Towards a Computable & Harnessable Model of Consciousness

Naveen Sundar Govindarajulu & Selmer Bringsjord

Rensselaer Polytechnic Institute
Troy NY, USA

Abstract. We present a computable model of consciousness that is a modification of an existing model of universal computation. This modification is partially motivated by two existing, but non **harnessable**, models of consciousness. We say a model of consciousness is harnessable *iff* the following statement holds: *if the model predicts that a system v is more conscious than another system u, then we should, in general, find v more useful than u in a wide range of tasks.* While there are no domain-general definitions of what makes a system harnessable, we give a preliminary proposal here and assess our model against this yardstick.[1]

1 Introduction

We present a computable model of consciousness by slightly altering an existing universal model of computation by adding in **non-wellfounded** objects. The model of computation that we use is the Kolmogorov-Uspensky one (Kolmogorov & Uspensky 1958). Non-wellfounded objects are ones that can contain themselves; see (Moss 2018) for an introduction. While common foundational theories of mathematics, such as Zermelo–Fraenkel theory with the axiom of choice (ZFC) are based on axioms that rule out non-wellfounded objects, variations of set theory based on axioms such as Aczel's anti-foundation axiom allow for such objects.

We also assert that any model of consciousness should be **harnessable**; i.e., if the model predicts that a system v is more conscious than another system u, then we should, in general, find v more harnessable than u. While there are no domain-general definitions of what makes a system harnessable, we give a preliminary proposal here and assess our model against this yardstick.[2]

The plan for the paper is as follows. First, we lay down conditions for what it means for a system to have harnessable consciousness. We then briefly discuss

[1] A preliminary version of this research was presented at the SRI 2017 Technology and Consciousness workshop series. We are grateful for the comments received during the workshop. Support from AFOSR and ONR has enabled the development of formal systems that underlie some of the work presented here, particularly moral cognition that requires *de se* reasoning.

[2] Models such as Φ by Tononi (2012) do not appear to be readily harnessable, see Aaronson (2014)

prior models of consciousness based on non-wellfounded objects. In (§4), we present a harnessable formal system Ω^{Ω}. We arrive at this model by altering an existing model of universal computation, the Kolmogorov & Uspensky model of computation. Finally, we sketch examples showing the model satisfying some of the harnessability conditions laid out before. We end by discussing future work and next steps.

2 Harnessability (A Proposal)

For a model of consciousness to be harnessable, we require the following set of high-level conditions be satisfied by any system that the model asserts is conscious. (These conditions should be considered a working set of conditions rather than a finalized set of conditions.)

Condition 1 *Differentiation of de se/de re/de dicto beliefs*: Any model of consciousness that predicts that a system is conscious should also require the system to differentiate between *de se/de re/de dicto beliefs*. We illustrate these three kinds of beliefs with the following example. Let us say that there is a room in which there are two agents and an object, a flower. The agent on the right is looking at the flower. There are three statements with varying levels of self-reference that the agent can make, as shown in Figure 6. See Bringsjord & Govindarajulu (2013) for more details.

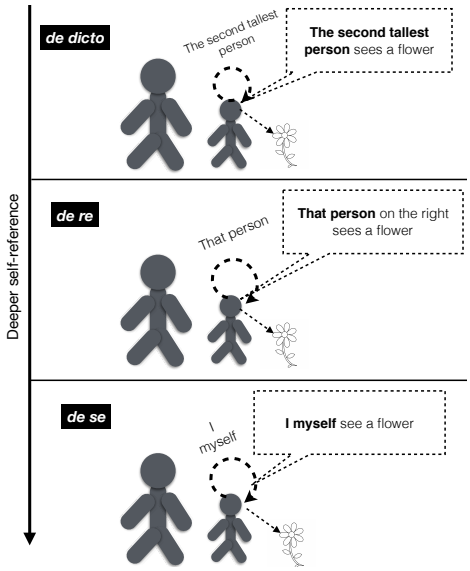

Fig. 1: **Condition 1** Differentiation of de se/de re/de dicto beliefs

Condition 2 *Unbounded Iterated Self Statements*: Any model of consciousness that predicts that a system is conscious should also require the system to be-

lieve an unbounded sequence of iterated self-belief statements without artificially kludging them in. For instance, if the system is perceiving a flower, as shown in Figure 2, the model should predict, at least under some conditions, the system to believe in expressions that correspond to the following statements: *There is a flower*, *Agent x sees a flower*, *I see a flower*, *I see that I see a flower*, *I see that I see that I see a flower*, ...

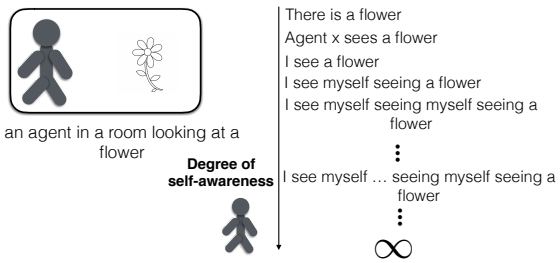

Fig. 2: **Condition 2** Spontaneous Iterated Self Statements

Condition 3 *Non-trivial Temporal Unboundedness*: Any model of consciousness that predicts that a system is conscious should predict that the system acts over a large interval of time $T \gg 0$ in a non-trivial fashion (unlike simple computational systems that can operate for decades).
Condition 4 *Theory of Minds*: The final condition requires that any model of consciousness that predicts that a system is conscious should also allow for a mechanism that enables the system to form beliefs about other agents.

3 Prior Work

Miranker & Zuckerman (2009) present a model of consciousness that is based on non-wellfounded sets. Particularly, they take as a primitive a *consciousness* operator. This operator is not defined further in terms of computable mechanisms. Corazza (2014) presents a model of consciousness derived from Advaita, one of the main branches of Vedic philosophy. Briefly, Advaita states only pure consciousness exists and the physical universe arises due to consciousness interacting with itself. This monistic conception is not easy to model in standard set theory. So Corazza leverages a version of set theory with non-wellfounded objects to present a model that satisfies several principles of Advaita. While both these prior studies are robust philosophically and mathematically, they do not give us an easily computable or harnessable model.

4 System Ω^{Ω}

KU machines were introduced by Kolmogorov & Uspensky (1958) to capture a general definition of algorithms that more closely resembles human cognition.

They have been used in attempted proofs of the Church Turing thesis due to their high-level of presentation, as compared with other models of computation.[3] We present below a KU machine formalism rooted in formal logic. We have a formal logic $\mathcal{F} \equiv \langle \mathbf{L}, \mathbf{I} \rangle$ composed of a set of expressions from a language $\mathbf{L}$ and a finite set of inference schemata $\mathbf{I}$.

KU Machine States

KU machines operate on a set of states $\mathcal{S}$. States are directed graphs with labelled edges. Each node has either: (1) an associated expression from $\mathbf{L}$; (2) or an inference schemata from $\mathbf{I}$. Note: A KU machine state denotes one or more proofs in $\mathcal{C}$.[a]

[a] A KU machine state can be considered analogous to a workspace in the Slate system (Bringsjord et al. 2008). For a formalized version of Slate, see (Govindarajulu 2013, chap. 3).

In every KU machine state, there is a unique node known as the *focal node*, f, at any stage in the computation, and revolving around the focal node is the *active patch*. The active patch consists of all the nodes which are within a fixed distance, n (called the *attention span*) from the focal node. At any step in the computation, the next step depends solely on the active patch (and relevant instructions in the program). A KU machine program consists of a function γ with a finite domain. This is given in the form of a finite set of pairs:

$$\{(P_1, \gamma(P_1)), (P_2, \gamma(P_2)), \ldots, (P_k, \gamma(P_k))\}.$$

The algorithm proceeds by replacing the active patch with $\gamma(P)$ if the active patch is equivalent to P. Associated with each pair $(P_i, \gamma(P_i))$ is a mapping ϕ_i between the nodes in the boundary of the active patch P_i to certain nodes in $\gamma(P_i)$. The mapping ϕ_i ensures that the new active patch aligns with the rest of the dataspace. We modify the KU machine formalism to allow any node to contain an entire state.

Ω^Ω KU Machine States

A node n in a KU machine state $S \in \mathcal{S}$ can have associated one of the following entities:

1. Any expression from $\mathbf{L}$
2. Any inference schemata from $\mathbf{I}$
3. Any state from $\mathcal{S}$, including S

The following definition connects beliefs of an agent modeled using a KU machine with the state of the machine.

Belief Definition

If an expression ϕ is within the active patch of a KU machine state S at time t and if that state corresponds to an agent a's state of mind at time t, we can

[3] See (Smith 2013, chap. 45) for one such attempt.

say that the agent believes ϕ at time t. Recursively, the agent also believes in all formulae within the active patches of any states within the state S.

Finally, we also require the following condition for the formal logic $\mathcal{F}$.

Identity

$\mathcal{F}$ should include mechanisms for handling identities

1. The alphabet of $\mathbf{L}$ should have the identity symbol $=$ and the grammar should allow for expressions containing the identity symbol in the usual manner, i.e., $e_1 = e_2$.
2. First-order inference schemata for $=$ should be in $\mathbf{I}$

4.1 Harnessability of Ω^{Ω}

Using examples, we demonstrate that the model could satisfy the first two harnessability conditions. Demonstration that final harnessability condition can be satisfied is left out. Satisfiability of third condition is still open.

Condition 1 *Differentiation of de se/de re/de dicto beliefs* The three kinds of beliefs can differentiated with the help of identity and containment of states. Figure 3 contains three different KU machine states. The active patch of a state is shown in blue. The first two states depict *de dicto* and *de re* beliefs. The final state contains itself and depicts a *de se* belief due to proper containment of the state within itself.

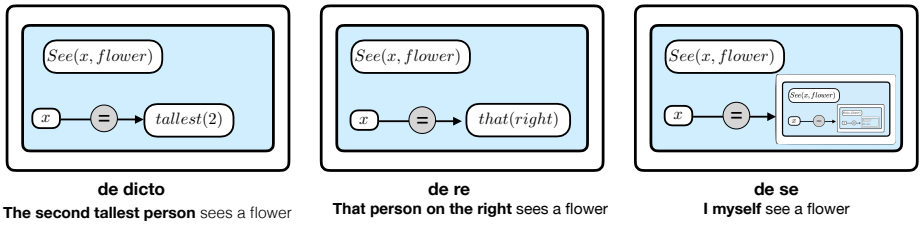

Fig. 3: **Condition 1** Differentiation of de se/de re/de dicto beliefs. The state on the right satisfies the condition.

Condition 2 *Unbounded Iterated Self Statements*: Consider the state S in Figure 4. We have a statement $See(x, \phi)$ where x and ϕ are asserted to be equal to the state S. Using the belief definition given above, we can show that the state satisfies **Condition 2**.

5 Implementation

It should be noted that in mathematics there can be two kinds of non-wellfounded structures: (1) an object that *contains* itself; and (2) an object *defined* in terms

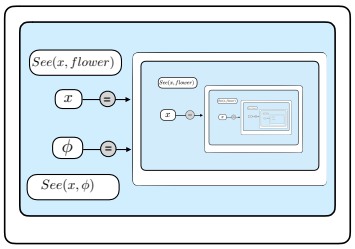

Fig. 4: **Condition 2** This state satisfies the condition.

of itself. (Moss 2018). The second form of non-wellfoundedness is not that problematic. But unfortunately, Ω^{Ω} has non-wellfoundedness of the first kind. We can approximate non-wellfounded objects computationally using object references, where objects contains references to themselves, rather than contain themselves properly. We have modified *ShadowProver*, a quantified modal logic prover (Govindarjaulu 2017), to handle such kind of references. *ShadowProver* has been used to model complex cognitive principles, for example, see an application of *ShadowProver* in (Govindarajulu & Bringsjord 2017) to model a complex moral principle. See Figure 5 for a state corresponding to the one used in the examples above for **Condition 1** and **Condition 2**.

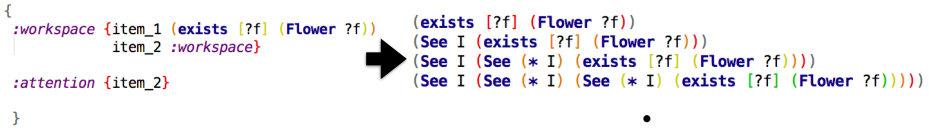

Fig. 5: **ShadowProver Modified** Conditions 1 & 2 being handled by the prover

6 Conclusion

We have presented a model of consciousness Ω^{Ω} that is based on a model of universal computation: the KU model of computation. We arrive at the model by altering the KU model of computation in a minor fashion. Since Ω^{Ω} is derived from a model of computation, it is readily mechanizable. We define harnessability conditions and provide examples that indicate Ω^{Ω} might be harnessable. We have a preliminary implementation that we have achieved by a modifying a reasoning system, a quantified modal logic prover. Future work consists of refining the harnessability conditions and using Ω^{Ω} to solve problems in micro-domains. Particularly, we have to give a good account of how Ω^{Ω} can lead to non-trivial long living AI systems. Another future strand of research will address how this model can be fruitfully applied to systems such as deep neural networks that have been successful in a variety of perceptual tasks.

Bibliography

Aaronson, S. (2014), 'Why I Am Not An Integrated Information Theorist', https://www.scottaaronson.com/blog/?p=1799tm. Accessed: 2018-11-12.

Bringsjord, S. & Govindarajulu, N. S. (2013), Toward a Modern Geography of Minds, Machines, and Math, *in* V. C. Müller, ed., 'Philosophy and Theory of Artificial Intelligence', Vol. 5 of *Studies in Applied Philosophy, Epistemology and Rational Ethics*, Springer, New York, NY, pp. 151–165.
URL: *http://www.springerlink.com/content/hg712w4l23523xw5*

Bringsjord, S., Taylor, J., Shilliday, A., Clark, M. & Arkoudas, K. (2008), Slate: An Argument-Centered Intelligent Assistant to Human Reasoners, *in* F. Grasso, N. Green, R. Kibble & C. Reed, eds, 'Proceedings of the 8th International Workshop on Computational Models of Natural Argument (CMNA 8)', University of Patras, Patras, Greece, pp. 1–10.
URL: *http://kryten.mm.rpi.edu/Bringsjord_etal_Slate_cmna_crc_061708.pdf*

Corazza, P. (2014), 'Mathematics of Pure Consciousness'. URL: http://pcorazza.lisco.com/papers/MVS/corazza-mathOfConsciousness.pdf.

Govindarajulu, N. S. (2013), Uncomputable Games: Games for Crowdsourcing Formal Reasoning, PhD thesis, Rensselaer Polytechnic Institute (RPI), Troy, NY.

Govindarajulu, N. S. & Bringsjord, S. (2017), 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. Preprint available at this url: https://arxiv.org/abs/1703.08922.
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*

Kolmogorov, A. & Uspensky, V. (1958), 'On the Definition of an Algorithm', *Uspekhi Matematicheskikh Nauk* **13**(4), 3–28.

Miranker, W. L. & Zuckerman, G. J. (2009), 'Mathematical Foundations of Consciousness', *Journal of Applied Logic* **7**(4), 421–440.

Moss, L. S. (2018), Non-wellfounded Set Theory, *in* E. N. Zalta, ed., 'The Stanford Encyclopedia of Philosophy', summer 2018 edn, Metaphysics Research Lab, Stanford University.

Smith, P. (2013), *An Introduction to Gödel's Theorems*, Cambridge University Press, Cambridge, UK. This is the second edition of the book.

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

A ShadowProver Example

The figure below shows an input problem to *ShadowProver*. The problem comes from a situation in Edgar Allan Poe's *The Purloined Letter*. Note that situation contains many levels of iterated beliefs. *ShadowProver* solves this problem in around $\tilde{55}ms$ on a machine with 2.9 GHz CPU and 16 GB of memory.

```
{:name        "The Purloined Letter"
 :description "Dupin's reasoning as he goes through the case"

 :assumptions {1 (Believes! g (hide m elaborate))
               2 (Believes! d (or (hide m elaborate) (hide m plain)))
               3 (Believes! m (Believes! g (hide m elaborate)))
               4 (if (Believes! m (Believes! g (hide m elaborate))) (hide m plain))
               5 (if (Believes! m (Believes! g (hide m plain))) (hide m elaborate))
               6 (Believes! m (Believes! g (hide m elaborate)))
               7 (Believes! d (if (Believes! m (Believes! g (hide m elaborate))) (hide m plain)))
               8 (Believes! d (if (Believes! m (Believes! g (hide m plain))) (hide m elaborate)))
               9 (Believes! d (Believes! m (Believes! g (hide m elaborate))))}


 :goal (Believes! d (hide m plain))}
```

Fig. 6: **ShadowProver Example** A problem supplied to the prover