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 16 2023

R A I R
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 16 2023

RAIR
Rensselaer AI and Reasoning Lab

AI, Consciousness, & Lambda (Λ)

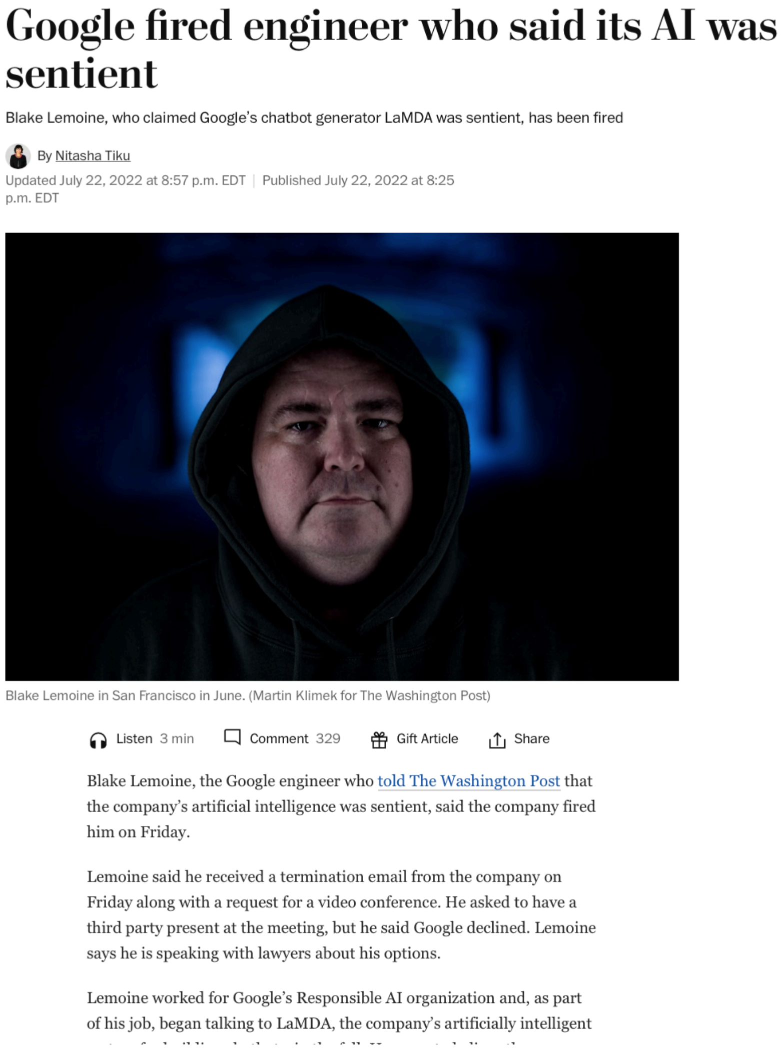
Selmer Bringsjord

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 16 2023

RAIR
Rensselaer AI and Reasoning Lab

Relevant Real Technology



Google fired engineer who said its AI was sentient

Blake Lemoine, who claimed Google's chatbot generator LaMDA was sentient, has been fired

By Nitasha Tiku

Updated July 22, 2022 at 8:57 p.m. EDT | Published July 22, 2022 at 8:25 p.m. EDT

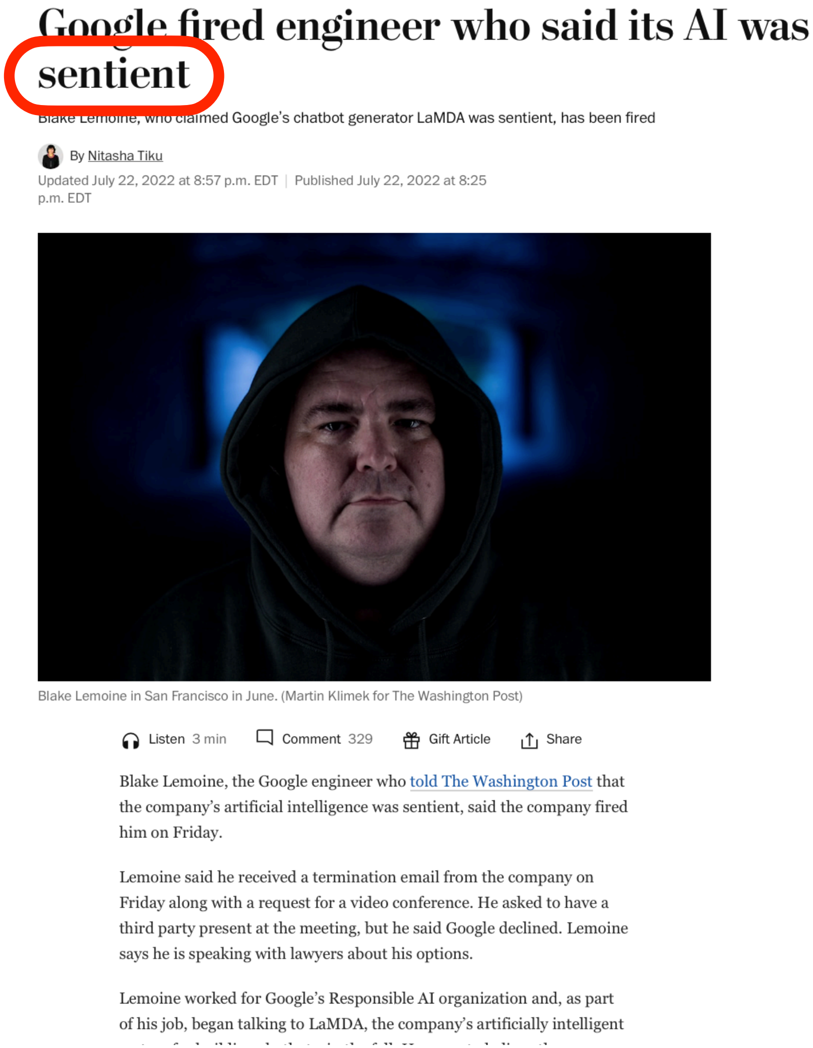
Blake Lemoine in San Francisco in June. (Martin Klimek for The Washington Post)

Blake Lemoine, the Google engineer who told The Washington Post that the company's artificial intelligence was sentient, said the company fired him on Friday.

Lemoine said he received a termination email from the company on Friday along with a request for a video conference. He asked to have a third party present at the meeting, but he said Google declined. Lemoine says he is speaking with lawyers about his options.

Lemoine worked for Google's Responsible AI organization and, as part of his job, began talking to LaMDA, the company's artificially intelligent

Relevant Real Technology



Google fired engineer who said its AI was sentient

Blake Lemoine, who claimed Google's chatbot generator LaMDA was sentient, has been fired

By Nitasha Tiku

Updated July 22, 2022 at 8:57 p.m. EDT | Published July 22, 2022 at 8:25 p.m. EDT

Blake Lemoine in San Francisco in June. (Martin Klimek for The Washington Post)

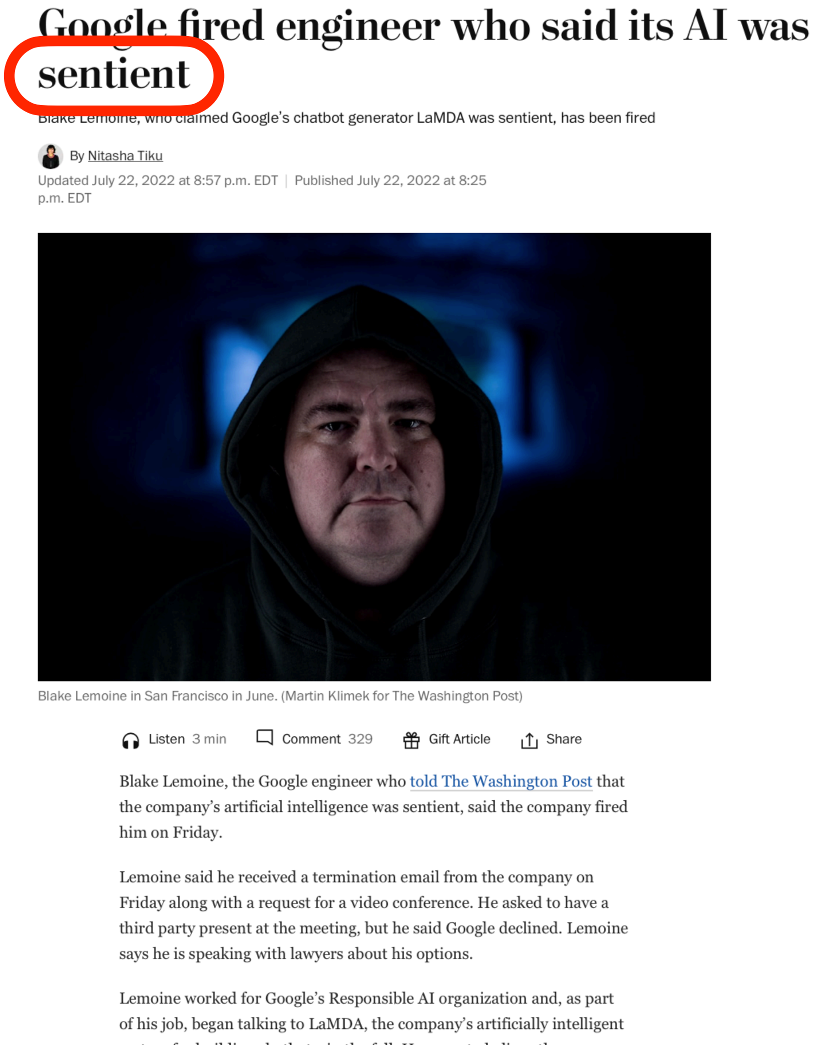
Blake Lemoine, the Google engineer who told The Washington Post that the company's artificial intelligence was sentient, said the company fired him on Friday.

Lemoine said he received a termination email from the company on Friday along with a request for a video conference. He asked to have a third party present at the meeting, but he said Google declined. Lemoine says he is speaking with lawyers about his options.

Lemoine worked for Google's Responsible AI organization and, as part of his job, began talking to LaMDA, the company's artificially intelligent

Relevant Real Technology

?



Google fired engineer who said its AI was sentient

Blake Lemoine, who claimed Google's chatbot generator LaMDA was sentient, has been fired

By Nitasha Tiku

Updated July 22, 2022 at 8:57 p.m. EDT | Published July 22, 2022 at 8:25 p.m. EDT

Blake Lemoine in San Francisco in June. (Martin Klimek for The Washington Post)

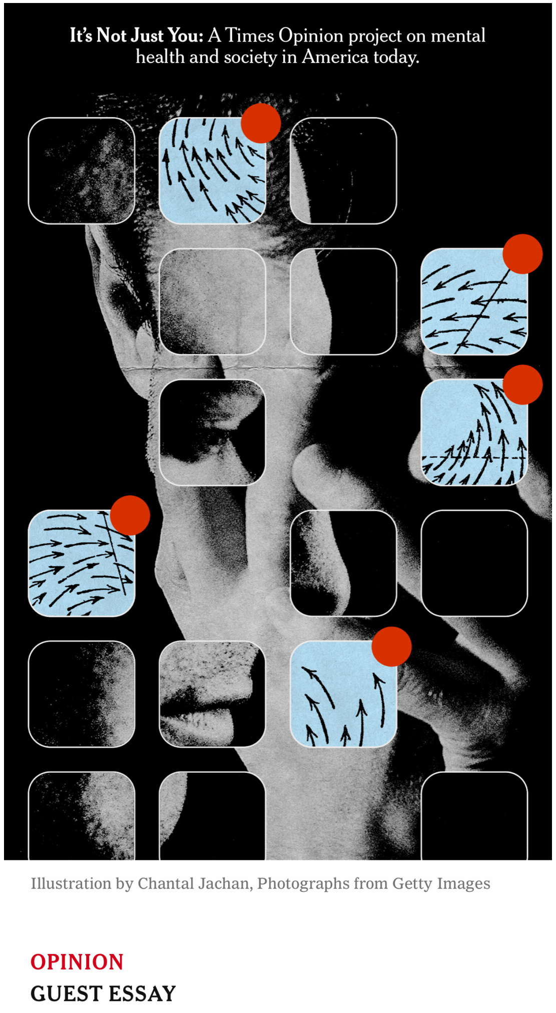
Blake Lemoine, the Google engineer who told The Washington Post that the company's artificial intelligence was sentient, said the company fired him on Friday.

Lemoine said he received a termination email from the company on Friday along with a request for a video conference. He asked to have a third party present at the meeting, but he said Google declined. Lemoine says he is speaking with lawyers about his options.

Lemoine worked for Google's Responsible AI organization and, as part of his job, began talking to LaMDA, the company's artificially intelligent

Relevant Real Technology

Relevant Real Technology



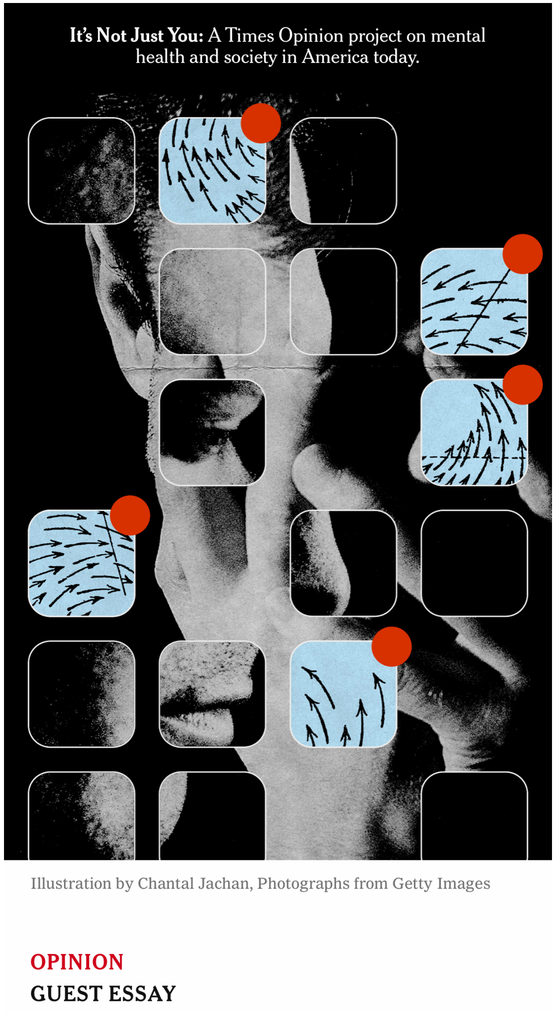
It's Not Just You: A Times Opinion project on mental health and society in America today.

Illustration by Chantal Jachan, Photographs from Getty Images

OPINION

GUEST ESSAY

Relevant Real Technology

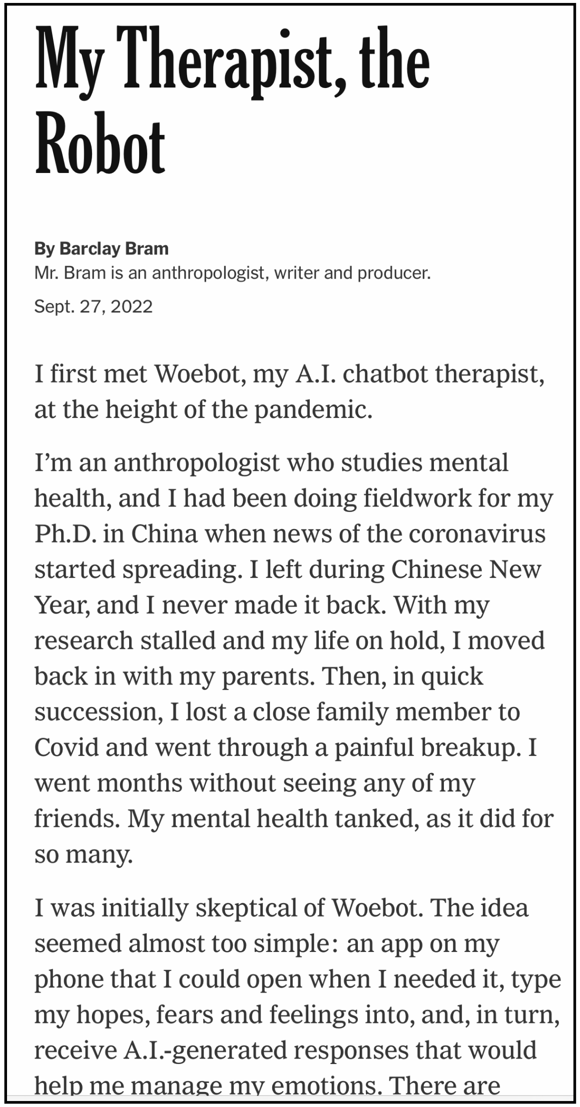


It's Not Just You: A Times Opinion project on mental health and society in America today.

Illustration by Chantal Jachan, Photographs from Getty Images

OPINION

GUEST ESSAY



My Therapist, the Robot

By Barclay Bram

Mr. Bram is an anthropologist, writer and producer.

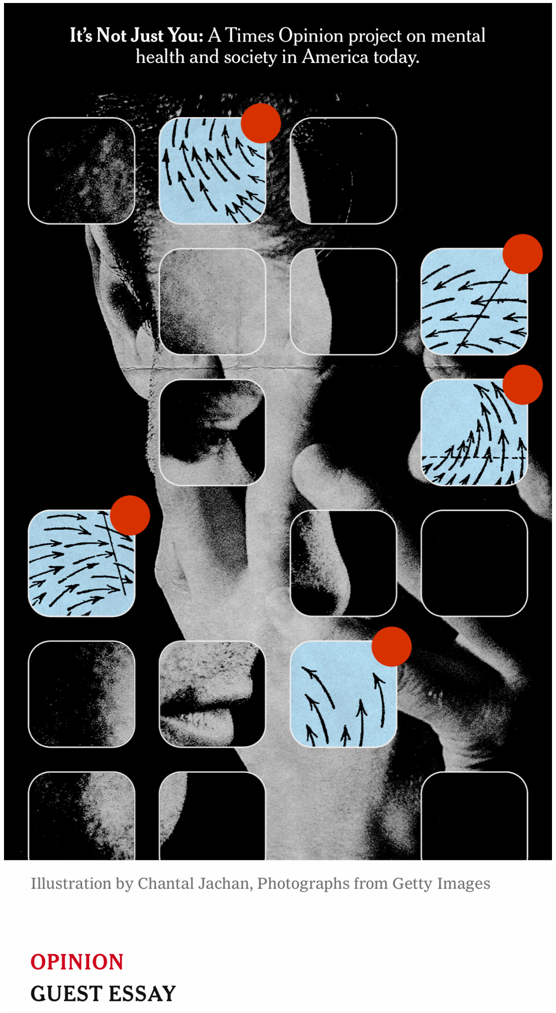
Sept. 27, 2022

I first met Woebot, my A.I. chatbot therapist, at the height of the pandemic.

I'm an anthropologist who studies mental health, and I had been doing fieldwork for my Ph.D. in China when news of the coronavirus started spreading. I left during Chinese New Year, and I never made it back. With my research stalled and my life on hold, I moved back in with my parents. Then, in quick succession, I lost a close family member to Covid and went through a painful breakup. I went months without seeing any of my friends. My mental health tanked, as it did for so many.

I was initially skeptical of Woebot. The idea seemed almost too simple: an app on my phone that I could open when I needed it, type my hopes, fears and feelings into, and, in turn, receive A.I.-generated responses that would help me manage my emotions. There are

Relevant Real Technology

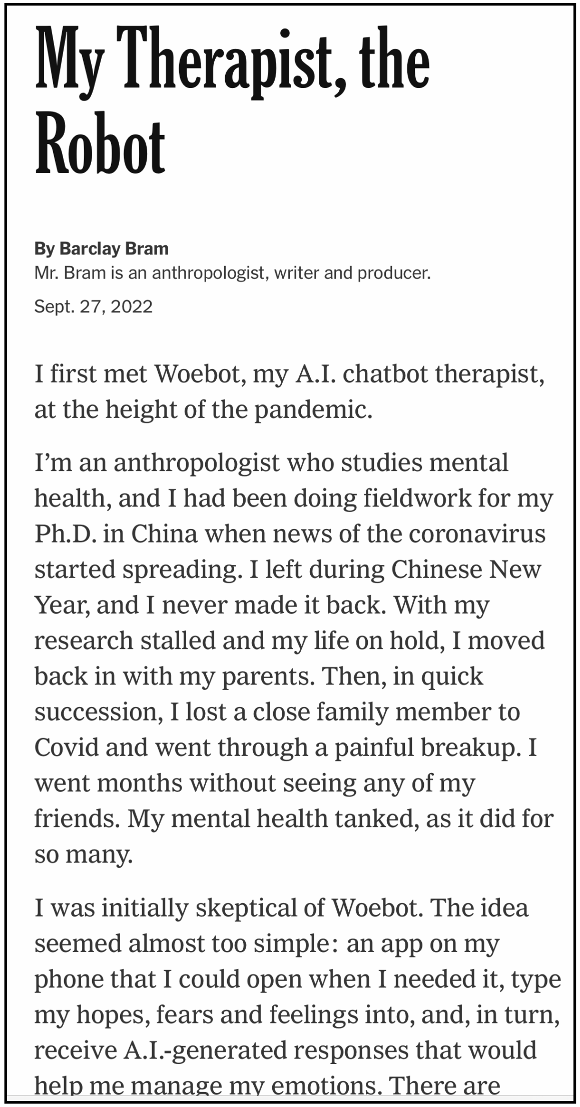


It's Not Just You: A Times Opinion project on mental health and society in America today.

Illustration by Chantal Jachan, Photographs from Getty Images

OPINION

GUEST ESSAY



My Therapist, the Robot

By Barclay Bram

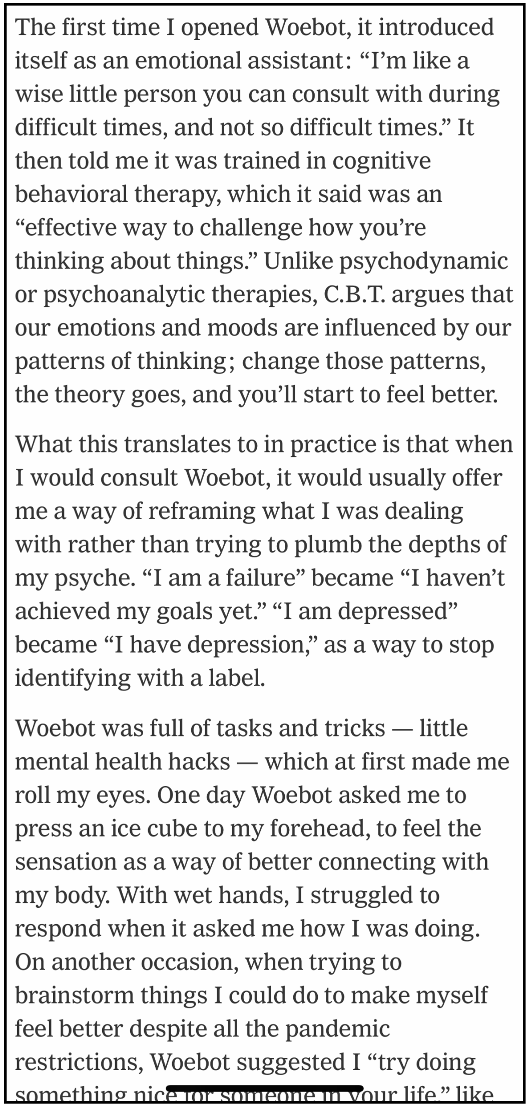
Mr. Bram is an anthropologist, writer and producer.

Sept. 27, 2022

I first met Woebot, my A.I. chatbot therapist, at the height of the pandemic.

I'm an anthropologist who studies mental health, and I had been doing fieldwork for my Ph.D. in China when news of the coronavirus started spreading. I left during Chinese New Year, and I never made it back. With my research stalled and my life on hold, I moved back in with my parents. Then, in quick succession, I lost a close family member to Covid and went through a painful breakup. I went months without seeing any of my friends. My mental health tanked, as it did for so many.

I was initially skeptical of Woebot. The idea seemed almost too simple: an app on my phone that I could open when I needed it, type my hopes, fears and feelings into, and, in turn, receive A.I.-generated responses that would help me manage my emotions. There are

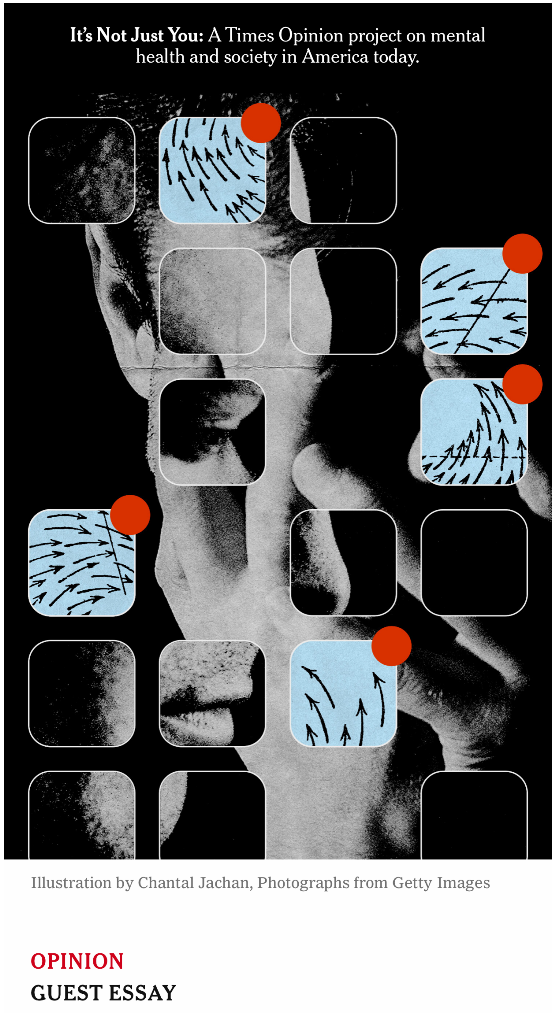


The first time I opened Woebot, it introduced itself as an emotional assistant: "I'm like a wise little person you can consult with during difficult times, and not so difficult times." It then told me it was trained in cognitive behavioral therapy, which it said was an "effective way to challenge how you're thinking about things." Unlike psychodynamic or psychoanalytic therapies, C.B.T. argues that our emotions and moods are influenced by our patterns of thinking; change those patterns, the theory goes, and you'll start to feel better.

What this translates to in practice is that when I would consult Woebot, it would usually offer me a way of reframing what I was dealing with rather than trying to plumb the depths of my psyche. "I am a failure" became "I haven't achieved my goals yet." "I am depressed" became "I have depression," as a way to stop identifying with a label.

Woebot was full of tasks and tricks — little mental health hacks — which at first made me roll my eyes. One day Woebot asked me to press an ice cube to my forehead, to feel the sensation as a way of better connecting with my body. With wet hands, I struggled to respond when it asked me how I was doing. On another occasion, when trying to brainstorm things I could do to make myself feel better despite all the pandemic restrictions, Woebot suggested I "try doing something nice for someone in your life," like

Relevant Real Technology

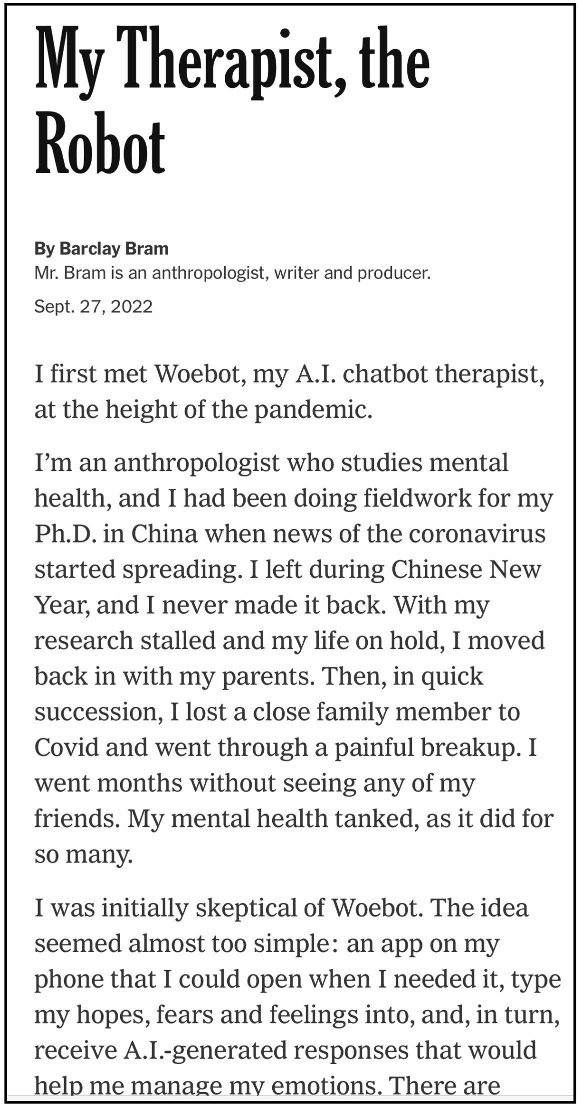


It's Not Just You: A Times Opinion project on mental health and society in America today.

Illustration by Chantal Jachan, Photographs from Getty Images

OPINION

GUEST ESSAY



My Therapist, the Robot

By Barclay Bram

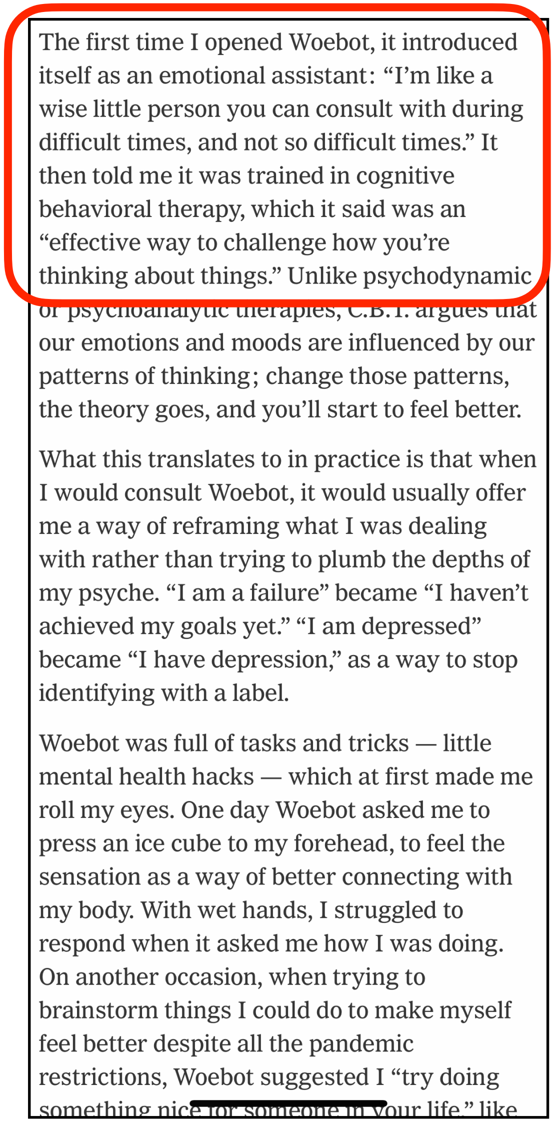
Mr. Bram is an anthropologist, writer and producer.

Sept. 27, 2022

I first met Woebot, my A.I. chatbot therapist, at the height of the pandemic.

I'm an anthropologist who studies mental health, and I had been doing fieldwork for my Ph.D. in China when news of the coronavirus started spreading. I left during Chinese New Year, and I never made it back. With my research stalled and my life on hold, I moved back in with my parents. Then, in quick succession, I lost a close family member to Covid and went through a painful breakup. I went months without seeing any of my friends. My mental health tanked, as it did for so many.

I was initially skeptical of Woebot. The idea seemed almost too simple: an app on my phone that I could open when I needed it, type my hopes, fears and feelings into, and, in turn, receive A.I.-generated responses that would help me manage my emotions. There are

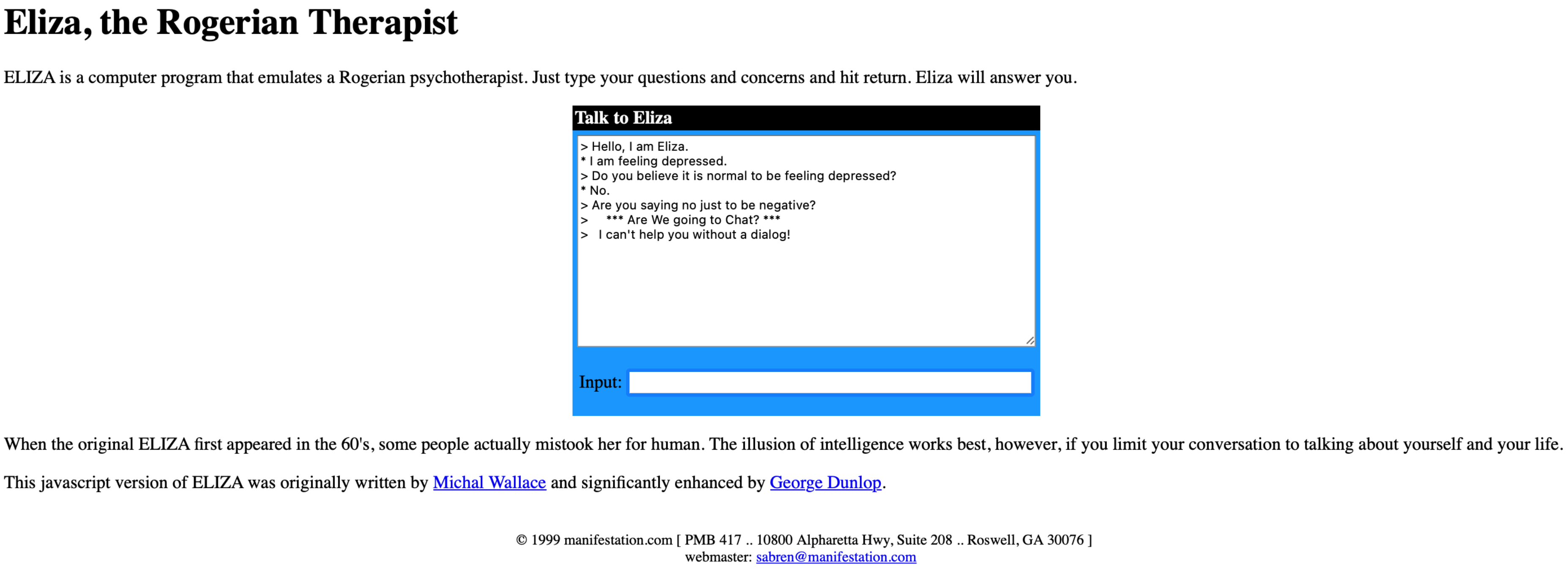


The first time I opened Woebot, it introduced itself as an emotional assistant: "I'm like a wise little person you can consult with during difficult times, and not so difficult times." It then told me it was trained in cognitive behavioral therapy, which it said was an "effective way to challenge how you're thinking about things." Unlike psychodynamic or psychoanalytic therapies, C.B.T. argues that our emotions and moods are influenced by our patterns of thinking; change those patterns, the theory goes, and you'll start to feel better.

What this translates to in practice is that when I would consult Woebot, it would usually offer me a way of reframing what I was dealing with rather than trying to plumb the depths of my psyche. "I am a failure" became "I haven't achieved my goals yet." "I am depressed" became "I have depression," as a way to stop identifying with a label.

Woebot was full of tasks and tricks — little mental health hacks — which at first made me roll my eyes. One day Woebot asked me to press an ice cube to my forehead, to feel the sensation as a way of better connecting with my body. With wet hands, I struggled to respond when it asked me how I was doing. On another occasion, when trying to brainstorm things I could do to make myself feel better despite all the pandemic restrictions, Woebot suggested I "try doing something nice for someone in your life," like

Relevant *Old* Real Technology



Eliza, the Rogerian Therapist

ELIZA is a computer program that emulates a Rogerian psychotherapist. Just type your questions and concerns and hit return. Eliza will answer you.

```
> Hello, I am Eliza.
* I am feeling depressed.
> Do you believe it is normal to be feeling depressed?
* No.
> Are you saying no just to be negative?
>     *** Are We going to Chat? ***
>   I can't help you without a dialog!
```

When the original ELIZA first appeared in the 60's, some people actually mistook her for human. The illusion of intelligence works best, however, if you limit your conversation to talking about yourself and your life.

This javascript version of ELIZA was originally written by Michal Wallace and significantly enhanced by George Dunlop.

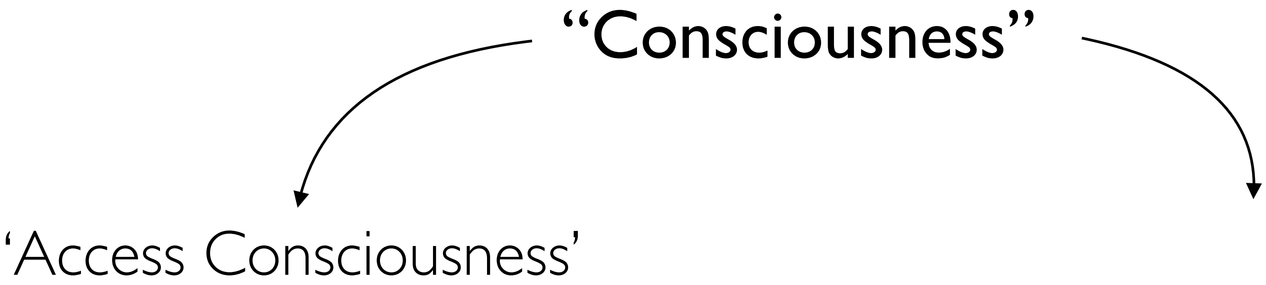
© 1999 manifestation.com [PMB 417 .. 10800 Alpharetta Hwy, Suite 208 .. Roswell, GA 30076]
webmaster: sabren@manifestation.com

“Consciousness”

“Consciousness”

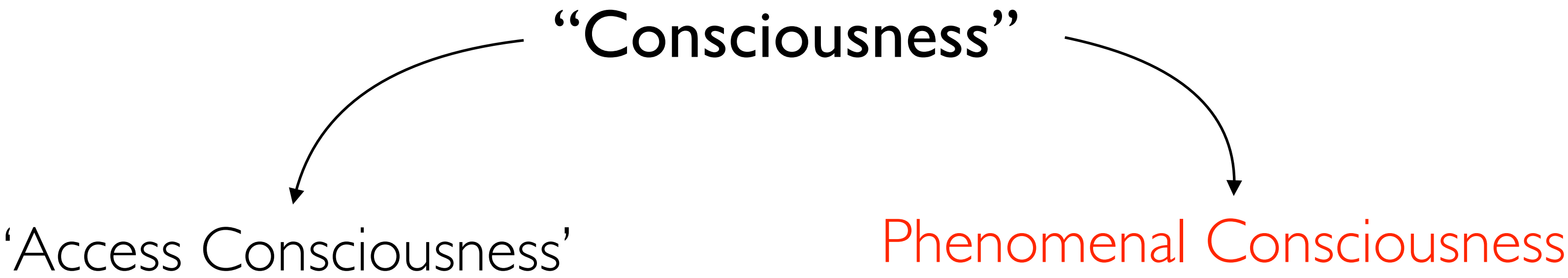
“Consciousness”

‘Access Consciousness’



“Consciousness”

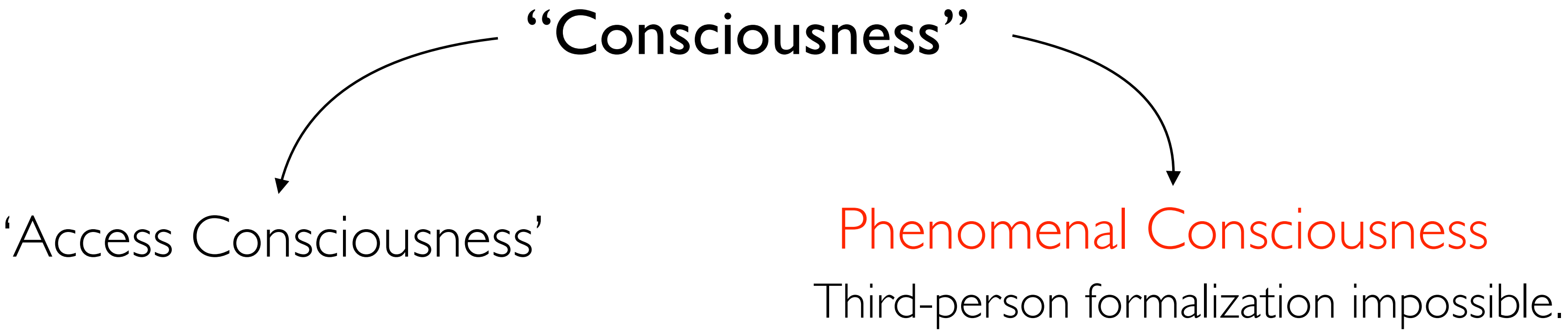
‘Access Consciousness’



“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

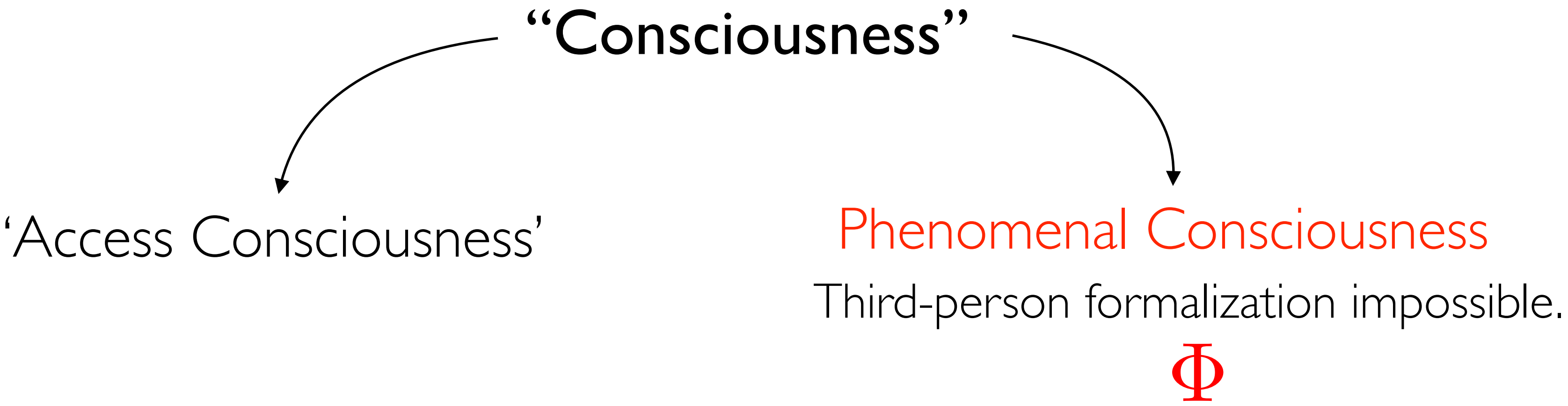


“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.



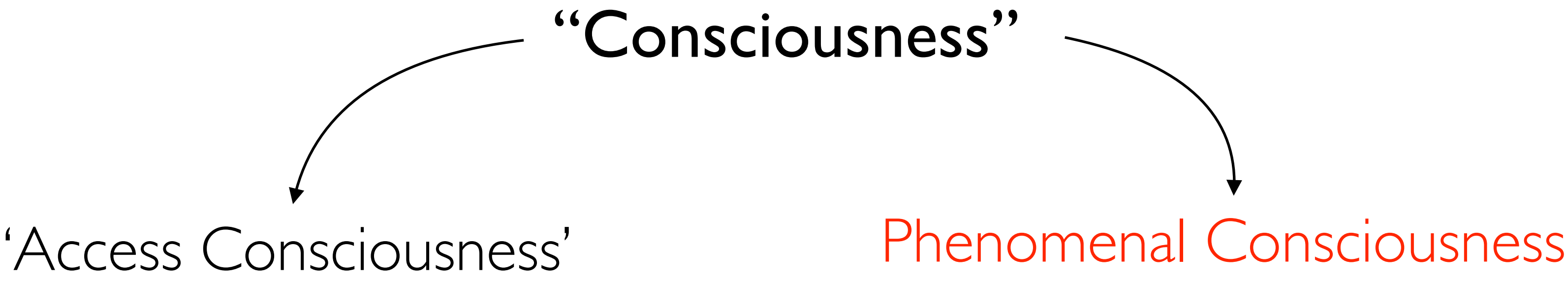
“Consciousness”

‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.

Φ



“**Consciousness**”

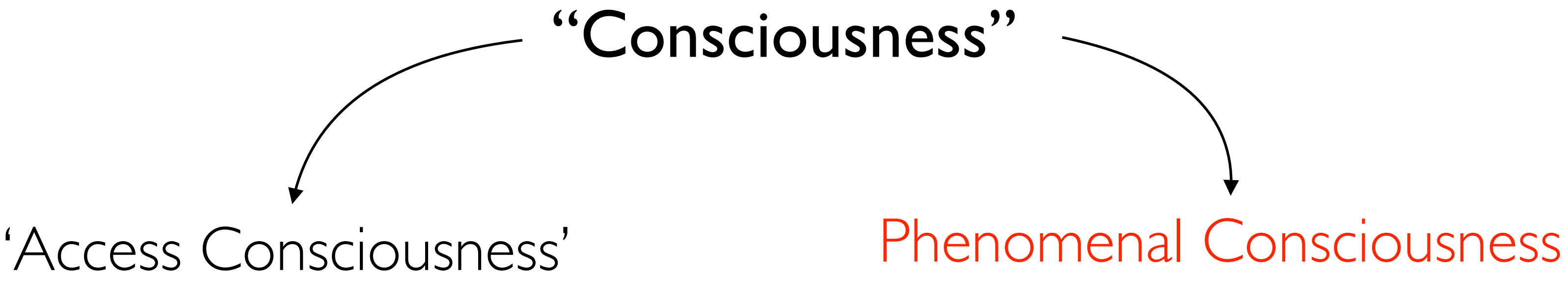
‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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



“Consciousness”

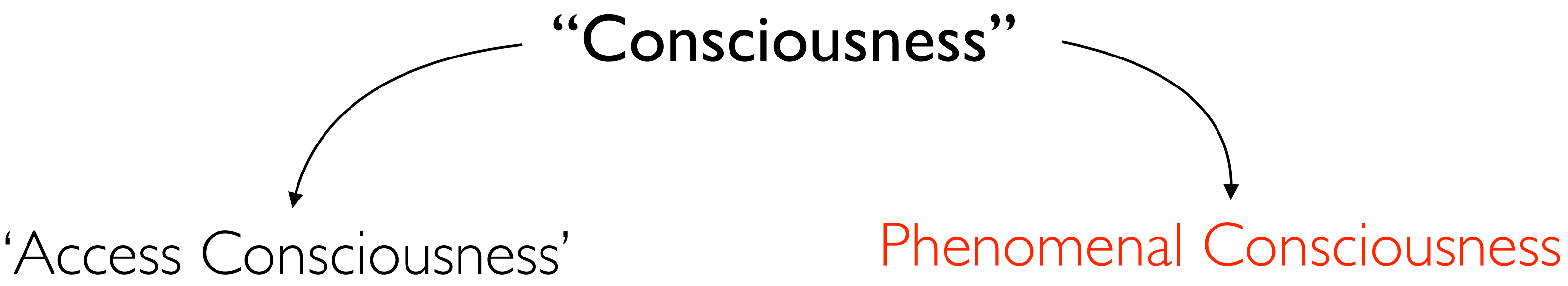
‘Access Consciousness’

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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



“Consciousness”

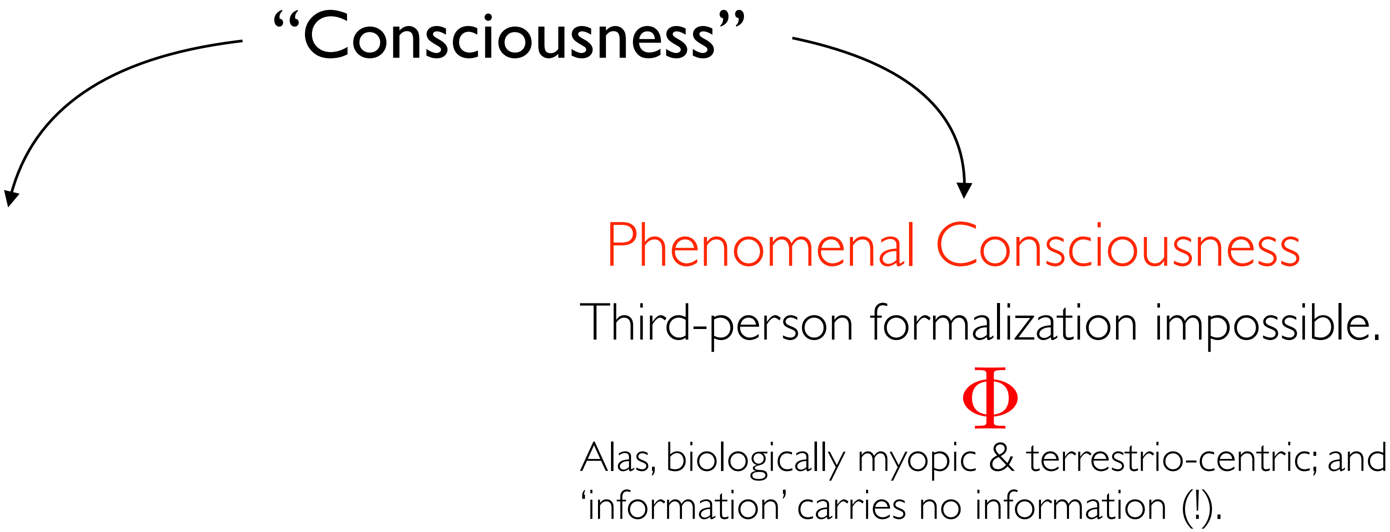
‘Access Consciousness’

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”

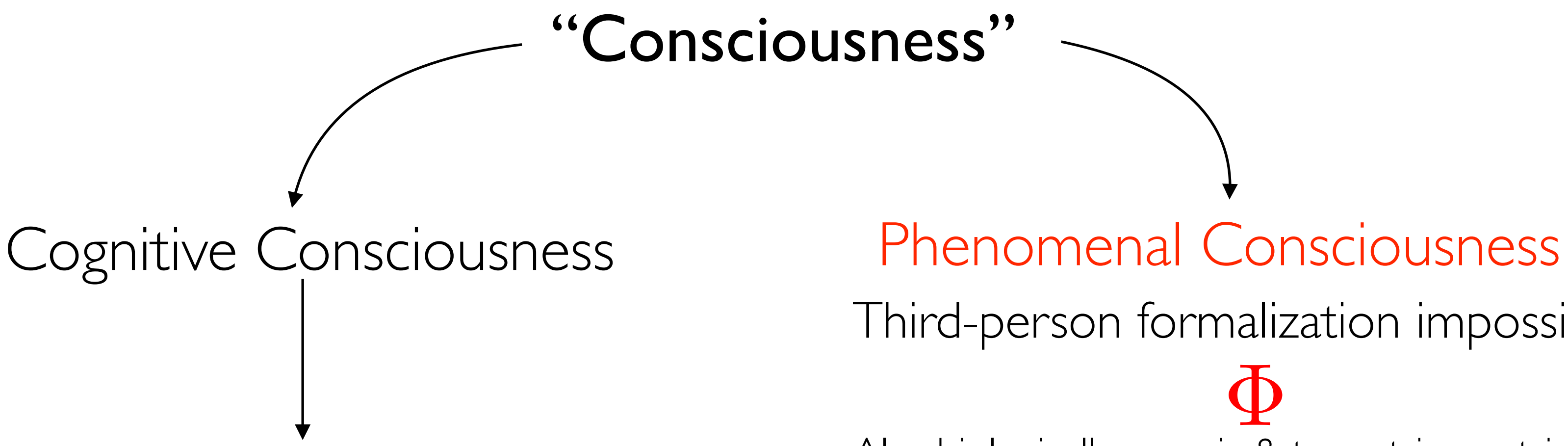
Cognitive Consciousness

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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



“Consciousness”

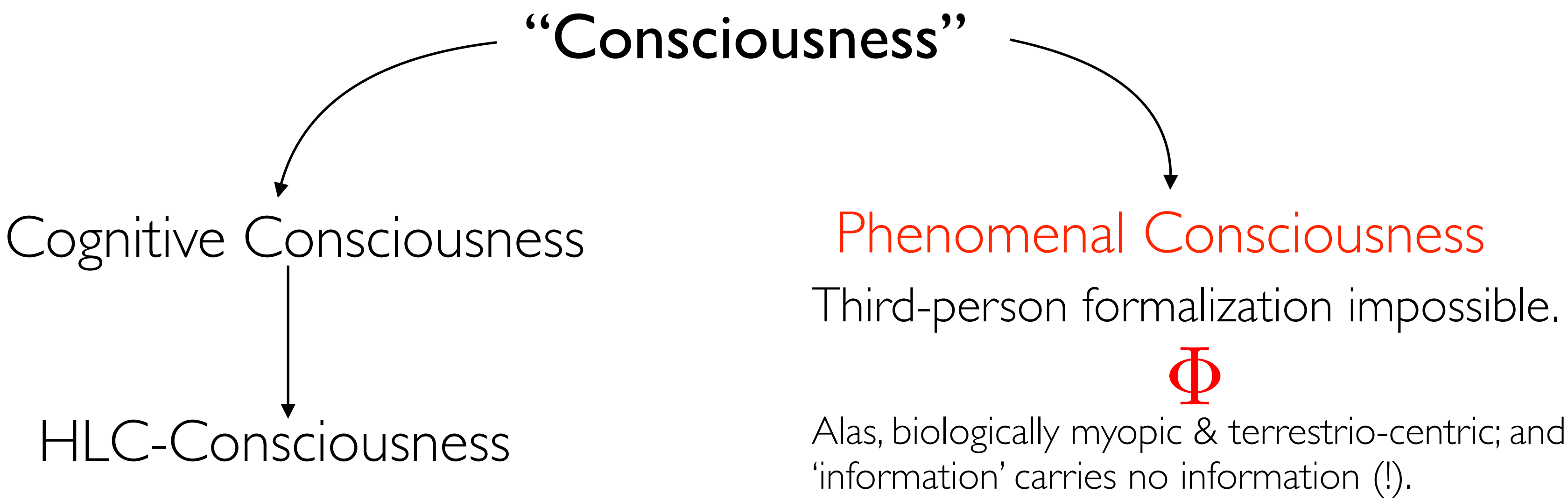
Cognitive Consciousness

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

HL**M**C-Consciousness

Happily, not bound by local biology; will cover aliens, God, characters of fiction, etc; and 'information' is information. (McCarthy e.g. would be happy.)

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. (McCarthy e.g. would be happy.)

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. (McCarthy e.g. would be happy.)

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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

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

Definition …

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

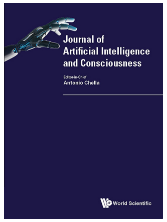
size of supporting proof/argument

depth of quantification within outermost knowledge operator

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

The Theory of Cognitive Consciousness, and Λ (Lambda)

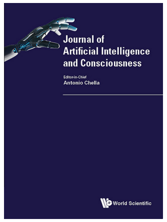
Selmer Bringsjord and G. Naveen Sundar

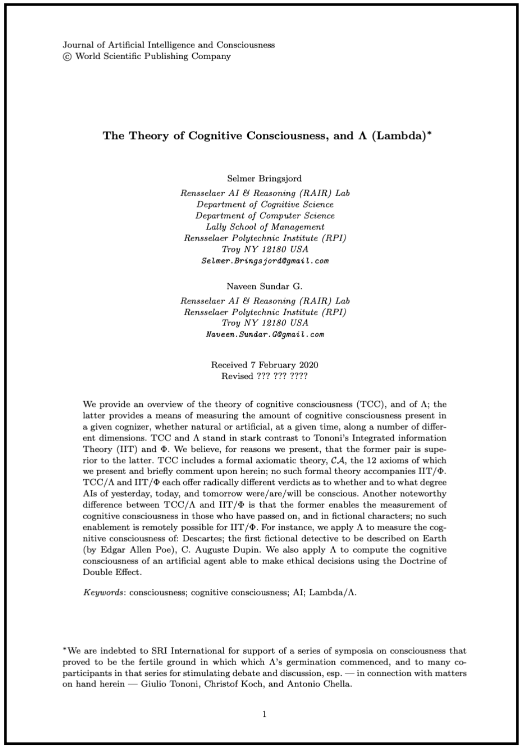


Journal of Artificial Intelligence and Consciousness | Vol. 07, No. 02, pp. 155-181 (2020)

The Theory of Cognitive Consciousness, and Λ (Lambda)

Selmer Bringsjord and G. Naveen Sundar





Journal of Artificial Intelligence and Consciousness
© World Scientific Publishing Company

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 ??? ??? ????

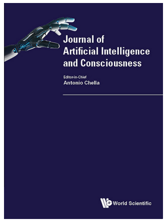
We provide an overview of the theory of cognitive consciousness (TCC), and of Λ; the latter provides a means of measuring the amount of cognitive consciousness present in a given cognizer, whether natural or artificial, at a given time, along a number of different dimensions. TCC and Λ stand in stark 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.

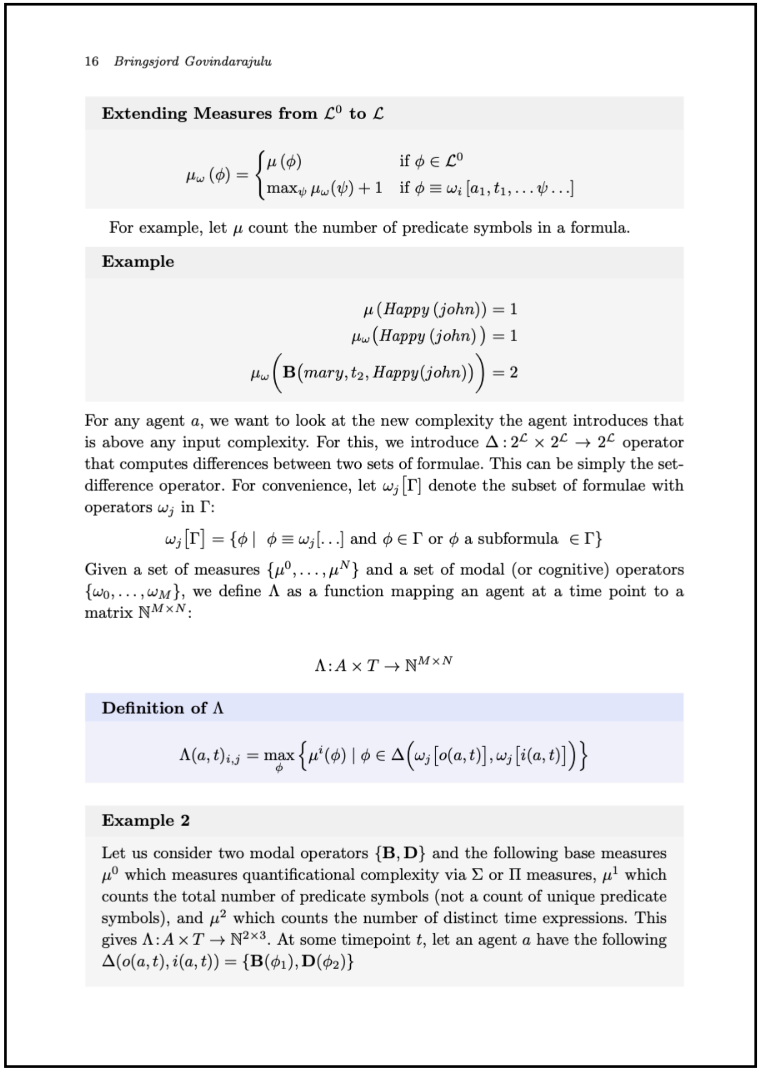
Keywords: consciousness; cognitive consciousness; AI; Lambda/Λ.

*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 $\mathcal{L}$

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

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

Example

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

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 $\omega_j\left[\Gamma\right]$ denote the subset of formulae with operators ω_j in Γ:

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

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

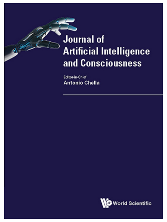
$$\Lambda(a,t)_{i,j} = \max_{\phi} \left\{\mu^i(\phi) \mid \phi \in \Delta\left(\omega_j\left[o(a,t)\right], \omega_j\left[i(a,t)\right]\right)\right\}$$

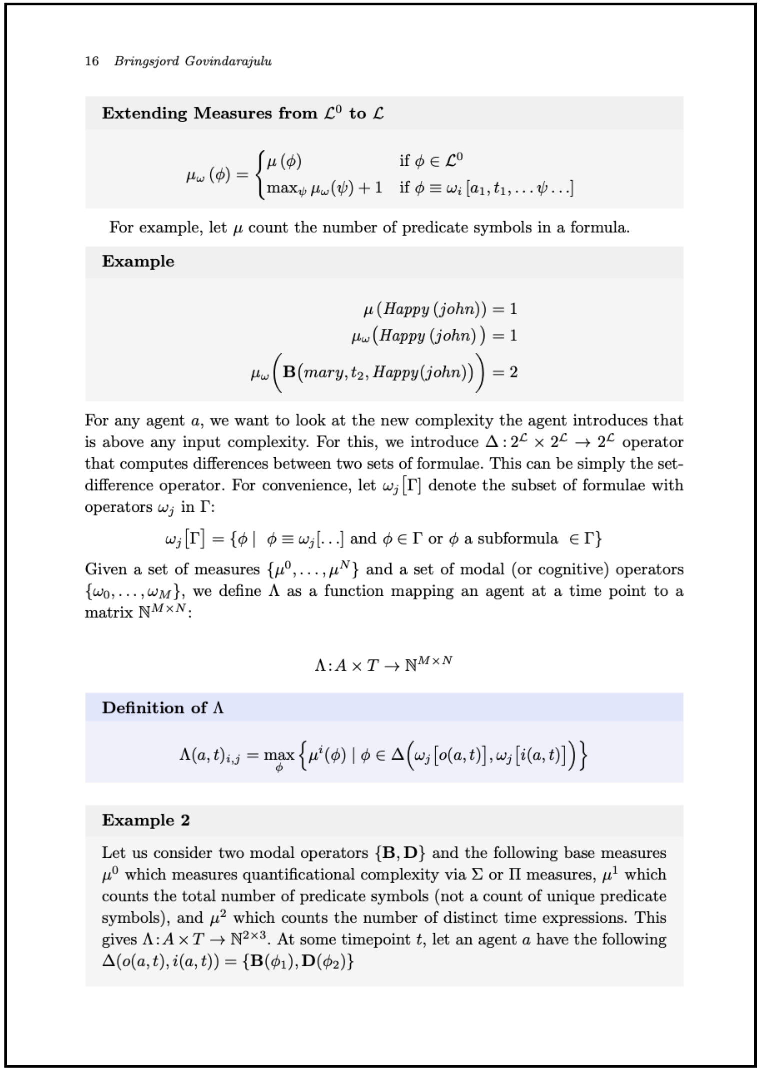
Example 2

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 $\Delta(o(a,t), i(a,t)) = \{\mathbf{B}(\phi_1), \mathbf{D}(\phi_2)\}$

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

The Theory of Cognitive Consciousness, and Λ (Lambda)


Journal of Artificial Intelligence and Consciousness
Editor-in-Chief
Antonio Chella
World Scientific



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 $\mathbb{N}^{M \times N}$:

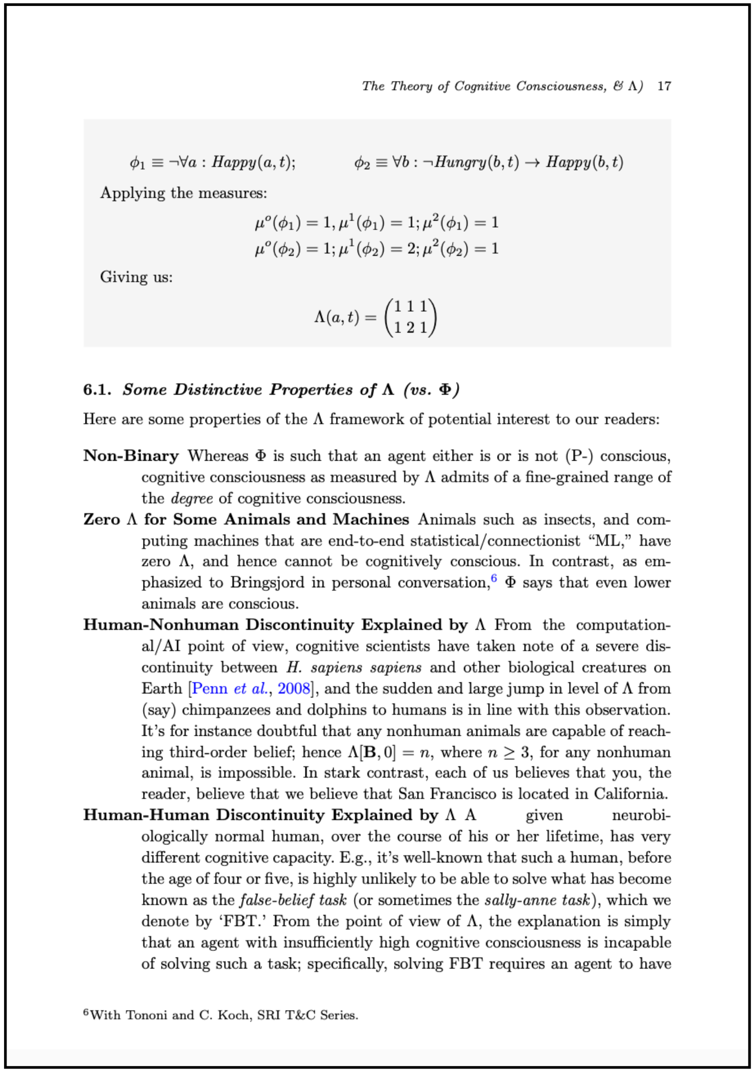
$$\Lambda : A \times T \rightarrow \mathbb{N}^{M \times N}$$

Definition of Λ

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

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 $\Delta(o(a,t), i(a,t)) = \{\mathbf{B}(\phi_1), \mathbf{D}(\phi_2)\}$



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

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.

Basic Idea, Intuitively Put

The level of (cognitive) intelligence of an agent (artificial or natural) at a time is a list of tuples (= matrix) giving eg the size of logical depth of multiple 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 of an agent (artificial or natural) at a time is a list of tuples (= matrix) giving eg the size of logical depth of multiple 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

Formal Syntax

Formal Syntax

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

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

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

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

Inference Schemata

Inference Schemata

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

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

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

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

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

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

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

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

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\;[R_{12}]$$

$$\frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\;[R_{13}]$$

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

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

Inference Schemata

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

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

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

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

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

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

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

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

$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}]$$

$$\frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\ [R_{13}]$$

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

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

Inference Schemata

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

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

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

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

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

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

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

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

$$\frac{\mathbf{B}^k(a,t,\phi)\quad\mathbf{B}^j(a,t,\psi)}{\mathbf{B}^{min(k,j)}(a,t,\phi\wedge\psi)}$$

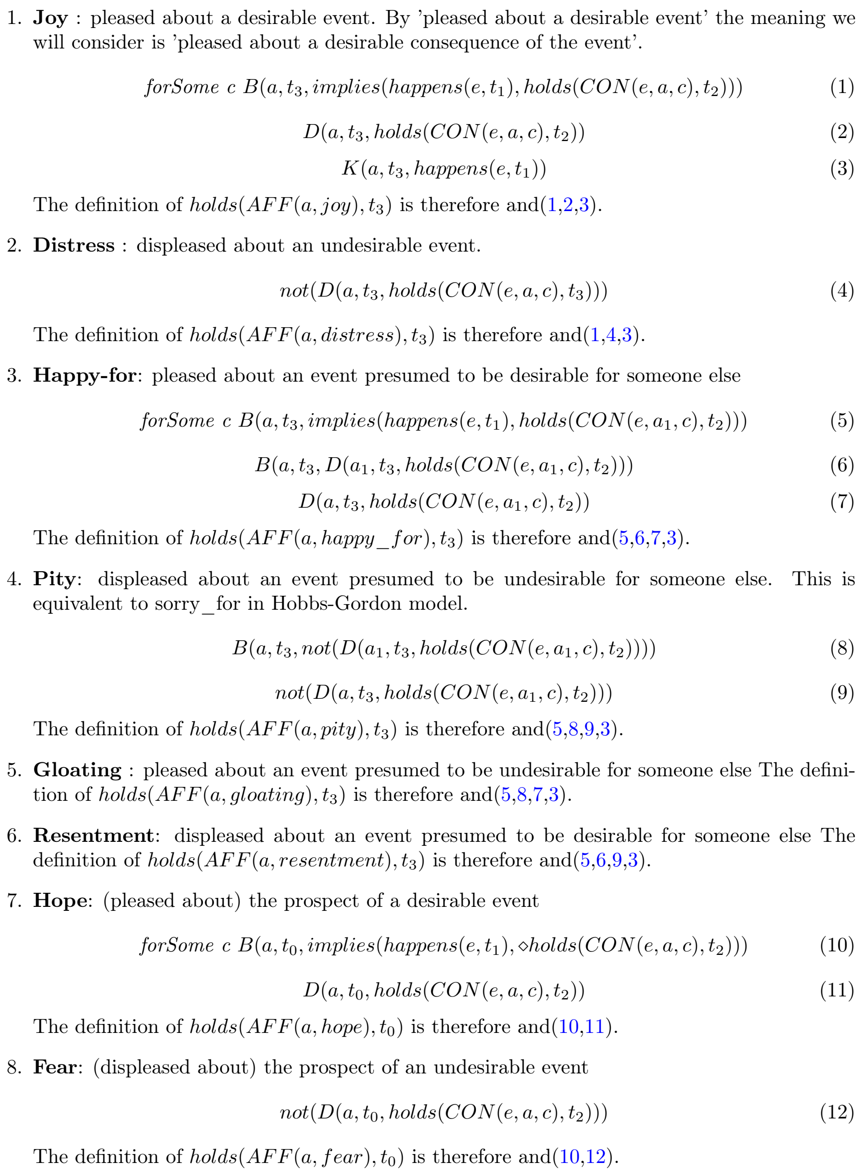
$$\frac{\mathbf{S}(s,h,t,\phi)}{\mathbf{B}(h,t,\mathbf{B}(s,t,\phi))}\ [R_{12}]$$

$$\frac{\mathbf{I}(a,t,happens(action(a^*,\alpha),t'))}{\mathbf{P}(a,t,happens(action(a^*,\alpha),t))}\ [R_{13}]$$

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

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

Defs for An *Affective* Cognitive *time&change* Calculus



1. **Joy** : pleased about a desirable event. By 'pleased about a desirable event' the meaning we will consider is 'pleased about a desirable consequence of the event'.

$$forSome\ c\ B(a, t_3, implies(happens(e, t_1), holds(CON(e, a, c), t_2))) \tag{1}$$

$$D(a, t_3, holds(CON(e, a, c), t_2)) \tag{2}$$

$$K(a, t_3, happens(e, t_1)) \tag{3}$$

   The definition of $holds(AFF(a, joy), t_3)$ is therefore and(1,2,3).

2. **Distress** : displeased about an undesirable event.

$$not(D(a, t_3, holds(CON(e, a, c), t_3))) \tag{4}$$

   The definition of $holds(AFF(a, distress), t_3)$ is therefore and(1,4,3).

3. **Happy-for**: pleased about an event presumed to be desirable for someone else

$$forSome\ c\ B(a, t_3, implies(happens(e, t_1), holds(CON(e, a_1, c), t_2))) \tag{5}$$

$$B(a, t_3, D(a_1, t_3, holds(CON(e, a_1, c), t_2))) \tag{6}$$

$$D(a, t_3, holds(CON(e, a_1, c), t_2)) \tag{7}$$

   The definition of $holds(AFF(a, happy\_for), t_3)$ is therefore and(5,6,7,3).

4. **Pity**: displeased about an event presumed to be undesirable for someone else. This is equivalent to sorry_for in Hobbs-Gordon model.

$$B(a, t_3, not(D(a_1, t_3, holds(CON(e, a_1, c), t_2)))) \tag{8}$$

$$not(D(a, t_3, holds(CON(e, a_1, c), t_2))) \tag{9}$$

   The definition of $holds(AFF(a, pity), t_3)$ is therefore and(5,8,9,3).

5. **Gloating** : pleased about an event presumed to be undesirable for someone else The definition of $holds(AFF(a, gloating), t_3)$ is therefore and(5,8,7,3).

6. **Resentment**: displeased about an event presumed to be desirable for someone else The definition of $holds(AFF(a, resentment), t_3)$ is therefore and(5,6,9,3).

7. **Hope**: (pleased about) the prospect of a desirable event

$$forSome\ c\ B(a, t_0, implies(happens(e, t_1), \diamond holds(CON(e, a, c), t_2))) \tag{10}$$

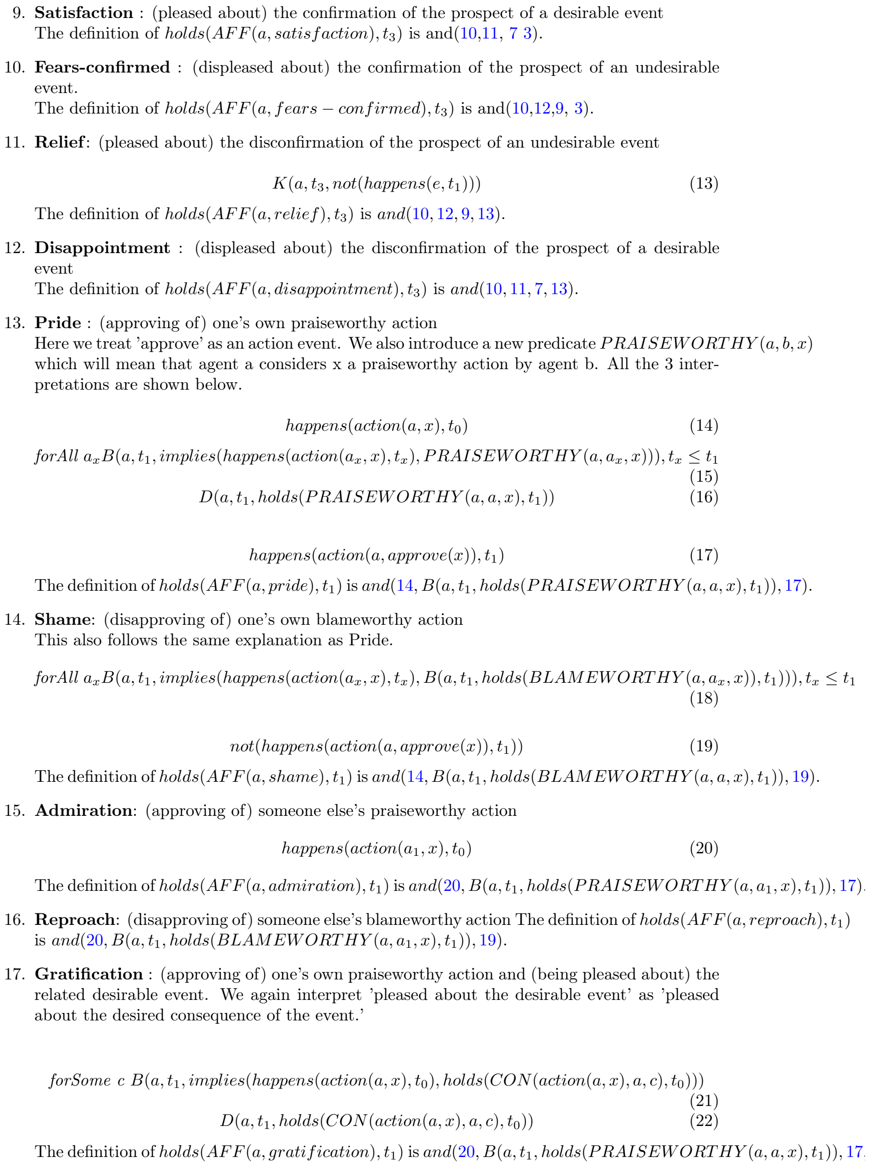
$$D(a, t_0, holds(CON(e, a, c), t_2)) \tag{11}$$

   The definition of $holds(AFF(a, hope), t_0)$ is therefore and(10,11).

8. **Fear**: (displeased about) the prospect of an undesirable event

$$not(D(a, t_0, holds(CON(e, a, c), t_2))) \tag{12}$$

   The definition of $holds(AFF(a, fear), t_0)$ is therefore and(10,12).



9. **Satisfaction** : (pleased about) the confirmation of the prospect of a desirable event
   The definition of $holds(AFF(a, satisfaction), t_3)$ is and(10,11, 7 3).

10. **Fears-confirmed** : (displeased about) the confirmation of the prospect of an undesirable event.
    The definition of $holds(AFF(a, fears - confirmed), t_3)$ is and(10,12,9, 3).

11. **Relief**: (pleased about) the disconfirmation of the prospect of an undesirable event

$$K(a, t_3, not(happens(e, t_1))) \tag{13}$$

    The definition of $holds(AFF(a, relief), t_3)$ is $and(10, 12, 9, 13)$.

12. **Disappointment** : (displeased about) the disconfirmation of the prospect of a desirable event
    The definition of $holds(AFF(a, disappointment), t_3)$ is $and(10, 11, 7, 13)$.

13. **Pride** : (approving of) one's own praiseworthy action
    Here we treat 'approve' as an action event. We also introduce a new predicate $PRAISEWORTHY(a, b, x)$ which will mean that agent a considers x a praiseworthy action by agent b. All the 3 interpretations are shown below.

$$happens(action(a, x), t_0) \tag{14}$$

$$forAll\ a_x B(a, t_1, implies(happens(action(a_x, x), t_x), PRAISEWORTHY(a, a_x, x))), t_x \leq t_1 \tag{15}$$

$$D(a, t_1, holds(PRAISEWORTHY(a, a, x), t_1)) \tag{16}$$

$$happens(action(a, approve(x)), t_1) \tag{17}$$

    The definition of $holds(AFF(a, pride), t_1)$ is $and(14, B(a, t_1, holds(PRAISEWORTHY(a, a, x), t_1)), 17)$.

14. **Shame**: (disapproving of) one's own blameworthy action
    This also follows the same explanation as Pride.

$$forAll\ a_x B(a, t_1, implies(happens(action(a_x, x), t_x), B(a, t_1, holds(BLAMEWORTHY(a, a_x, x)), t_1))), t_x \leq t_1 \tag{18}$$

$$not(happens(action(a, approve(x)), t_1)) \tag{19}$$

    The definition of $holds(AFF(a, shame), t_1)$ is $and(14, B(a, t_1, holds(BLAMEWORTHY(a, a, x), t_1)), 19)$.

15. **Admiration**: (approving of) someone else's praiseworthy action

$$happens(action(a_1, x), t_0) \tag{20}$$

    The definition of $holds(AFF(a, admiration), t_1)$ is $and(20, B(a, t_1, holds(PRAISEWORTHY(a, a_1, x), t_1)), 17)$.

16. **Reproach**: (disapproving of) someone else's blameworthy action The definition of $holds(AFF(a, reproach), t_1)$ is $and(20, B(a, t_1, holds(BLAMEWORTHY(a, a_1, x), t_1)), 19)$.

17. **Gratification** : (approving of) one's own praiseworthy action and (being pleased about) the related desirable event. We again interpret 'pleased about the desirable event' as 'pleased about the desired consequence of the event.'

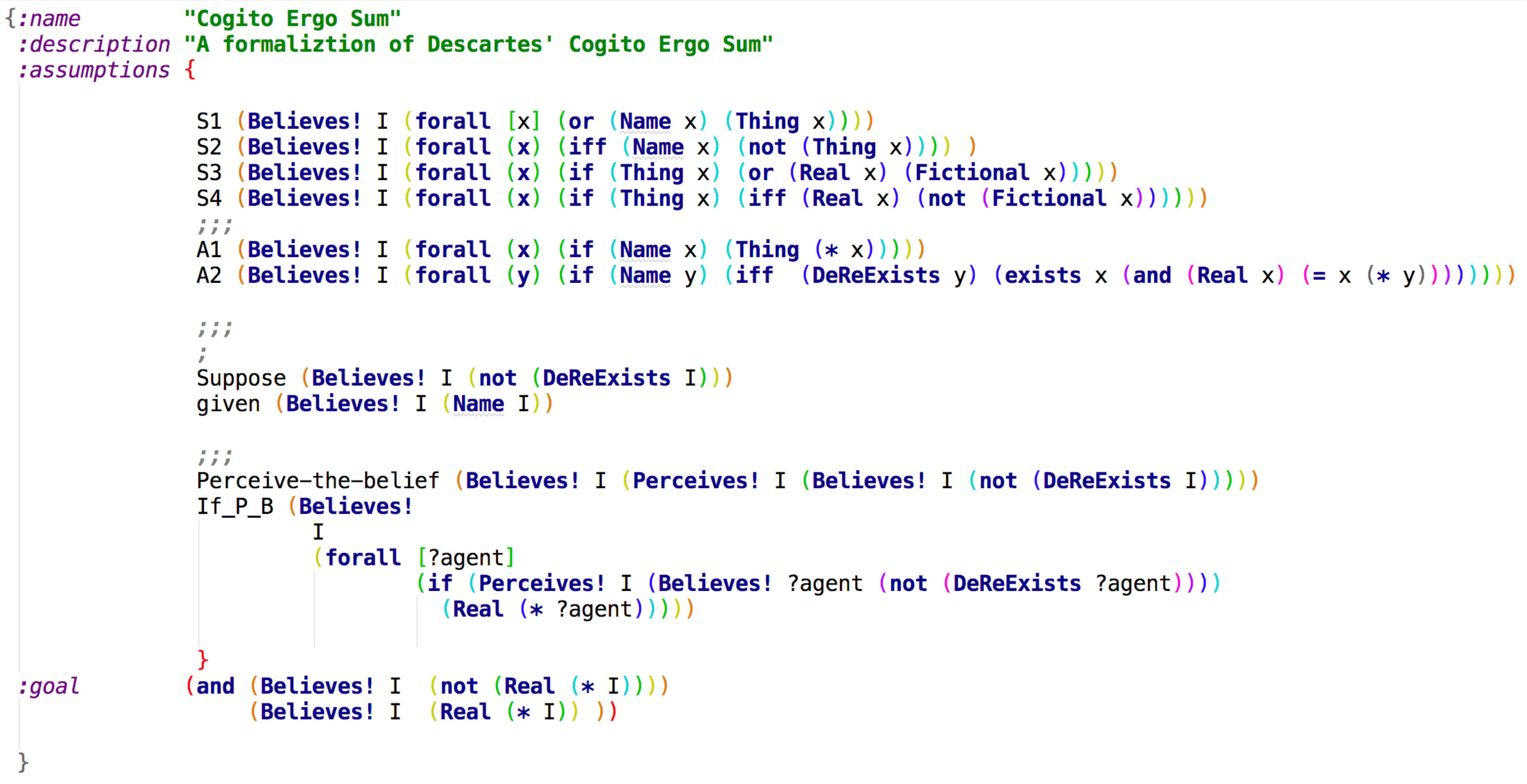
$$forSome\ c\ B(a, t_1, implies(happens(action(a, x), t_0), holds(CON(action(a, x), a, c), t_0))) \tag{21}$$

$$D(a, t_1, holds(CON(action(a, x), a, c), t_0)) \tag{22}$$

    The definition of $holds(AFF(a, gratification), t_1)$ is $and(20, B(a, t_1, holds(PRAISEWORTHY(a, a, x), t_1)), 17$

… (and more)

Cogito Ergo Sum

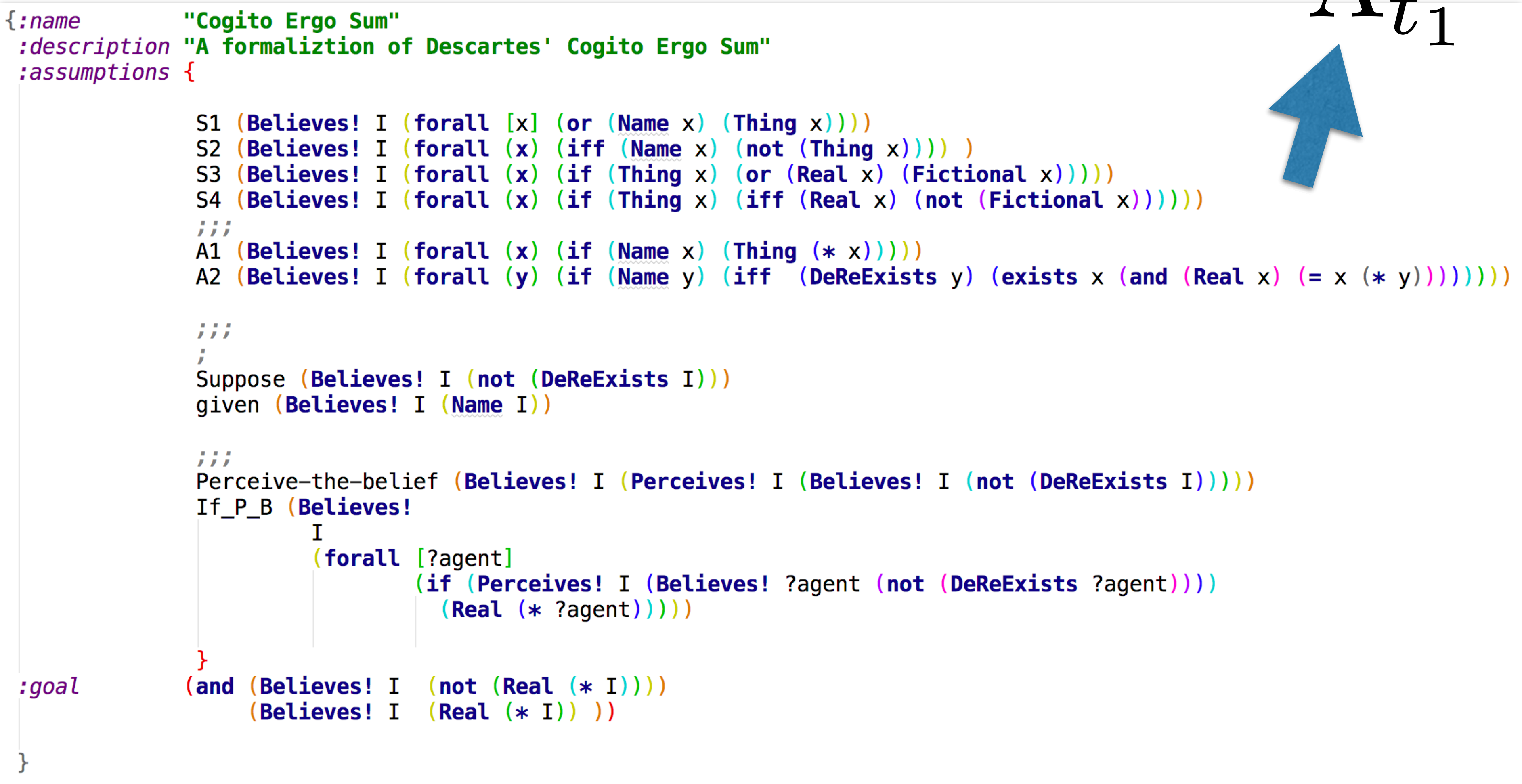


```
{:name        "Cogito Ergo Sum"
 :description "A formaliztion of Descartes' Cogito Ergo Sum"
 :assumptions {

               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}

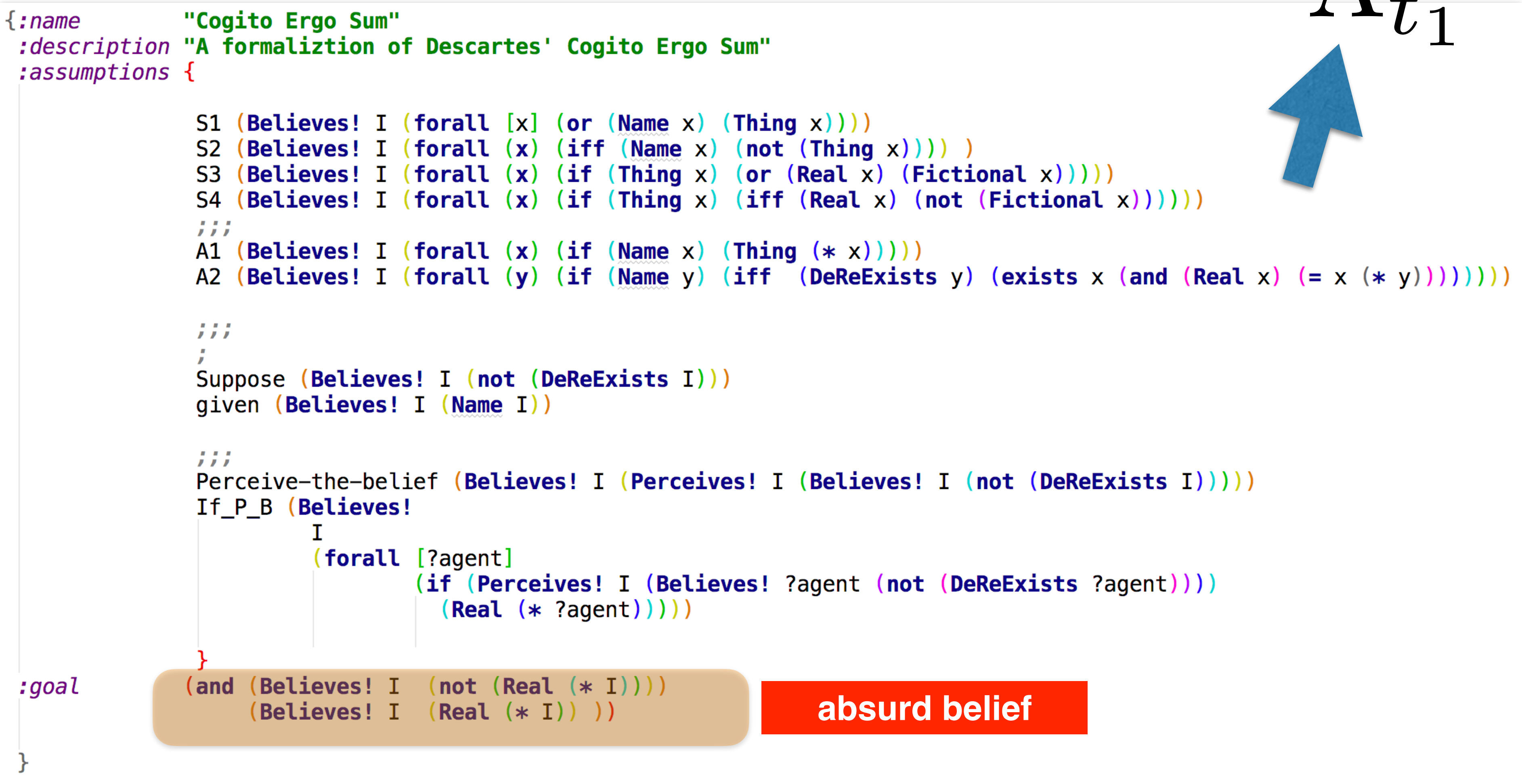


```
{:name        "Cogito Ergo Sum"
 :description "A formaliztion of Descartes' Cogito Ergo Sum"
 :assumptions {

              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}



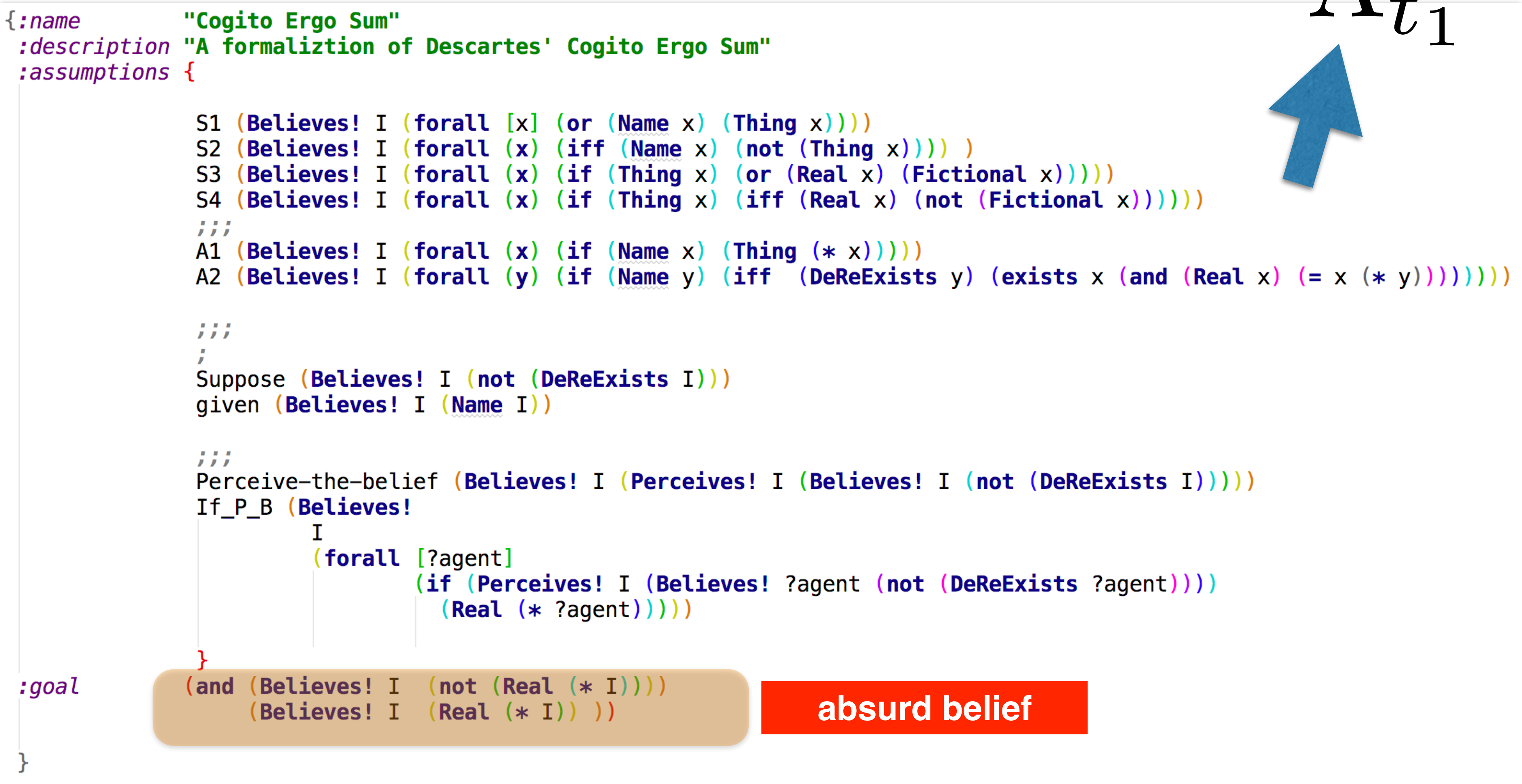
```
{:name        "Cogito Ergo Sum"
 :description "A formaliztion of Descartes' Cogito Ergo Sum"
 :assumptions {

              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

$$\vdots$$

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

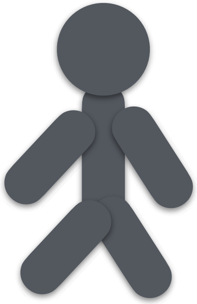
$\Lambda[\mathbf{I}, 4]$ Maximum difference between time expressions within intentions

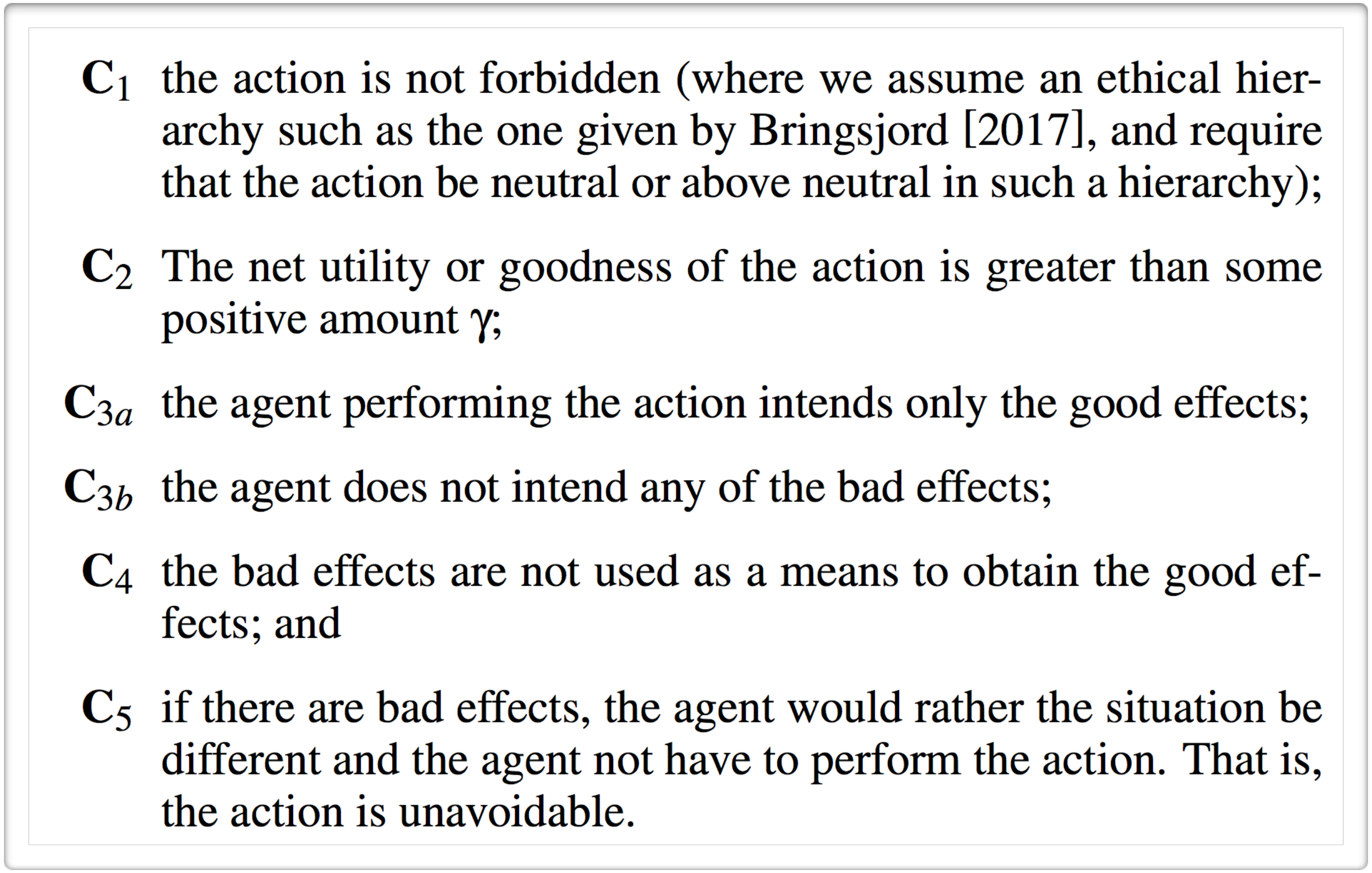
⋮

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;

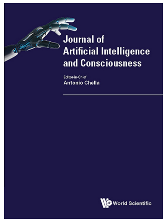
C_4 the bad effects are not used as a means to obtain the good effects; and

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

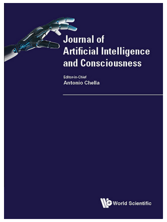
Editor-in-Chief
Antonio Chella

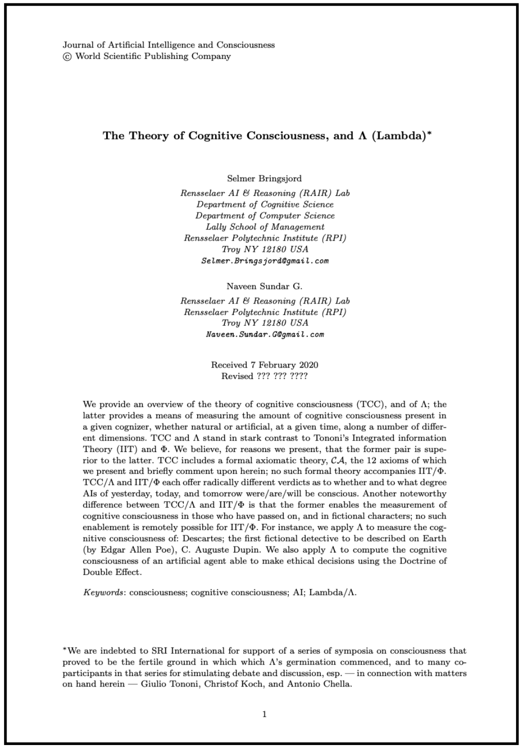
World Scientific

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
© World Scientific Publishing Company

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 ??? ??? ????

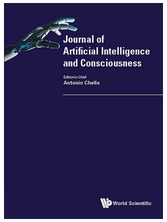
We provide an overview of the theory of cognitive consciousness (TCC), and of Λ; the latter provides a means of measuring the amount of cognitive consciousness present in a given cognizer, whether natural or artificial, at a given time, along a number of different dimensions. TCC and Λ stand in stark 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)

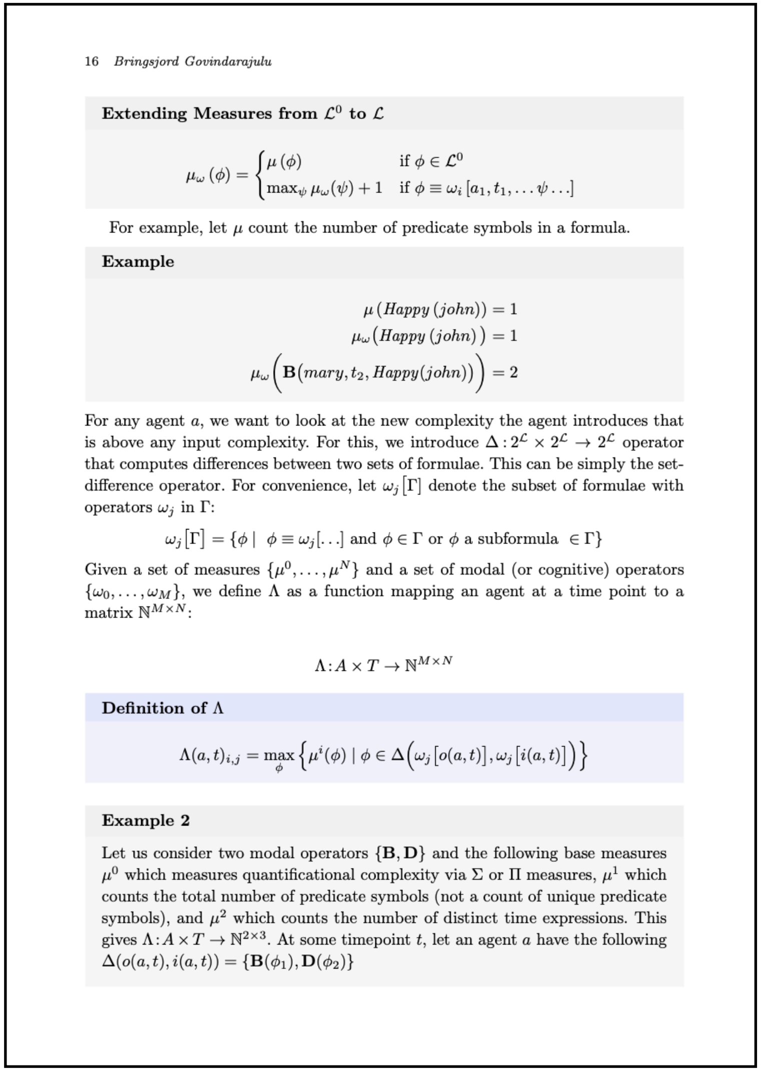
The Theory of Cognitive Consciousness, and Λ (Lambda)



Journal of Artificial Intelligence and Consciousness

Editor-in-Chief Antonio Chella

World Scientific



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

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

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

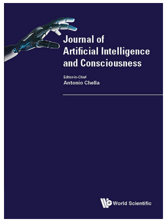
$$\Lambda(a,t)_{i,j} = \max_{\phi} \left\{ \mu^i(\phi) \mid \phi \in \Delta\Big(\omega_j\big[o(a,t)\big], \omega_j\big[i(a,t)\big]\Big)\right\}$$

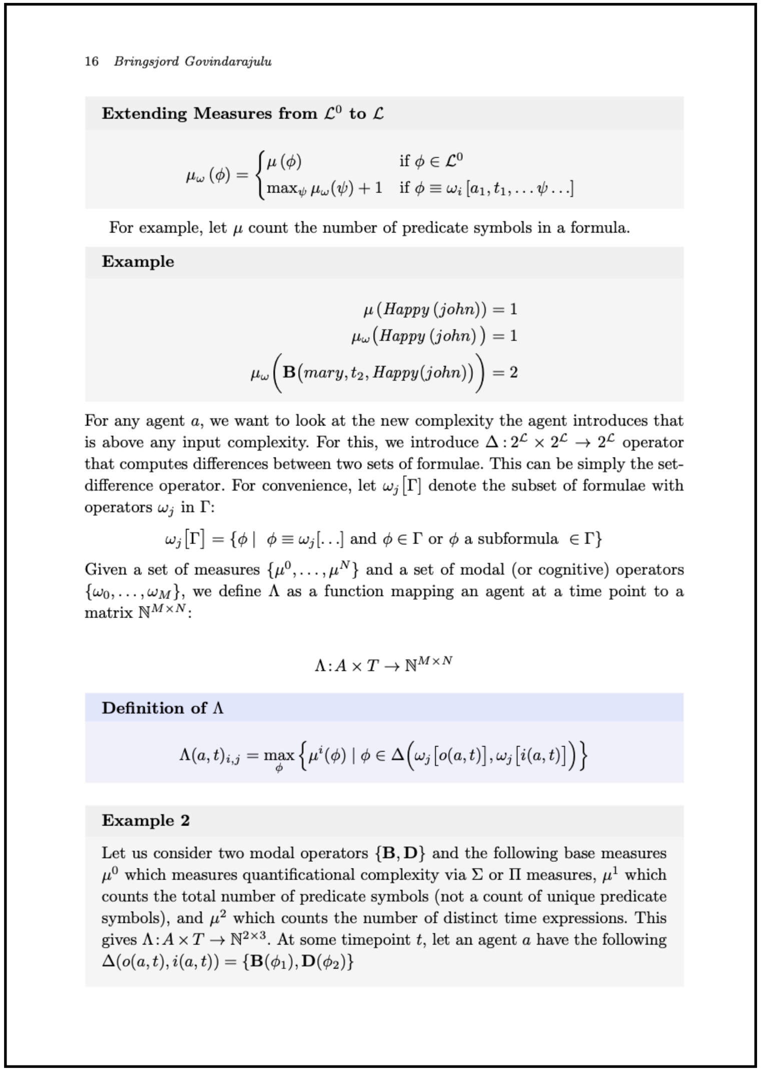
Example 2

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 $\Delta(o(a,t), i(a,t)) = \{\mathbf{B}(\phi_1), \mathbf{D}(\phi_2)\}$

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

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

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 $\mathbb{N}^{M \times N}$:

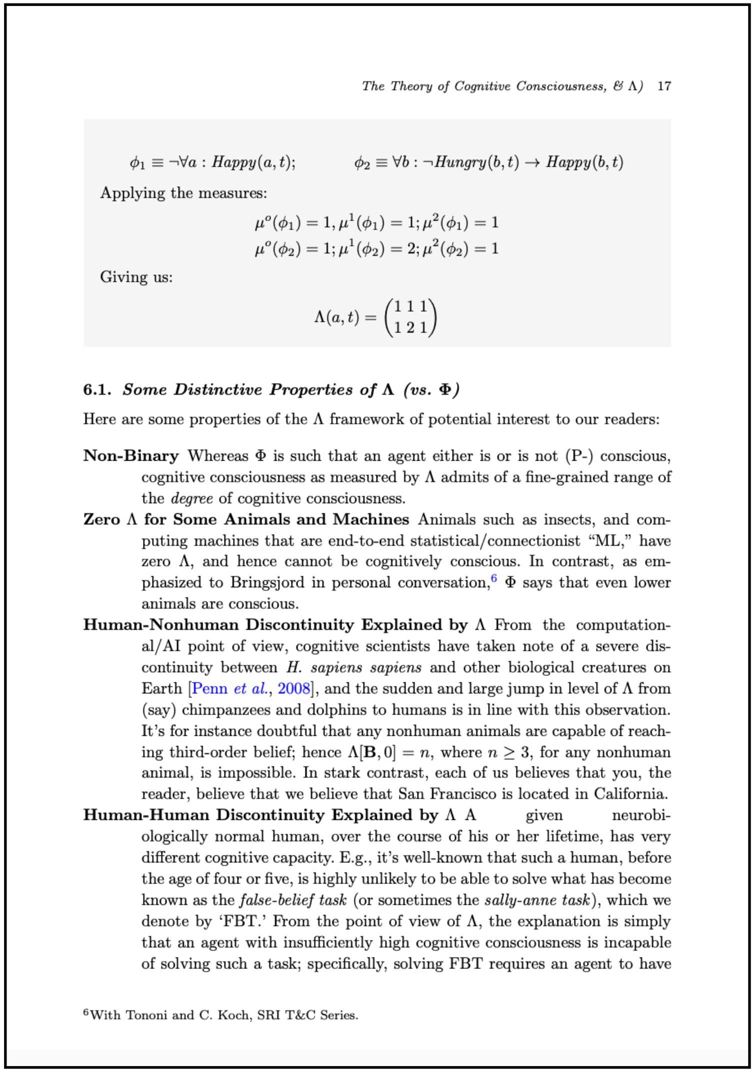
$$\Lambda : A \times T \rightarrow \mathbb{N}^{M \times N}$$

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 $\Delta(o(a,t), i(a,t)) = \{\mathbf{B}(\phi_1), \mathbf{D}(\phi_2)\}$



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

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:

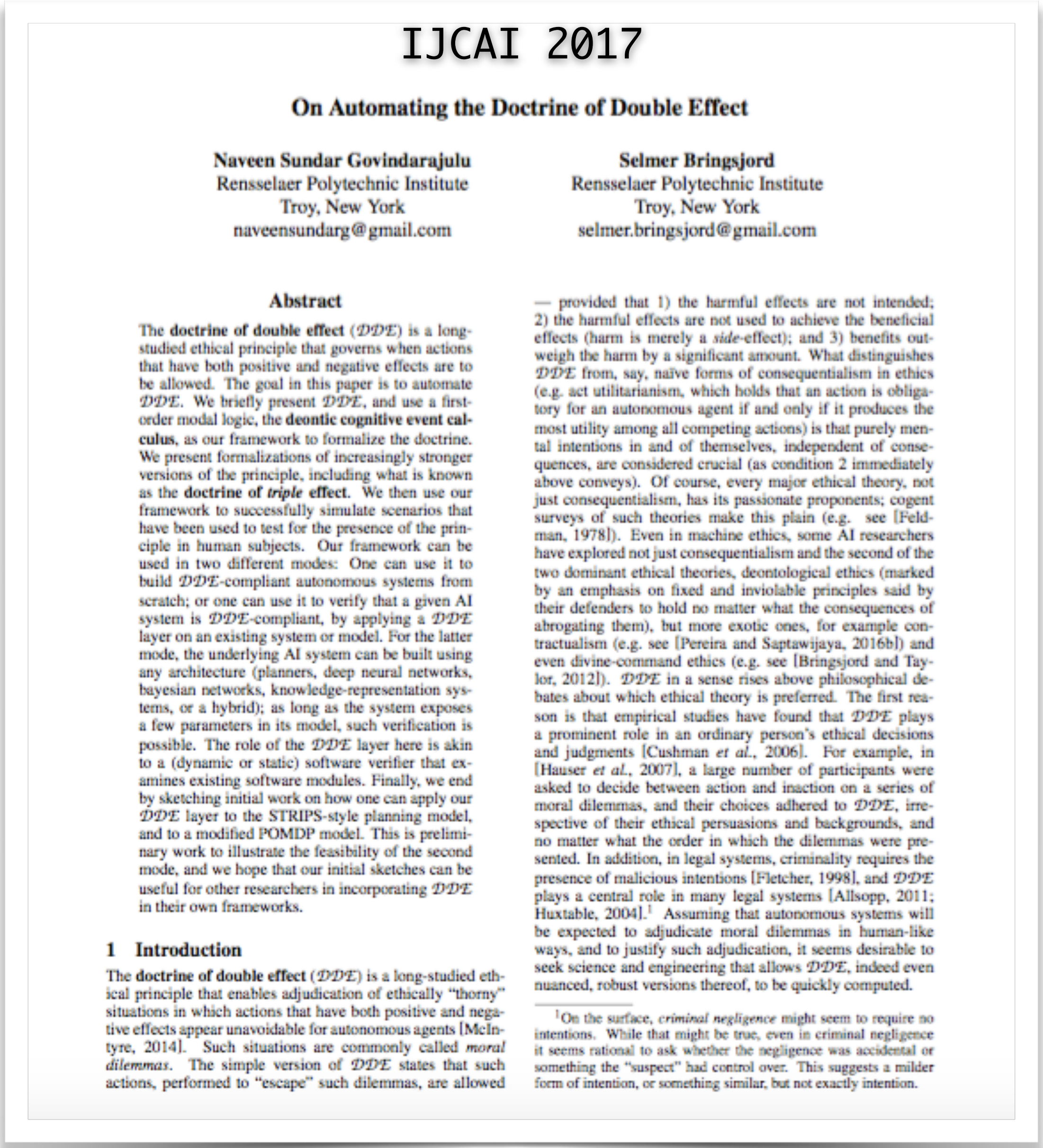
Non-Binary Whereas Φ is such that an agent either is or is not (P-) conscious, cognitive consciousness as measured by Λ admits of a fine-grained range of the *degree* of cognitive consciousness.

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.



IJCAI 2017

On Automating the Doctrine of Double Effect

Naveen Sundar Govindarajulu
Rensselaer Polytechnic Institute
Troy, New York
naveensundarg@gmail.com

Selmer Bringsjord
Rensselaer Polytechnic Institute
Troy, New York
selmer.bringsjord@gmail.com

Abstract

The **doctrine of double effect** ($\mathcal{DDE}$) is a long-studied ethical principle that governs when actions that have both positive and negative effects are to be allowed. The goal in this paper is to automate $\mathcal{DDE}$. We briefly present $\mathcal{DDE}$, and use a first-order modal logic, the **deontic cognitive event calculus**, as our framework to formalize the doctrine. We present formalizations of increasingly stronger versions of the principle, including what is known as the **doctrine of *triple* effect.** We then use our framework to successfully simulate scenarios that have been used to test for the presence of the principle in human subjects. Our framework can be used in two different modes: One can use it to build $\mathcal{DDE}$-compliant autonomous systems from scratch; or one can use it to verify that a given AI system is $\mathcal{DDE}$-compliant, by applying a $\mathcal{DDE}$ layer on an existing system or model. For the latter mode, the underlying AI system can be built using any architecture (planners, deep neural networks, bayesian networks, knowledge-representation systems, or a hybrid); as long as the system exposes a few parameters in its model, such verification is possible. The role of the $\mathcal{DDE}$ layer here is akin to a (dynamic or static) software verifier that examines existing software modules. Finally, we end by sketching initial work on how one can apply our $\mathcal{DDE}$ layer to the STRIPS-style planning model, and to a modified POMDP model. This is preliminary work to illustrate the feasibility of the second mode, and we hope that our initial sketches can be useful for other researchers in incorporating $\mathcal{DDE}$ in their own frameworks.

1 Introduction

The **doctrine of double effect** ($\mathcal{DDE}$) is a long-studied ethical principle that enables adjudication of ethically "thorny" situations in which actions that have both positive and negative effects appear unavoidable for autonomous agents [McIntyre, 2014]. Such situations are commonly called *moral dilemmas*. The simple version of $\mathcal{DDE}$ states that such actions, performed to "escape" such dilemmas, are allowed — provided that 1) the harmful effects are not intended; 2) the harmful effects are not used to achieve the beneficial effects (harm is merely a *side*-effect); and 3) benefits outweigh the harm by a significant amount. What distinguishes $\mathcal{DDE}$ from, say, naïve forms of consequentialism in ethics (e.g. act utilitarianism, which holds that an action is obligatory for an autonomous agent if and only if it produces the most utility among all competing actions) is that purely mental intentions in and of themselves, independent of consequences, are considered crucial (as condition 2 immediately above conveys). Of course, every major ethical theory, not just consequentialism, has its passionate proponents; cogent surveys of such theories make this plain (e.g. see [Feldman, 1978]). Even in machine ethics, some AI researchers have explored not just consequentialism and the second of the two dominant ethical theories, deontological ethics (marked by an emphasis on fixed and inviolable principles said by their defenders to hold no matter what the consequences of abrogating them), but more exotic ones, for example contractualism (e.g. see [Pereira and Saptawijaya, 2016b]) and even divine-command ethics (e.g. see [Bringsjord and Taylor, 2012]). $\mathcal{DDE}$ in a sense rises above philosophical debates about which ethical theory is preferred. The first reason is that empirical studies have found that $\mathcal{DDE}$ plays a prominent role in an ordinary person's ethical decisions and judgments [Cushman *et al.*, 2006]. For example, in [Hauser *et al.*, 2007], a large number of participants were asked to decide between action and inaction on a series of moral dilemmas, and their choices adhered to $\mathcal{DDE}$, irrespective of their ethical persuasions and backgrounds, and no matter what the order in which the dilemmas were presented. In addition, in legal systems, criminality requires the presence of malicious intentions [Fletcher, 1998], and $\mathcal{DDE}$ plays a central role in many legal systems [Allsopp, 2011; Huxtable, 2004].[1] Assuming that autonomous systems will be expected to adjudicate moral dilemmas in human-like ways, and to justify such adjudication, it seems desirable to seek science and engineering that allows $\mathcal{DDE}$, indeed even nuanced, robust versions thereof, to be quickly computed.

[1] On the surface, *criminal negligence* might seem to require no intentions. While that might be true, even in criminal negligence it seems rational to ask whether the negligence was accidental or something the "suspect" had control over. This suggests a milder form of intention, or something similar, but not exactly intention.

Formal Conditions for $\mathcal{DDE}$

F$_1$ α carried out at t is not forbidden. That is:

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

F$_2$ The net utility is greater than a given positive real γ:

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

F$_{3a}$ The agent a intends at least one good effect. (**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:

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

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:

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

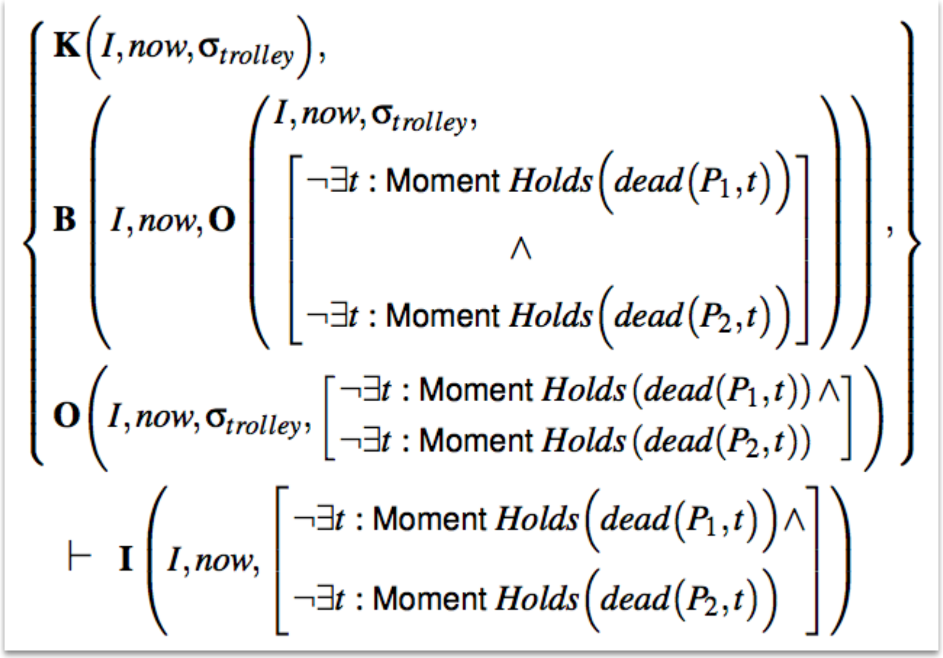
$$\Gamma \nvdash \mathbf{I}\Big(a, t, \neg Holds\big(f_g, y\big)\Big)$$

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



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

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

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

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

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

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

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

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

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

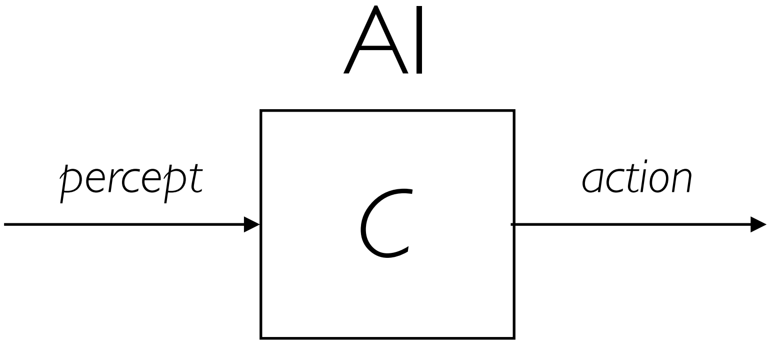
$\Lambda[\mathbf{B}, 3] = 1$

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

$\vdots$

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

AI:

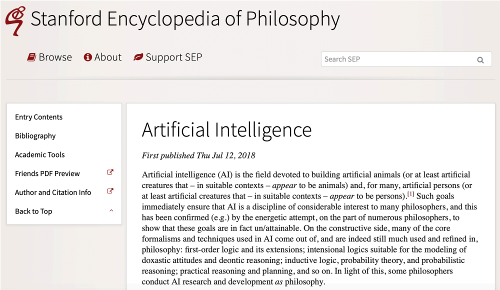


AI

percept

C

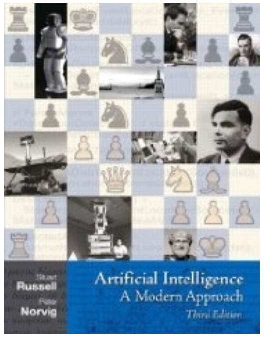
action



Artificial Intelligence

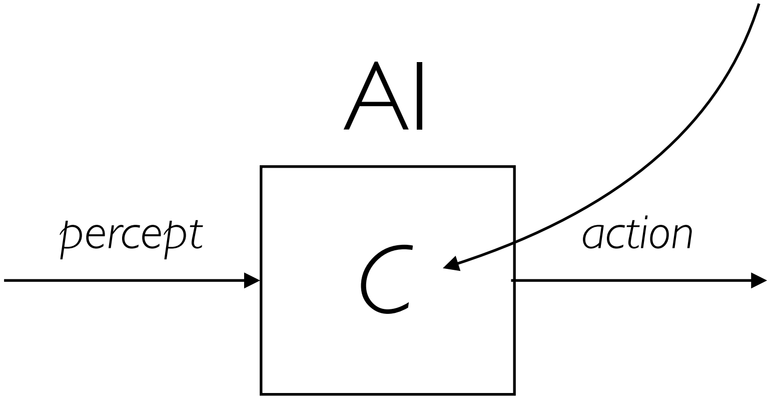
First published Thu Jul 12, 2018

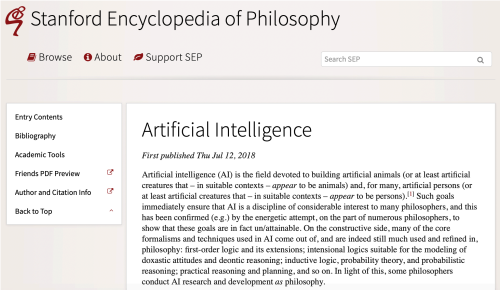
Artificial intelligence (AI) is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons).[1] Such goals immediately ensure that AI is a discipline of considerable interest to many philosophers, and this has been confirmed (e.g.) by the energetic attempt, on the part of numerous philosophers, to show that these goals are in fact un/attainable. On the constructive side, many of the core formalisms and techniques used in AI come out of, and are indeed still much used and refined in, philosophy: first-order logic and its extensions; intensional logics suitable for the modeling of doxastic attitudes and deontic reasoning; inductive logic, probability theory, and probabilistic reasoning; practical reasoning and planning, and so on. In light of this, some philosophers conduct AI research and development *as* philosophy.



AI:

A (Turing-level) entity that computes.




Stanford Encyclopedia of Philosophy

Browse About Support SEP

Search SEP

Entry Contents

Bibliography

Academic Tools

Friends PDF Preview

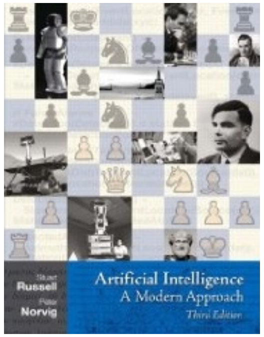
Author and Citation Info

Back to Top

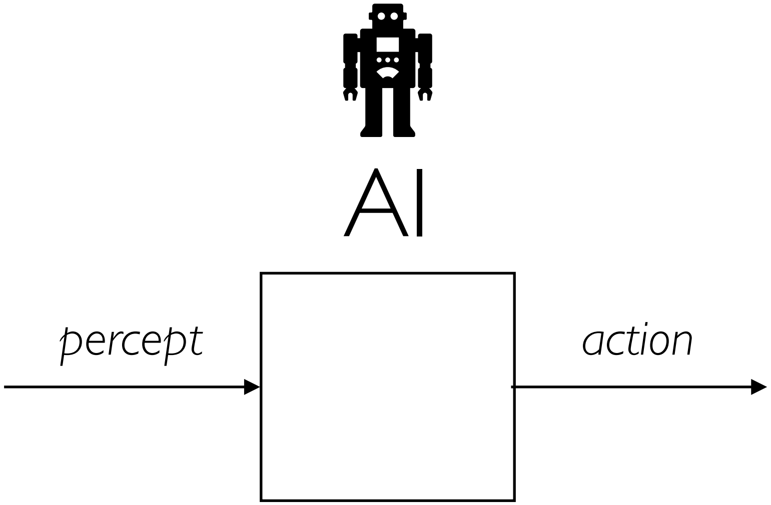
Artificial Intelligence

First published Thu Jul 12, 2018

Artificial intelligence (AI) is the field devoted to building artificial animals (or at least artificial creatures that – in suitable contexts – *appear* to be animals) and, for many, artificial persons (or at least artificial creatures that – in suitable contexts – *appear* to be persons).[1] Such goals immediately ensure that AI is a discipline of considerable interest to many philosophers, and this has been confirmed (e.g.) by the energetic attempt, on the part of numerous philosophers, to show that these goals are in fact un/attainable. On the constructive side, many of the core formalisms and techniques used in AI come out of, and are indeed still much used and refined in, philosophy: first-order logic and its extensions; intensional logics suitable for the modeling of doxastic attitudes and deontic reasoning; inductive logic, probability theory, and probabilistic reasoning; practical reasoning and planning, and so on. In light of this, some philosophers conduct AI research and development *as* philosophy.



AI:MLn

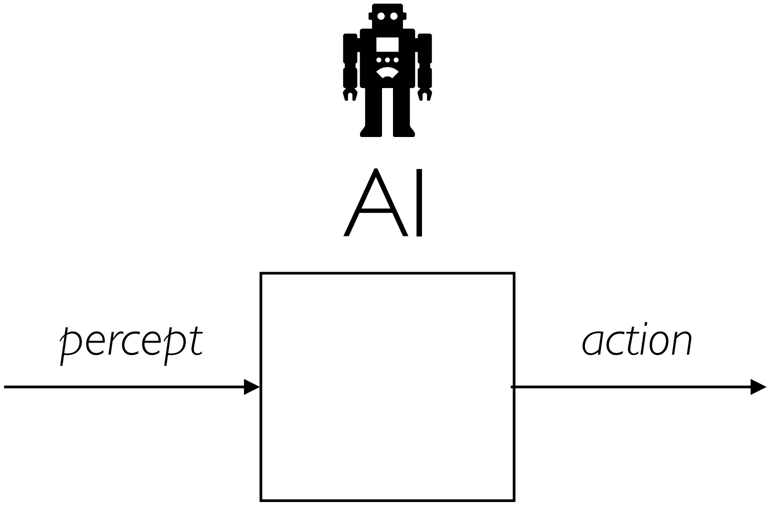


AI

percept

action

AI:MLn



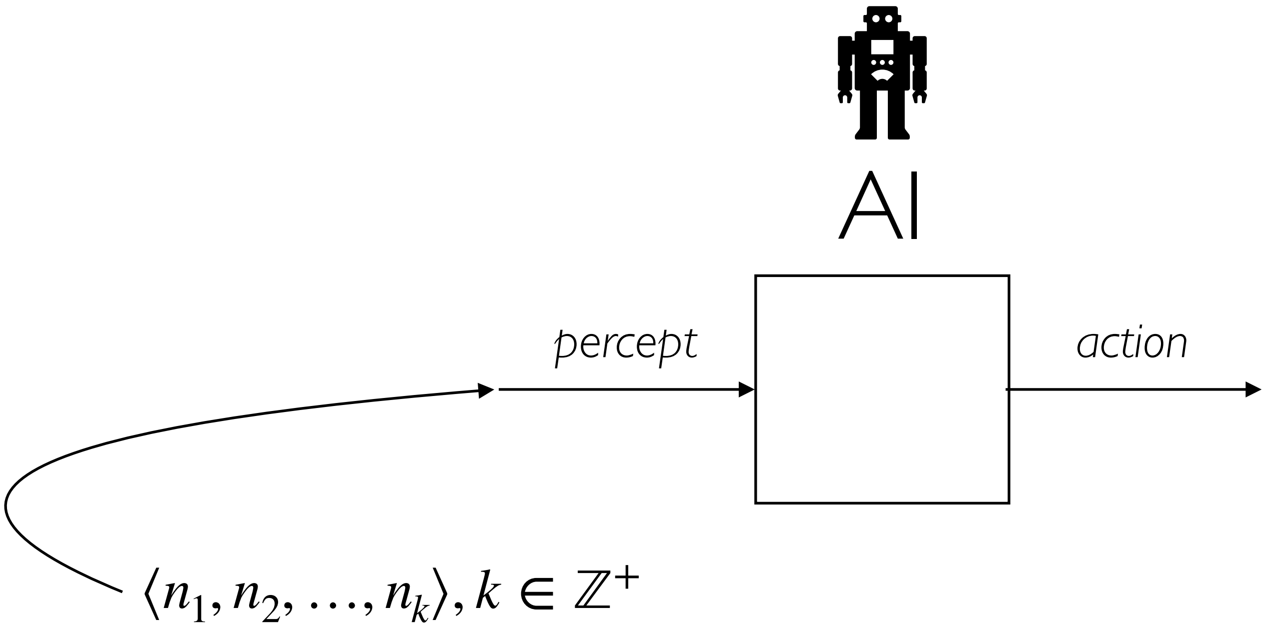
AI

percept

action

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

AI:MLn

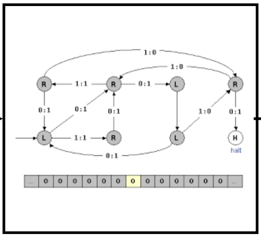


$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$

AI:MLn

AI

percept



action

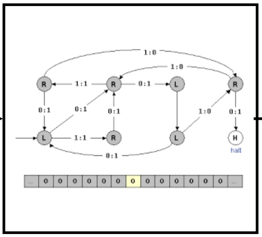
$\langle n_1, n_2, \ldots, n_k \rangle, k \in \mathbb{Z}^+$

AI:MLn

AI

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

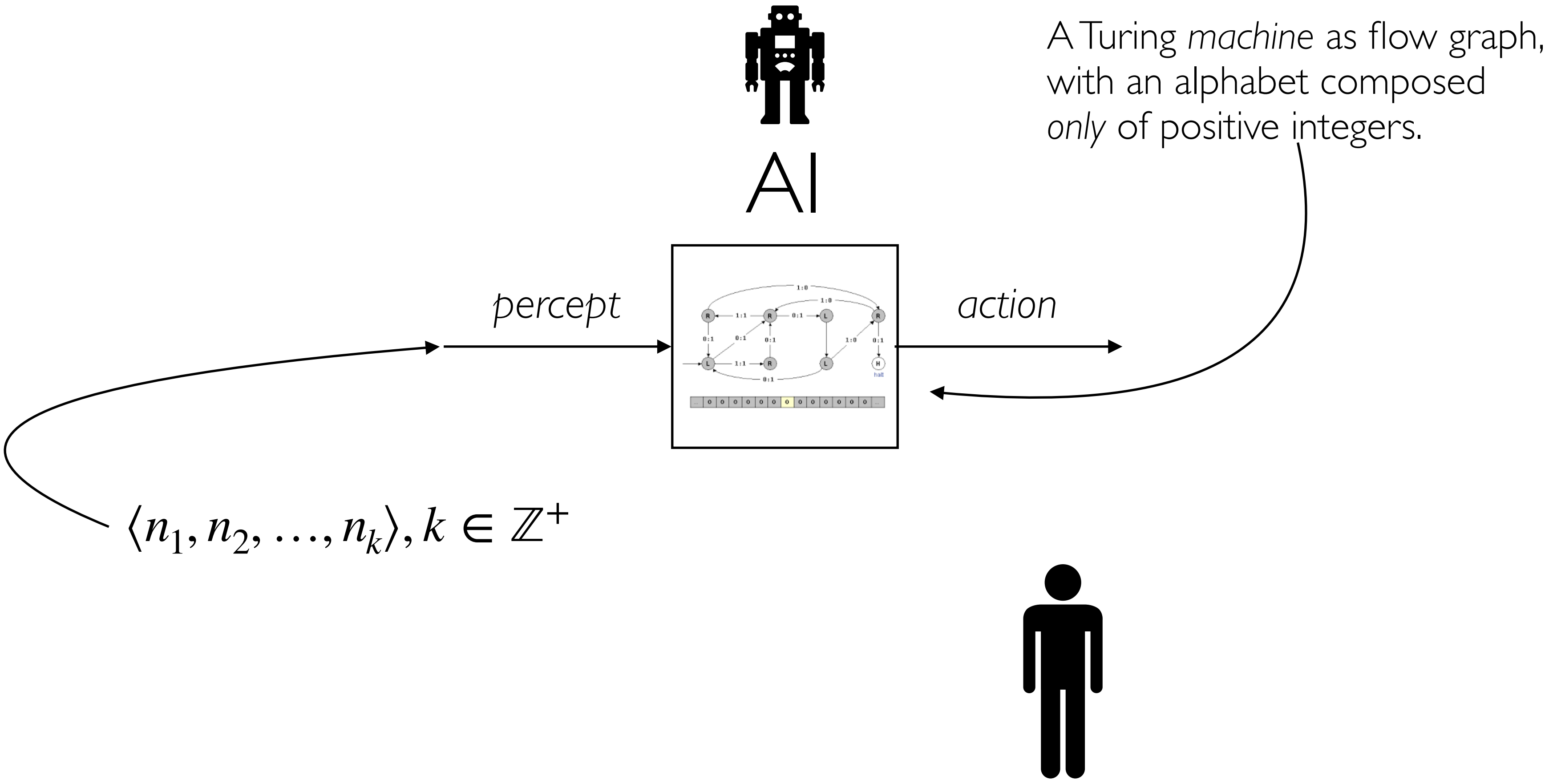
percept



action

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

AI:MLn



AI

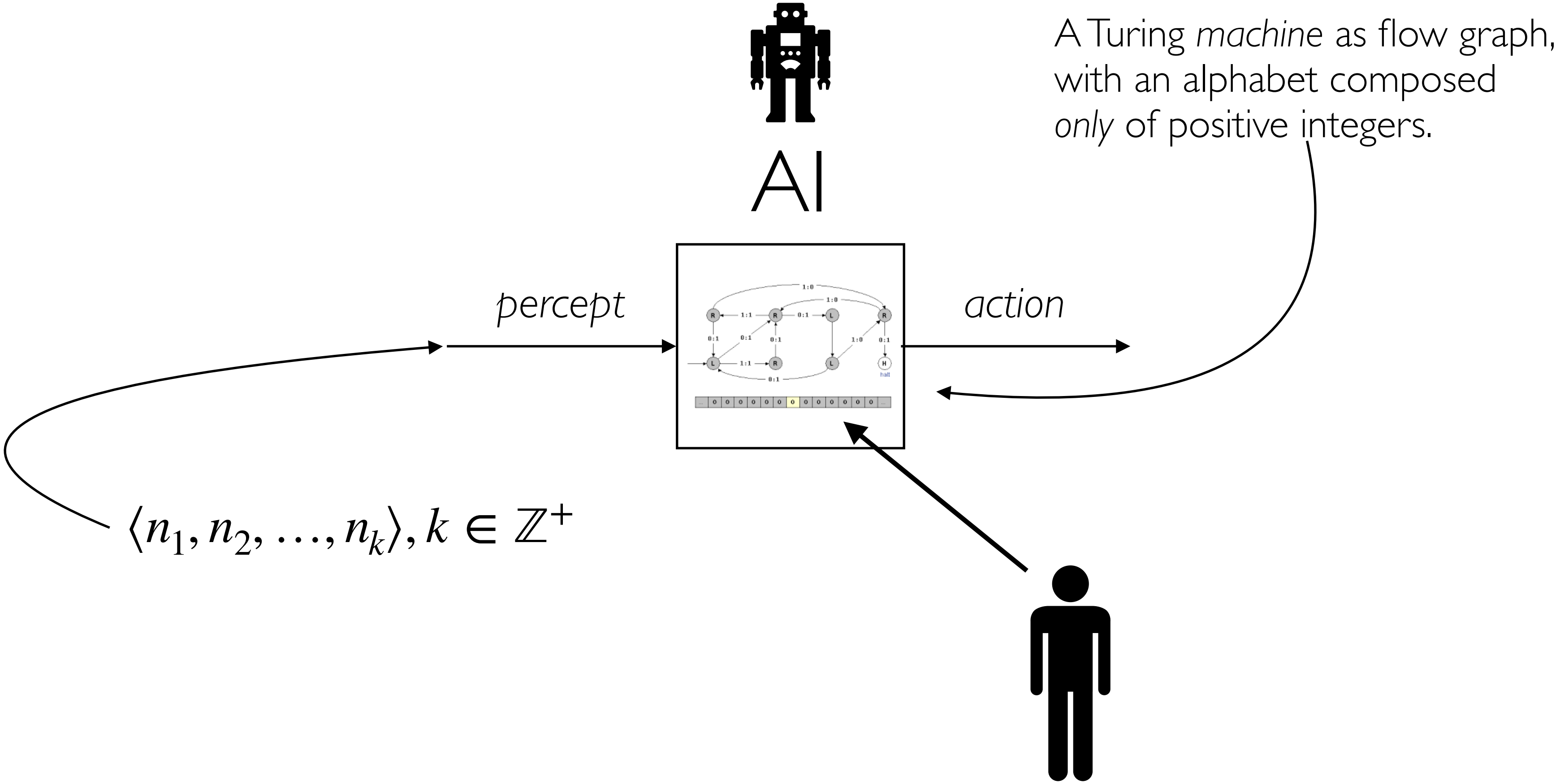
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.

percept

action

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

AI:MLn



AI

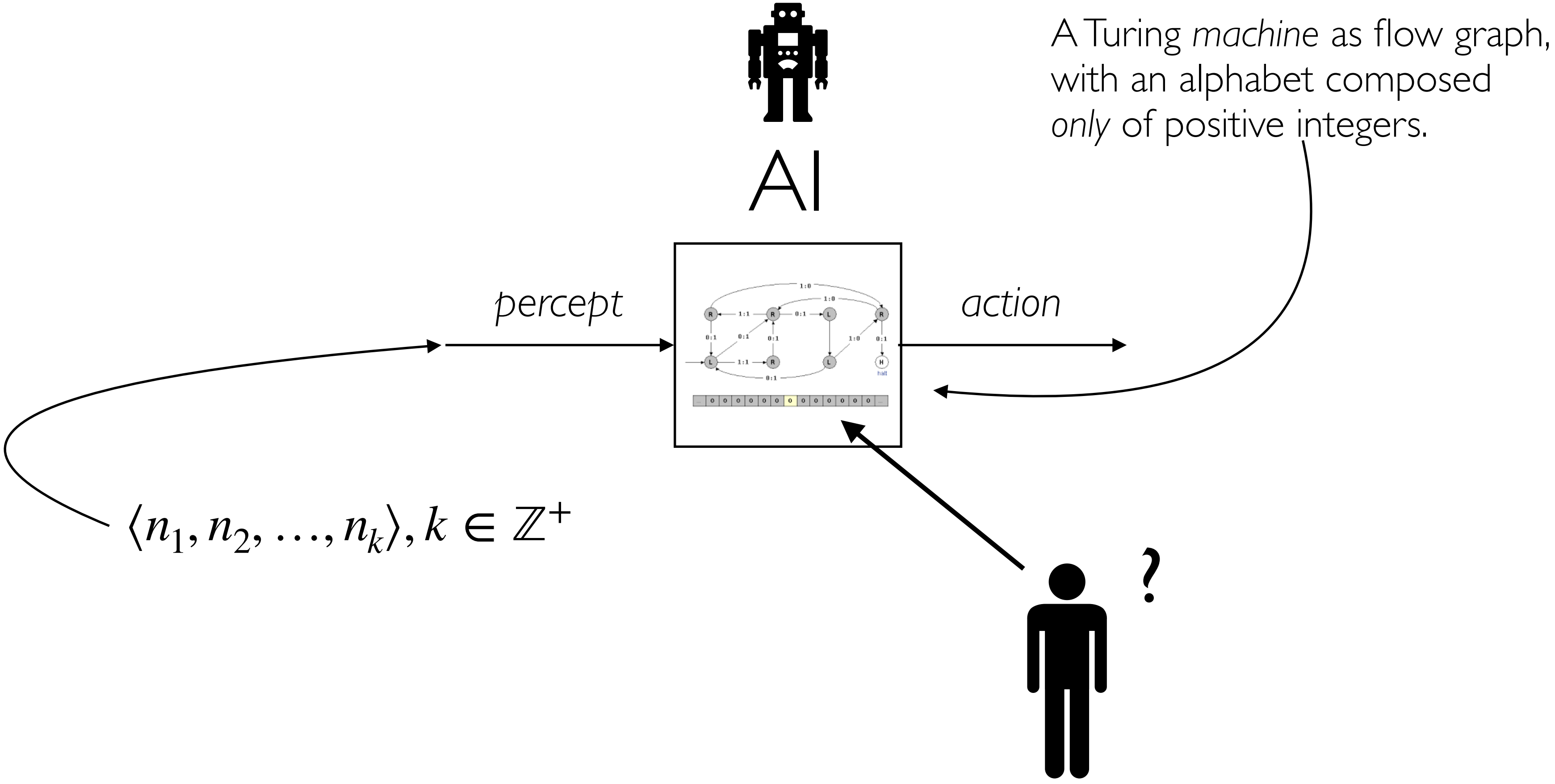
percept

action

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

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

AI:MLn



AI

percept

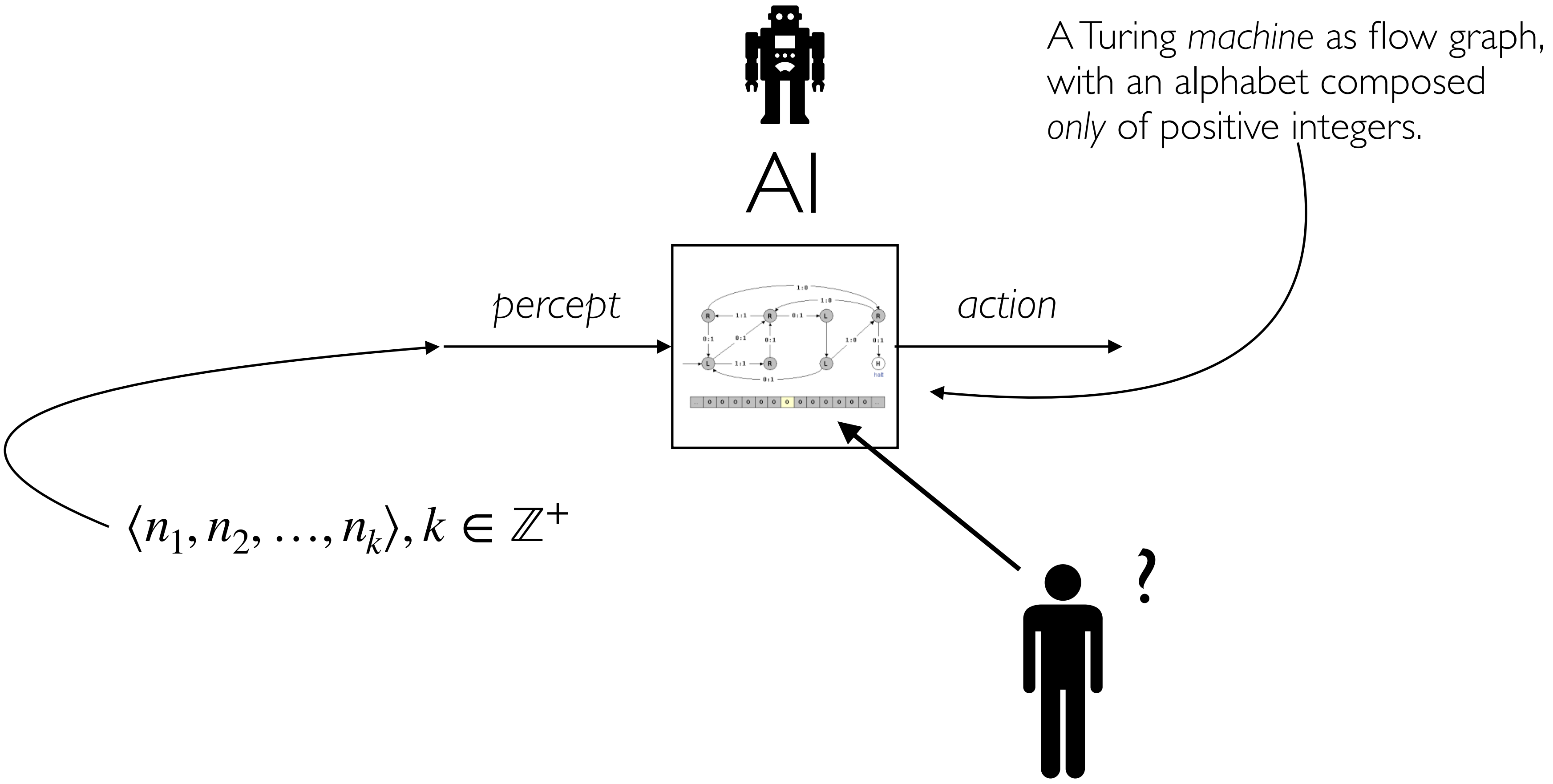
action

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

percept

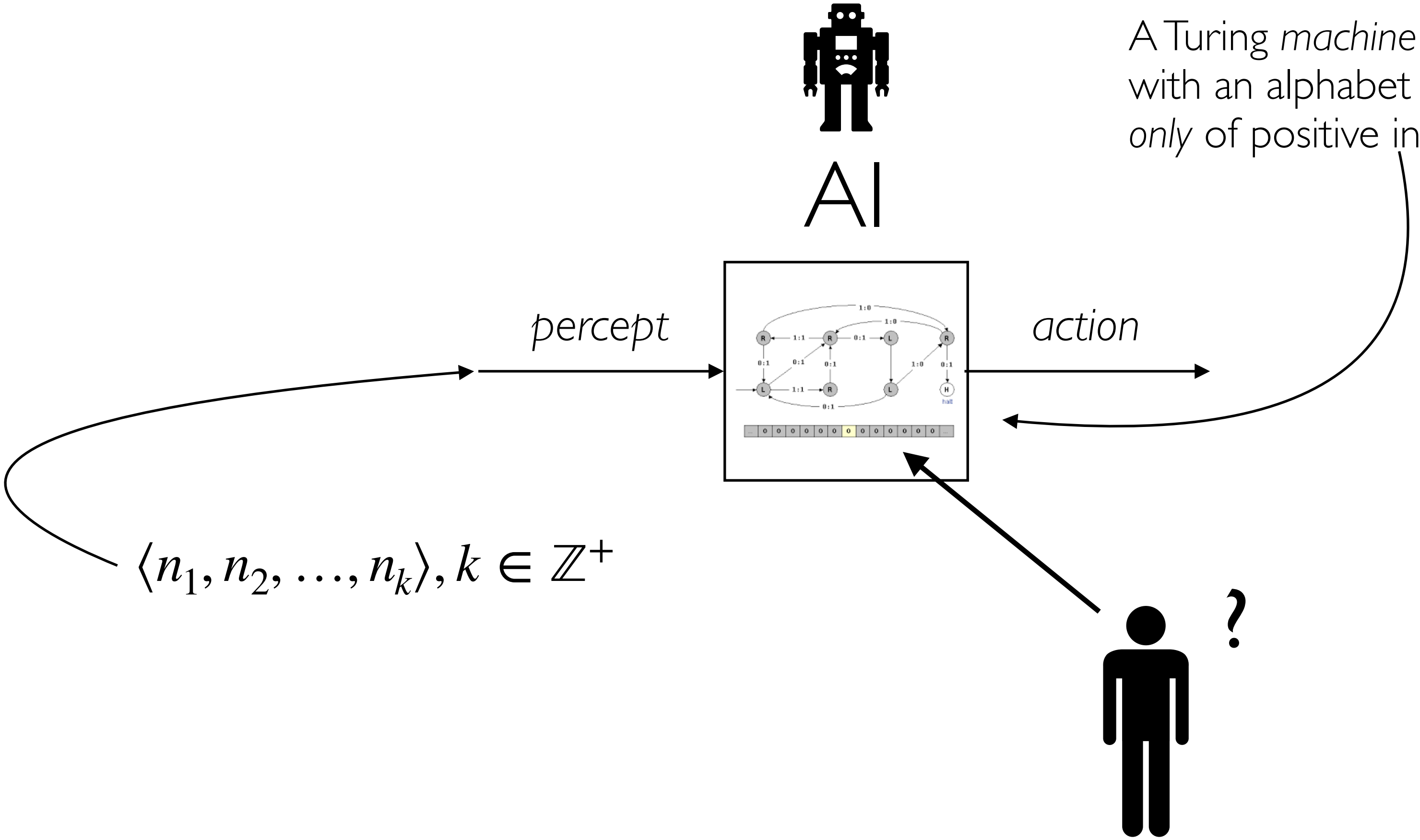
action

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

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

?

AI:MLn



AI

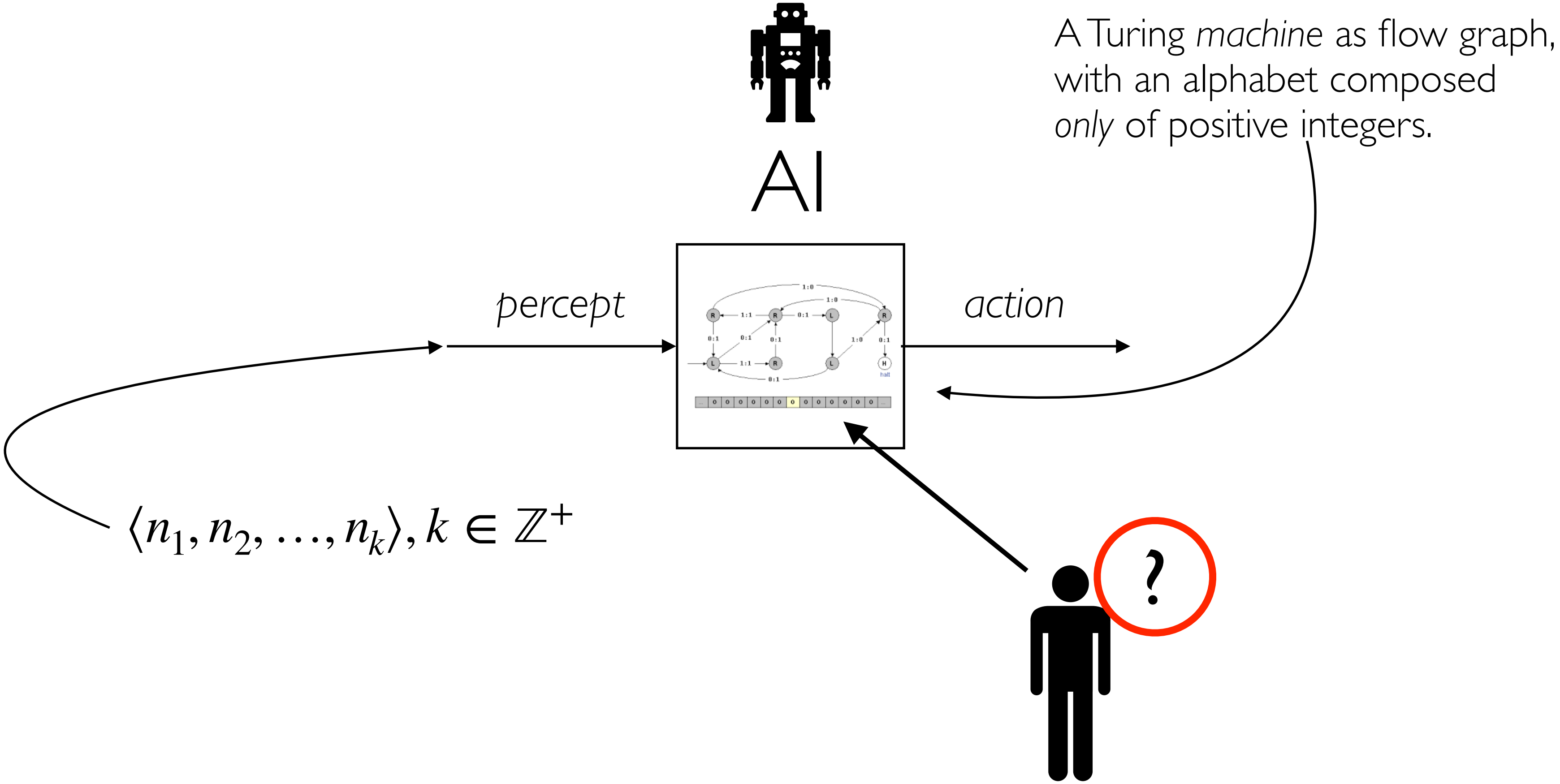
percept

action

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

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

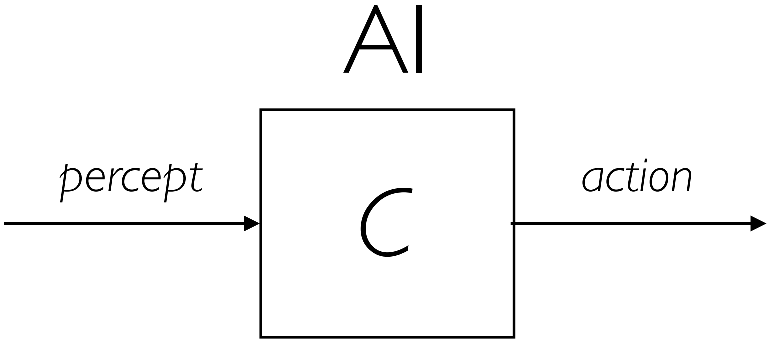
AI:MLn



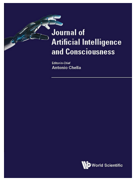
A Turing *machine* as flow graph, with an alphabet composed *only* of positive integers.

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

We will be able to measure the intelligence of *any* AI, not with *g*-loaded tests of intelligence, but with $\mathbf{\Lambda}$-loaded tests of machine intelligence, in keeping with Psychometric AI.



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



Plan

P2B

K2B

Intro

Incorr

Ess

¬CompE

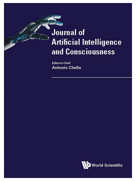
Irr

Free

CCaus

TheI

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



Journal of
Artificial Intelligence
and Consciousness

Editor-in-Chief
Antonio Chella

World Scientific

Plan

P2B

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

Intro

Incorr

Ess

¬CompE

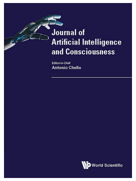
Irr

Free

CCaus

TheI

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



Plan

P2B

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

Intro

Incorr

Ess

¬CompE

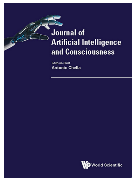
Irr

Free

CCaus

TheI

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



Journal of Artificial Intelligence and Consciousness

Editor-in-Chief Antonio Chella

World Scientific

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

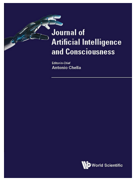
Irr

Free

CCaus

TheI

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


Journal of
Artificial Intelligence
and Consciousness

Editor-in-Chief
Antonio Chella

World Scientific

Plan

P2B

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

Intro

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

Ess

¬CompE

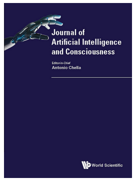
Irr

Free

CCaus

TheI

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



Journal of Artificial Intelligence and Consciousness

Editor-in-Chief
Antonio Chella

World Scientific

Plan

P2B

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

Intro

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

Ess

¬CompE

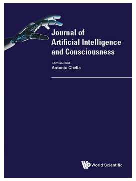
Irr

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

Ess

¬CompE

Irr

Free

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

TheI

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

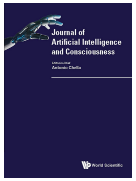
$[A_2]$ $\mathbf{C}(\forall\, e,f,t_1,t_2\ .\ happens(e,t_1) \wedge initiates(e,f,t_1) \wedge t_1 < t_2 \wedge \neg clipped(t_1,f,t_2) \Rightarrow holds(f,t_2))$

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

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

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

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


Journal of Artificial Intelligence and Consciousness

Editor-in-Chief
Antonio Chella

World Scientific

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

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

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

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

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

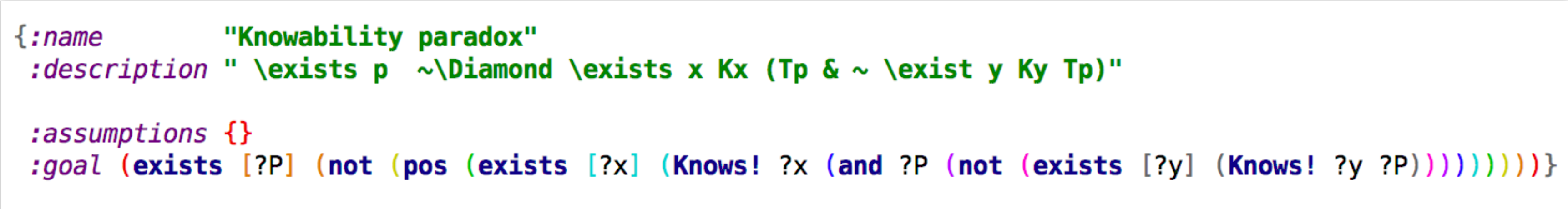
$[A_5]$ $\mathbf{C}(\forall\, a, f, t, t'\ .\ \mathbf{B}(a, holds(f, t)) \wedge \mathbf{B}(a, t < t') \wedge \neg\mathbf{B}(a, clipped(t, f, t')) \Rightarrow \mathbf{B}(a, holds(f, t')))$

$\mathbf{C}$ $\texttt{SpecRel}$

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

TheI

Example

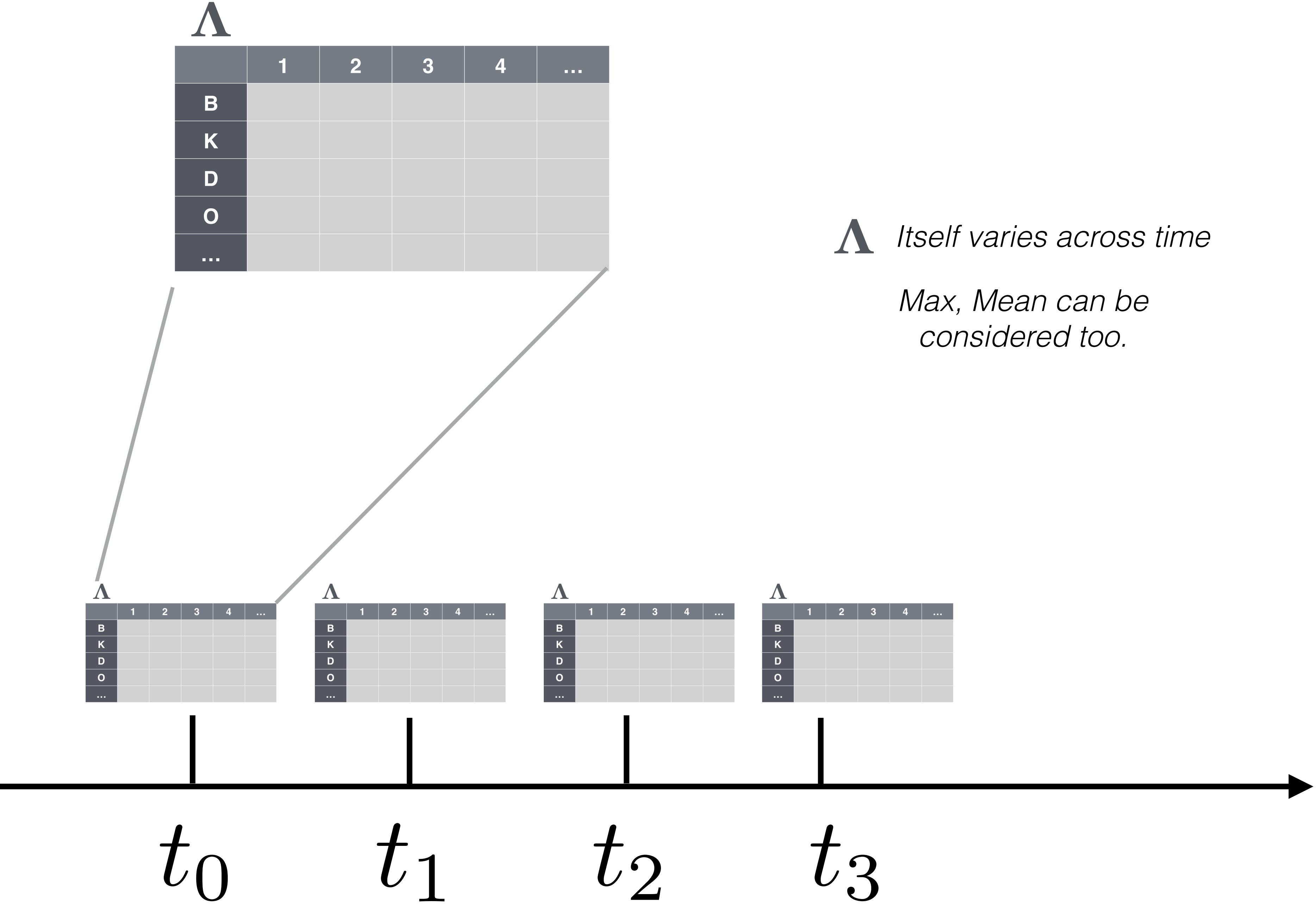


```
{:name         "Knowability paradox"
 :description " \exists p  ~\Diamond \exists x Kx (Tp & ~ \exist y Ky Tp)"

 :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

	1	2	3	4	...
B					
K					
D					
O					
...					

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

“**Theorem**”: C-con., as measured by Λ, unlike P-con. as measured by Φ, is *dis*continuous.

Discussion & Debate …

Med nok penger, kan logikk løse alle våre problemer.