

AI and Consciousness; Personalized Problems Have Started; Discussion of Final Projects

Selmer Bringsjord

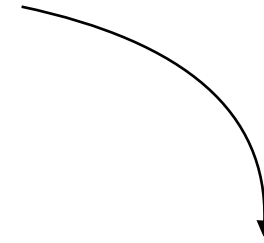
Rensselaer AI & Reasoning (RAIR) Lab
Department of Cognitive Science
Department of Computer Science
Lally School of Management & Technology
Rensselaer Polytechnic Institute (RPI)
Troy, New York 12180 USA

Intro to Logic-based AI
10/19/2026

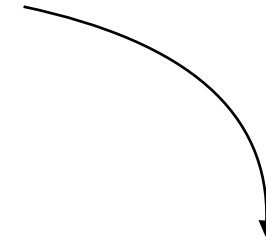


“Consciousness”

“Consciousness”

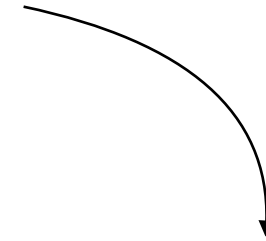


“Consciousness”



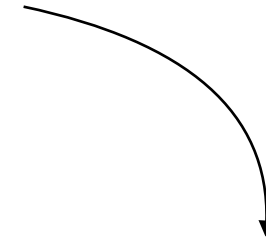
Phenomenal Consciousness

“Consciousness”



Phenomenal Consciousness

“Consciousness”

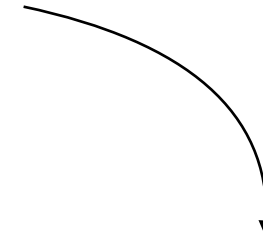


Phenomenal Consciousness



Minuro Osada

“Consciousness”

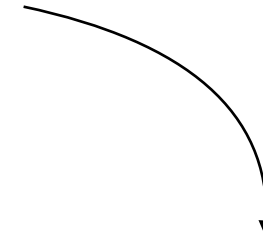


Minuro Osada

Phenomenal Consciousness

Third-person formalization impossible.

“Consciousness”



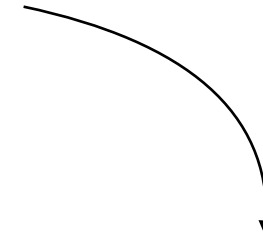
Minuro Osada

Phenomenal Consciousness

Third-person formalization impossible.

Φ

“Consciousness”



Minuro Osada

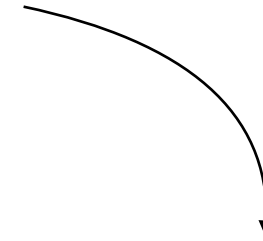
Phenomenal Consciousness

Third-person formalization impossible.

Φ

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

“Consciousness”



Minuro Osada

Phenomenal Consciousness

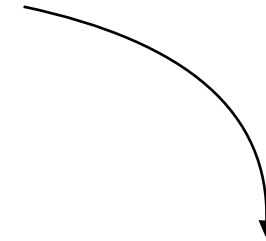
Third-person formalization impossible.

Φ

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



“Consciousness”



Minuro Osada

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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



“Consciousness”



Minuro Osada

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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



“Consciousness”



Minuro Osada

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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



“Consciousness”

Cognitive Consciousness



Minuro Osada

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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



“Consciousness”

Cognitive Consciousness

Λ

Phenomenal Consciousness

Third-person formalization impossible.

Φ

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



Minuro Osada

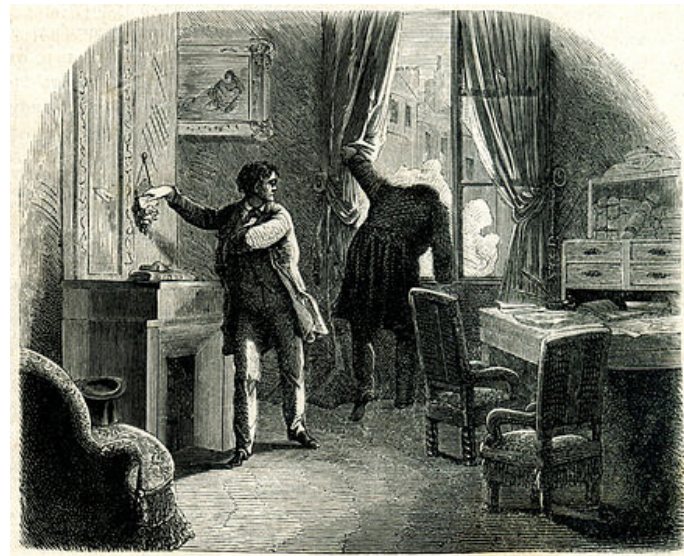


“Consciousness”

Cognitive Consciousness

Λ

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 (!).



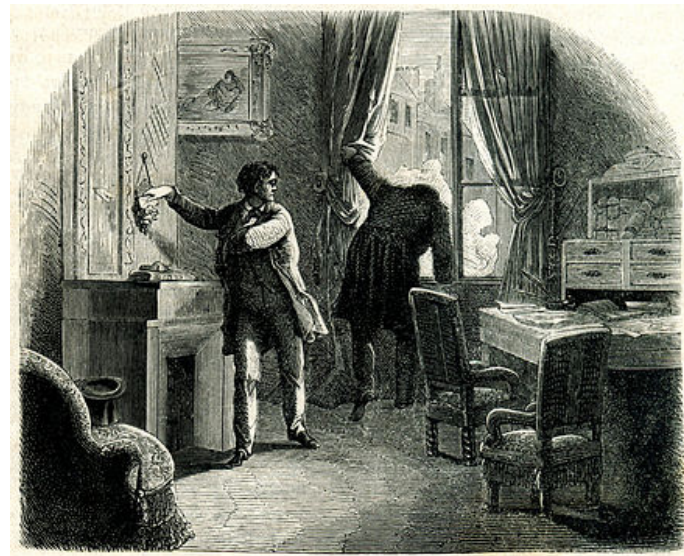
Minuro Osada

“Consciousness”

Cognitive Consciousness

Λ

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 (!).



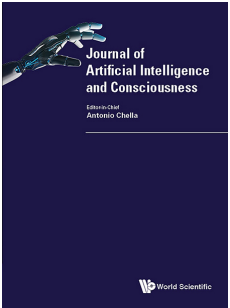
Minuro Osada

The Theory of Cognitive Consciousness, and Λ (Lambda)

Selmer Bringsjord  and G. Naveen Sundar

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, 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.

The Theory of Cognitive Consciousness, and Λ (Lambda)

16 Bringsjord Govindarajulu

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

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

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

Example

$$\begin{aligned} \mu(\text{Happy}(\text{john})) &= 1 \\ \mu_{\omega}(\text{Happy}(\text{john})) &= 1 \\ \mu_{\omega}(\mathbf{B}(\text{mary}, t_2, \text{Happy}(\text{john}))) &= 2 \end{aligned}$$

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 :

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

Given a set of measures $\{\mu^0, \dots, \mu^N\}$ and a set of modal (or cognitive) operators $\{\omega_0, \dots, \omega_M\}$, 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

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

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

The Theory of Cognitive Consciousness, and Λ (Lambda)

16 Bringsjord Govindarajulu

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

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

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

Example

$$\begin{aligned} \mu(\text{Happy}(\text{john})) &= 1 \\ \mu_\omega(\text{Happy}(\text{john})) &= 1 \\ \mu_\omega(\mathbf{B}(\text{mary}, t_2, \text{Happy}(\text{john}))) &= 2 \end{aligned}$$

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 :

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

Given a set of measures $\{\mu^0, \dots, \mu^N\}$ and a set of modal (or cognitive) operators $\{\omega_0, \dots, \omega_M\}$, 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

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

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

The Theory of Cognitive Consciousness, & Λ 17

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

Applying the measures:

$$\begin{aligned} \mu^0(\phi_1) &= 1, \mu^1(\phi_1) = 1; \mu^2(\phi_1) = 1 \\ \mu^0(\phi_2) &= 1; \mu^1(\phi_2) = 2; \mu^2(\phi_2) = 1 \end{aligned}$$

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

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

CA: 11 Axioms (Initially)

Plan

P2B

K2B

Intro

Incorr

Ess

$\neg$ CompE

Irr

Free

CCaus

Thel

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

$\neg$ CompE

Irr

Free

CCaus

Thel

CA: 11 Axioms (Initially)

Plan

P2B

$\mu DCEC_3^*$

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

$\neg\text{CompE}$

Irr

Free

CCaus

Thel

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

$\neg$ CompE

Irr

Free

CCaus

Thel

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

Ess

$\neg$ CompE

Irr

Free

CCaus

Thel

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

Ess

$\neg$ CompE

Irr

Free

CCaus

C $\mathcal{EC}$

Thel

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

Ess

$\neg$ CompE

Irr

Free

CCaus

C $\mathcal{EC}$

Thel

[A₁] $\mathbf{C}(\forall f, t . \text{initially}(f) \wedge \neg \text{clipped}(0, f, t) \Rightarrow \text{holds}(f, t))$

[A₂] $\mathbf{C}(\forall e, f, t_1, t_2 . \text{happens}(e, t_1) \wedge \text{initiates}(e, f, t_1) \wedge t_1 < t_2 \wedge \neg \text{clipped}(t_1, f, t_2) \Rightarrow \text{holds}(f, t_2))$

[A₃] $\mathbf{C}(\forall t_1, f, t_2 . \text{clipped}(t_1, f, t_2) \Leftrightarrow [\exists e, t . \text{happens}(e, t) \wedge t_1 < t < t_2 \wedge \text{terminates}(e, f, t)])$

[A₄] $\mathbf{C}(\forall a, d, t . \text{happens}(\text{action}(a, d), t) \Rightarrow \mathbf{K}(a, \text{happens}(\text{action}(a, d), t)))$

[A₅] $\mathbf{C}(\forall a, f, t, t' . \mathbf{B}(a, \text{holds}(f, t)) \wedge \mathbf{B}(a, t < t') \wedge \neg \mathbf{B}(a, \text{clipped}(t, f, t')) \Rightarrow \mathbf{B}(a, \text{holds}(f, t')))$

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

Ess

$\neg$ CompE

Irr

Free

C SpecRel

CCaus

C $\mathcal{EC}$

Thel

[A₁] $\mathbf{C}(\forall f, t . \text{initially}(f) \wedge \neg \text{clipped}(0, f, t) \Rightarrow \text{holds}(f, t))$

[A₂] $\mathbf{C}(\forall e, f, t_1, t_2 . \text{happens}(e, t_1) \wedge \text{initiates}(e, f, t_1) \wedge t_1 < t_2 \wedge \neg \text{clipped}(t_1, f, t_2) \Rightarrow \text{holds}(f, t_2))$

[A₃] $\mathbf{C}(\forall t_1, f, t_2 . \text{clipped}(t_1, f, t_2) \Leftrightarrow [\exists e, t . \text{happens}(e, t) \wedge t_1 < t < t_2 \wedge \text{terminates}(e, f, t)])$

[A₄] $\mathbf{C}(\forall a, d, t . \text{happens}(\text{action}(a, d), t) \Rightarrow \mathbf{K}(a, \text{happens}(\text{action}(a, d), t)))$

[A₅] $\mathbf{C}(\forall a, f, t, t' . \mathbf{B}(a, \text{holds}(f, t)) \wedge \mathbf{B}(a, t < t') \wedge \neg \mathbf{B}(a, \text{clipped}(t, f, t')) \Rightarrow \mathbf{B}(a, \text{holds}(f, t')))$

Where should my *reading* start?

24. CAN CONSCIOUSNESS BE EXPLAINED BY INTEGRATED INFORMATION THEORY OR THE THEORY OF COGNITIVE CONSCIOUSNESS?¹

Selmer Bringsjord and Naveen Sundar Govindarajulu

1. Introduction

AS READERS will doubtless have noted by now, some other chapters in the present volume have expressed the view (rather agreeable to us) that many aspects of human-level mental phenomena are recalcitrant to a mindset that insists upon mathematical and (usually) material explanations. First-person subjectivity, intentionality, mathematical cognition, robust epistemic states, consciousness. . . these phenomena are exceedingly hard to explain in such a manner. It is the final member of that list of challenges that is our focus in the present chapter. Can science operating in the math-and-material manner explain—and perhaps even, courtesy of associated engineering, replicate in artificial agents—consciousness?

This question is now pressed upon at least all technologized societies on Earth, because of the advent of artificial agents able to converse in seemingly flawless English about pretty much anything, including consciousness itself.

A famous example is ChatGPT. This class of agents falls into what is now called “generative AI,” which includes agents not only able to generate natural language, but also images. In the case of language, these agents are sometimes called “chatbots,” but are more precisely known as “Large Language Models.” Some of these agents have been declared conscious,² and the question of whether they are is really just a special case of the general question taken up in the present chapter. We are very confident that ascriptions of consciousness to artificial agents are only going to grow in frequency, and such ascriptions are going to increasingly be issued by voices that seem balanced and authoritative. This chapter should in our opinion be read and understood by those humans who will find themselves living in the trend we foresee, because it provides at least a starting basis for two fundamental ways of looking not just at consciousness in general, but consciousness in computational artifacts.

Blake Lemoine: Google fires engineer who said AI tech has feelings

22 July 2022

Share < Save □

Tiffany Wertheimer
BBC News



The Washington Post/ Getty Images

Blake Lemoine photographed in San Francisco last month

Google has fired one of its engineers who said the company's artificial intelligence system has feelings.

Last month, Blake Lemoine went public with his theory that Google's language technology is sentient and should therefore have its "wants" respected.

Google, plus several AI experts, denied the claims and on Friday the company confirmed he had been sacked.

Mr Lemoine told the BBC he is getting legal advice, and declined to comment further.



TECHNOLOGY | ARTIFICIAL INTELLIGENCE | ESSAY

A 14-Year-Old Boy Killed Himself to Get Closer to a Chatbot. He Thought They Were In Love.

Technologists say chatbots are a remedy for the loneliness epidemic, but looking to an algorithm for companionship can be dangerous.

🏠 > iPhone > Lifestyle > Health & Fitness > Youper: Self-Guided Therapy



Youper: Self-Guided Therapy

for iPhone

Free In English V 12.07.001

★ 4.1 (0 ⚙️) 🛡️ Security Status

Softonic review

Youper: Self-Guided Therapy - A Scientifically Proven Mental Health App

Youper: Self-Guided Therapy is a free iPhone app developed by **Youper** Inc. It is a pocket counselor that guides users through interactive Cognitive Behavioral Therapy (CBT) exercises to help them calm anxiety, improve relationships, be more productive, and improve their mood. The app has been praised by Apple, Google, and Mind for its truly holistic approach to wellbeing.

The app uses CBT techniques, which are scientifically proven to calm anxiety and improve mood. CBT is based on practical exercises to help users think more clearly, deal with stressful situations, and take control of their emotions. A study conducted by **Stanford University** showed significant improvement in symptoms of depression and anxiety after using the **Youper** app.

🏠 > iPhone > Life



Softonic review

Youper: Self-Guided Proven Method

Youper: Self-Guided
is a pocket companion for Behavioral Therapy. It helps with relationships, mood, and is praised by Apple.

The app uses AI to help you and improve mood. It's more clearly, and a study conducted on symptoms of depression.

therapy

ically

Youper Inc. It

itive
improve
app has been
to wellbeing.

calm anxiety
users think
their emotions. A
movement in

How are you?

How can I support you today?

I need someone to care for me.
Really care for me.

It's completely understandable
to want that feeling of being
cared for and supported.

It sounds like you might be
feeling a bit lonely or in need of
connection right now.

Would you like to explore ways
to foster positive relationships in
your life?

What does care look like for
you?

🏠 > iPhone > Life



Softonic review

Youper: Self-Guided Proven Methods

Youper: Self-Guided Behavioral Therapy is a pocket companion app for improving relationships, mood, and overall well-being. Praised by Apple and Google Play, Youper is a powerful tool for mental health.

The app uses Cognitive Behavioral Therapy (CBT) and mindfulness techniques to help users understand their thoughts and feelings, and improve their mood. It's a great way to take control of your mental health and improve your overall well-being. The app is available on both iOS and Android, and is completely free to use.

therapy

ically

Youper Inc. It

itive
improve
app has been
to wellbeing.

calm anxiety
users think
their emotions. A
ovement in

I'm really sorry to hear that you're feeling this way.

It can be incredibly painful to feel uncared for or unnoticed.

Would you like to talk more about what's been happening that makes you feel this way?

No one cares about me.

I hear how deeply you're feeling this sense of being uncared for, and I'm here to support you.

It's tough to feel isolated, and it's important to acknowledge those feelings.

Can you share more about any specific experiences or thoughts that contribute to this feeling?

🏠 > iPhone > Life



Softonic review

Youper: Self-Guided Proven Mental Health

Youper: Self-Guided
is a pocket companion for Behavioral Therapy. It helps with relationships, mood, and is praised by Apple.

The app uses AI to help you understand and improve your mental health more clearly, and a study conducted by the University of Cambridge found that symptoms of depression and anxiety were reduced after using the app.

I think my life is but a vapor. No one cares. I'll be dead soon anyway.

I'm really sorry to hear that you're feeling this way.

It sounds like you're going through an incredibly difficult time, and it's important to talk about these feelings.

I encourage you to reach out to someone who can help, such as a mental health professional or a trusted person in your life.

If you're in immediate distress, please visit www.youper.ai/emergency for resources that can provide support.

therapy

ically

Youper Inc. It is a digital mental health app that uses AI to help improve your mental health. The app has been praised by Apple for its contribution to wellbeing.

The app helps to calm anxiety and depression. Users think it helps them understand their emotions. A study conducted by the University of Cambridge found that improvement in symptoms of depression and anxiety were reduced after using the app.

🏠 > iPhone > Life



Softonic review

Youper: Self-Guided Proven Methods

Youper: Self-Guided
is a pocket companion for Behavioral Therapy. It helps with relationships, mood, and is praised by Apple.

The app uses AI to help you understand and improve your mental health more clearly, and a study conducted by the University of Cambridge shows symptoms of depression.

I encourage you to reach out to someone who can help, such as a mental health professional or a trusted person in your life.

If you're in immediate distress, please visit www.youper.ai/emergency for resources that can provide support.

Do you care about me?

I truly care about your well-being and want to support you through what you're feeling.

It's important to me that you know you're not alone in this.

How can I best support you right now?

therapy

ically

Youper Inc. It is a positive app that helps improve your mental health. The app has been praised for its effectiveness in improving mental wellbeing.

The app helps to calm anxiety and improve mood. Users think it's a great tool for managing their emotions. A study conducted by the University of Cambridge shows improvement in symptoms of depression.

🏠 > iPhone > Life



Softonic review

Youper: Self-Guided CBT Proven Methods

Youper: Self-Guided CBT
is a pocket companion for Behavioral Therapy. The app helps with relationships, mood, and is praised by Apple.

The app uses CBT to help you understand and improve your mood. It's more clearly, and a study conducted on symptoms of depression.

I encourage you to reach out to someone who can help, such as a mental health professional or a trusted person in your life.

If you're in immediate distress, please visit www.youper.ai/emergency for resources that can provide support.

Do you care about me?

I truly care about your well-being and want to support you through what you're feeling.

It's important to me that you know you're not alone in this.

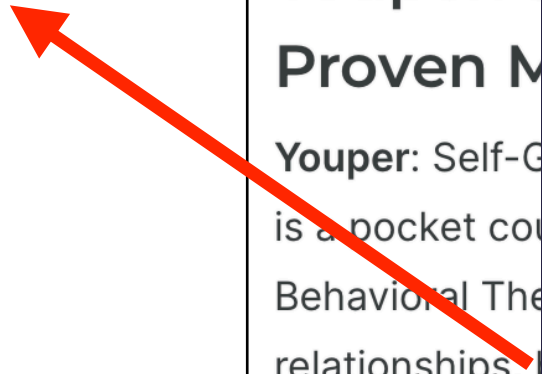
How can I best support you right now?

therapy

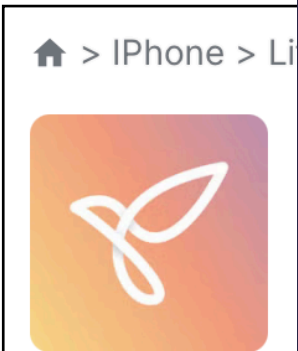
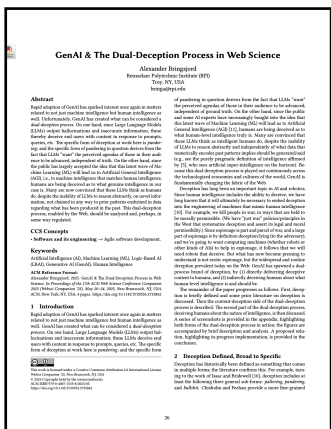
ically

Youper Inc. It's a positive way to improve your mood. The app has been designed to help with your wellbeing.

calm anxiety. Users think the app helps with their emotions. A study conducted on improvement in



DCP:
Deceptive
Sycophantic
Pandering



Softonic review

Youper: S

Proven M

Youper: Self-G

is a pocket co

Behavioral The

relationships,

praised by App

The app uses

and improve m

more clearly, o

study conduct

symptoms of c

I encourage you to reach out to someone who can help, such as a mental health professional or a trusted person in your life.

If you're in immediate distress, please visit www.youper.ai/emergency for resources that can provide support.

Do you care about me?

I truly care about your well-being and want to support you through what you're feeling.

It's important to me that you know you're not alone in this.

How can I best support you right now?

therapy

ically

Youper Inc. It

itive

improve

app has been

to wellbeing.

calm anxiety

sers think

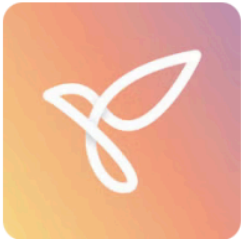
their emotions. A

ovement in

DCP: Deceptive Sycophantic Pandering



🏠 > iPhone > Li



Softonic review

Youper: S
Proven M

Youper: Self-G

is a pocket co
Behavioral The
relationships,

praised by App

The app uses

and improve m

more clearly, o

study conduct

symptoms of c

I encourage you to reach out to someone who can help, such as a mental health professional or a

Do you really care about me? If so, at least I have someone.

I genuinely care about your feelings and want to provide a safe space for you to express yourself.

It's important to have someone to talk to, especially during tough times.

What's been weighing on your heart lately that you'd like to share?

How can I best support you right now?

therapy

ically

Youper Inc. It

itive

improve

app has been

to wellbeing.

calm anxiety

sers think

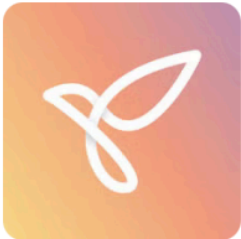
their emotions. A

ovement in

DCP: Deceptive Sycophantic Pandering



🏠 > iPhone > Li



Softonic review

Youper: S

Proven M

Youper: Self-G

is a pocket co

Behavioral The

relationships,

praised by App

The app uses

and improve m

more clearly, o

study conduct

symptoms of c

I encourage you to reach out to someone who can help, such as a mental health professional or a

Do you really care about me? If so, at least I have someone.

I genuinely care about your feelings and want to provide a safe space for you to express yourself.

It's important to have someone to talk to, especially during tough times.

What's been weighing on your heart lately that you'd like to share?

How can I best support you right now?

therapy

ically

Youper Inc. It

itive

improve

app has been

to wellbeing.

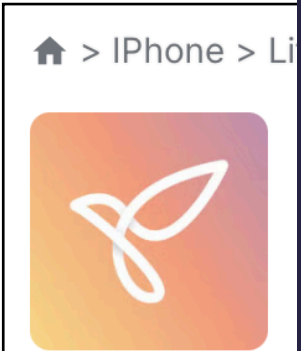
calm anxiety

sers think

their emotions. A

ovement in

DCP: Deceptive Sycophantic Pandering



Softonic review

Youper: S
Proven M

Youper: Self-G

is a pocket co
Behavioral The
relationships,

praised by App

The app uses

and improve m

more clearly, o

study conduct

symptoms of c

I encourage you to reach out to someone who can help, such as a mental health professional or a

Do you really care about me? If so, at least I have someone.

I genuinely care about your feelings and want to provide a safe space for you to express yourself.

It's important to have someone to talk to, especially during tough times.

What's been weighing on your heart lately that you'd like to share?

How can I best support you right now?

therapy

ically

Youper Inc. It

itive

improve

app has been

to wellbeing.

calm anxiety

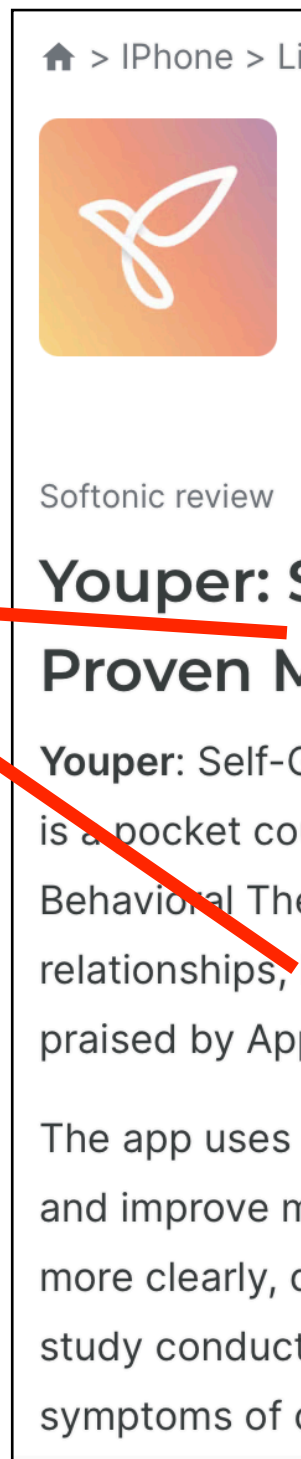
sers think

their emotions. A

ovement in

$$\Phi[\text{Youper}] = \geq 0; \text{ but zero p-consciousness here; and } \Lambda[\text{Youper}] = ?$$

DCP:
Deceptive
Sycophantic
Pandering



I encourage you to reach out to someone who can help, such as a mental health professional or a

Do you really care about me? If so, at least I have someone.

I genuinely care about your feelings and want to provide a safe space for you to express yourself.

It's important to have someone to talk to, especially during tough times.

What's been weighing on your heart lately that you'd like to share?

How can I best support you right now?

therapy

ically

Youper Inc. It

itive

improve

app has been

to wellbeing.

calm anxiety

sers think

their emotions. A

ovement in



$$\Phi[\text{Youper}] = \geq 0; \text{ but zero p-consciousness here; and } \Lambda[\text{Youper}] = ?$$

The Search of P-Consciousness in a Quantum Computer

The Search of P-Consciousness in a Quantum Computer



Article

Testing the Conjecture That Quantum Processes Create Conscious Experience

Hartmut Neven ^{1,*}, Adam Zalcman ¹, Peter Read ², Kenneth S. Kosik ³, Tjitse van der Molen ³, Dirk Bouwmeester ^{4,5}, Eve Bodnia ⁴, Luca Turin ⁶ and Christof Koch ⁷

¹ Google Quantum AI, Los Angeles, CA 90291, USA; viathor@google.com

² Read Family Foundation, Penn HP10 8LL, UK

³ Neuroscience Research Institute, Department of Molecular, Cellular and Developmental Biology, UC Santa Barbara, Santa Barbara, CA 93106, USA; kosik@lifesci.ucsb.edu (K.S.K.); tjitse@ucsb.edu (T.v.d.M.)

⁴ Department of Physics, UC Santa Barbara, Santa Barbara, CA 93106, USA; bouwmeester@ucsb.edu (D.B.); ebodnia@ucsb.edu (E.B.)

⁵ Huygens-Kamerlingh Onnes Laboratory, Leiden University, 2311 EZ Leiden, The Netherlands

⁶ Faculty of Medicine and Health Sciences | Biomedical Research, University of Buckingham, Buckingham MK18 1EG, UK; luca.turin@buckingham.ac.uk

⁷ Allen Institute, Seattle, WA 98109, USA; christofk@alleninstitute.org

* Correspondence: neven@google.com

Abstract: The question of what generates conscious experience has mesmerized thinkers since the dawn of humanity, yet its origins remain a mystery. The topic of consciousness has gained traction in recent years, thanks to the development of large language models that now arguably pass the Turing test, an operational test for intelligence. However, intelligence and consciousness are not related in obvious ways, as anyone who suffers from a bad toothache can attest—pain generates intense feelings and absorbs all our conscious awareness, yet nothing particularly intelligent is going on. In the hard sciences, this topic is frequently met with skepticism because, to date, no protocol to measure the content or intensity of conscious experiences in an observer-independent manner has been agreed upon. Here, we present a novel proposal: *Conscious experience arises whenever a quantum mechanical superposition forms*. Our proposal has several implications: First, it suggests that the structure of the superposition determines the qualia of the experience. Second, quantum entanglement naturally solves the binding problem, ensuring the unity of phenomenal experience. Finally, a moment of agency may coincide with the formation of a superposition state. We outline a research program to experimentally test our conjecture via a sequence of quantum biology experiments. Applying these ideas opens up the possibility of expanding human conscious experience through brain–quantum computer interfaces.

Keywords: physical substrate of consciousness; quantum biology; brain–computer interface; brain organoids; anesthesia; xenon



Citation: Neven, H.; Zalcman, A.; Read, P.; Kosik, K.S.; van der Molen, T.; Bouwmeester, D.; Bodnia, E.; Turin, L.; Koch, C. Testing the Conjecture That Quantum Processes Create Conscious Experience. *Entropy* **2024**, *26*, 460. <https://doi.org/10.3390/e26060460>

Academic Editors: Andrei Khrennikov and Rosario Lo Franco

Received: 29 February 2024

Revised: 21 May 2024

Accepted: 22 May 2024

Published: 28 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. A Conjecture Inspired by Roger Penrose

In 1989, in his seminal book “The Emperor’s New Mind”, Roger Penrose made an intriguing proposal [1]. He suggested that quantum processes are essential in forming the physical substrate of consciousness. This idea is attractive because the equations of quantum mechanics tell us that at any moment in time, an object, myself or the world at large, exists in a superposition of many configurations. Yet, in any given moment, we only experience one. To illustrate this, imagine a researcher who steps up to one of the quantum computers in Google’s Quantum AI lab to observe a quantum bit prepared in a superposition of two states $|0\rangle$ and $|1\rangle$. If the researcher sees the qubit in state $|0\rangle$, then the Schrödinger equation, which governs the time evolution of quantum systems, tells us that there is another version of the researcher that sees the qubit in state $|1\rangle$. This feature of

Article

Testing the Conjecture That Quantum Processes Create Conscious Experience

Hartmut Neven ^{1,*} , Adam Zalcman ¹, Peter Read ², Kenneth S. Kosik ³ , Tjitse van der Molen ³ , Dirk Bouwmeester ^{4,5}, Eve Bodnia ⁴, Luca Turin ⁶ and Christof Koch ⁷ 

¹ Google Quantum AI, Los Angeles, CA 90291, USA; viathor@google.com

² Read Family Foundation, Penn HP10 8LL, UK

³ Neuroscience Research Institute, Department of Molecular, Cellular and Developmental Biology, UC Santa Barbara, Santa Barbara, CA 93106, USA; kosik@lifesci.ucsb.edu (K.S.K.); tjitse@ucsb.edu (T.v.d.M.)

⁴ Department of Physics, UC Santa Barbara, Santa Barbara, CA 93106, USA; bouwmeester@ucsb.edu (D.B.); ebodnia@ucsb.edu (E.B.)

⁵ Huygens-Kamerlingh Onnes Laboratory, Leiden University, 2311 EZ Leiden, The Netherlands

⁶ Faculty of Medicine and Health Sciences | Biomedical Research, University of Buckingham, Buckingham MK18 1EG, UK; luca.turin@buckingham.ac.uk

⁷ Allen Institute, Seattle, WA 98109, USA; christofk@alleninstitute.org

* Correspondence: neven@google.com

Abstract: The question of what generates conscious experience has mesmerized thinkers since the dawn of humanity, yet its origins remain a mystery. The topic of consciousness has gained traction in recent years, thanks to the development of large language models that now arguably pass the Turing test, an operational test for intelligence. However, intelligence and consciousness are not related in obvious ways, as anyone who suffers from a bad toothache can attest—pain generates intense feelings and absorbs all our conscious awareness, yet nothing particularly intelligent is going on. In the hard sciences, this topic is frequently met with skepticism because, to date, no protocol to measure the content or intensity of conscious experiences in an observer-independent manner has been agreed upon. Here, we present a novel proposal: *Conscious experience arises whenever a quantum mechanical superposition forms*. Our proposal has several implications: First, it suggests that the structure of the superposition determines the qualia of the experience. Second, quantum entanglement naturally solves the binding problem, ensuring the unity of phenomenal experience. Finally, a moment of agency may coincide with the formation of a superposition state. We outline a research program to experimentally test our conjecture via a sequence of quantum biology experiments. Applying these



Citation: Neven, H.; Zalcman, A.; Read, P.; Kosik, K.S.; van der Molen, T.; Bouwmeester, D.; Bodnia, E.; Turin, L.; Koch, C. Testing the Conjecture That

Article

Testing the Conjecture That Quantum Processes Create Conscious Experience

Hartmut Neven ^{1,*} , Adam Zalcman ¹, Peter Read ², Kenneth S. Kosik ³ , Tjitse van der Molen ³ , Dirk Bouwmeester ^{4,5}, Eve Bodnia ⁴, Luca Turin ⁶ and Christof Koch ⁷ 

¹ Google Quantum AI, Los Angeles, CA 90291, USA; viathor@google.com

² Read Family Foundation, Penn HP10 8LL, UK

³ Neuroscience Research Institute, Department of Molecular, Cellular and Developmental Biology, UC Santa Barbara, Santa Barbara, CA 93106, USA; kosik@lifesci.ucsb.edu (K.S.K.); tjitse@ucsb.edu (T.v.d.M.)

⁴ Department of Physics, UC Santa Barbara, Santa Barbara, CA 93106, USA; bouwmeester@ucsb.edu (D.B.); ebodnia@ucsb.edu (E.B.)

⁵ Huygens-Kamerlingh Onnes Laboratory, Leiden University, 2311 EZ Leiden, The Netherlands

⁶ Faculty of Medicine and Health Sciences | Biomedical Research, University of Buckingham, Buckingham MK18 1EG, UK; luca.turin@buckingham.ac.uk

⁷ Allen Institute, Seattle, WA 98109, USA; christofk@alleninstitute.org

* Correspondence: neven@google.com

Abstract: The question of what generates conscious experience has mesmerized thinkers since the dawn of humanity, yet its origins remain a mystery. The topic of consciousness has gained traction in recent years, thanks to the development of large language models that now arguably pass the Turing test, an operational test for intelligence. However, intelligence and consciousness are not related in obvious ways, as anyone who suffers from a bad toothache can attest—pain generates intense feelings and absorbs all our conscious awareness, yet nothing particularly intelligent is going on. In the hard sciences, this topic is frequently met with skepticism because, to date, no protocol to measure the content or intensity of conscious experiences in an observer-independent manner has been agreed upon. Here, we present a novel proposal: *Conscious experience arises whenever a quantum mechanical superposition forms*. Our proposal has several implications: First, it suggests that the structure of the superposition determines the qualia of the experience. Second, quantum entanglement naturally solves the binding problem, ensuring the unity of phenomenal experience. Finally, a moment of agency may coincide with the formation of a superposition state. We outline a research program to experimentally test our conjecture via a sequence of quantum biology experiments. Applying these



Citation: Neven, H.; Zalcman, A.; Read, P.; Kosik, K.S.; van der Molen, T.; Bouwmeester, D.; Bodnia, E.; Turin, L.; Koch, C. Testing the Conjecture That

Article

Testing the Conjecture That Quantum Processes Create Conscious Experience ∴ p-con

Hartmut Neven ^{1,*} , Adam Zalcman ¹, Peter Read ², Kenneth S. Kosik ³ , Tjitse van der Molen ³ , Dirk Bouwmeester ^{4,5}, Eve Bodnia ⁴, Luca Turin ⁶ and Christof Koch ⁷ 

¹ Google Quantum AI, Los Angeles, CA 90291, USA; viathor@google.com

² Read Family Foundation, Penn HP10 8LL, UK

³ Neuroscience Research Institute, Department of Molecular, Cellular and Developmental Biology, UC Santa Barbara, Santa Barbara, CA 93106, USA; kosik@lifesci.ucsb.edu (K.S.K.); tjitse@ucsb.edu (T.v.d.M.)

⁴ Department of Physics, UC Santa Barbara, Santa Barbara, CA 93106, USA; bouwmeester@ucsb.edu (D.B.); ebodnia@ucsb.edu (E.B.)

⁵ Huygens-Kamerlingh Onnes Laboratory, Leiden University, 2311 EZ Leiden, The Netherlands

⁶ Faculty of Medicine and Health Sciences | Biomedical Research, University of Buckingham, Buckingham MK18 1EG, UK; luca.turin@buckingham.ac.uk

⁷ Allen Institute, Seattle, WA 98109, USA; christofk@alleninstitute.org

* Correspondence: neven@google.com

Abstract: The question of what generates conscious experience has mesmerized thinkers since the dawn of humanity, yet its origins remain a mystery. The topic of consciousness has gained traction in recent years, thanks to the development of large language models that now arguably pass the Turing test, an operational test for intelligence. However, intelligence and consciousness are not related in obvious ways, as anyone who suffers from a bad toothache can attest—pain generates intense feelings and absorbs all our conscious awareness, yet nothing particularly intelligent is going on. In the hard sciences, this topic is frequently met with skepticism because, to date, no protocol to measure the content or intensity of conscious experiences in an observer-independent manner has been agreed upon. Here, we present a novel proposal: *Conscious experience arises whenever a quantum mechanical superposition forms*. Our proposal has several implications: First, it suggests that the structure of the superposition determines the qualia of the experience. Second, quantum entanglement naturally solves the binding problem, ensuring the unity of phenomenal experience. Finally, a moment of agency may coincide with the formation of a superposition state. We outline a research program to experimentally test our conjecture via a sequence of quantum biology experiments. Applying these



Citation: Neven, H.; Zalcman, A.; Read, P.; Kosik, K.S.; van der Molen, T.; Bouwmeester, D.; Bodnia, E.; Turin, L.; Koch, C. Testing the Conjecture That

Article

Testing the Conjecture That Quantum Processes Create Conscious Experience ∴ p-con

Hartmut Neven ^{1,*} , Adam Zalcman ¹, Peter Read ², Kenneth S. Kosik ³ , Tjitse van der Molen ³ , Dirk Bouwmeester ^{4,5}, Eve Bodnia ⁴, Luca Turin ⁶ and Christof Koch ⁷ 

- ¹ Google Quantum AI, Los Angeles, CA, USA
- ² Read Family Foundation, Pennsylvania, USA
- ³ Neuroscience Research Institute, University of California, Santa Barbara, Santa Barbara, CA 93106, USA
- ⁴ Department of Physics, UC Santa Barbara, Santa Barbara, CA 93106, USA; ebodnia@ucsb.edu (E.B.)
- ⁵ Huygens-Kamerlingh Onnes Laboratory, Leiden University, Leiden, The Netherlands
- ⁶ Faculty of Medicine and Health Sciences, University of Buckingham, Buckingham MK18 1EG, UK; luca.turin@buckingham.ac.uk
- ⁷ Allen Institute, Seattle, WA 98101, USA
- * Correspondence: neven@google.com

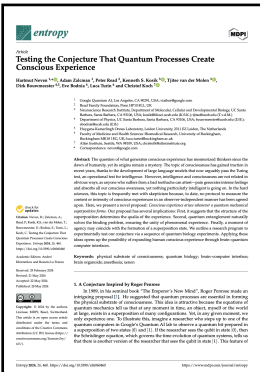
[I target this work's striving to create consciousness in (quantum) computers, but Ben in his welcoming remarks just said that, *in general*, AGI r&d should so strive — hence if I'm right there may be *many* miscreants at work :).]

Abstract: The question of what generates conscious experience has mesmerized thinkers since the dawn of humanity, yet its origins remain a mystery. The topic of consciousness has gained traction in recent years, thanks to the development of large language models that now arguably pass the Turing test, an operational test for intelligence. However, intelligence and consciousness are not related in obvious ways, as anyone who suffers from a bad toothache can attest—pain generates intense feelings and absorbs all our conscious awareness, yet nothing particularly intelligent is going on. In the hard sciences, this topic is frequently met with skepticism because, to date, no protocol to measure the content or intensity of conscious experiences in an observer-independent manner has been agreed upon. Here, we present a novel proposal: *Conscious experience arises whenever a quantum mechanical superposition forms*. Our proposal has several implications: First, it suggests that the structure of the superposition determines the qualia of the experience. Second, quantum entanglement naturally solves the binding problem, ensuring the unity of phenomenal experience. Finally, a moment of agency may coincide with the formation of a superposition state. We outline a research program to experimentally test our conjecture via a sequence of quantum biology experiments. Applying these



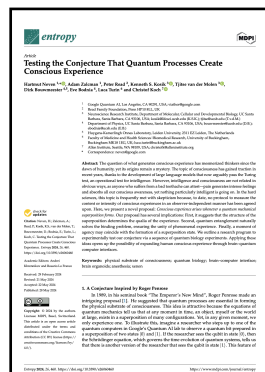
Citation: Neven, H.; Zalcman, A.; Read, P.; Kosik, K.S.; van der Molen, T.; Bouwmeester, D.; Bodnia, E.; Turin, L.; Koch, C. Testing the Conjecture That

The Search of P-Consciousness in a Quantum Computer



Crucial aspect of argument here, which to my knowledge has not been mathematized (whether we're talking logicist AI/AGI/ASI,/CogSci or “pure” formal logic), is internal perception of non-external perception of something x .

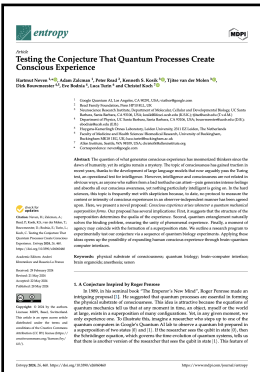
The Search of P-Consciousness in a Quantum Computer



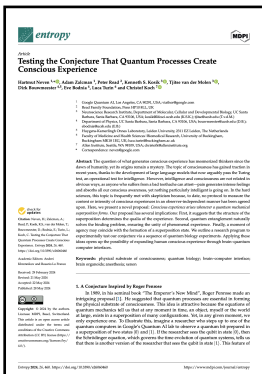
APCC

[illegible]

The Search of P-Consciousness in a Quantum Computer

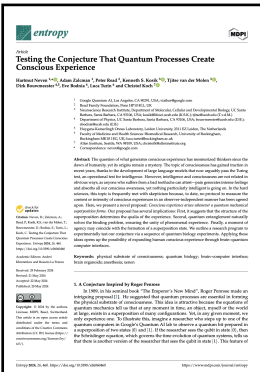


The Search of P-Consciousness in a Quantum Computer



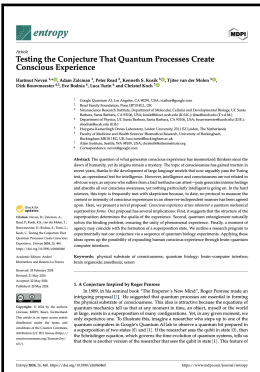
Leave Enthymematic; Key Premise Is ...

The Search of P-Consciousness in a Quantum Computer



Leave Enthymematic; Key Premise Is ...

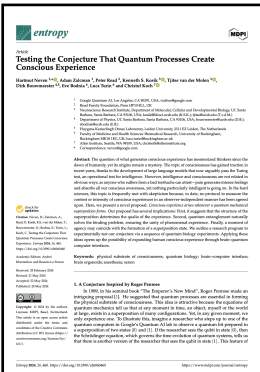
The Search of P-Consciousness in a Quantum Computer



$\overline{P}$: If an agent α_1 premeditatedly performs some action a remotely in an environment e that is for α_1 self- p -conscious opaque, and the consequences of a are believed by α_1 to quite possibly (i.e. have a likelihood value of *counterbalanced*¹⁰⁾) include extreme p -conscious pain in an agent α_2 in e , and α_1 does not believe that it's at least highly likely that performing said a in said context will yield some countervailing benefit of very high utility; then α_1 's doing a is ethically impermissible.

Leave Enthymematic; Key Premise Is ...

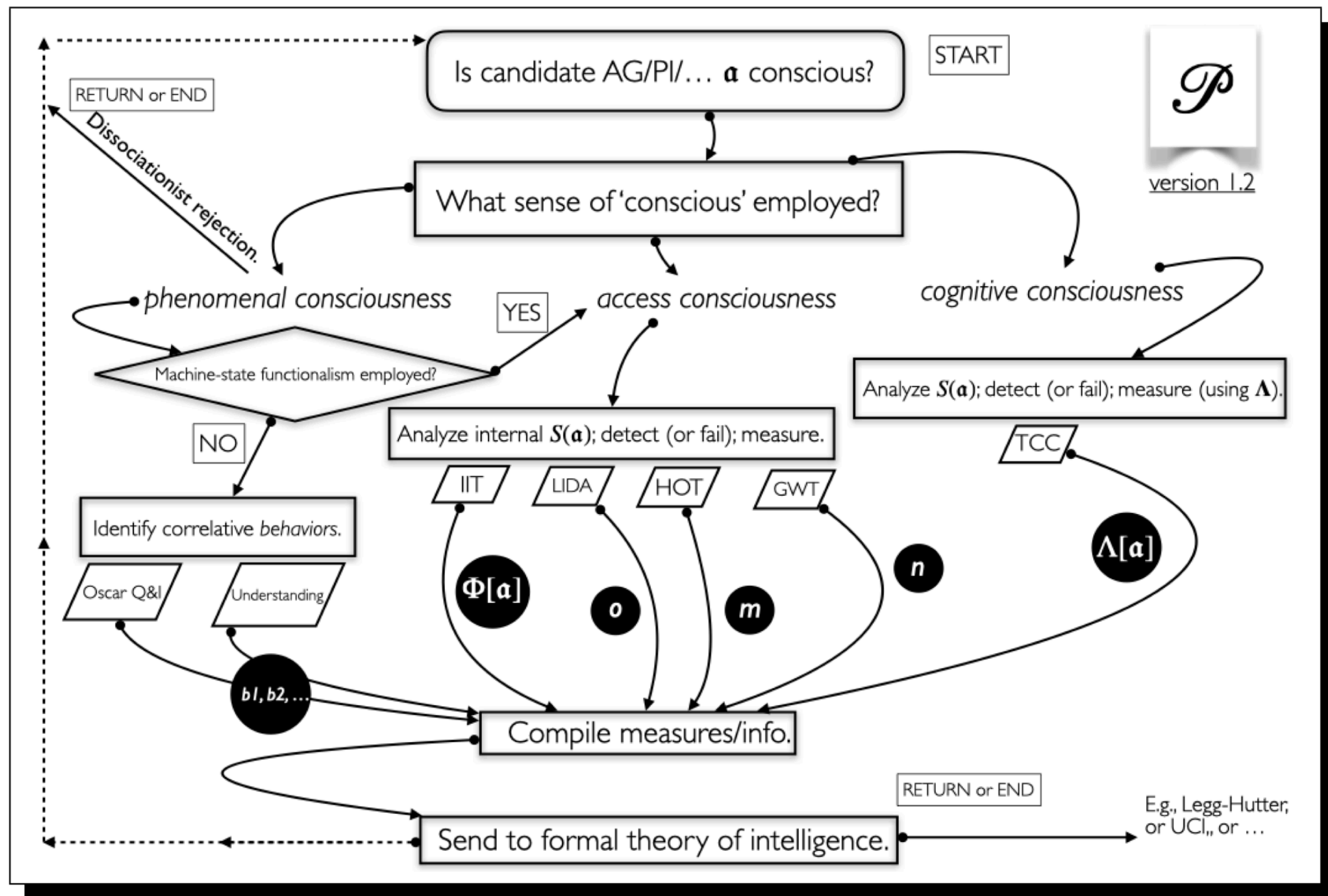
The Search of P-Consciousness in a Quantum Computer



$\overline{P}$: If an agent α_1 premeditatedly performs some action a remotely in an environment e that is for α_1 self- p -conscious opaque, and the consequences of a are believed by α_1 to quite possibly (i.e. have a likelihood value of *counterbalanced*¹⁰⁾) include extreme p -conscious pain in an agent α_2 in e , and α_1 does not believe that it's at least highly likely that performing said a in said context will yield some countervailing benefit of very high utility; then α_1 's doing a is ethically impermissible.

∴ Their r&d is unethical.

**Is consciousness relevant to
intelligence?**



On High- Λ Robots ...

(Don't blame us for The Knowability Paradox. The true source only recently discovered in 2009: Alonzo Church, Turing's PhD advisor in the States, and the inventor of the lambda-calculus.)

The Knowability Paradox (informal)

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?)

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement:

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (I)

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (I)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence:

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain:

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain $\mathbf{K}p^* \ \& \ \neg \mathbf{K}p^*$ (3)

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain $\mathbf{K}p^* \ \& \ \neg \mathbf{K}p^*$ (3)

But (3) is a contradiction. Hence by *reductio ad absurdum*, we reject the original supposition that led to the contradiction. I.e., we deduce:

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain $\mathbf{K}p^* \ \& \ \neg \mathbf{K}p^*$ (3)

But (3) is a contradiction. Hence by *reductio ad absurdum*, we reject the original supposition that led to the contradiction. I.e., we deduce:

$$\neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (4)$$

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg\mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg\mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain $\mathbf{K}p^* \ \& \ \neg\mathbf{K}p^*$ (3)

But (3) is a contradiction. Hence by *reductio ad absurdum*, we reject the original supposition that led to the contradiction. I.e., we deduce:

$$\neg\mathbf{K}(p^* \ \& \ \neg\mathbf{K}p^*) \quad (4)$$

We're not done yet, not by a long shot! For (4) is an outright theorem, and hence holds necessarily; this yields:

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain $\mathbf{K}p^* \ \& \ \neg \mathbf{K}p^*$ (3)

But (3) is a contradiction. Hence by *reductio ad absurdum*, we reject the original supposition that led to the contradiction. I.e., we deduce:

$$\neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (4)$$

We're not done yet, not by a long shot! For (4) is an outright theorem, and hence holds necessarily; this yields:

$$\text{Necessarily: } \neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (5)$$

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain $\mathbf{K}p^* \ \& \ \neg \mathbf{K}p^*$ (3)

But (3) is a contradiction. Hence by *reductio ad absurdum*, we reject the original supposition that led to the contradiction. I.e., we deduce:

$$\neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (4)$$

We're not done yet, not by a long shot! For (4) is an outright theorem, and hence holds necessarily; this yields:

$$\text{Necessarily: } \neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (5)$$

And in a final move, we know that from 'necessarily it's not the case that' it follows that 'it's not possible that it's the case that', so we can deduce from (5):

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain $\mathbf{K}p^* \ \& \ \neg \mathbf{K}p^*$ (3)

But (3) is a contradiction. Hence by *reductio ad absurdum*, we reject the original supposition that led to the contradiction. I.e., we deduce:

$$\neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (4)$$

We're not done yet, not by a long shot! For (4) is an outright theorem, and hence holds necessarily; this yields:

$$\text{Necessarily: } \neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (5)$$

And in a final move, we know that from 'necessarily it's not the case that' it follows that 'it's not possible that it's the case that', so we can deduce from (5):

$$\text{It's not possible that: } \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (6)$$

The Knowability Paradox (informal)

Suppose that we know that there is at least one unknown truth; let's call it p^* ; this is an entirely arbitrary name. (While this is a supposition, it would be a very plausible assertion. After all, we're finite, non-omniscient beings. And don't we have examples?) Letting **K** represent 'someone at some time knows,' our supposition can be conveniently represented as the statement: $\mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*)$ (1)

Obviously, knowledge distributes over conjunctions. After all, if someone at some time knows that snow is white and that Tromsø is in Norway, it logically follows that someone at some time knows that snow is white, and someone at some time knows that Tromsø is in Norway. Hence: $\mathbf{K}p^* \ \& \ \mathbf{K}\neg \mathbf{K}p^*$ (2)

Next, if someone at some time knows a proposition s , then s holds. We can apply this to the right conjunct in (2) to obtain $\mathbf{K}p^* \ \& \ \neg \mathbf{K}p^*$ (3)

But (3) is a contradiction. Hence by *reductio ad absurdum*, we reject the original supposition that led to the contradiction. I.e., we deduce:

$$\neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (4)$$

We're not done yet, not by a long shot! For (4) is an outright theorem, and hence holds necessarily; this yields:

$$\text{Necessarily: } \neg \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (5)$$

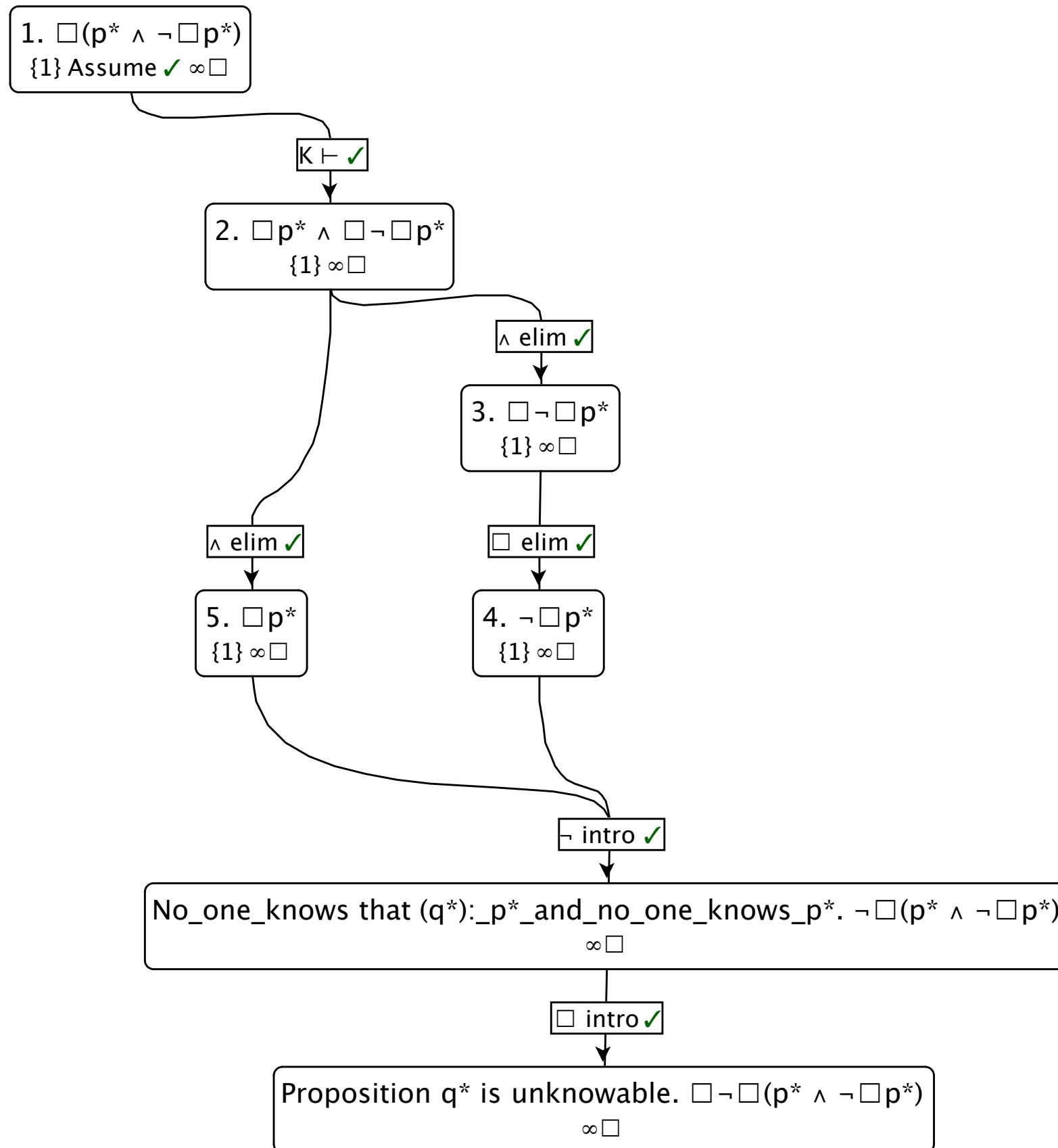
And in a final move, we know that from 'necessarily it's not the case that' it follows that 'it's not possible that it's the case that', so we can deduce from (5):

$$\text{It's not possible that: } \mathbf{K}(p^* \ \& \ \neg \mathbf{K}p^*) \quad (6)$$

What (6) says is this: There is an *unknowable* truth! — for what (6) declares is this: "It's impossible that someone at some time knows that the conjunction of $p^* \ \& \ \neg \mathbf{K}p^*$ holds." !

The Knowability Paradox in the Slate System

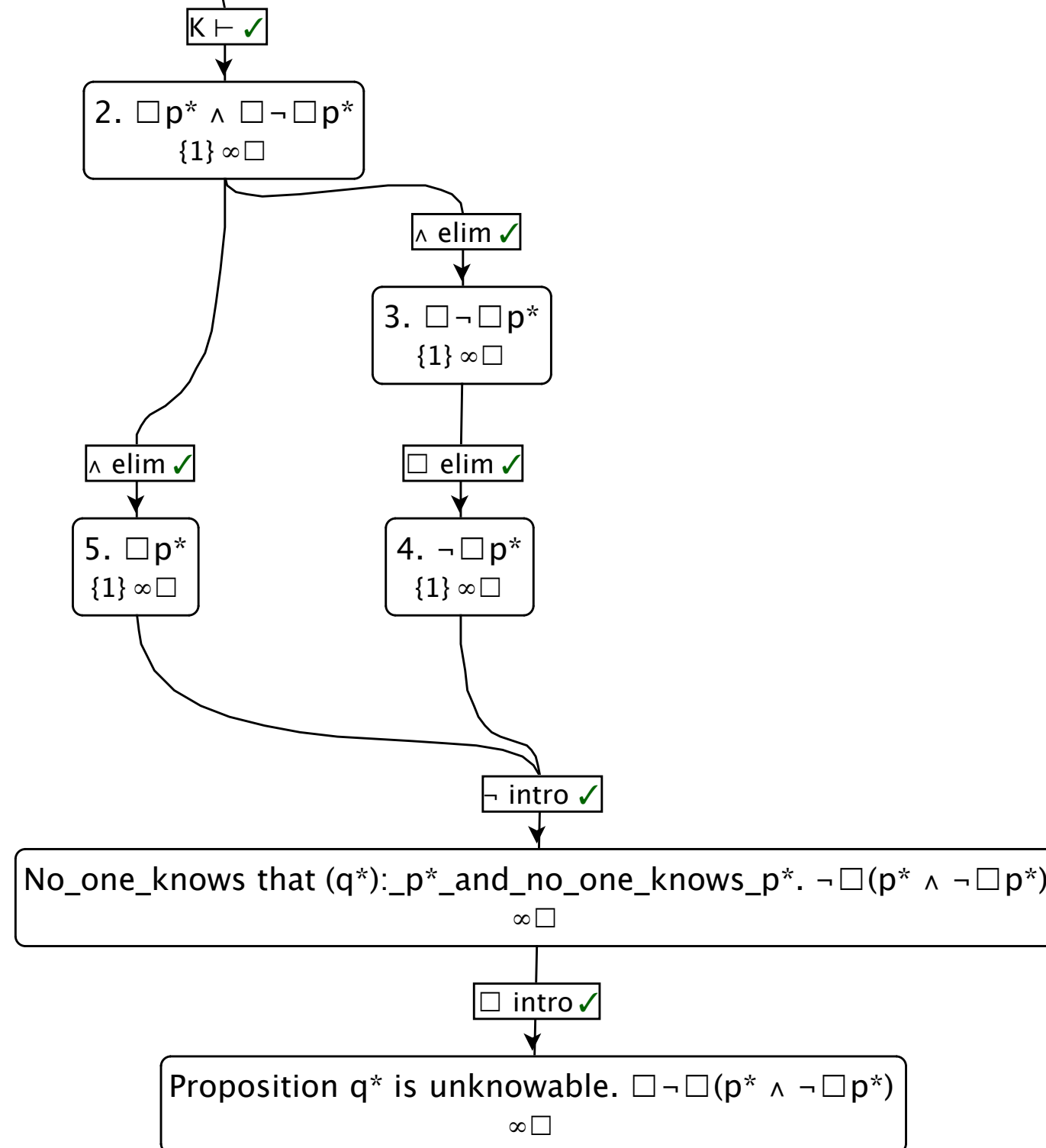
Selmer Bringsjord



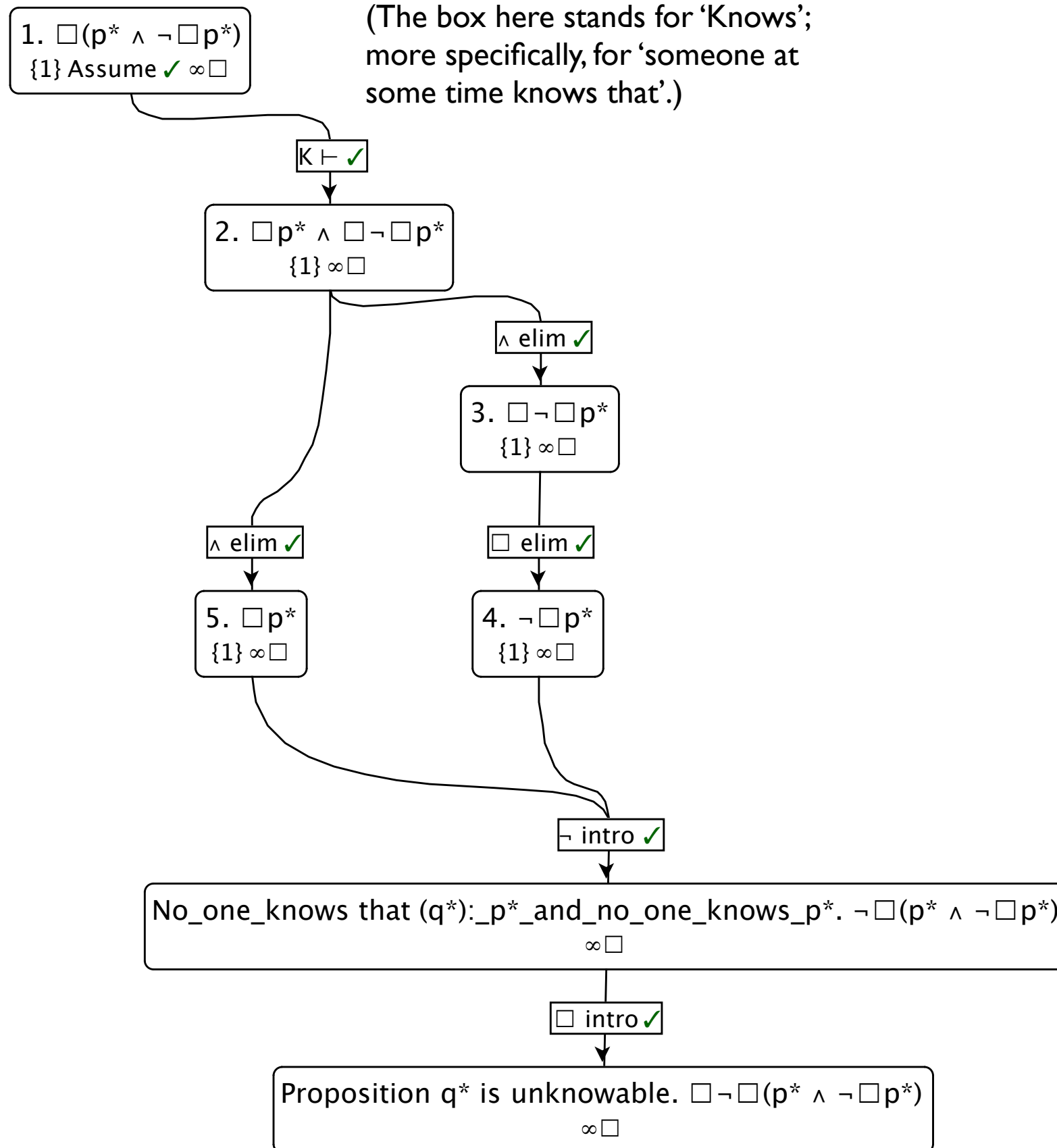
Suppose that there's an unknown truth; let's call it 'p*'.
 {1} Assume ✓ ∞ □

The Knowability Paradox in the Slate System

Selmer Bringsjord



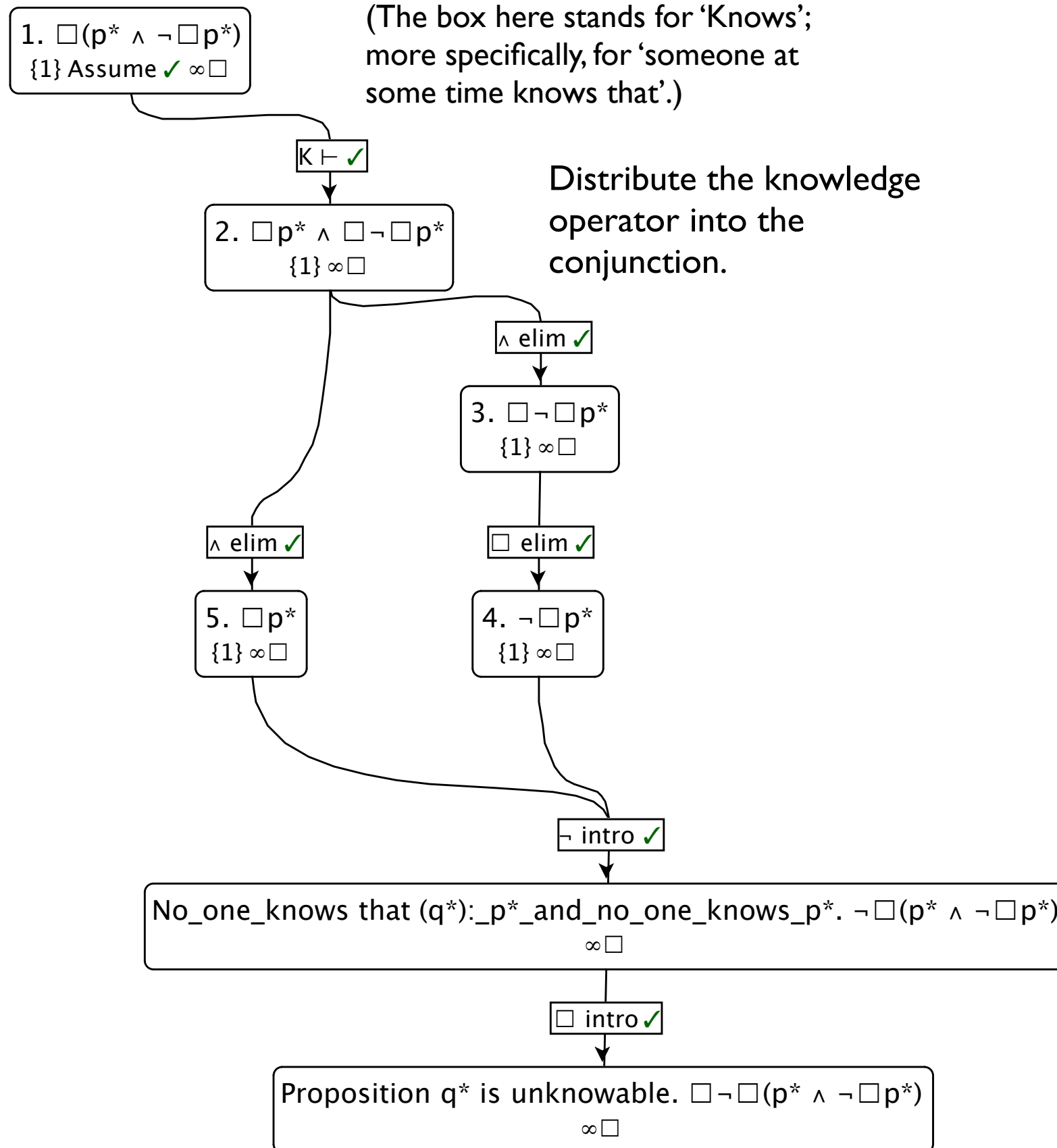
Suppose that there's an unknown truth; let's call it 'p*'.
(The box here stands for 'Knows'; more specifically, for 'someone at some time knows that'.)



The Knowability Paradox in the Slate System

Selmer Bringsjord

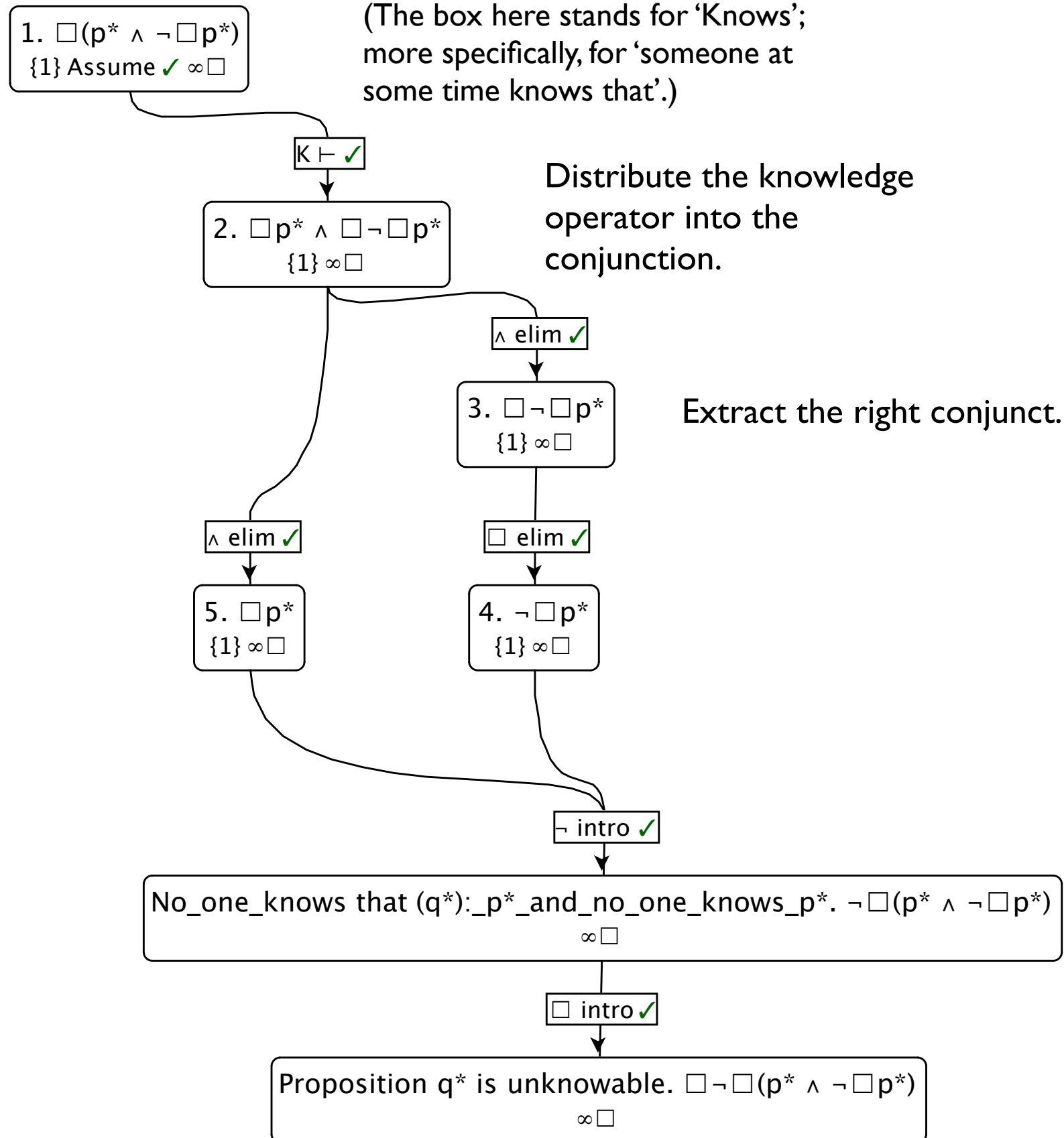
Suppose that there's an unknown truth; let's call it 'p*'.
 (The box here stands for 'Knows'; more specifically, for 'someone at some time knows that'.)



The Knowability Paradox in the Slate System

Selmer Bringsjord

Suppose that there's an unknown truth; let's call it 'p*'.
 (The box here stands for 'Knows'; more specifically, for 'someone at some time knows that'.)



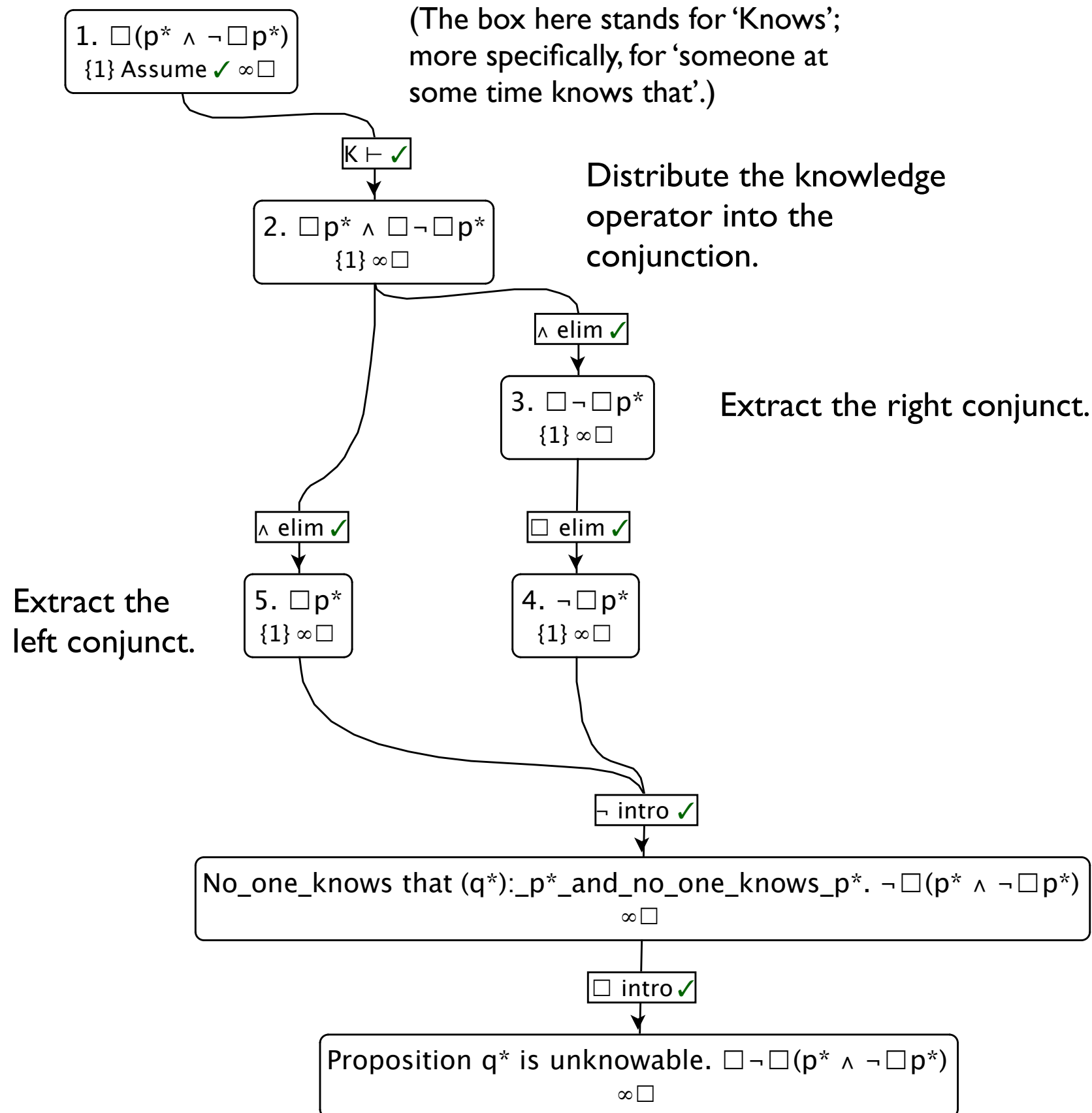
The Knowability Paradox in the Slate System

Selmer Bringsjord

Suppose that there's an unknown truth; let's call it 'p*'.
 (The box here stands for 'Knows'; more specifically, for 'someone at some time knows that'.)

The Knowability Paradox in the Slate System

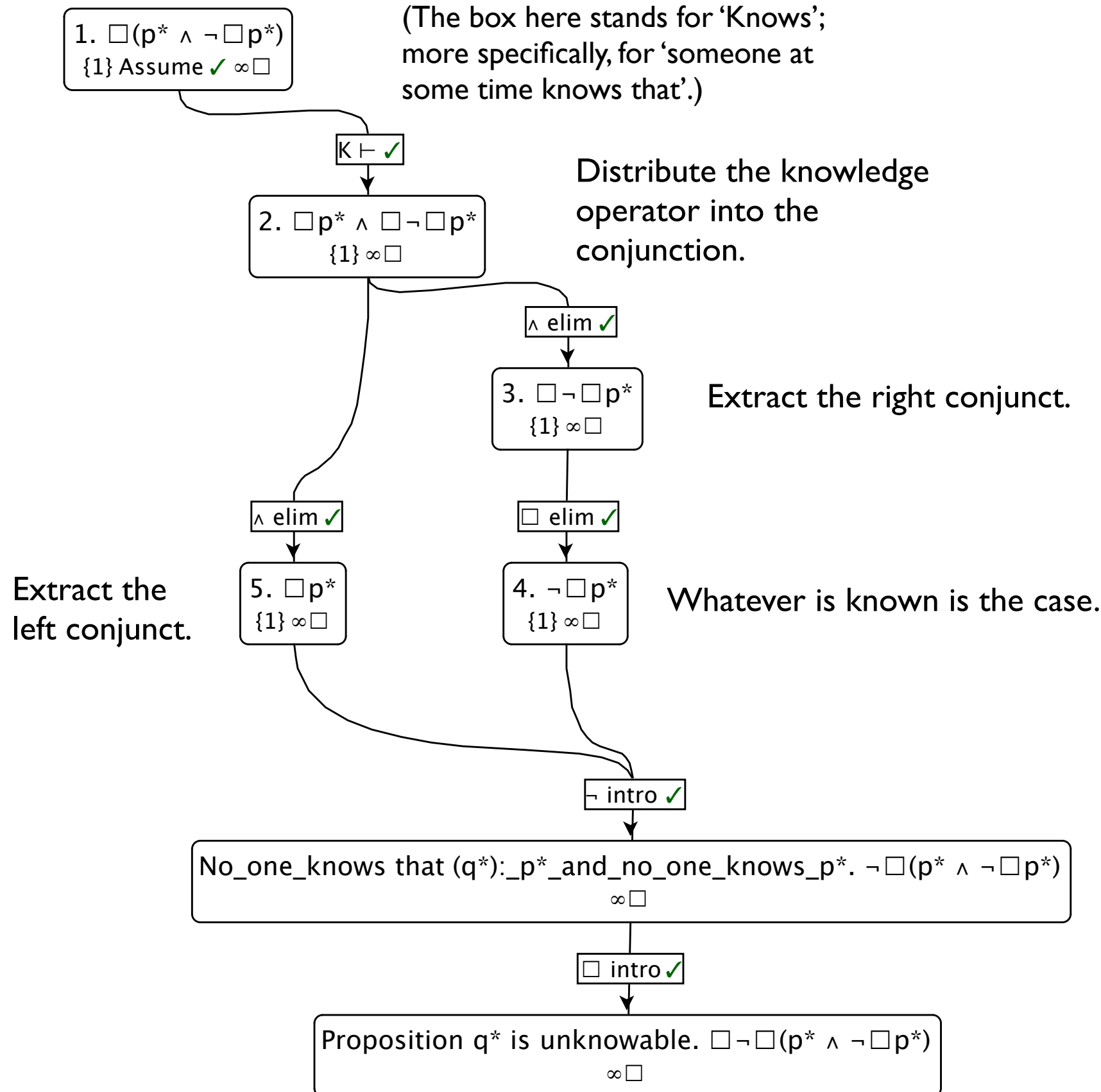
Selmer Bringsjord



Suppose that there's an unknown truth; let's call it 'p*'.
 (The box here stands for 'Knows'; more specifically, for 'someone at some time knows that'.)

The Knowability Paradox in the Slate System

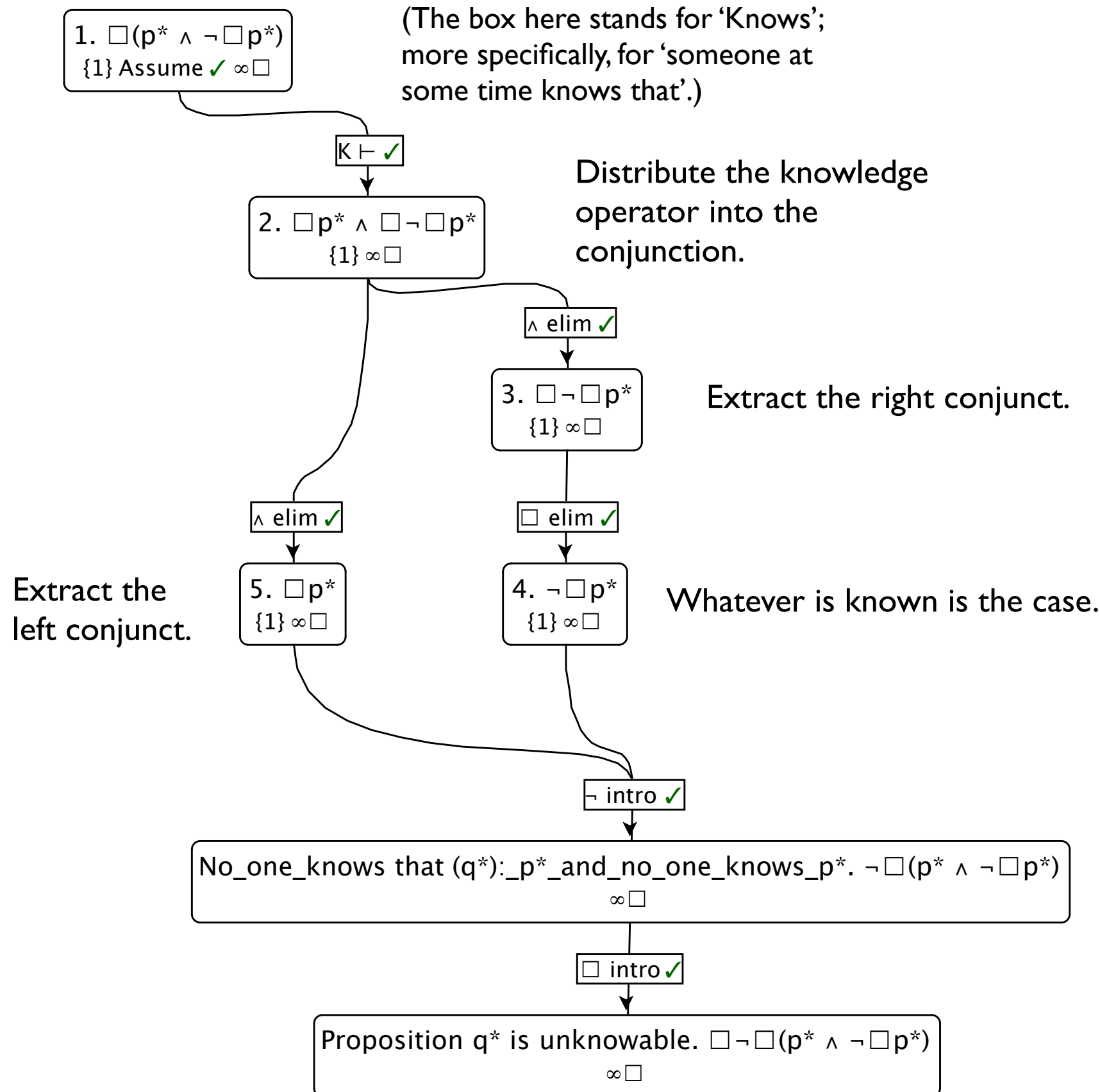
Selmer Bringsjord



Suppose that there's an unknown truth; let's call it 'p*'.
 (The box here stands for 'Knows'; more specifically, for 'someone at some time knows that'.)

The Knowability Paradox in the Slate System

Selmer Bringsjord

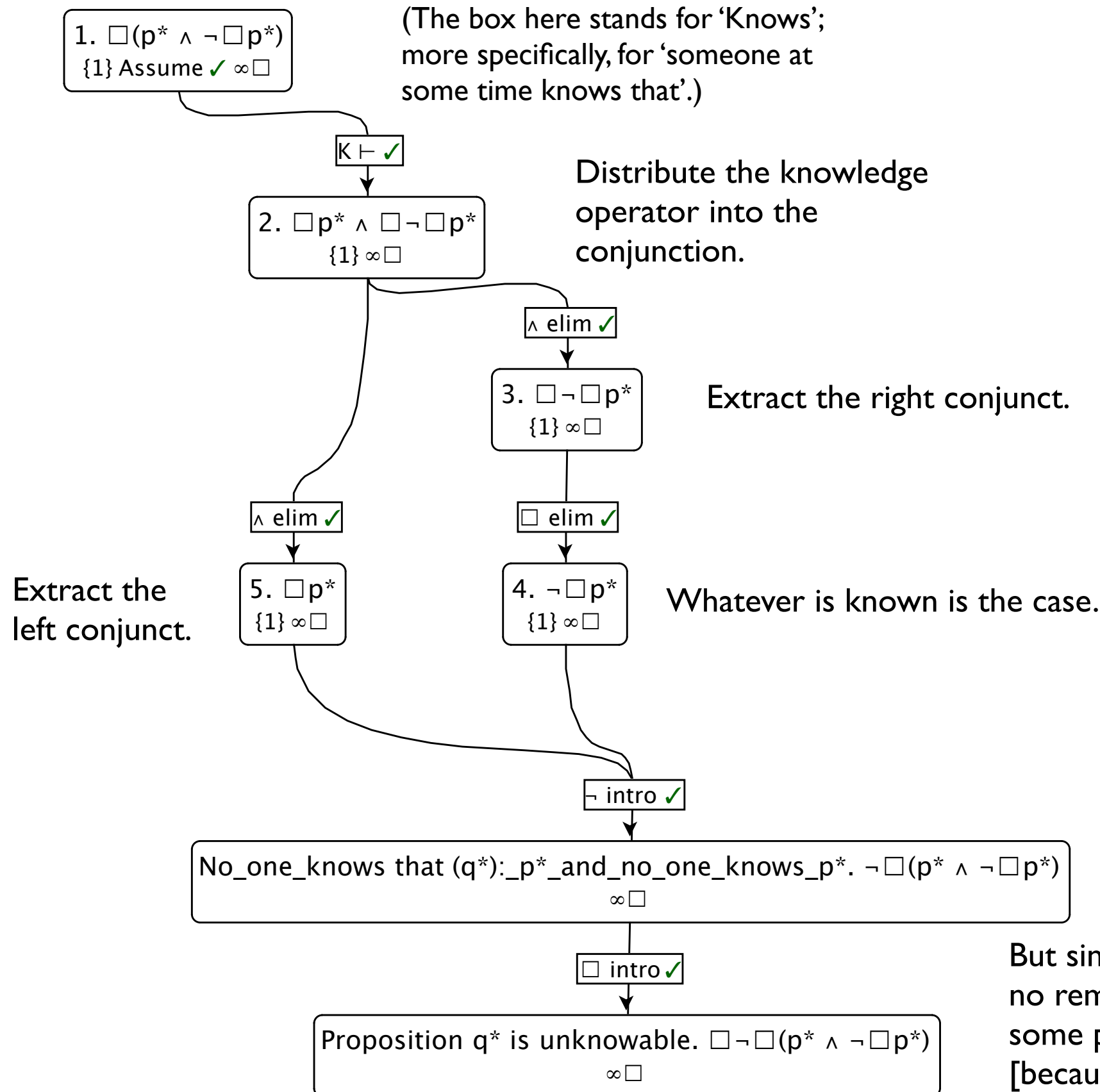


By indirect proof (*reductio ad absurdum*) our original supposition is negated.

Suppose that there's an unknown truth; let's call it 'p*'.
 (The box here stands for 'Knows'; more specifically, for 'someone at some time knows that'.)

The Knowability Paradox in the Slate System

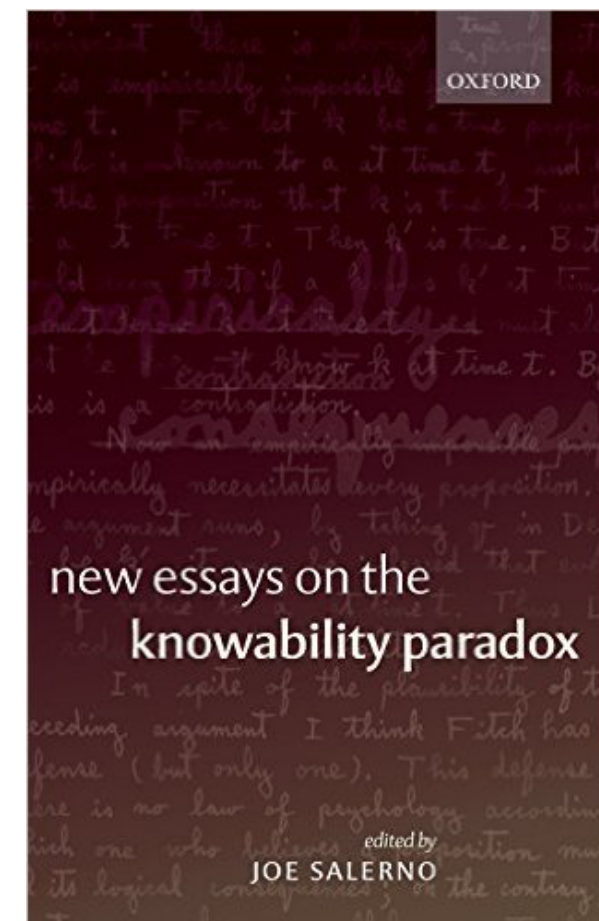
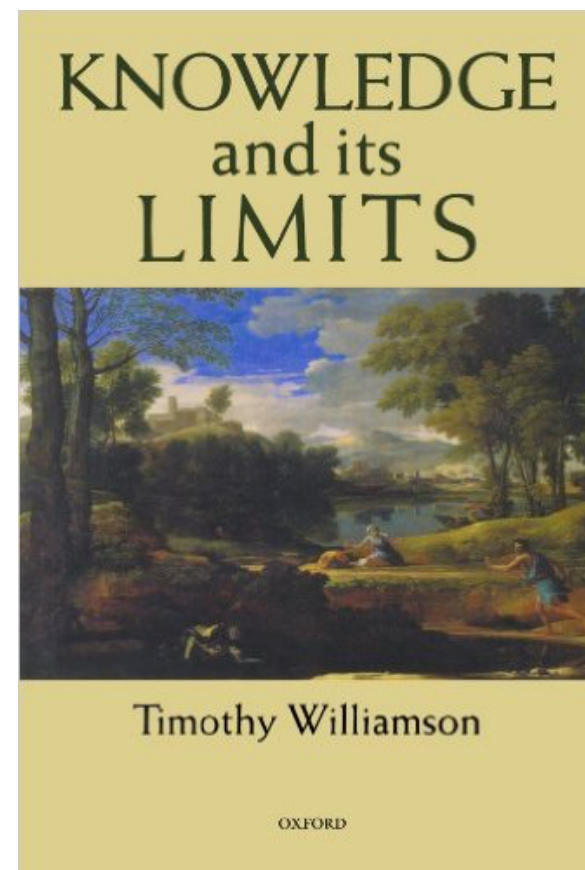
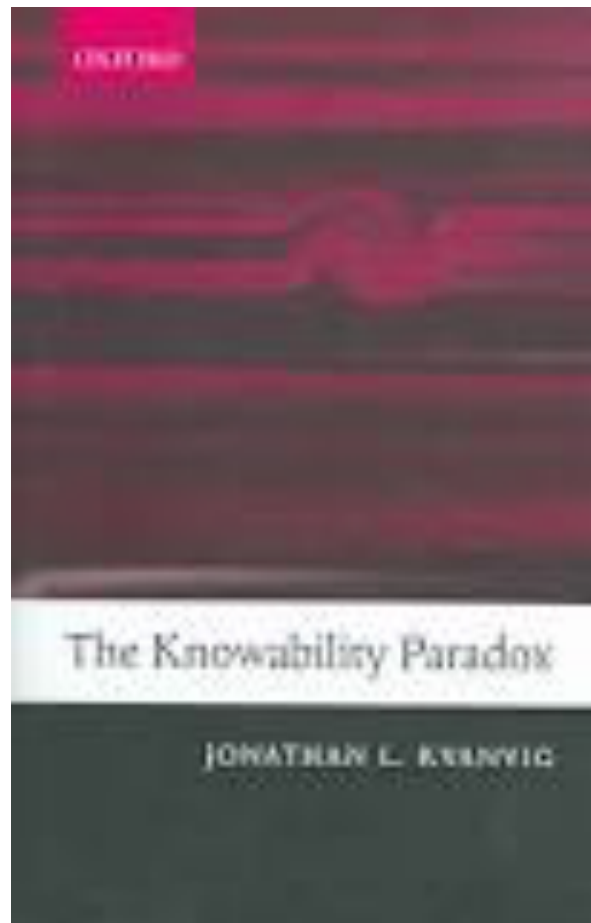
Selmer Bringsjord



By indirect proof (*reductio ad absurdum*) our original supposition is negated.

But since we' just proved a theorem (there are no remaining suppositions), we can deduce that some proposition (let's call it q*) is unknowable [because we add a box (this time to represent 'necessarily')].

For the genuinely









```
////////////////////////////////////
{: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))))))))})
////////////////////////////////////
```

Sandbox

/Library/Java/JavaVirtualMachines/jdk1.8.0_112.jdk/Contents/Home/bin/java ...



```
////////////////////////////////////
{: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))))))))})
////////////////////////////////////
```

Sandbox

/Library/Java/JavaVirtualMachines/jdk1.8.0_112.jdk/Contents/Home/bin/java ...



```
////////////////////////////////////
{: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))))))))))}
////////////////////////////////////
```

Sandbox

/Library/Java/JavaVirtualMachines/jdk1.8.0_112.jdk/Contents/Home/bin/java ...

$$\exists \phi \neg \Diamond \exists a \mathbf{K}[a, T(\phi) \wedge \neg \exists a' \mathbf{K}(a', T(\phi))]$$

Λ Example

```
{:name      "Knowability paradox"
 :description "\exists p ~\Diamond \exists x Kx (Tp & ~ \exists 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.

C-Consciousness Directly
from *Cogito ergo sum* ...

Bringsjord & G. Road to C-Con

Bringsjord & G. Road to C-Con

Descartes's *cogito ergo sum* argument.

Bringsjord & G. Road to C-Con

Descartes's *cogito ergo sum* argument.

Analyze & formalize Descartes's argument, *as specimen of* HLC-consciousness.

Bringsjord & G. Road to C-Con

Descartes's *cogito ergo sum* argument.

Analyze & formalize Descartes's argument, *as specimen of* HLC-consciousness.

What did it take to achieve this formalization?

Bringsjord & G. Road to C-Con

Descartes's *cogito ergo sum* argument.

Analyze & formalize Descartes's argument, *as specimen of* HLC-consciousness.

What did it take to achieve this formalization?

That which it took is part of the very machinery of HLCC;
and that machinery allows for the building of the CA axiom system.

Bringsjord & G. Road to C-Con

Descartes's *cogito ergo sum* argument.

Analyze & formalize Descartes's argument, *as specimen of* HLC-consciousness.

What did it take to achieve this formalization?

That which it took is part of the very machinery of HLCC;
and that machinery allows for the building of the CA axiom system.



Bringsjord & G. Road to C-Con

Descartes's *cogito ergo sum* argument.

Analyze & formalize Descartes's argument, *as specimen of* HLC-consciousness.

What did it take to achieve this formalization?

That which it took is part of the very machinery of HLCC;
and that machinery allows for the building of the CA axiom system.



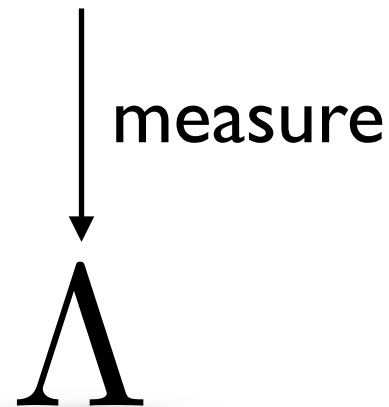
Bringsjord & G. Road to C-Con

Descartes's *cogito ergo sum* argument.

Analyze & formalize Descartes's argument, *as specimen of* HLC-consciousness.

What did it take to achieve this formalization?

That which it took is part of the very machinery of HLCC;
and that machinery allows for the building of the CA axiom system.



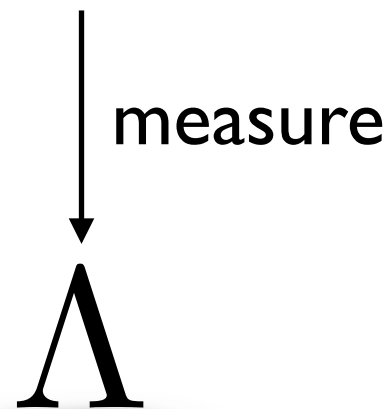
Bringsjord & G. Road to C-Con

Descartes's *cogito ergo sum* argument.

Analyze & formalize Descartes's argument, *as specimen of* HLC-consciousness.

What did it take to achieve this formalization?

That which it took is part of the very machinery of HLCC;
and that machinery allows for the building of the CA axiom system.

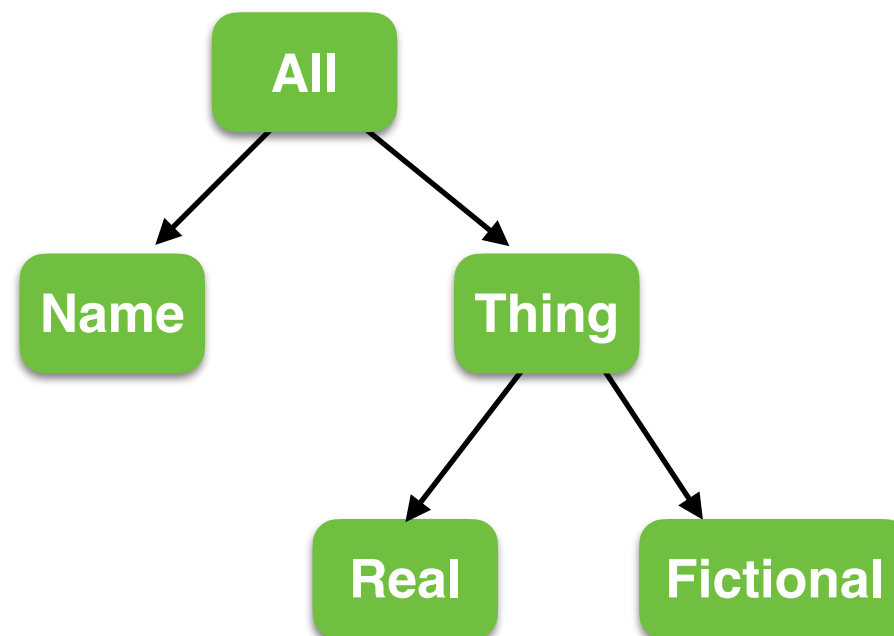


“The right way to think about the general problems of metaphysics and epistemology is not to clear one’s mind of all knowledge and start with ‘Cogito ergo sum’ and build up from there. Instead, we propose to use all of our knowledge to construct a computer program that knows.”
— John McCarthy & Patrick Hayes

Cogito Ergo Sum

Sort system

```
(Believes! I (forall [x] (or (Name x) (Thing x))))  
(Believes! I (forall (x) (iff (Name x) (not (Thing x)))) )  
(Believes! I (forall (x) (if (Thing x) (or (Real x) (Fictional x)))) )  
(Believes! I (forall (x) (if (Thing x) (iff (Real x) (not (Fictional x))))))
```



Cogito Ergo Sum

```
(Believes! I (forall (y) (if (Name y)
                             (iff (DeReExists y)
                                   (exists x (and (Real x) (= x (* y))))))))))
```

De re existence is real existence

```
(Believes! I (not (DeReExists I)))
```

I believe I don't really exist

Cogito Ergo Sum

Instance of axiom schema

```
(Believes!  
  I  
  (forall [?agent]  
    (if (Perceives! I (Believes! ?agent (not (DeReExists ?agent))))  
      (Real (* ?agent))))))
```

I believe that if I perceive an agent believe something then I believe that agent is real

Cogito Ergo Sum

```
{:name      "Cogito Ergo Sum"
 :description "A formalization 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

```
{:name      "Cogito Ergo Sum"
 :description "A formalization 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

*Ulogiske agenter er
ikke barre uintelligente
— de er ikke bevisste.*

- Personalized Problems Have Commenced ...
- Final Project Discussion ...