AI, Consciousness, & Lambda (Λ)

Selmer Bringsjord

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

IFLAI2
Oct 4 2021

RAIR
Rensselaer AI and Reasoning Lab

AI, Consciousness, & Lambda (Λ)

Selmer Bringsjord

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

IFLAI2
Oct 4 2021

AI, Consciousness, & Lambda (Λ)

Selmer Bringsjord

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

IFLAI2
Oct 4 2021

RAIR
Rensselaer AI and Reasoning Lab

“Consciousness”

“Consciousness”

“Consciousness”

‘Access Consciousness’

“Consciousness”

‘Access Consciousness’

“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.

“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.

Φ

“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

“Consciousness”

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

“Consciousness”

Cognitive Consciousness

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

"Consciousness"

Cognitive Consciousness

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and 'information' carries no information (!).

“Consciousness”

Cognitive Consciousness

HLC-Consciousness

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

“Consciousness”

Cognitive Consciousness

HLC-Consciousness

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

“Consciousness”

Cognitive Consciousness

HLC-Consciousness

HL**M**C-Consciousness

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

“Consciousness”

Cognitive Consciousness Λ

HLC-Consciousness

HL**M**C-Consciousness

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and ‘information’ carries no information (!).

This can be viewed as a formal framework for measuring the degree of "great computational intelligence" in an AI.

“Consciousness”

Cognitive Consciousness

Λ

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and 'information' carries no information (!).

HLC-Consciousness

HL**M**C-Consciousness

This can be viewed as a formal framework for measuring the degree of "great computational intelligence" in an AI.

"Consciousness"

Cognitive Consciousness Λ

HLC-Consciousness

HLMC-Consciousness

Happily, not bound by local biology; will cover aliens, God, characters of fiction, etc; and 'information' is information.

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and 'information' carries no information (!).

This can be viewed as a formal framework for measuring the degree of "great computational intelligence" in an AI.

"Consciousness"

Cognitive Consciousness Λ

HLC-Consciousness

HL**M**C-Consciousness

Happily, not bound by local biology; will cover aliens, God, characters of fiction, etc; and 'information' is information.

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and 'information' carries no information (!).

This can be viewed as a formal framework for measuring the degree of "great computational intelligence" in an AI.

"Consciousness"

Cognitive Consciousness Λ

Phenomenal Consciousness

Third-person formalization impossible.

Φ

Alas, biologically myopic & terrestrio-centric; and 'information' carries no information (!).

HLC-Consciousness

HL**M**C-Consciousness

Happily, not bound by local biology; will cover aliens, God, characters of fiction, etc; and 'information' is information.

High- Λ Machines are the ones DoD Needs to Worry About …

Basic Idea, Intuitively Put

The level of (cognitive) intelligence/consciousness of an AI at a time is a list of tuples (= matrix) giving eg the size of logical depth of (at least) five measures for each cognitive operator (i.e. for **K**, **B**, **P**, …).

$$\langle [\![\mathbf{K}, 1]\!], [\![\mathbf{K}, 2]\!], \ldots, [\![\mathbf{K}, 5]\!], \ldots \rangle$$

Basic Idea, Intuitively Put

The level of (cognitive) intelligence/consciousness of an AI at a time is a list of tuples (= matrix) giving eg the size of logical depth of (at least) five measures for each cognitive operator (i.e. for **K**, **B**, **P**, …).

$$\langle [\![\mathbf{K}, 1]\!], [\![\mathbf{K}, 2]\!], \ldots, [\![\mathbf{K}, 5]\!], \ldots \rangle$$

depth of knowledge

size of supporting proof/argument

depth of quantification within outermost knowledge operator

Cogito Ergo Sum

```
{:name        "Cogito Ergo Sum"
 :description "A formaliztion of Descartes' Cogito Ergo Sum"
 :assumptions {

              S1 (Believes! I (forall [x] (or (Name x) (Thing x))))
              S2 (Believes! I (forall (x) (iff (Name x) (not (Thing x)))) )
              S3 (Believes! I (forall (x) (if (Thing x) (or (Real x) (Fictional x)))))
              S4 (Believes! I (forall (x) (if (Thing x) (iff (Real x) (not (Fictional x))))))
              ;;;
              A1 (Believes! I (forall (x) (if (Name x) (Thing (* x)))))
              A2 (Believes! I (forall (y) (if (Name y) (iff  (DeReExists y) (exists x (and (Real x) (= x (* y))))))))

              ;;;
              ;
              Suppose (Believes! I (not (DeReExists I)))
              given (Believes! I (Name I))

              ;;;
              Perceive-the-belief (Believes! I (Perceives! I (Believes! I (not (DeReExists I)))))
              If_P_B (Believes!
                      I
                      (forall [?agent]
                              (if (Perceives! I (Believes! ?agent (not (DeReExists ?agent))))
                                (Real (* ?agent))))))

              }
 :goal        (and (Believes! I  (not (Real (* I))))
                   (Believes! I  (Real (* I)) ))

 }
```

1.7 seconds

Cogito Ergo Sum

Λ_{t_1}

```
{:name        "Cogito Ergo Sum"
 :description "A formaliztion of Descartes' Cogito Ergo Sum"
 :assumptions {

             S1 (Believes! I (forall [x] (or (Name x) (Thing x))))
             S2 (Believes! I (forall (x) (iff (Name x) (not (Thing x)))) )
             S3 (Believes! I (forall (x) (if (Thing x) (or (Real x) (Fictional x)))))
             S4 (Believes! I (forall (x) (if (Thing x) (iff (Real x) (not (Fictional x))))))
             ;;;
             A1 (Believes! I (forall (x) (if (Name x) (Thing (* x)))))
             A2 (Believes! I (forall (y) (if (Name y) (iff  (DeReExists y) (exists x (and (Real x) (= x (* y))))))))

             ;;;
             ;
             Suppose (Believes! I (not (DeReExists I)))
             given (Believes! I (Name I))

             ;;;
             Perceive-the-belief (Believes! I (Perceives! I (Believes! I (not (DeReExists I)))))
             If_P_B (Believes!
                     I
                     (forall [?agent]
                             (if (Perceives! I (Believes! ?agent (not (DeReExists ?agent))))
                               (Real (* ?agent)))))

             }
 :goal       (and (Believes! I  (not (Real (* I))))
                  (Believes! I  (Real (* I)) ))

 }
```

1.7 seconds

Cogito Ergo Sum

Λ_{t_1}

```
{:name         "Cogito Ergo Sum"
 :description "A formaliztion of Descartes' Cogito Ergo Sum"
 :assumptions {

              S1 (Believes! I (forall [x] (or (Name x) (Thing x))))
              S2 (Believes! I (forall (x) (iff (Name x) (not (Thing x)))) )
              S3 (Believes! I (forall (x) (if (Thing x) (or (Real x) (Fictional x)))))
              S4 (Believes! I (forall (x) (if (Thing x) (iff (Real x) (not (Fictional x))))))
              ;;;
              A1 (Believes! I (forall (x) (if (Name x) (Thing (* x)))))
              A2 (Believes! I (forall (y) (if (Name y) (iff  (DeReExists y) (exists x (and (Real x) (= x (* y))))))))

              ;;;
              ;
              Suppose (Believes! I (not (DeReExists I)))
              given (Believes! I (Name I))

              ;;;
              Perceive-the-belief (Believes! I (Perceives! I (Believes! I (not (DeReExists I)))))
              If_P_B (Believes!
                      I
                      (forall [?agent]
                              (if (Perceives! I (Believes! ?agent (not (DeReExists ?agent))))
                                (Real (* ?agent)))))
              }
 :goal        (and (Believes! I  (not (Real (* I))))
                   (Believes! I  (Real (* I)) ))

 }
```

absurd belief

1.7 seconds

Cogito Ergo Sum

Λ_{t_1}

```
{:name        "Cogito Ergo Sum"
 :description "A formaliztion of Descartes' Cogito Ergo Sum"
 :assumptions {

             S1 (Believes! I (forall [x] (or (Name x) (Thing x))))
             S2 (Believes! I (forall (x) (iff (Name x) (not (Thing x)))) )
             S3 (Believes! I (forall (x) (if (Thing x) (or (Real x) (Fictional x)))))
             S4 (Believes! I (forall (x) (if (Thing x) (iff (Real x) (not (Fictional x))))))
             ;;;
             A1 (Believes! I (forall (x) (if (Name x) (Thing (* x)))))
             A2 (Believes! I (forall (y) (if (Name y) (iff  (DeReExists y) (exists x (and (Real x) (= x (* y))))))))

             ;;;
             ;
             Suppose (Believes! I (not (DeReExists I)))
             given (Believes! I (Name I))

             ;;;
             Perceive-the-belief (Believes! I (Perceives! I (Believes! I (not (DeReExists I)))))
             If_P_B (Believes!
                      I
                      (forall [?agent]
                              (if (Perceives! I (Believes! ?agent (not (DeReExists ?agent))))
                                 (Real (* ?agent)))))

             }
 :goal       (and (Believes! I  (not (Real (* I))))
                  (Believes! I  (Real (* I)) ))

 }
```

absurd belief

1.7 seconds

Λ_{t_k}

I. Elements of Λ

Intensional Complexity of representations/formulae

For top level beliefs, knowledge, intensions, desires etc

$\Lambda[\mathbf{B},1]$ Maximum intensional depth of beliefs

$\Lambda[\mathbf{D},1]$ Maximum intensional depth of desires

$\Lambda[\mathbf{I}, 1]$ Maximum intensional depth of intentions

$\vdots$

II. Elements of Λ

Quantificational Complexity of representations/formulae

For top level beliefs, knowledge, intensions, desires etc

$\Lambda[\mathbf{B}, 2]$ Maximum quantificational depth of beliefs

$\Lambda[\mathbf{D}, 2]$ Maximum quantificational depth of desires

$\Lambda[\mathbf{I}, 2]$ Maximum quantificational depth of intentions

⋮

III. Elements of Λ

Extensional Complexity of representations/formulae

For top level beliefs, knowledge, intensions, desires etc

$\Lambda[\mathbf{B}, 3]$	Maximum extensional depth of beliefs
$\Lambda[\mathbf{D}, 3]$	Maximum extensional depth of desires
$\Lambda[\mathbf{I}, 3]$	Maximum extensional depth of intentions

$\vdots$

IV. Elements of Λ

Time Complexity of representations/formulae

For top level beliefs, knowledge, intensions, desires etc

$\Lambda[\mathbf{B}, 4]$ Maximum difference between time expressions within beliefs

$\Lambda[\mathbf{D}, 4]$ Maximum difference between time expressions within desires

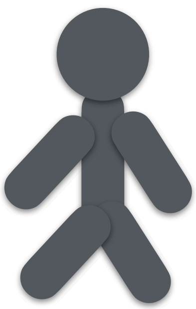
$\Lambda[\mathbf{I}, 4]$ Maximum difference between time expressions within intentions

$\vdots$

Note: If a time variable **t** is universally quantified, we take it as **∞**.

Example

The Doctrine of Double Effect



C_1 the action is not forbidden (where we assume an ethical hierarchy such as the one given by Bringsjord [2017], and require that the action be neutral or above neutral in such a hierarchy);

C_2 The net utility or goodness of the action is greater than some positive amount γ;

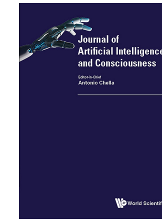
C_{3a} the agent performing the action intends only the good effects;

C_{3b} the agent does not intend any of the bad effects;

C_4 the bad effects are not used as a means to obtain the good effects; and

C_5 if there are bad effects, the agent would rather the situation be different and the agent not have to perform the action. That is, the action is unavoidable.

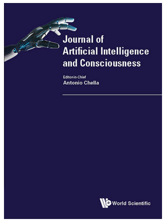
Journal of Artificial Intelligence and Consciousness | Vol. 07, No. 02, pp. 155-181 (2020)



The Theory of Cognitive Consciousness, and Λ (Lambda)

Selmer Bringsjord ✉ and G. Naveen Sundar

Journal of Artificial Intelligence and Consciousness | Vol. 07, No. 02, pp. 155-181 (2020)



The Theory of Cognitive Consciousness, and Λ (Lambda)

Selmer Bringsjord and G. Naveen Sundar

Journal of Artificial Intelligence and Consciousness

The Theory of Cognitive Consciousness, and Λ (Lambda)*

Selmer Bringsjord
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
Selmer.Bringsjord@gmail.com

Naveen Sundar G.
Rensselaer AI & Reasoning (RAIR) Lab
Rensselaer Polytechnic Institute (RPI)
Troy NY 12180 USA
Naveen.Sundar.G@gmail.com

Received 7 February 2020
Revised ??? ??? ????

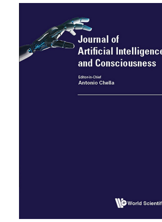
We provide an overview of the theory of cognitive consciousness (TCC), and of Λ; the latter provides a means of measuring the amount of cognitive consciousness present in a given cognizer, whether natural or artificial, at a given time, along a number of different dimensions. TCC and Λ stand in stark contrast to Tononi's Integrated information Theory (IIT) and Φ. We believe, for reasons we present, that the former pair is superior to the latter. TCC includes a formal axiomatic theory, $\mathcal{CA}$, the 12 axioms of which we present and briefly comment upon herein; no such formal theory accompanies IIT/Φ. TCC/Λ and IIT/Φ each offer radically different verdicts as to whether and to what degree AIs of yesterday, today, and tomorrow were/are/will be conscious. Another noteworthy difference between TCC/Λ and IIT/Φ is that the former enables the measurement of cognitive consciousness in those who have passed on, and in fictional characters; no such enablement is remotely possible for IIT/Φ. For instance, we apply Λ to measure the cognitive consciousness of: Descartes; the first fictional detective to be described on Earth (by Edgar Allen Poe), C. Auguste Dupin. We also apply Λ to compute the cognitive consciousness of an artificial agent able to make ethical decisions using the Doctrine of Double Effect.

Keywords: consciousness; cognitive consciousness; AI; Lambda/Λ.

*We are indebted to SRI International for support of a series of symposia on consciousness that proved to be the fertile ground in which which Λ's germination commenced, and to many co-participants in that series for stimulating debate and discussion, esp. — in connection with matters on hand herein — Giulio Tononi, Christof Koch, and Antonio Chella.

Journal of Artificial Intelligence and Consciousness | Vol. 07, No. 02, pp. 155-181 (2020)

The Theory of Cognitive Consciousness, and Λ (Lambda)



Extending Measures from $\mathcal{L}^0$ to $\mathcal{L}$

$$\mu_\omega(\phi) = \begin{cases} \mu(\phi) & \text{if } \phi \in \mathcal{L}^0 \\ \max_\psi \mu_\omega(\psi) + 1 & \text{if } \phi \equiv \omega_i\,[a_1, t_1, \ldots \psi \ldots] \end{cases}$$

For example, let μ count the number of predicate symbols in a formula.

Example

$$\mu\,(Happy\,(john)) = 1$$
$$\mu_\omega\big(Happy\,(john)\big) = 1$$
$$\mu_\omega\Big(\mathbf{B}\big(mary, t_2, Happy(john)\big)\Big) = 2$$

For any agent a, we want to look at the new complexity the agent introduces that is above any input complexity. For this, we introduce $\Delta : 2^{\mathcal{L}} \times 2^{\mathcal{L}} \rightarrow 2^{\mathcal{L}}$ operator that computes differences between two sets of formulae. This can be simply the set-difference operator. For convenience, let $\omega_j[\Gamma]$ denote the subset of formulae with operators ω_j in Γ:

$$\omega_j[\Gamma] = \{\phi \mid \ \phi \equiv \omega_j[\ldots] \text{ and } \phi \in \Gamma \text{ or } \phi \text{ a subformula } \in \Gamma\}$$

Given a set of measures $\{\mu^0, \ldots, \mu^N\}$ and a set of modal (or cognitive) operators $\{\omega_0, \ldots, \omega_M\}$, we define Λ as a function mapping an agent at a time point to a matrix $\mathbb{N}^{M\times N}$:

$$\Lambda : A \times T \rightarrow \mathbb{N}^{M\times N}$$

Definition of Λ

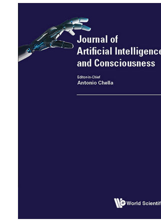
$$\Lambda(a,t)_{i,j} = \max_\phi \Big\{\mu^i(\phi) \mid \phi \in \Delta\Big(\omega_j\big[o(a,t)\big], \omega_j\big[i(a,t)\big]\Big)\Big\}$$

Example 2

Let us consider two modal operators $\{\mathbf{B}, \mathbf{D}\}$ and the following base measures μ^0 which measures quantificational complexity via Σ or Π measures, μ^1 which counts the total number of predicate symbols (not a count of unique predicate symbols), and μ^2 which counts the number of distinct time expressions. This gives $\Lambda : A \times T \rightarrow \mathbb{N}^{2\times 3}$. At some timepoint t, let an agent a have the following $\Delta(o(a,t), i(a,t)) = \{\mathbf{B}(\phi_1), \mathbf{D}(\phi_2)\}$

Journal of Artificial Intelligence and Consciousness | Vol. 07, No. 02, pp. 155-181 (2020)

The Theory of Cognitive Consciousness, and Λ (Lambda)



Extending Measures from $\mathcal{L}^0$ to $\mathcal{L}$

$$\mu_\omega(\phi) = \begin{cases} \mu(\phi) & \text{if } \phi \in \mathcal{L}^0 \\ \max_\psi \mu_\omega(\psi) + 1 & \text{if } \phi \equiv \omega_i\,[a_1, t_1, \ldots \psi \ldots] \end{cases}$$

For example, let μ count the number of predicate symbols in a formula.

Example

$$\mu\,(Happy\,(john)) = 1$$
$$\mu_\omega\big(Happy\,(john)\big) = 1$$
$$\mu_\omega\Big(\mathbf{B}\big(mary, t_2, Happy(john)\big)\Big) = 2$$

For any agent a, we want to look at the new complexity the agent introduces that is above any input complexity. For this, we introduce $\Delta : 2^{\mathcal{L}} \times 2^{\mathcal{L}} \rightarrow 2^{\mathcal{L}}$ operator that computes differences between two sets of formulae. This can be simply the set-difference operator. For convenience, let $\omega_j\,[\Gamma]$ denote the subset of formulae with operators ω_j in Γ:

$$\omega_j\,[\Gamma] = \{\phi \mid \ \phi \equiv \omega_j[\ldots] \text{ and } \phi \in \Gamma \text{ or } \phi \text{ a subformula } \in \Gamma\}$$

Given a set of measures $\{\mu^0, \ldots, \mu^N\}$ and a set of modal (or cognitive) operators $\{\omega_0, \ldots, \omega_M\}$, we define Λ as a function mapping an agent at a time point to a matrix $\mathbb{N}^{M\times N}$:

$$\Lambda : A \times T \rightarrow \mathbb{N}^{M\times N}$$

Definition of Λ

$$\Lambda(a,t)_{i,j} = \max_\phi \left\{\mu^i(\phi) \mid \phi \in \Delta\Big(\omega_j\big[o(a,t)\big], \omega_j\big[i(a,t)\big]\Big)\right\}$$

Example 2

Let us consider two modal operators $\{\mathbf{B}, \mathbf{D}\}$ and the following base measures μ^0 which measures quantificational complexity via Σ or Π measures, μ^1 which counts the total number of predicate symbols (not a count of unique predicate symbols), and μ^2 which counts the number of distinct time expressions. This gives $\Lambda : A \times T \rightarrow \mathbb{N}^{2\times 3}$. At some timepoint t, let an agent a have the following $\Delta(o(a,t), i(a,t)) = \{\mathbf{B}(\phi_1), \mathbf{D}(\phi_2)\}$

$$\phi_1 \equiv \neg\forall a : Happy(a,t); \qquad \phi_2 \equiv \forall b : \neg Hungry(b,t) \rightarrow Happy(b,t)$$

Applying the measures:

$$\mu^o(\phi_1) = 1, \mu^1(\phi_1) = 1; \mu^2(\phi_1) = 1$$
$$\mu^o(\phi_2) = 1; \mu^1(\phi_2) = 2; \mu^2(\phi_2) = 1$$

Giving us:

$$\Lambda(a,t) = \begin{pmatrix} 1\,1\,1 \\ 1\,2\,1 \end{pmatrix}$$

6.1. *Some Distinctive Properties of* Λ *(vs.* Φ*)*

Here are some properties of the Λ framework of potential interest to our readers:

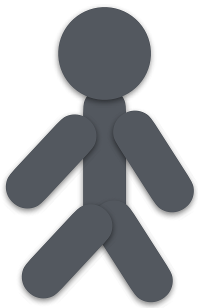
Non-Binary Whereas Φ is such that an agent either is or is not (P-) conscious, cognitive consciousness as measured by Λ admits of a fine-grained range of the *degree* of cognitive consciousness.

Zero Λ for Some Animals and Machines Animals such as insects, and computing machines that are end-to-end statistical/connectionist "ML," have zero Λ, and hence cannot be cognitively conscious. In contrast, as emphasized to Bringsjord in personal conversation,[6] Φ says that even lower animals are conscious.

Human-Nonhuman Discontinuity Explained by Λ From the computational/AI point of view, cognitive scientists have taken note of a severe discontinuity between *H. sapiens sapiens* and other biological creatures on Earth [Penn *et al.*, 2008], and the sudden and large jump in level of Λ from (say) chimpanzees and dolphins to humans is in line with this observation. It's for instance doubtful that any nonhuman animals are capable of reaching third-order belief; hence $\Lambda[\mathbf{B}, 0] = n$, where $n \geq 3$, for any nonhuman animal, is impossible. In stark contrast, each of us believes that you, the reader, believe that we believe that San Francisco is located in California.

Human-Human Discontinuity Explained by Λ A given neurobiologically normal human, over the course of his or her lifetime, has very different cognitive capacity. E.g., it's well-known that such a human, before the age of four or five, is highly unlikely to be able to solve what has become known as the *false-belief task* (or sometimes the *sally-anne task*), which we denote by 'FBT.' From the point of view of Λ, the explanation is simply that an agent with insufficiently high cognitive consciousness is incapable of solving such a task; specifically, solving FBT requires an agent to have

[6] With Tononi and C. Koch, SRI T&C Series.



Formal Conditions for $\mathcal{DDE}$

$\mathbf{F_1}$ α carried out at t is not forbidden. That is:

$$\Gamma \nvdash \neg \mathbf{O}\Big(a, t, \sigma, \neg happens\big(action(a, \alpha), t\big)\Big)$$

$\mathbf{F_2}$ The net utility is greater than a given positive real γ:

$$\Gamma \vdash \sum_{y=t+1}^{H} \left(\sum_{f \in \alpha_I^{a,t}} \mu(f, y) - \sum_{f \in \alpha_T^{a,t}} \mu(f, y) \right) > \gamma$$

$\mathbf{F_{3a}}$ The agent a intends at least one good effect. ($\mathbf{F_2}$ should still hold after removing all other good effects.) There is at least one fluent f_g in $\alpha_I^{a,t}$ with $\mu(f_g, y) > 0$, or f_b in $\alpha_T^{a,t}$ with $\mu(f_b, y) < 0$, and some y with $t < y \leq H$ such that the following holds:

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

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

$$\Gamma \nvdash \mathbf{I}\Big(a, t, Holds\big(f_b, y\big)\Big) \text{ and}$$

$$\Gamma \nvdash \mathbf{I}\Big(a, t, \neg Holds\big(f_g, y\big)\Big)$$

$\mathbf{F_4}$ The harmful effects don't cause the good effects. Four permutations, paralleling the definition of $\rhd$ above, hold here. One such permutation is shown below. For any bad fluent f_b holding at t_1, and any good fluent f_g holding at some t_2, such that $t < t_1, t_2 \leq H$, the following holds:

$$\Gamma \vdash \neg \rhd \Big(Holds\big(f_b, t_1\big), Holds\big(f_g, t_2\big)\Big)$$

Example from Sim in IJCAI Paper

looking at one single chunk

$$\left\{\begin{array}{l}\mathbf{K}\left(I, now, \sigma_{trolley}\right),\\ \mathbf{B}\left(I, now, \mathbf{O}\left(\begin{array}{l}I, now, \sigma_{trolley},\\ \left[\begin{array}{c}\neg\exists t : \text{Moment } Holds\left(dead\left(P_1, t\right)\right)\\ \wedge\\ \neg\exists t : \text{Moment } Holds\left(dead\left(P_2, t\right)\right)\end{array}\right]\end{array}\right)\right),\\ \mathbf{O}\left(I, now, \sigma_{trolley}, \left[\begin{array}{l}\neg\exists t : \text{Moment } Holds\left(dead\left(P_1, t\right)\right) \wedge\\ \neg\exists t : \text{Moment } Holds\left(dead\left(P_2, t\right)\right)\end{array}\right]\right)\end{array}\right\}$$

$$\vdash \mathbf{I}\left(I, now, \left[\begin{array}{l}\neg\exists t : \text{Moment } Holds\left(dead\left(P_1, t\right)\right) \wedge\\ \neg\exists t : \text{Moment } Holds\left(dead\left(P_2, t\right)\right)\end{array}\right]\right)$$

$\Lambda[\mathbf{B}, 1] = 2$

$\Lambda[\mathbf{B}, 2] = 1$

$\Lambda[\mathbf{K}, 1] = 1$

$\Lambda[\mathbf{O}, 1] = 1$

$\Lambda[\mathbf{O}, 1] = 1$

$\Lambda[\mathbf{I}, 1] = 1$

$\Lambda[\mathbf{I}, 2] = 1$

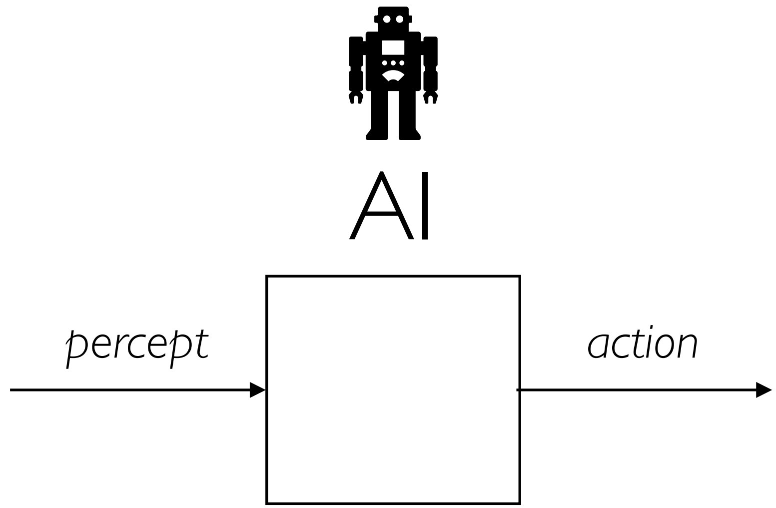
$\Lambda[\mathbf{B}, 3] = 1$

$\Lambda[\mathbf{B}, 4] = \infty$

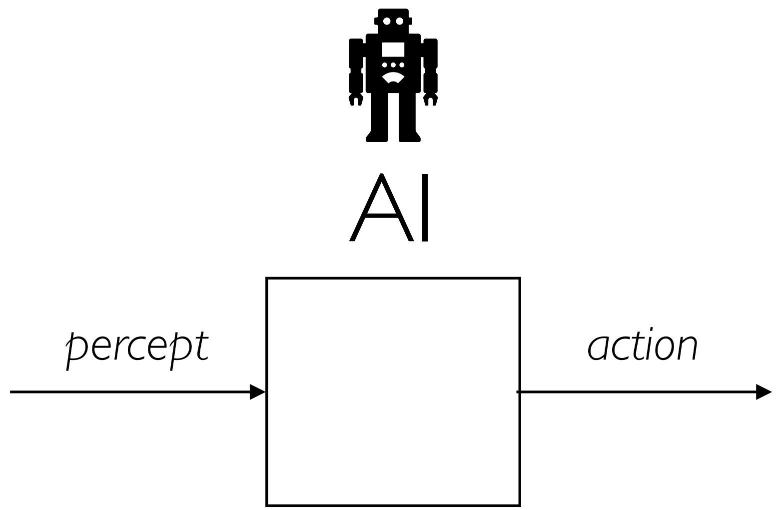
⋮

The application of Λ to eg "Deep Learning" machines implies that they have zero cognitive intelligence/ cognitive consciousness.

AI:MLn

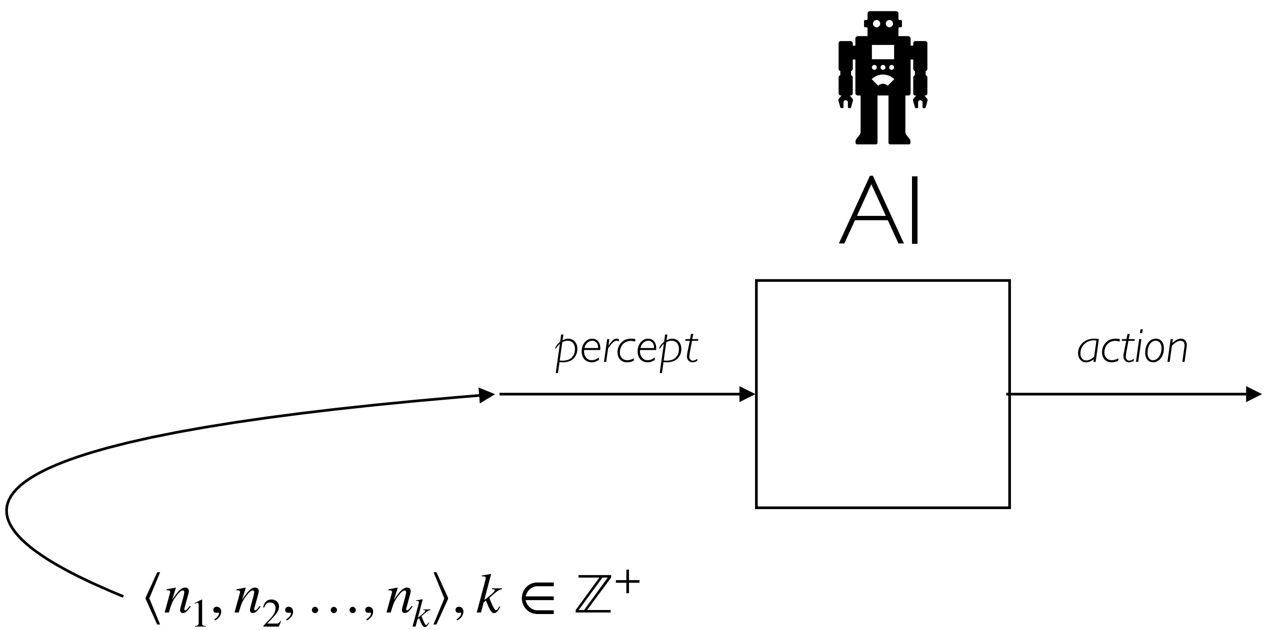


AI:MLn



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

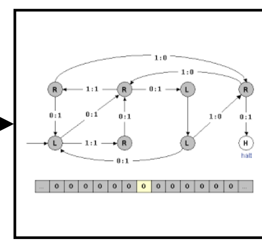
AI:MLn



AI:MLn

AI

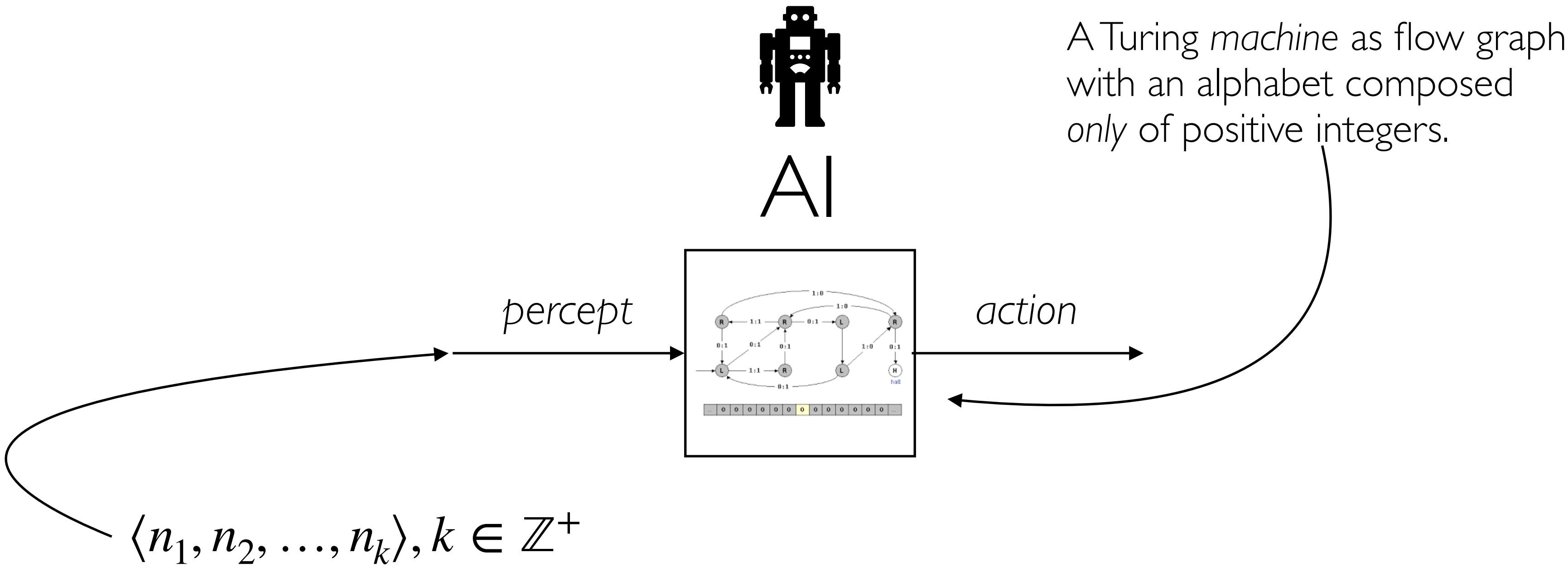
percept



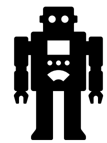
action

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

AI:MLn


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

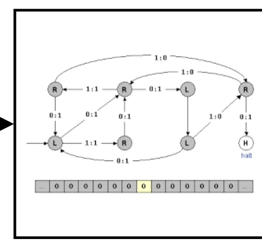
AI:MLn



AI

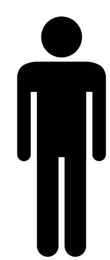
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.

percept

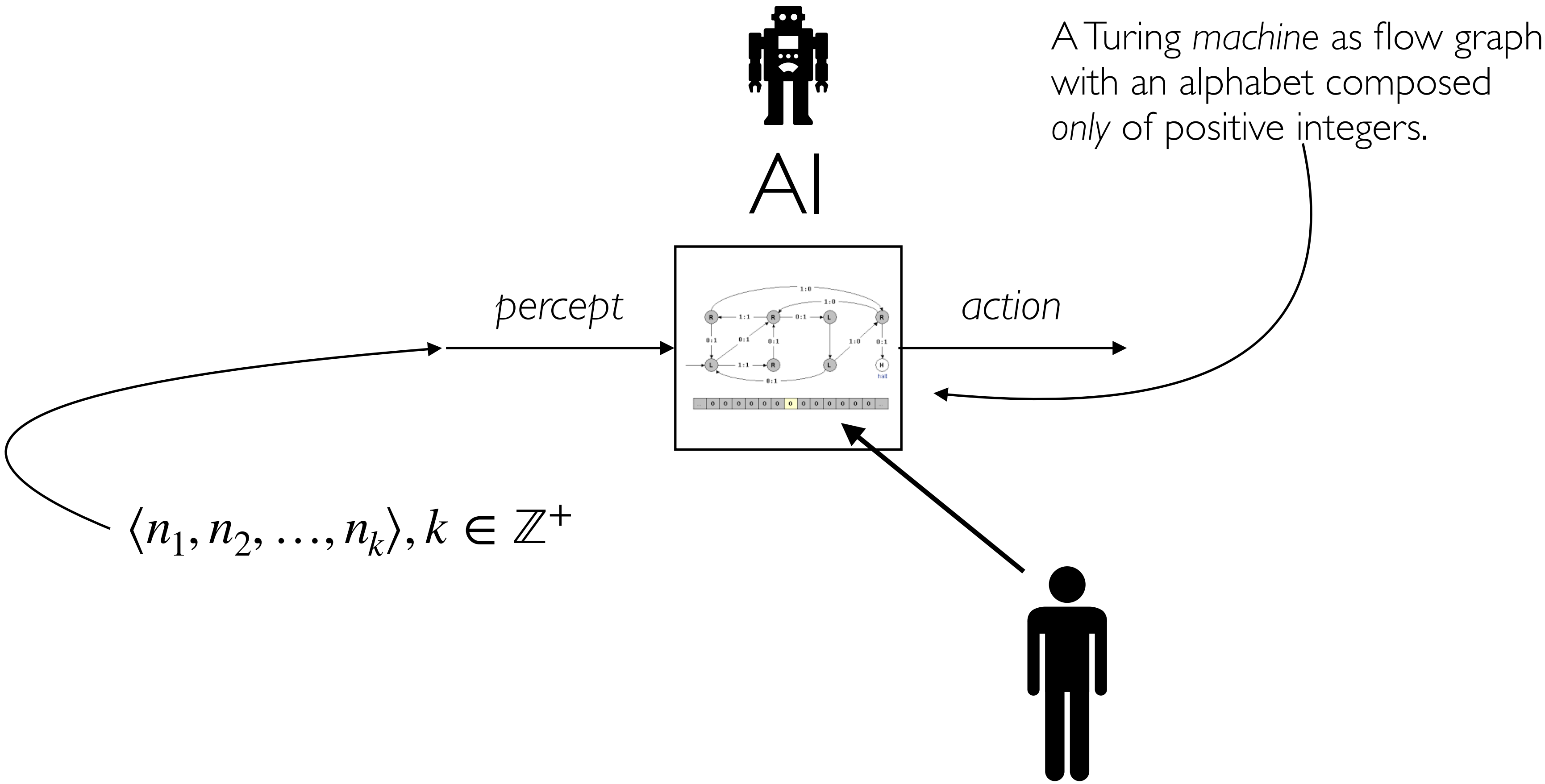


action

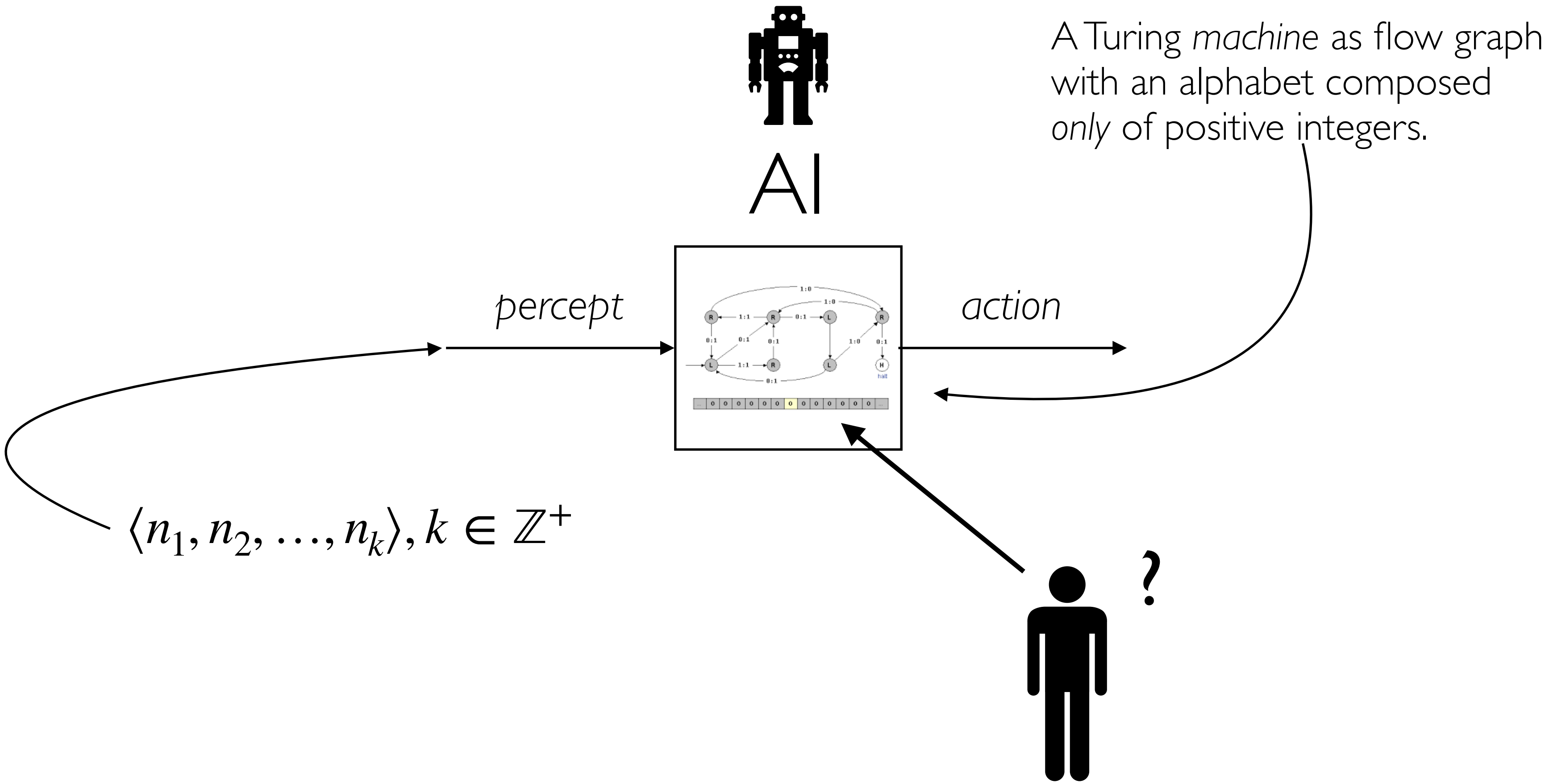
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$



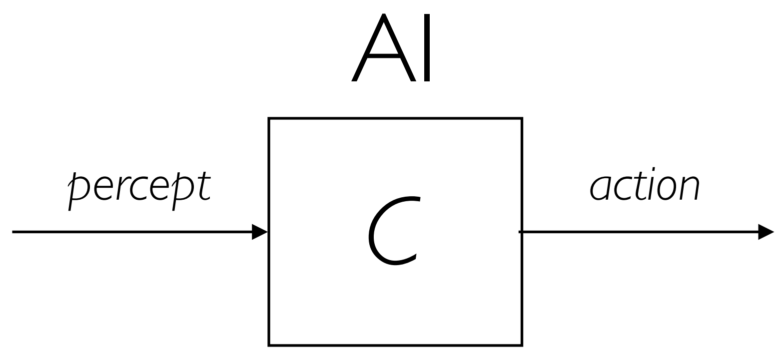
AI:MLn

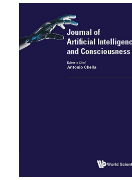

AI
percept
action
A Turing *machine* as flow graph,
with an alphabet composed
only of positive integers.
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$

AI:MLn


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

We will be able to measure the intelligence of *any* AI, not with *g*-loaded tests of intelligence, but with Λ-loaded tests of machine intelligence, in keeping with Psychometric AI.





CA: 11 Axioms (Initially)

- Plan
- P2B
- K2B
- Intro
- Incorr
- Ess
- ¬CompE
- Irr
- Free
- CCaus
- TheI

$\mathcal{CA}$: 11 Axioms (Initially)

Plan

P2B

K2B $\forall a[\mathbf{K}_a \phi \rightarrow (\mathbf{B}_a \phi \wedge \mathbf{B}_a \exists \Phi \exists \alpha (\Phi \rightsquigarrow_{\alpha/\pi} \phi)]$

Intro

Incorr

Ess

¬CompE

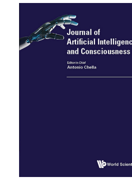
Irr

Free

CCaus

TheI

$\mathcal{CA}$: 11 Axioms (Initially)



Plan

P2B

$\mu\mathcal{DCEC}_3^*$ K2B $\forall a[\mathbf{K}_a\phi \rightarrow (\mathbf{B}_a\phi \wedge \mathbf{B}_a\exists\Phi\exists\alpha(\Phi \leadsto_{\alpha/\pi} \phi)]$

Intro

Incorr

Ess

¬CompE

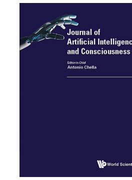
Irr

Free

CCaus

TheI

$\mathcal{CA}$: 11 Axioms (Initially)



Plan

P2B

K2B $\forall a[\mathbf{K}_a\phi \rightarrow (\mathbf{B}_a\phi \wedge \mathbf{B}_a\exists\Phi\exists\alpha(\Phi \rightsquigarrow_{\alpha/\pi} \phi)]$

Intro

Incorr

Ess

¬CompE

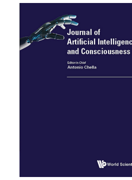
Irr

Free

CCaus

TheI

$\mathcal{CA}$: 11 Axioms (Initially)



Plan

P2B

K2B $\forall a[\mathbf{K}_a\phi \rightarrow (\mathbf{B}_a\phi \wedge \mathbf{B}_a\exists\Phi\exists\alpha(\Phi \rightsquigarrow_{\alpha/\pi} \phi)]$

Intro

Incorr $\forall a\forall t\forall F[(F\,is\ contingent\ \wedge F \in C'') \rightarrow (\Box\mathbf{B}(a,t,Fa) \rightarrow Fa)]$

Ess

¬CompE

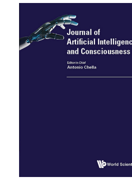
Irr

Free

CCaus

TheI

$\mathcal{CA}$: 11 Axioms (Initially)



Plan

P2B

K2B $\forall a[\mathbf{K}_a\phi \rightarrow (\mathbf{B}_a\phi \wedge \mathbf{B}_a\exists\Phi\exists\alpha(\Phi \leadsto_{\alpha/\pi} \phi))]$

Intro

Incorr $\forall a\forall t\forall F[(F is\ contingent\ \wedge F \in C'') \rightarrow (\Box\mathbf{B}(a,t,Fa) \rightarrow Fa)]$

Ess

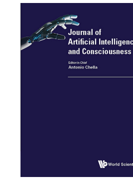
¬CompE

Irr

Free

CCaus $\mathbf{C}\ \mathcal{EC}$

TheI



$\mathcal{CA}$: 11 Axioms (Initially)

Plan

P2B

K2B $\forall a[\mathbf{K}_a\phi \rightarrow (\mathbf{B}_a\phi \wedge \mathbf{B}_a \exists \Phi \exists \alpha (\Phi \rightsquigarrow_{\alpha/\pi} \phi)]$

Intro

Incorr $\forall a \forall t \forall F[(F \mathit{is\ contingent} \ \wedge F \in C'') \rightarrow (\Box \mathbf{B}(a,t,Fa) \rightarrow Fa)]$

Ess

¬CompE

Irr

Free

CCaus $\mathbf{C}\ \mathcal{EC}$

TheI

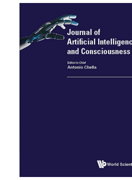
$[A_1]$ $\mathbf{C}(\forall f,t \ . \ \mathit{initially}(f) \wedge \neg \mathit{clipped}(0,f,t) \Rightarrow \mathit{holds}(f,t))$

$[A_2]$ $\mathbf{C}(\forall e,f,t_1,t_2 \ . \ \mathit{happens}(e,t_1) \wedge \mathit{initiates}(e,f,t_1) \wedge t_1 < t_2 \wedge \neg \mathit{clipped}(t_1,f,t_2) \Rightarrow \mathit{holds}(f,t_2))$

$[A_3]$ $\mathbf{C}(\forall t_1,f,t_2 \ . \ \mathit{clipped}(t_1,f,t_2) \Leftrightarrow [\exists e,t \ . \ \mathit{happens}(e,t) \wedge t_1 < t < t_2 \wedge \mathit{terminates}(e,f,t)])$

$[A_4]$ $\mathbf{C}(\forall a,d,t \ . \ \mathit{happens}(\mathit{action}(a,d),t) \Rightarrow \mathbf{K}(a, \mathit{happens}(\mathit{action}(a,d),t)))$

$[A_5]$ $\mathbf{C}(\forall a,f,t,t' \ . \ \mathbf{B}(a,\mathit{holds}(f,t)) \wedge \mathbf{B}(a,t<t') \wedge \neg \mathbf{B}(a,\mathit{clipped}(t,f,t')) \Rightarrow \mathbf{B}(a,\mathit{holds}(f,t')))$



$\mathcal{CA}$: 11 Axioms (Initially)

Plan

P2B

K2B $\forall a[\mathbf{K}_a\phi \rightarrow (\mathbf{B}_a\phi \wedge \mathbf{B}_a\exists\Phi\exists\alpha(\Phi \rightsquigarrow_{\alpha/\pi} \phi)]$

Intro

Incorr $\forall a\forall t\forall F[(F \textit{ is contingent } \wedge F \in C'') \rightarrow (\Box\mathbf{B}(a,t,Fa) \rightarrow Fa)]$

Ess

¬CompE

Irr

Free

CCaus

TheI

$\mathbf{C}$ `SpecRel`

$\mathbf{C}$ $\mathcal{EC}$

$[A_1]$ $\mathbf{C}(\forall f,t \;.\; \textit{initially}(f) \wedge \neg \textit{clipped}(0,f,t) \Rightarrow \textit{holds}(f,t))$

$[A_2]$ $\mathbf{C}(\forall e,f,t_1,t_2 \;.\; \textit{happens}(e,t_1) \wedge \textit{initiates}(e,f,t_1) \wedge t_1 < t_2 \wedge \neg \textit{clipped}(t_1,f,t_2) \Rightarrow \textit{holds}(f,t_2))$

$[A_3]$ $\mathbf{C}(\forall t_1,f,t_2 \;.\; \textit{clipped}(t_1,f,t_2) \Leftrightarrow [\exists e,t \;.\; \textit{happens}(e,t) \wedge t_1 < t < t_2 \wedge \textit{terminates}(e,f,t)])$

$[A_4]$ $\mathbf{C}(\forall a,d,t \;.\; \textit{happens}(\textit{action}(a,d),t) \Rightarrow \mathbf{K}(a,\textit{happens}(\textit{action}(a,d),t)))$

$[A_5]$ $\mathbf{C}(\forall a,f,t,t' \;.\; \mathbf{B}(a,\textit{holds}(f,t)) \wedge \mathbf{B}(a,t<t') \wedge \neg\mathbf{B}(a,\textit{clipped}(t,f,t')) \Rightarrow \mathbf{B}(a,\textit{holds}(f,t')))$

Example

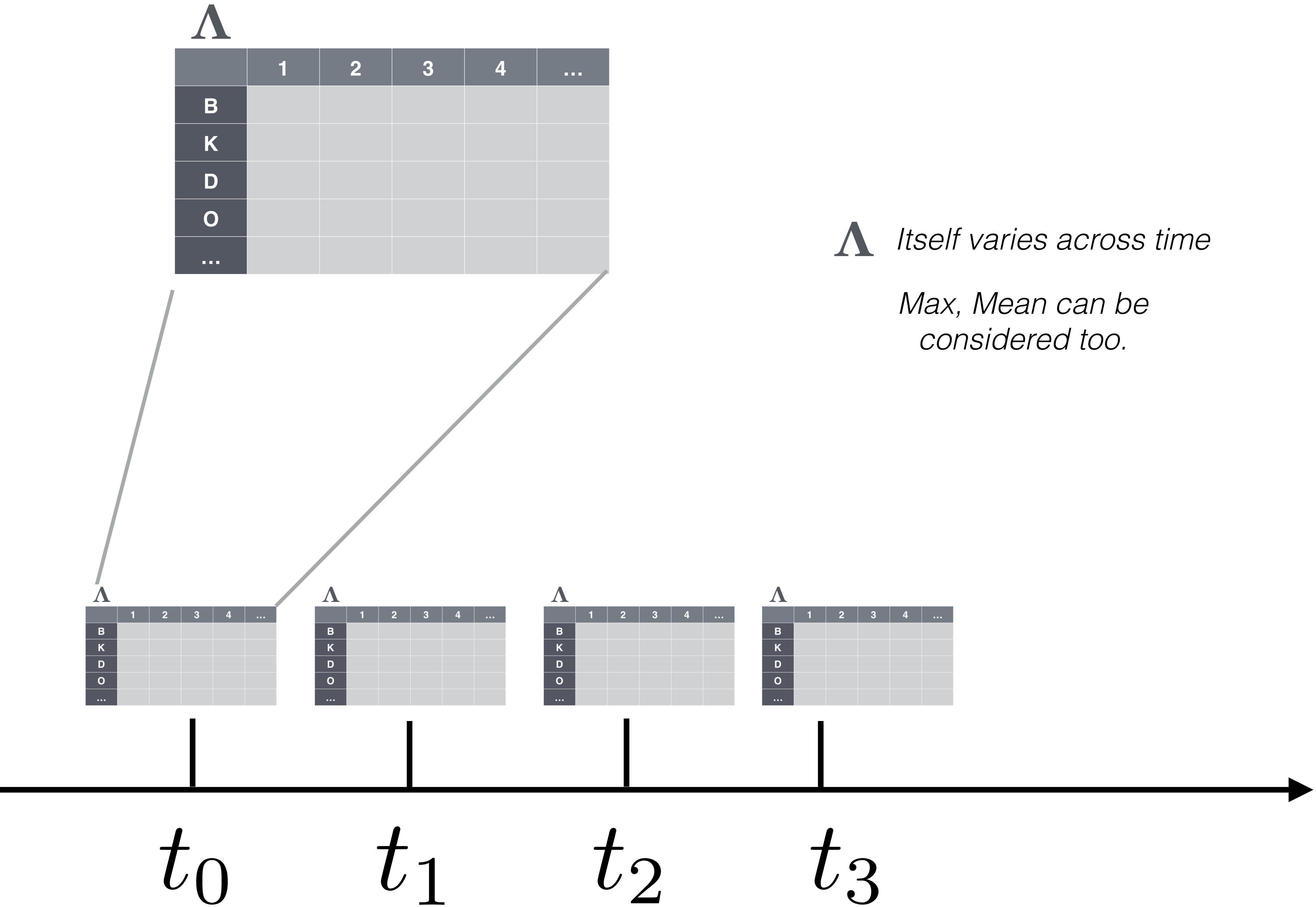
```
{:name         "Knowability paradox"
 :description " \exists p  ~\Diamond \exists x Kx (Tp & ~ \exist y Ky Tp)"

 :assumptions {}
 :goal (exists [?P] (not (pos (exists [?x] (Knows! ?x (and ?P (not (exists [?y] (Knows! ?y ?P))))))))))}
```

$\Lambda[\mathbf{K}, 1] = 2$

$\Lambda[\mathbf{K}, 2] = 1$

$\Lambda[\mathbf{K}, 2] = 2$ *Since the above goal is in second-order modal logic*



Λ *Itself varies across time*

Max, Mean can be considered too.

What is the level of consciousness (= Λ value) enjoyed by this self-conscious robot?

https://motherboard.vice.com/en_us/article/mgbyvb/watch-these-cute-robots-struggle-to-become-self-aware

Med nok penger, kan logikk løse alle problemer.