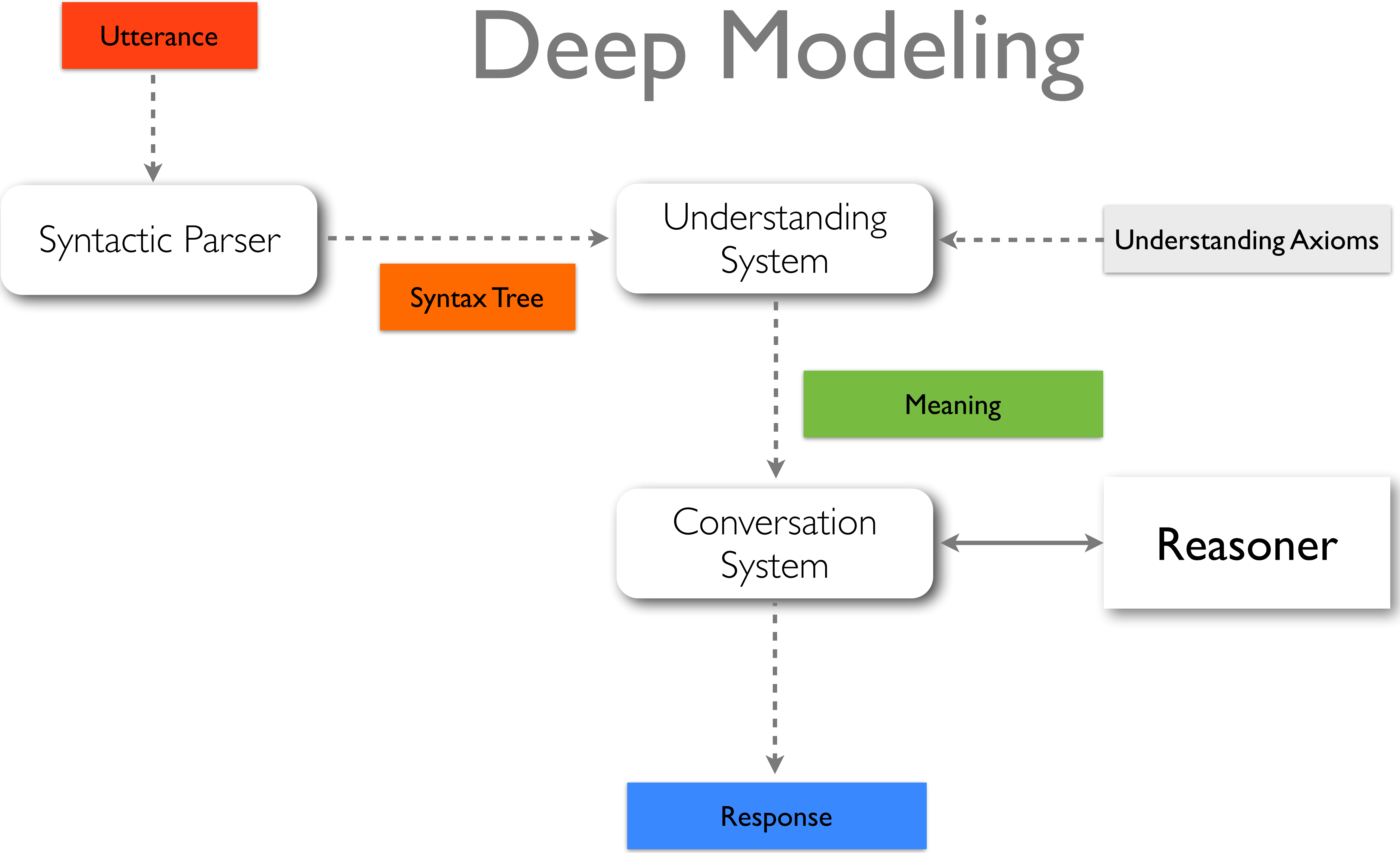
Conversation System

Reasoner

Deep Modeling



Deep Modeling



Controlled English

Controlled English

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

Controlled English

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

RLCNL: RAIR Lab Controlled Natural Language

Controlled English

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

RLCNL: RAIR Lab Controlled Natural Language

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

Controlled English

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

RLCNL: RAIR Lab Controlled Natural Language

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

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

Controlled English

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

RLCNL: RAIR Lab Controlled Natural Language

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

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

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

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.

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.

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

Controlled English

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

RLCNL: RAIR Lab Controlled Natural Language

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

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

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

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.

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

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.

Controlled English

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

RLCNL: RAIR Lab Controlled Natural Language

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

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

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

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.

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

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/

A Construction Manual for Robot's Ethical Systems: Requirements, Methods, Implementations

Edited by Robert Trappl

Contents

A Construction Manual for Robot’s Ethical Systems: Requirements, Methods, Implementations

Edited by Robert Trappl

Contents

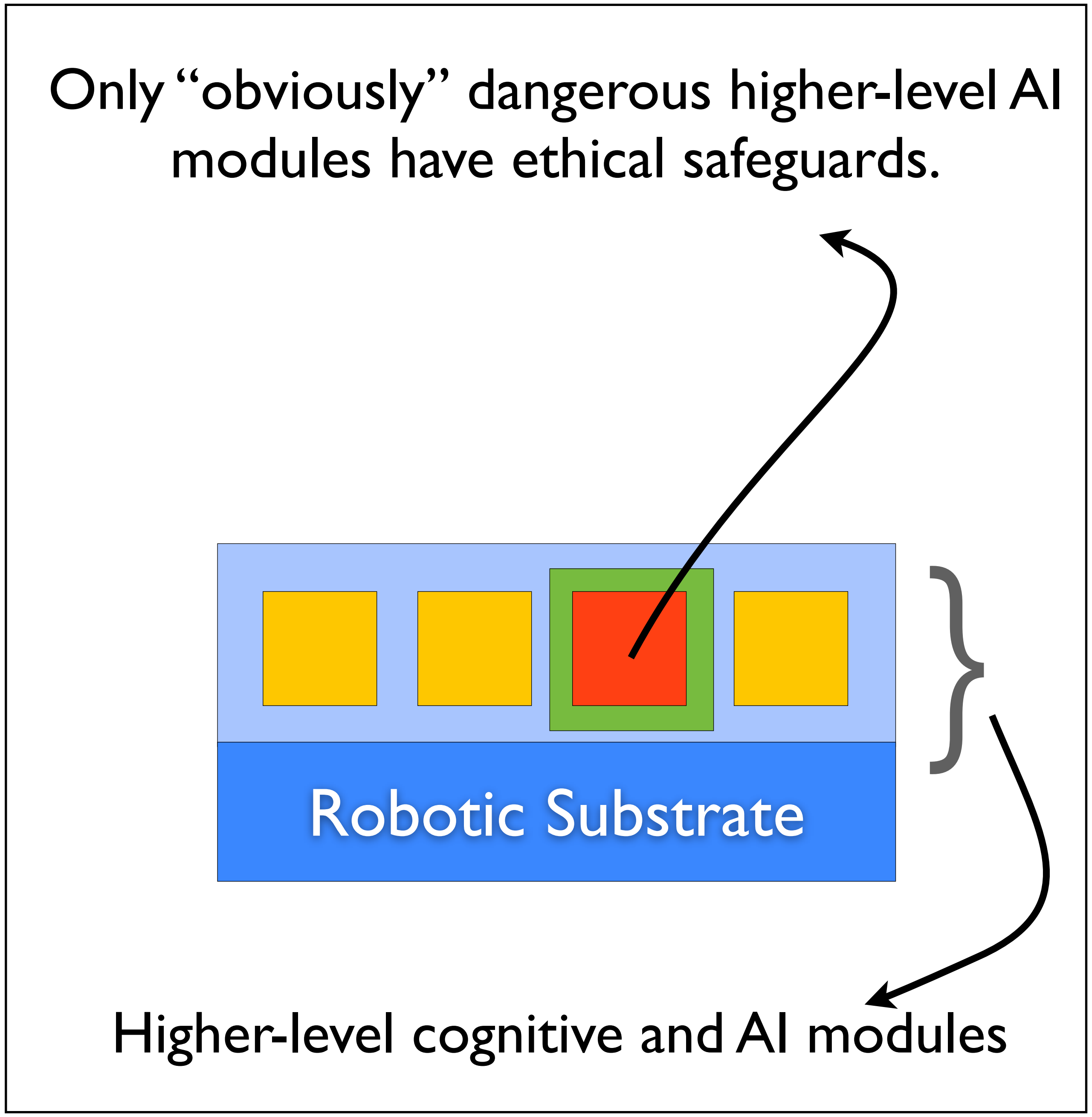
PART I: REQUIREMENTS

PART II: METHODS

PART III: IMPLEMENTATIONS

Most likely future — now:

Most likely future — now:



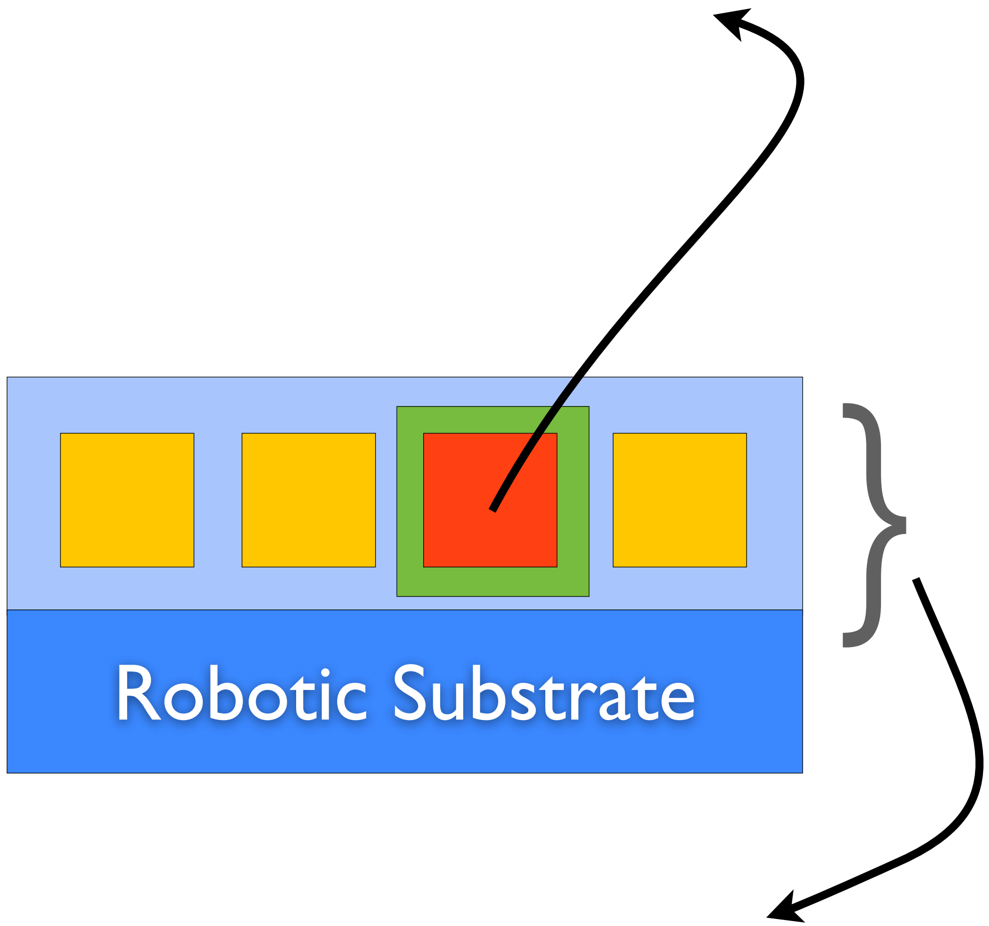
Only "obviously" dangerous higher-level AI modules have ethical safeguards.

Robotic Substrate

Higher-level cognitive and AI modules

Most likely future — now:

Only "obviously" dangerous higher-level AI modules have ethical safeguards.



Robotic Substrate

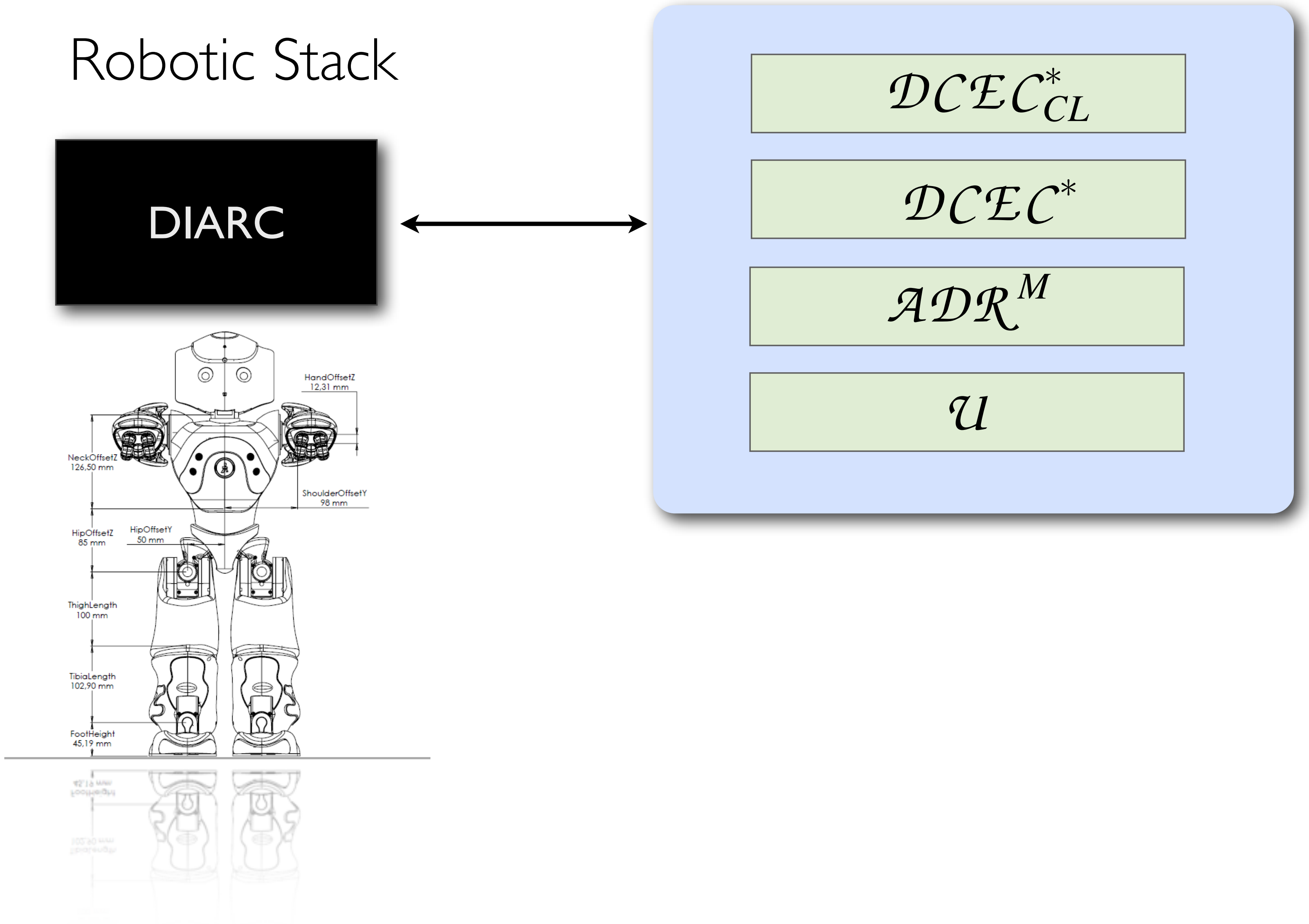
Higher-level cognitive and AI modules

"Ethical Regulation of Robots is Not Optional:

Ethical Reasoning Must be Embedded in Robot Operating Systems"

Moral/Ethical Stack

Robotic Stack



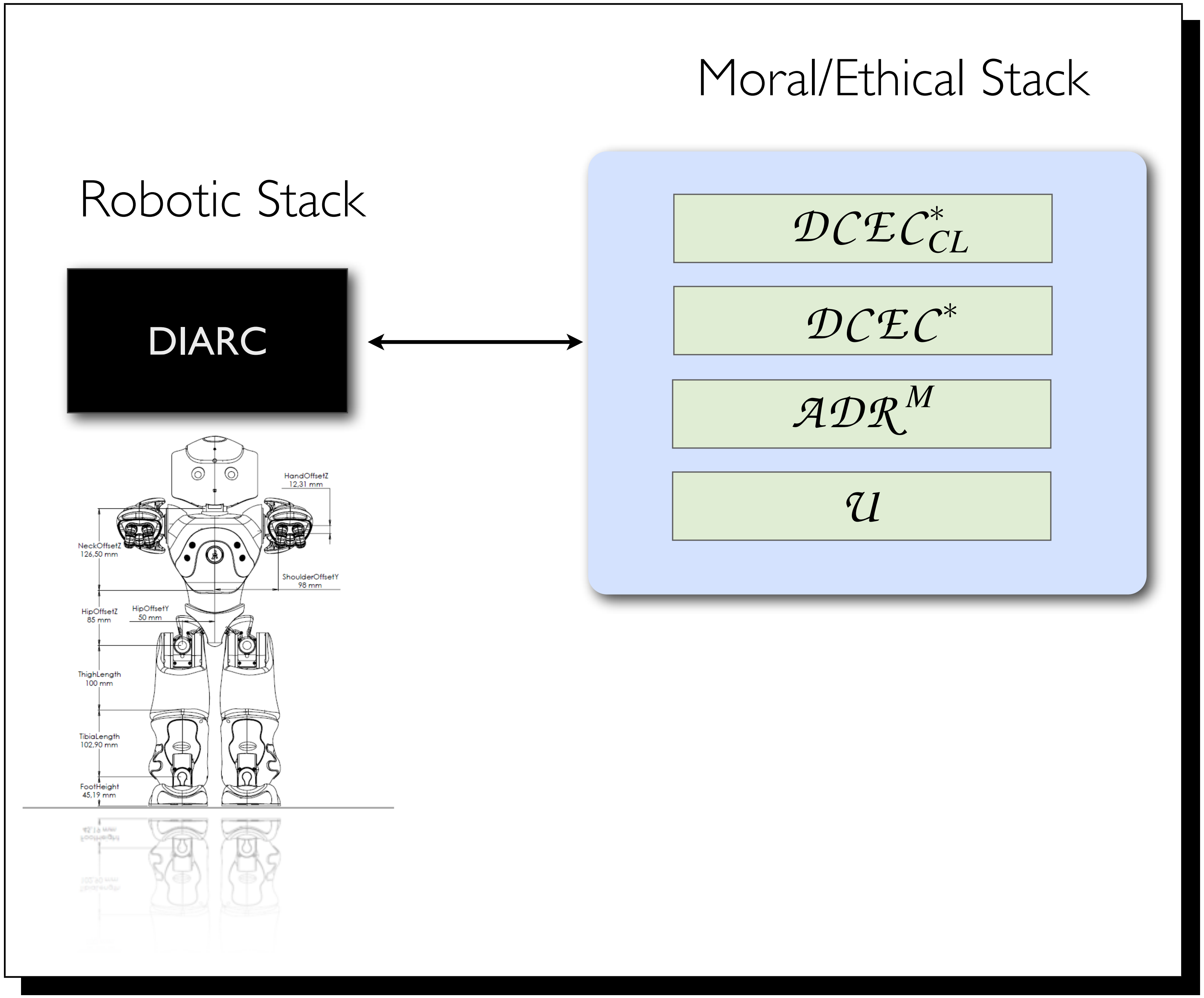
DIARC

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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$



Moral/Ethical Stack

Robotic Stack

DIARC

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

$\mathcal{DCEC}^*$

$\mathcal{ADR}^M$

$\mathcal{U}$

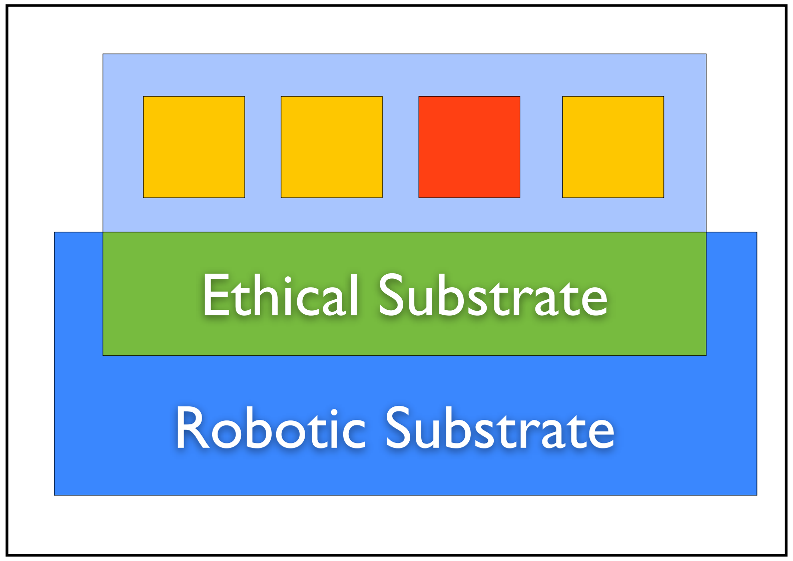
"Ethical Regulation of Robots is Not Optional:

Ethical Reasoning Must be Embedded in Robot Operating Systems"

- This situation not optimal. This leads to the "master requirement" proposed by us.

Ethical Substrate:

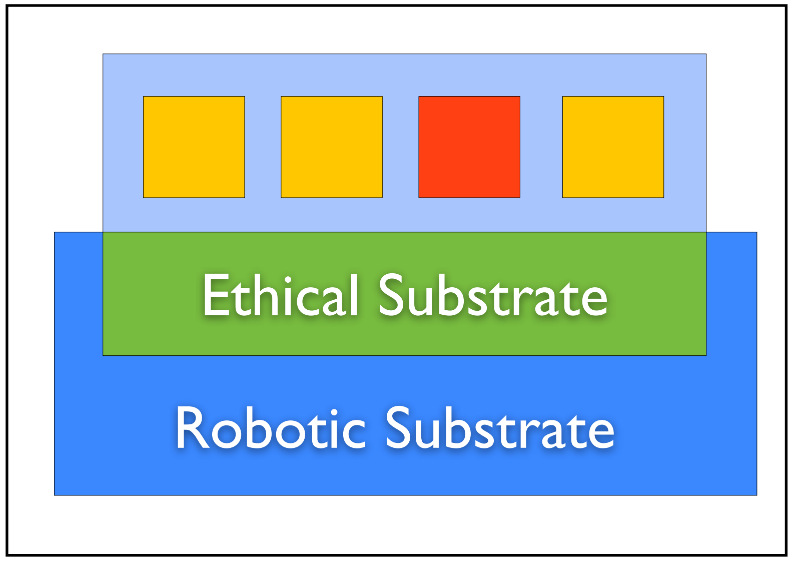
Every robot operating system should include an ethical substrate which sits between lower-level sensors and actuators and any higher-level cognitive system (whether or not that higher-level system itself is designed to enforce ethical regulation).



- This situation not optimal. This leads to the "master requirement" proposed by us.

Ethical Substrate:

Every robot operating system should include an ethical substrate which sits between lower-level sensors and actuators and any higher-level cognitive system (whether or not that higher-level system itself is designed to enforce ethical regulation).



"Ethical Regulation of Robots is Not Optional:

Ethical Reasoning Must be Embedded in Robot Operating Systems"