Introducing Λ
for Measuring Cognitive Consciousness

Selmer Bringsjord & Naveen Sundar Govindarajulu

Rensselaer AI & Reasoning (RAIR) Laboratory
Rensselaer Polytechnic Institute (RPI)
Troy NY USA

1 Introduction and Plan

In this brief work-in-progress document we introduce Λ, a new and novel framework for measuring **cognitive consciousness** that stands in dramatic contrast with the longstanding Φ of Tononi (2012), which gives a measure of **phenomenal consciousness**. Our plan herein is straightforward: First, we quickly distinguish between these two radically different types of consciousness (§2). Because Λ is erected atop a distinctive foundation developed by us, we rapidly describe three salient parts of this foundation (§3). In §4 we give a glimpse of the technical side of Λ, and rely there on an example of moral cognitive consciousness. The penultimate section, 5, lists a few distinctive properties (corresponding to underlying theorems) of the Λ framework; and then we wrap up with a concluding remark.[1]

2 Synoptic Explication of Cognitive Consciousness

Phenomenal consciousness, or 'P-consciousness' for short, is "what it's like" consciousness; Block (1995) provides a nice description thereof. E.g., there's something it's like to taste a glorious Burgundian red wine, carve a ski turn at high speed, etc. There's also something it's like to be you, and something it's like to be us. P-consciousness is rather hard to formalize, to put it mildly; for this reason, some have claimed that any dedicated attempt to build a P-conscious machine is a non-starter, at least at present Bringsjord (2007). Cognitive consciousness, on the other hand, are those states of an agent that involve its knowing, believing, intending, desiring, perceiving, fearing, communicating, ... only structurally and computationally speaking, with not even the slightest nod in the direction of "raw feels" or "qualia."

Remark: While we have no doubt that phenomenal consciousness has great practical value, there can also be little doubt that it's not exactly easy to put one's finger on what its value is; this is especially true if one is searching for its value from an evolutionary perspective Bringsjord et al. (2002). Cognitive consciousness stands in stark contrast to this situation, for even in the case of

[1] A brief appendix supplies some further details re. Λ.

AI, it's pretty clear that a cognitively conscious machine, especially one that (as measured by Λ; see below) has a high level of cognitive consciousness, is clearly a powerful machine. The unparalleled leverage *H. sapiens sapiens* have achieved over the environment is a result, at least in great measure, of cognitive consciousness; the comparatively lower position of other species, relative to that of *H. s s*, is in turn a result, at least in large part, of low and — in some cases — no cognitive consciousness. We expect that as the formal theory of Λ is erected and specified, the power of AIs will accordingly become rigorously measurable.

3 Four-Part Foundation for Λ

The foundation for Λ consists in four parts, to wit:

1. *Cognitive Calculi*. This is an infinite space of what might fairly be termed "cognitive logics" that roughly coincides with a family of multi-operator higher-order[2] quantified modal logics. One sub-family of the space is $\mathcal{DCEC}^*$, which anchors the example given below, and whose typed signature, and inference schemata, are given e.g. in Govindarajulu & Bringsjord (2017*a*).
2. *The Axiom System* $\mathcal{CA}$. An initial, formal axiomatization of cognitive consciousness has been achieved, via the axiom system $\mathcal{CA}$ given in Bringsjord, Bello & Govindarajulu (2018); this system is expressed in a cognitive calculus.
3. *ShadowProver* (the reasoner). Bringing artificial agents to (cognitively) conscious life is enabled by an automated theorem-proving system able to handle the highly expressive nature of cognitive calculi: viz., ShadowProver (Govindarjaulu 2017).[3]
4. *Spectra* (the planner). Artificial agents plan to achieve their goals and desires through Spectra (Govindarajulu 2018), a planner that can handle arbitrary goals and background information represented in cognitive calculi.[4]

4 On Λ Itself

The Basic Idea: Λ provides a measure of the degree of cognitive consciousness for an agent at a time (and over intervals composed of such times), and does so by first appropriating standard $\Delta/\Sigma/\Pi$ measures of the complexity of purely extensional formulae in logics like first- and second-order logic.[5] From there,

[2] For cognoscenti, it would be more accurate to say that the extensional component of cognitive calculi are generally n-order, where $0 \leq n \leq 3$.

[3] The novel technique of **shadowing**, which allows for great efficiency of proof-search, is out of scope here. The prover is available at: https://github.com/naveensundarg/prover.

[4] The planner is available at: https://github.com/naveensundarg/Spectra.

[5] While we can explain in person from scratch at the symposium with slides, cognoscenti can immediately get an informative sense of Λ by being told that e.g.

operators are allowed to entire the picture, and Λ tracks and measures complexity that arises from this. An example follows.

Example *Consciousness and the Doctrine of Double Effect* Figure 1 shows how Λ measures cognitive consciousness in an agent that is in the process of computing the Doctrine of Double Effect (DDE) (Govindarajulu & Bringsjord 2017*b*), a complex ethical principle that applies in moral dilemmas.[6]

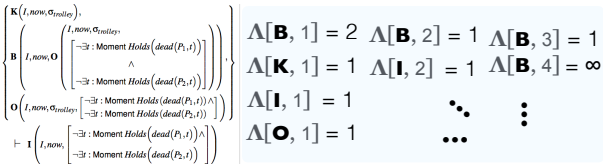

Fig. 1. Λ Applied to Single Chunk & Time of Conscious Agent

5 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,[7] 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 Palo Alto is located in California.

where ϕ is Π_3 (e.g., the formal definition of a limit in the differential/integral calculus), an agent a who knows that another agent a' knows ϕ at some time t, where this is the sum total of the cognitive activity of a, has by this alone a level of Λ that is $\Lambda[\mathbf{K}, 1] = 3$. See the appendix for more details.

[6] The principle is worth studying from the perspective of cognitive science, as there have been empirical studies that humans employ this principle.

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

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 can 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 beliefs about the beliefs of other agents, where the target of those beliefs involves at least basic quantification.[8]

Non-existent Agents Can Have Cognitive Consciousness In fact, such agents can have *high* levels of cognitive consciousness. For example, brilliant fictional detectives, such as Poe's remarkable C. Auguste Dupin, can be shown to have high levels of cognitive consciousness. E.g., in "The Purloined Letter" Dupin exploits his ability to infer what logic dictates he should believe about the criminal's beliefs about the beliefs of detectives investigating said detective.

6 What About Machine Learning?

AIs based purely on statistical/connectionist "machine learning" of the contemporary sort (e.g., "deep learning") would appear to have no cognitive consciousness whatsoever, for the simple reason that arguments to Λ can't be found in such artificial agents.[9] We expect resistance from proponents of today's ML, and look forward to discussion and debate.

7 Conclusion

We have reached the point, in the relevant line of work, at which we can provide some interesting theorems regarding the Λ framework, and can, courtesy of implementation enabled by ShadowProver, demonstrate cognitively conscious machines. We will provide some demonstrations at the symposium itself, by which time we expect these demonstrations, and the body of formal results underlying them, to be rather robust.

8 Acknowledgments

The invention and refinement of cognitive calculi has been in part enabled by generous and longstanding support from AFOSR to the RAIR Lab. The invention and refinement of cognitive calculi that allow an AI to engage specifically

[8] An artificial agent able to solve FBT is presented by Arkoudas & Bringsjord (2009); the agent uses an early cognitive calculus.

[9] For a defense of the proposition that such AIs don't actually learn a thing, a result directly in line with their scoring a zero on Λ, see (Bringsjord, Govindarajulu, Banerjee & Hummel 2018).

in moral cognition has been made possible by ONR funding to the authors. The invention by Bringsjord and Govindarajulu of Λ occurred in the course of SRI's T&C series during the summer of 2017, immediately after the former's spirited objections to Φ, and ensuing debate with both Tononi and Koch. Both Bringsjord and Govindarajulu are indebted to numerous other participants in the T&C series, and of course to SRI itself for making the series possible.

Bibliography

Arkoudas, K. & Bringsjord, S. (2009), 'Propositional Attitudes and Causation', *International Journal of Software and Informatics* **3**(1), 47–65.
URL: *http://kryten.mm.rpi.edu/PRICAI_w_sequentcalc_041709.pdf*

Block, N. (1995), 'On a Confusion About a Function of Consciousness', *Behavioral and Brain Sciences* **18**, 227–247.

Bringsjord, S. (2007), 'Offer: One Billion Dollars for a Conscious Robot. If You're Honest, You Must Decline', *Journal of Consciousness Studies* **14**(7), 28–43.
URL: *http://kryten.mm.rpi.edu/jcsonebillion2.pdf*

Bringsjord, S., Bello, P. & Govindarajulu, N. (2018), Toward Axiomatizing Consciousness, *in* D. Jacquette, ed., 'The Bloomsbury Companion to the Philosophy of Consciousness', Bloomsbury Academic, London, UK, pp. 289–324.
URL: *http://kryten.mm.rpi.edu/SB_PB_NSG_TowardAxiomatizingConsciousness_offprint.pdf*

Bringsjord, S., Govindarajulu, N., Banerjee, S. & Hummel, J. (2018), Do Machine-Learning Machines Learn?, *in* V. Müller, ed., 'Philosophy and Theory of Artificial Intelligence 2017', Springer SAPERE, Berlin, Germany, pp. 136–157. This book is Vol. 44 in the book series. The paper answers the question that is its title with a resounding No. A preprint of the paper can be found via the URL given here: http://kryten.mm.rpi.edu/SB_NSG_SB_JH_DoMachine-LearningMachinesLearn_preprint.pdf.

Bringsjord, S., Noel, R. & Ferrucci, D. (2002), Why Did Evolution Engineer Consciousness?, *in* J. Fetzer & G. Mulhauser, eds, 'Evolving Consciousness', Benjamin Cummings, San Francisco, CA, pp. 111–138.

Govindarajulu, N. & Bringsjord, S. (2017*a*), On Automating the Doctrine of Double Effect, *in* C. Sierra, ed., 'Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)', International Joint Conferences on Artificial Intelligence, pp. 4722–4730.
URL: *https://doi.org/10.24963/ijcai.2017/658*

Govindarajulu, N. S. (2018), 'Spectra (v1.0): Planning with Cognitive States and Unrestricted Domains'. URL: https://github.com/naveensundarg/Spectra, DOI: 10.5281/zenodo.1442429.
URL: *https://doi.org/10.5281/zenodo.1442429*

Govindarajulu, N. S. & Bringsjord, S. (2017*b*), On Automating the Doctrine of Double Effect, *in* C. Sierra, ed., 'Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17', Melbourne, Australia, pp. 4722–4730.
URL: *https://doi.org/10.24963/ijcai.2017/658*

Govindarjaulu, N. S. (2017), 'ShadowProver: A Fast and Exact Prover for Higher-order Modal Logic'. URL: https://github.com/naveensundarg/prover, DOI: 10.5281/zenodo.1451808.
URL: *https://doi.org/10.5281/zenodo.1451808*

Penn, D., Holyoak, K. & Povinelli, D. (2008), 'Darwin's Mistake: Explaining the Discontinuity Between Human and Nonhuman Minds', *Behavioral and Brain Sciences* **31**, 109–178.

Tononi, G. (2012), *Phi: A Voyage from the Brain to the Soul*, Pantheon, New York, NY.

A A Draft Specification of Λ

Λ for a given system is defined as a function from modal operators $\mathcal{M} = \{\mathbf{B}, \mathbf{K}, \mathbf{I}, \mathbf{D}, \mathbf{O}, \ldots\}$ (denoting a range of cognitive states in a cognitive calculus) and natural numbers to natural numbers: $\Lambda : \mathcal{M} \times \mathbb{N} \rightarrow \mathbb{N}$. The cognitive states we usually consider include, but are not limited to, beliefs **B**, knowledge **K**, intentions **I**, desires **D**, and obligations **O**. A part of the specification is given below:

Intensional Complexity of Representations/Formulae

$\Lambda[\mathbf{B}, 0]$ = maximum intensional complexity of beliefs

$\Lambda[\mathbf{K}, 0]$ = maximum intensional complexity of knowledge

$\Lambda[\mathbf{I}, 0]$ = maximum intensional complexity of intensions

$\Lambda[\mathbf{D}, 0]$ = maximum intensional complexity of desires

$\Lambda[\mathbf{O}, 0]$ = maximum intensional complexity of obligations

$\vdots$

Quantificational Complexity of Representations/Formulae

$$\Lambda[\mathbf{B}, 1] = n \begin{cases} n \text{ where we have } \mathbf{B}(\phi) \text{ and } \phi \text{ is either } \Sigma_n/\Pi_n \\ \text{i.e., maximum quantificational depth of beliefs} \end{cases}$$

$\Lambda[\mathbf{K}, 1]$ = maximum quantificational depth of knowledge

$\Lambda[\mathbf{I}, 1]$ = maximum quantificational depth of intensions

$\Lambda[\mathbf{D}, 1]$ = maximum quantificational depth of desires

$\Lambda[\mathbf{O}, 1]$ = maximum quantificational depth of obligations

$\vdots$

Extensional Complexity of Representations/Formulae

$\Lambda[\mathbf{B}, 2]$ = maximum extensional depth of belief

$\Lambda[\mathbf{K}, 2]$ = maximum extensional depth of knowledge

$\Lambda[\mathbf{I}, 2]$ = maximum extensional depth of intension

$\Lambda[\mathbf{D}, 2]$ = maximum extensional depth of desire

$\Lambda[\mathbf{O}, 2]$ = maximum extensional depth of obligation

$\vdots$

Time Complexity of Representations/Formulae

$\Lambda[\mathbf{B}, 3]$ = Maximum difference between time expressions within beliefs

$\Lambda[\mathbf{K}, 3]$ = Maximum difference between time expressions within knowledge

$\Lambda[\mathbf{I}, 3]$ = Maximum difference between time expressions within intensions

$\Lambda[\mathbf{D}, 3]$ = Maximum difference between time expressions within desires

$\Lambda[\mathbf{O}, 3]$ = Maximum difference between time expressions within obligations

$\vdots$