Why can the brain (and not a computer) make sense of the liar paradox?

Patrick Fraser*,[1] Ricard Solé,[2,3,4] and Gemma de las Cuevas[5]
[1] Department of Philosophy, University of Toronto, Toronto, Ontario, Canada
[2] ICREA-Complex Systems Lab, UPF-PRBB, Dr. Aiguader 80, 08003 Barcelona
[3] Institut de Biologia Evolutiva, CSIC-UPF, Passeig Maritim de la Barceloneta 37, 08003 Barcelona
[4] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501, USA.
[5] Institute for Theoretical Physics, Techinkerstr. 21a, A-6020 Innsbruck, Austria

Ordinary computing machines prohibit self-reference because it leads to logical inconsistencies and undecidability. In contrast, the human mind can understand self-referential statements without necessitating physically impossible brain states. Why can the brain make sense of self-reference? Here, we address this question by defining the Strange Loop Model, which features causal feedback between two brain modules, and circumvents the paradoxes of self-reference and negation by unfolding the inconsistency in time. We also argue that the metastable dynamics of the brain inhibit and terminate unhalting inferences. Finally, we show that the representation of logical inconsistencies in the Strange Loop Model leads to causal incongruence between brain subsystems in Integrated Information Theory.

Keywords: Self-reference, cognition, consciousness, computation, causal structure, integrated information theory

I. INTRODUCTION

Are brains like computers? Can technological metaphors provide satisfactory explanations for the complexity of human brains (and brains in general)? Before electronic computers became a reality, some versions of the previous questions had always been there. In the seventeenth century, the development of mechanical clocks and later on mechanical automata led to questions with far-reaching philosophical implications, such as the possibility of creating a mechanical human and an artificial mind (by René Descartes and others (Wood, 2002)). Later, brains and machines were compared to electric batteries (since it became clear that electricity was involved in brain processes), and early works by visionaries such as Alfred Smee represented brains and the activity of thinking in terms of networks of connected batteries (Smee, 1850). Other network-level metaphors of the brain such as telegraphs and telephone webs replaced the old ones, until the metaphor of the computer prevailed in the 1950s (Cobb, 2020).

The computer was apparently the right metaphor: It could store large amounts of data, manipulate them and perform complex input-output tasks that involved information processing. Additionally, the new wave of computing machines provided an appropriate technological context to simulate logical elements similar to those present in nervous systems. Theoretical developments within mathematical biology by (McCulloch and Pitts, 1943) revealed one first major result: The units of cognition—neurons—could be described with a formal framework. Formal neurons were described in terms of threshold units, largely inspired by the state-of-the-art knowledge of real neurons (Rashevsky, 1960). Over the last decades, major quantitative advances have been obtained by combining neuron-inspired models with multilayer architecture (LeCun et al., 2015) and physics of neuromoprphic computing (Indiveri and Liu, 2015; Markovi et al., 2020). These developments are largely grounded in early theories (Fukushima, 1988; Rumelhart et al., 1986) with novel hardware improvements and a massive use of training data.

Despite the obvious success of computing and information technology, we are still far from the dream of building or simulating a truly intelligent system. To begin with, computers and their abstract representation in terms of Turing machines are highly modular, programmable and sequential (Arbib, 2012) (see Figure 1). Instead, neural systems are the result of evolutionary tinkering and selection that favoured exploiting redundancy and parallelism (Allman, 1999; Martinez and Sprecher, 2020). That does not prohibit the existence of interesting links that help make sense of brain in terms of Turing machines: Many functional responses of brains are essentially sequential in nature, despite the highly parallel integration that feeds serial (and slow) cognitive task production (Zylberberg et al., 2011). Yet, the most remarkable departure of brains from computers is probably the presence of re-entrant circuits, i.e. the recursive exchange of signals across multiple, parallel and reciprocal connections (Edelman, 1992). Indeed, some authors have posited that closed feedback loops are crucial for conscious experience (Hofstadter, 1979; Oizumi et al., 2014). Are closed feedback loops the key for a formal differentiation between brains and computers? Closed feedback loops can allow for self-reference (Grim, 1993), and the human brain is capable of self-referential inference. So this begs the question: Why can the brain make sense of self-reference, whereas a computer can't?

We address this question by considering paradoxes of self-reference and negation (Prokopenko et al., 2019). Studies in logic, linguistics, and general philosophy for many centuries have illustrated that when statements negatively refer to their own features, contradictions follow in short order. This is made clear from sentences such as:

The sentence presently being uttered is false. (1)

Taking this sentence at its word—supposing it to be true—we

*Corresponding author

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

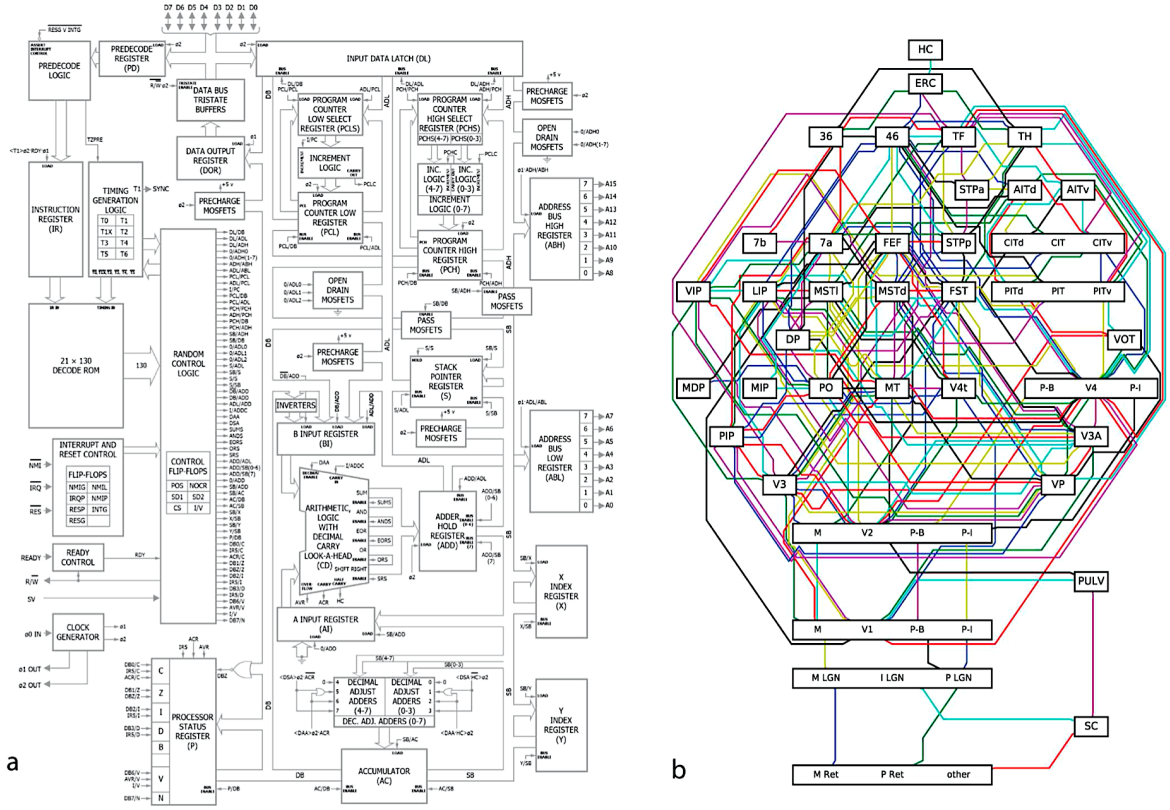

FIG. 1 Computer versus brain architecture. A topological analysis of (a) computer chips and (b) brains (visual cortex organization) reveals fundamental dissimilarities. These include the strict modular organization of the former contrasted with the highly parallel, integrated architecture of the latter. The circuits responsible for higher-order cognitive brain tasks display re-entrant feedback loops that are absent on the in-silico counterparts (compare with Figure 2). Image adapted from (Jonas and Kording, 2017).

find out it is false. However, taking it to be false, we are forced to conclude that it is true. When we assign truth values to sentences, we classically assume that truth and falsity are mutually exclusive and exhaustive, yet self-referential sentences appear to have *over-determined* truth values (Priest, 2006, pp. 14–15): We are obliged to evaluate them simultaneously as true and false, a contradiction. This may compel the logician to use formal languages that block such self-referential constructions to preserve their consistency at the cost of limiting their expressiveness. Such a pursuit of consistency is perhaps well-motivated in purely formal settings such as mathematics, but self-reference is readily available within natural language, and human minds are capable of formulating and thinking about self-referential paradoxes and becoming aware of their inconsistency.

Computers are incapable of resolving a paradox such as (1)—they get caught in endless loops—, whereas the brain can 'reason' about this paradox. Let us examine the latter statement by bringing forward some basic facts about the workings of the brain. In the ordinary course of experience, our state of mind may possess many subtle and composite features, but we only ever occupy one such mental state at a time: There are no 'superpositions' of mental states. Furthermore, if we take mental states to be somehow derivative of brain states (by whatever account of the emergence of consciousness one prefers), the deterministic or unitary evolution of physical systems given by our best physical theories suggests that our brains only ever occupy a single physical state.[1] Whatever the mechanism responsible for the emergence of mental states from brain states is, surely the brain state that grounds the awareness of some fact is different from the brain state that grounds the awareness of its negation. Thus, occupying a mental state corresponding to awareness of a contradiction would *seem* to be a physical impossibility *par excellence* insofar as it would necessitate one's brain to be in two distinct states at once. Yet, upon interpreting the sentence 1, the reader comes to think about a self-referential statement and understand its contradictory nature, and so the cognitive processing of self-referential statements is clearly *not* a physical impossibility (nor do they get stuck in an unhalting cycle of thoughts one might expect of a machine tasked with deciding the truth value of such a sentence). How is this possible?

[1] Since the brain is fundamentally a quantum system, this physical state *could* be in a superposition, but as (Tegmark, 2000) has shown, even neurons are sufficiently macroscopic systems that decoherence would likely prevent quantum effects from being relevant.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

In this paper, we address this question by constructing a high-level model of the brain, termed a Strange Loop Model (II), from which we conclude that:

1. The brain makes sense of self-reference by spreading out inconsistent truth values in time, thereby avoiding physically impossible states (III.A).

2. The representation of logical inconsistencies in the brain leads to causal incongruence between brain subsystems (III.C).

3. The metastable dynamics of the brain and its interactions with external stimuli inhibit and terminate unhalting inferences (III.D).

Statement 1 says that the brain represents and processes self-referential sentences by treating their truth values as dynamical quantities. It follows that the resulting contradictions are unfolded in time, and thus do not require physically impossible brain states. Statement 2 describes how this 'unfolding' works: Different parts of the brain yield disagreeing predictions about the brain's future states, and this disagreement is made apparent by analysing the causal feedback between these parts. This disagreement is known in Integrated Information Theory (IIT) as *incongruence* (Albantakis and Tononi, 2019). This causal feedback is not encountered in Turing machines because they are feed-forward systems. Statement 3 claims that the brain does not succumb to halting problems when processing statements whose truth values are undecidable, because the metastable nature of brain dynamics precludes falling into lock-in states (Tognoli and Kelso, 2014).

This paper is structured as follows. We present the Strange Loop Model (SLM) of the brain (II), and we use it to represent self-referential inferences in the brain (III). Finally we conclude and discuss further directions (IV).

II. THE STRANGE LOOP MODEL

Here we present a high-level model of the brain by describing it as a discrete dynamical system (II.A), partitioning it into functionally distinct modules (II.B), and investigating their causal structure (II.C). The name originates from (Hofstadter, 1979, 2007): Strange loops involve arise when, by moving only upwards (or downwards) in a hierarchy, one encounters oneself at the same place where one started.

A. Discrete dynamics of brain modules

Here we describe the brain as a discrete dynamical network of connectomic units (Sporns et al., 2005). We consider that n such units (indexed $i = 1, \dots, n$), evolving in discrete time $t \in \mathbb{Z}$, and denote the state of unit i at time t by $x_i^t \in \Sigma_i$, where Σ_i is a finite state space. The state of the 'brain' in the SLM at time t is denoted

$$B^t = (x_1^t, \dots, x_n^t) \in \Sigma_1 \times \dots \times \Sigma_n =: \Sigma.$$

The dynamics of such a system are given by a transition function $\mathcal{T} : \Sigma \to \Sigma$ so that $B^{t+1} = \mathcal{T}(B^t)$ and we denote ith component of $\mathcal{T}(B^t)$ by $\mathcal{T}_i(B^t) := x_i^{t+1}$.

We consider a probability distribution p on Σ. For any $z \in \Sigma_i$, the conditional probability (also denoted p) is defined as

$$p(z|B^t) = \begin{cases} 1 & \text{if } z = \mathcal{T}_i(B^t) \\ 0 & \text{else .} \end{cases}$$

We suppose that all units are conditionally independent at any given time $t \in \mathbb{Z}$, so they satisfy:

$$p(B^{t+1}|B^t) = \prod_{i=1}^{n} p(x_i^{t+1}|B^t). \tag{2}$$

Additionally, we suppose that the future state of the brain depends only on the immediately preceding state (Markovianity), so that if $t_1 < t_2 < \dots < T$, the joint probability distribution factors as

$$p(B^{t_1}, B^{t_2}, \dots, B^T) = p(B^{t_1}) \prod_{n=1}^{T-1} p(B^{t_{n+1}}|B^{t_n}). \tag{3}$$

With this setup one may use the intervention calculus from probabilistic causal modelling (e.g. as elaborated by (Pearl, 2009)) to understand how connectomic units causally influence each other. Following the exposition in (Krohn and Ostwald, 2017), given any two subsystems $X, Y \subseteq B$, one defines the *effect probability* p_{e}, the joint *cause-effect probability* p_{ce}, and the *cause probability* p_{c} to be:

$$p_{\mathrm{e}}(Y^t|X^{t-1}) := p(Y^t|X^{t-1}) \tag{4}$$

$$p_{\mathrm{ce}}(Y^t, X^{t-1}) := q(X^{t-1})p(Y^t|X^{t-1}) \tag{5}$$

$$p_{\mathrm{c}}(Y^{t-1}|X^t) := \frac{p_{\mathrm{ce}}(Y^{t-1}, X^t)}{\sum_{Y^{t-1} \in \Sigma} p_{\mathrm{ce}}(Y^{t-1}, X^t)} \tag{6}$$

where $q(Y^{t-1})$ is the uniform distribution over the state space of Y. The distribution $p_{\mathrm{e}}(Y^t|X^{t-1})$ indicates the extent to which the current state of Y is an effect caused the previous state of X. Likewise, $p_{\mathrm{c}}(Y^{t-1}|X^t)$ indicates the extent to which the previous state of Y was a cause of the current state of X.

B. Brain process modules

The brain carries out a wide array of distinct, though integrated processes. While it is difficult to list and classify all of them, they may be roughly partitioned into three general interconnected categories: (i) pre-conscious processes, (ii) conscious processes, and (iii) post-conscious processes.

Pre-conscious processes are those which occur independent of conscious experience. The activity of the autonomic nervous system is paradigmatic of this category. Though extremely important for sustaining life, these functions are somewhat irrelevant to our considerations and shall hence be ignored in what follows.

Conscious processes are those which directly give rise to conscious experience; that is, they govern the dynamics of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

neural correlates of consciousness, and include those responsible for perception, the categorical discrimination thereof, awareness, and short-term memory recall, among other things. They are not to be conflated with the first-person subjective conscious *experiences* to which these correlates are thought to somehow give rise. At the physiological level, all we are concerned with are the neural correlates of conscious experience and awareness; we are agnostic as to *how* the mental states are determined by these correlates, and therefore do not commit to any view about the origins of consciousness as such.

Post-conscious processes are those which are not the primary basis for conscious experience, but still depend on the correlates of consciousness such as language processing and inference-making. This class of brain functions is roughly equivalent to cognitive processes.[2]

Each of these classes of brain processes has a reasonably well-defined collection of physiological regions in the brain which carry them out. Hence it is possible for us to conceptually partition the brain into three physical 'modules.' The important feature of these modules is that they are deeply interconnected. While it is hard to cleanly demarcate their physiological boundaries, what is important for our purposes is not how to carve up the brain into these modules, but the causal relations *between* them.

In the SLM (cf. II.A), we shall denote the 'consciousness' module by $X_{\text{Con}} \subseteq B$ and the individual connectomic units that compose it by $\{x_i\}$. Likewise, we shall denote the 'cognition' module by $Y_{\text{Cog}} \subseteq B$ and the connectomic units that compose it by $\{y_i\}$. The region of the brain that is relevant for our purposes is the joint system $X_{\text{Con}} \cup Y_{\text{Cog}}$.

A concrete realisation of the Strange Loop Model

To instantiate the SLM, suppose first that a mental state amounts to the awareness of some sentence in a formal language L. Such sentences carry an internal time index τ: at physical time t, one may occupy a mental state of *remembering* some sentence ϕ at an earlier time (i.e. $\tau < t$), they may *anticipate* being aware of ϕ in the future (i.e. $\tau > t$), or they may be aware of ϕ as a feature of the present experience (i.e. $\tau = t$). We suppose that every pair (ϕ, τ) is represented by a unit of Y_{Cog}. The mental state determined by the state of X_{Con} is simulated by the elements of Y_{Cog} via an injective map $S : \Sigma_X \to \{(\phi, \tau) | \phi \in L, \tau \in \mathbb{Z}\}$ where Σ_X is the state space of X_{Con}. That is, S takes the state of X_{Con} to the unit of Y_{Cog} that represents the corresponding mental state. The state of each unit $y = (\phi, \tau) \in Y_{\text{Cog}}$ at time t is given by $y^t = (a^t(y), s^t(y)) \in \{0, 1\} \times \{0, 1\}$, where $a^t(y) = 1$ if the thinking subject is consciously aware of y at time t, and it is 0 otherwise, and $s^t(y) = 1$ if the thinking subject assigns truth to ϕ at time τ (i.e. if they think ϕ was/is/will be true at time τ), and it is 0 otherwise. The state of Y_{Cog} at time t is determined by:

$$y^t = (a^t(y), s^t(y)) = \begin{cases} (1, 1) & \text{if } y = S(X^t_{\text{Con}}) \\ (0, s^{t-1}(y)) & \text{if } y \neq S(X^t_{\text{Con}}). \end{cases}$$

That is, to be aware of (ϕ, τ) at time t is to think it to be true, and to think ϕ is false is to be aware of the truth of $\neg\phi$. The state of Y_{Cog} at time $t+1$ is determined by the application of some inferential mechanism by Y_{Cog}. If the thinking subject applies a rule of inference of the form $\{\sigma_1, \ldots, \sigma_k\} \vdash \psi$, $a^t(y)$ is updated so that one is only presently aware of ψ, namely $a^{t+1}(\psi, t+1) = 1$, and $a^{t+1}(\phi, \tau) = 0$ for $\phi \neq \psi$ and any τ. The transition rule for s^t is

$$s^{t+1}(\psi, t+1) = \prod_{i=1}^{k} s^t(\sigma_i, t)$$

and $s^{t+1}(\phi, \tau) = s^t(\phi, \tau)$ for all τ when ϕ is independent of ψ, and $s^{t+1}(\xi, \tau) = s^t(\xi, \tau)$ for all $\tau \neq t+1$ and any ξ. Sentences containing ψ have their truth values adjusted according with the change in the truth value of ψ, for example $s^{t+1}(\neg\psi, t+1) = 1 - s^{t+1}(\psi, t+1)$ and $s^{t+1}(\phi \wedge \psi, t+1) = s^{t+1}(\phi, t+1) \cdot s^{t+1}(\psi, t+1)$ and so on. We do not fully specify the transition rule for X_{Con}, but we require that it be such that after such an inference, $S(X^{t+1}_{\text{Con}}) = (\psi, t+1)$.

The self-referential paradox arises when one may assert that $(a^t(\phi, t_1), s^t(\phi, t_1)) = (a^t(\neg\phi, t_2), s^t(\neg\phi, t_2))$ for $t_1 \neq t_2$. But, as shown in III.A, this scenario is not challenging to understand; these are two different nodes of Y^t_{Cog}, and there is no consistency requirement preventing this as a value assignment. Even if one imposes consistency conditions at equal times, since these are unequal-time units, such conditions need not prohibit this behaviour.

C. Causal feedback

We now argue that the brain modules X_{Con} and Y_{Cog} mutually exhibit causal feedback.

To see that X_{Con} causally influences Y_{Cog}, note that cognitive tasks are like computational tasks (broadly construed) which take as their inputs the correlates of consciousness. For instance, learning is a cognitive process that is informed by sensory stimuli. Likewise, language processing is a cognitive process that begins with a more abstract input of which the cognizing subject is usually consciously aware. More generally, changing what a person perceives or is conscious of af-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

fects how they make sense of their perceptions and what sorts of inferences they will draw.

What does the causal relation from X_{Con} to Y_{Cog} look like? It is known that a single neuron may participate in bringing about many sorts of perceptions and experiences, and many different neuronal states may correspond to one and the same perceptual experience (as there is great degeneracy). Hence, one cannot easily reduce a correlate of consciousness to an arrangement of neurons. That is, the correlates of consciousness are not identical to the state of X_{Con}—they are only determined by X_{Con}. More specifically, the intrinsic network of causal influences within X_{Con} determines these neural correlates (see (Edelman, 2005; Park and Friston, 2013; Tononi and Edelman, 1998) for discussion).[3] In order for cognition to take the correlates of consciousness as inputs, the system Y_{Cog} must be connected to system X_{Con} in such a way that the internal causal structure of X_{Con} is 'read off' of its state and encoded directly into the states of the neurons of Y_{Cog}, which must encode features of the probability distributions p_{e}, p_{ce}, and p_{c} of the subsystem X_{Con}. Since we shall establish that there are causal relations in both directions, to prevent circularity, we suppose that Y_{Cog} represents the intrinsic causal structure of X_{Con} as it appears when marginalized to X_{Con} (i.e. ignoring correlations with Y_{Cog}). Determining exactly how this translation could be carried out would require a full account of the emergence of conscious experience from the relevant causal information which we do not have. However, one may view the units of Y_{Cog} as 'simulating' the intrinsic causal structure of X_{Con}, and then carrying out an effective computing procedure on this simulation—this simulation could be modelled with ideas from hierarchical predictive processing which adopts a similar organizational structuring of the brain (cf. (Clark, 2013; Friston, 2005; Friston and Kiebel, 2009)). In summary, the causal relation $X_{\text{Con}} \to Y_{\text{Cog}}$ is highly non-trivial.

What does the causal relation from Y_{Cog} to X_{Con} look like? On its own, the system X_{Con} gives rise to the moment-by-moment passive perceptions present in the thinking subject's conscious experience. However, the content of conscious experience—at least for humans—is not merely a passive stream of perception; there is further underlying semantic content within these perceptions of which we come to be aware by carrying out cognitive tasks. While our perceptual apparatus may be capable of carrying out discrimination tasks to categorize our perceptions (e.g. such that we may become aware of the presence of 'pain' or 'blue' and so on within a given experience), we also come to be consciously aware of much richer structural and abstract features as well. Deprived of all sensory input, the mathematician may still prove complex theorems structured by a sophisticated underlying mathematical grammar and logic, but only if they are consciously aware

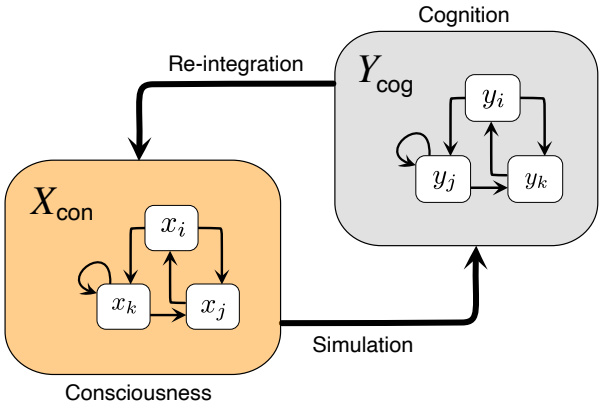

FIG. 2 The causal relations $X_{\text{Con}} \to Y_{\text{Cog}}$ needed to simulate perceptions for inference-making, and $Y_{\text{Cog}} \to X_{\text{Con}}$ manifest in the awareness of the outcome of cognitive processing.

that they are doing so. To the extent that the thinking subject may be conscious of the *outcomes* of their cognition—which they certainly are in many cases—we see that there must exist some non-trivial causal relation between Y_{Cog} and X_{Con} in which the former causally influences the latter.

More specifically, acts of cognition may change the content of conscious experience such that we may acquire *understanding* of our perceptions, for instance by giving them grammatical structure (over and above merely discriminating qualia), or by carrying out introspection or higher inference-making. It is through this process that one may go from a state of mind of the form 'it is the case that ϕ' to the state of mind 'I *know* that it is the case that ϕ.' Likewise, it is through this process that one may go from the state of mind that 'it is the case that ϕ and $\phi \to \psi$' to the state of mind 'it is the case that ψ' (via inference by *modus ponens*). In short, the outcomes of cognitive processes are *re-integrated* back into the correlates of consciousness. This causal feedback via simulation and re-integration between modules in illustrated in Figure 2.

We have established that cognition causally influences the content of conscious experience, and vice versa. This is not to say, however, that cognition is itself 'perceived.' In everyday life, the content of our experience forms the basis of some cognitive inference we may make and we become aware of the outcome of this inference, but we never perceive the inference itself. Indeed, even when one is proving mathematical theorems, at most one is aware of what cognitive rules they are applying when carrying out a deduction: they do not, however, experience the application of these rules as such. This illustrates that, while we argue that cognition causally influences the course of conscious experience in a very strong way, it is not itself *directly* responsible for conscious experience; the neuronal basis for cognition is not itself populated with correlates of consciousness, it merely interacts with these correlates in a reentrant manner. In this sense, we may faithfully view the cognitive module Y_{Cog} as implementing feed-forward computing procedures, e.g. through a neural network that is reintegrated with X_{Con} (such that inference making *in its entirety* is not merely a computing procedure).

[3] While the dynamical evolution of the brain may be reduced to a description of its individual neurons, and while its intrinsic causal structure is grounded in the interactions of these neurons, the intrinsic causal structure is not robust against small changes to the network architecture. It is in this way that neural correlates of consciousness are not 'reducible' to individual neurons.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

Formally, since both X_{Con} and Y_{Cog} causally influence one another in a highly non-trivial manner, we expect that

$$p(X^{t+1}_{\text{Con}}|Y^{t}_{\text{Cog}}) \neq p(X^{t+1}_{\text{Con}}|X^{t}_{\text{Con}}) \tag{7}$$

$$p(Y^{t+1}_{\text{Cog}}|Y^{t}_{\text{Cog}}) \neq p(Y^{t+1}_{\text{Cog}}|X^{t}_{\text{Con}}). \tag{8}$$

Thus, the simulation of X_{Con} encoded in Y_{Cog} will generally not be a faithful predictor of the future behaviour of X_{Con}, since it ignores its own causal influence on this behaviour. This is the reason we suppose that Y_{Cog} simulates the causal structure of X_{Con} as marginalized to X_{Con}. In Box 1 we provide a concrete realisation of the SLM presented above model, as well as its application to self-reference.

III. SELF-REFERENCE IN THE STRANGE LOOP MODEL

Here we use the SLM to investigate how to make sense of self-reference by unfolding the inconsistency in time (III.A) and provide some clarifying remarks (III.B). Then we show how logical inconsistency is transformed to incongruence (III.C), and argue that the brain does not get caught in endless loops (III.D).

Diagonalization, self-reference and paradoxes

While self-reference and its paradoxical consequences arise in a wide range of settings, the construction of the self-referential statement leading towards contradiction typically has a standard form, termed the Inclosure Schema; cf. (Priest, 2002, Chapter 9.4)). At a higher level of abstraction, this may be viewed as an instance of Lawvere's Theorem (Lawvere, 2006; Roberts, 2021; Yanofsky, 2003).

In plain language, the relevant actresses of the Inclosure Schema are the following. A predicate is a property that elements of a set may possess, and we identify the predicate with its extension, i.e. with the set of elements that instantiate it. For example, the predicate 'odd' of the set of natural numbers is the set $\{1, 3, 5, 7, 9, \ldots\}$. If a set x has property P, we write $P(x)$, meaning that $P(x)$ is true, i.e. x is in the extension of P. We will consider the collection of all sets V, and a function $\Delta : V \to V$.

More formally, let φ and ψ denote two predicates that may apply to arbitrary sets (where 'set' is meant in the sense of natural language, which is more expressive than formal set theory at the cost of being inconsistent), and let Δ be a function on sets. Then self-reference occurs when:

1. $E_\varphi = \{y|\varphi(y)\}$ is a set, and $\psi(E_\varphi)$
2. If $x \subseteq E_\varphi$ such that $\psi(x)$, then $\Delta(x) \notin x$ and $\Delta(x) \in E_\varphi$

Statement 1 says that the extension of the predicate φ is a set and is called E_φ, and that E_φ has property ψ. Statement 2 defines the features of Δ, namely Δ takes sets with property ψ whose *elements* all have property φ to sets whose elements have property φ but are not contained in the original set. The contradiction associated with self-reference appears when one applies condition 2 to the maximal subset, namely, E_φ itself, from which it follows that $\Delta(E_\varphi) \in E_\varphi$ and $\Delta(E_\varphi) \notin E_\varphi$; a contradiction.

Let us see this argument in action by considering Russell's paradox. In naïve set theory, the extension of any predicate is a set. Russell's paradox is as follows: suppose X is the set of all sets that do not contain themselves. Then if $X \in X$, by definition it follows that $X \notin X$. However, if $X \notin X$, then since X is the set of all sets that do not contain themselves, we find $X \in X$; a contradiction. On the Inclosure Schema this paradox may be recast as follows. First, ψ is the predicate 'is a set,' φ is the predicate 'does not contain itself', and define $\Delta(x) = \{y \in x | y \notin y\}$, i.e. as the set of all sets in x that do not contain themselves. Since ψ is a predicate in naïve set theory, 1 is true and asserts that E_φ exists and is the notorious 'set of all sets that do not contain themselves.' Then if x is a set, clearly $\Delta(x) \in E_\varphi$. Likewise, if $\Delta(x) \in \Delta(x)$, then by definition of Δ, $\Delta(x) \notin \Delta(x)$ and so we must conclude $\Delta(x) \notin x$. Thus 2 is also satisfied. But then setting $x = E_\varphi$, this implies simultaneously that $\Delta(E_\varphi) \in E_\varphi$ and $\Delta(E_\varphi) \notin E_\varphi$; a contradiction. This contradiction historically called for the reformulation set theory and was one of the many factors leading to modern-day ZF axiomatic set theory. All other famous self-reference paradoxes may be articulated using this Inclosure Schema.

A. Unfolding self-reference in time

We now analyze how the intrinsic thought process of an agent carrying out a self-referential deduction as given by the Inclosure Schema (Box 2) would appear in the dynamical behaviour of the joint system $X_{\text{Con}} \cup Y_{\text{Cog}}$. In formal logic, a deduction in a given formal system is a sequence of grammatically well-formed strings of symbols such that each string is either an instance of an assumed axiom or premise, or is the result of the application of a permitted rule of inference to previous lines in the sequence. If one views a deduction as a dynamical time-dependent thought process in which each line in the deduction corresponds to some fact about which the thinking subject is aware, the sequential ordering of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

lines of the deduction may be interpreted as the time ordering of a series of mental states (and thus, a constraint of the compatible dynamics of the underlying brain states).

Given some statement ϕ, to say that an agent is aware of ϕ at time t is to say that the *physical* state of X^t_{Con} grounds the *mental* state of being aware of ϕ. One can actively perceive ϕ by occupying such a mental state, or one can remember having perceived ϕ at a previous time. Thus, there is an internal time index $\tau \leq t$ that tracks the time at which ϕ was perceived that may differ from the time index of the state of X_{Con}. If we denote that class of all brain states that give rise to this mental state by $[\phi]$, and index the time at which ϕ is thought to be (or have been) perceived by $[\phi^\tau]$, we thus have $X^t_{\text{Con}} \in [\phi^t]$ if the thinking subject is actively thinking about ϕ, and $X^t_{\text{Con}} \in [\phi^\tau]$ for $\tau < t$ if they are recalling having thought about ϕ previously.

If ϕ and ψ are two formulas that are not logically equivalent to one another, one might suppose that $[\phi^\tau] \cap [\psi^\tau] = \emptyset$. This very general claim may be objected to in principle by noting, for instance, that if ϕ and ψ are sufficiently complex, the thinking subject may not always be immediately aware of their logical (in)equivalence.[4] Nevertheless, it should be agreeable that there are no brain states that are simultaneously neural correlates of the awareness of ϕ and also neural correlates of the awareness of $\neg\phi$. This weaker hypothesis is all we shall require. Then, if the thinking agent carries out a deductive inference whose sequential lines are denoted $\{\phi_n\}$, this corresponds to their brain undergoing a dynamical evolution of the form:

$$X^t_{\text{Con}} \in [\phi_0^{\tau(t)}] \to X^{t+1}_{\text{Con}} \in [\phi_1^{\tau(t+1)}] \to X^{t+2}_{\text{Con}} \in [\phi_2^{\tau(t+2)}] \to \cdots \to X^{t+n}_{\text{Con}} \in [\phi_n^{\tau(t+n)}] \tag{9}$$

where $\tau : \mathbb{Z} \to \mathbb{Z}$ satisfies $\tau(t) \leq t$. While the individual lines of a deduction correspond to mental states (and thus restricted classes of brain states), the axioms and rules of inference from which subsequent lines are produced do not reflect processes of which one is consciously aware during such a thought process. Rather, they reflect the cognitive rules that the thinking agent's brain may apply to the content of their experience in order to bring about their subsequent mental states. In this way, the axioms and rules of inference that enable one to formalize a given deduction correspond in the underlying thought process to processes implementations of cognitive processes via Y_{Cog} (see Box 1) for a concrete realisation thereof).

To illustrate this, let us consider a simple example. Suppose one sees a green apple before them. This perception, and the discrimination of various features of this perception are grounded in neural correlates that reside physiologically in the brain module X_{Con} at the present time t. Suppose, subsequently (say, at time $t + 1$), that one remembers from their past experiences that essentially all green apples have a sour taste. (Of course, the inductive formation of such a generalized belief from past memories is non-trivial, but it nevertheless happens.) This association, then, of sour flavour with green apples in general is something about which the thinking subject becomes consciously aware, and hence forms part of their conscious experience. Therefore, it is likewise encoded in the neural correlates of consciousness present in X_{Con} at time $t + 1$. From these two perceptions, the thinking subject may apply *modus ponens* to conclude that the apple they saw at time t would likely have had a sour taste were they to eat it. The general rule of *modus ponens*, however, is not something of which one has direct perception when it is being implemented; making such inferences is a higher cognitive process. The implementation of *modus ponens*, therefore, is a process carried out by the brain module Y_{Cog}. Importantly, once this inference has been carried out, the subject becomes aware of its outcome. Namely, at a subsequent time (say, $t + 2$), they become consciously aware that, had they eaten the apple, it would likely have tasted sour. This is the general manner in which deduction may be realized as thought processes implemented within our brain model.

We now apply this perspective to the linguistic processing of self-referential statements via the Inclosure Schema (see Box 2). The idea is to distinguish between the abstract logical results and the thought processes obtained when a thinking subject confronts an instance of self-reference and thinks about it over a finite period of time. Logically speaking, the contradiction arising from a diagonalization argument is absolute; we do not contest this. However, when we infer this contradiction—i.e. when the dynamical behaviour of a subject's brain implements the thought process that yields this contradiction—using diagonalization, we do so in two temporally separate parts; first, we prove that $\Delta(x) \notin x$ and conclude that $\Delta(E_\varphi) \notin E_\varphi$. Then, at a later time, we conclude that $\Delta(E_\varphi) \in E_\varphi$. The contradiction arises when we remember at a third time that we had proven both of these two facts separately.

Let us look at Tarski's paradox to see this play out concretely, following the exposition by (Priest, 2002). To begin, let T be a 'truth' predicate on sentences, i.e. for any sentence x, $T(x)$ is true if and only if x is true (this is called Tarski's *T-schema*). Let ψ denote definability such that $\psi(X)$ is true for any set of sentences X just in case there exists a sentence x which defines X as a set (of sentences). If X is any definable set of sentences, let $\Delta(X) = \alpha$ where $\alpha = \langle \alpha \notin X \rangle$ (here $\langle\cdot\rangle$ is used to denote the proper name of a sentence). That is, $\Delta(X)$ is the sentence α which expresses that α is not an element of the set of sentences X. Clearly, α is self-referential. If an agent thinks about the T-schema, their thought process might look like the following. First, one supposes that the totality of all true sentences exists and is definable, that is, that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

$\mathrm{Tr} := \{x \mid T(x)\}$ is a set that may be defined by some sentence. If X is definable (whence $\psi(X)$ is true) and if $X \subseteq \mathrm{Tr}$, we have in the temporal framework described:

Time	Inference	Rule
$t = 0$	$\Delta(X) \in X \rightarrow \langle \alpha \notin X \rangle \in X$	Definition of Δ
$t = 1$	$\langle \alpha \notin X \rangle \in X \rightarrow \langle \alpha \notin X \rangle \in \mathrm{Tr}$	Comprehension in ZF
$t = 2$	$\langle \alpha \notin X \rangle \in \mathrm{Tr} \rightarrow \alpha \notin X$	T-Schema
$t = 3$	$\alpha \notin X \rightarrow \Delta(X) \notin X$	Definition of Δ
$t = 4$	$\Delta(X) \in X \rightarrow \Delta(X) \notin X$	Modus ponens (three times)
$t = 5$	$\Delta(X) \notin X \rightarrow \Delta(X) \notin X$	Tautology
$t = 6$	$(\Delta(X) \in X) \vee (\Delta(X) \notin X) \rightarrow \Delta(X) \notin X$	Propositional logic
$t = 7$	$(\Delta(X) \in X) \vee (\Delta(X) \notin X)$	Excluded middle
$t = 8$	$\Delta(X) \notin X$	Modus ponens on $t = 6$ and $t = 7$
$t = 9$	$\Delta(\mathrm{Tr}) \notin \mathrm{Tr}$	Substitution of $X = \mathrm{Tr}$ to $t = 8$
$t = 10$	$\Delta(\mathrm{Tr}) \in \mathrm{Tr}$	Substitution of $X = \mathrm{Tr}$ to $t = 1$
$t = 11$	$(\Delta(\mathrm{Tr}) \in \mathrm{Tr}) \wedge (\Delta(\mathrm{Tr}) \notin \mathrm{Tr})$	Propositional logic

Let us now look at the brain states that could in principle produce the mental states associated with each line of this deduction. We may rewrite the above inference as follows:

$$
\begin{aligned}
&X^0_{\mathrm{Con}} \in [(\Delta(X) \in X \rightarrow \langle \alpha \notin X \rangle \in X)^0] \\
&X^1_{\mathrm{Con}} \in [(\langle \alpha \notin X \rangle \in X \rightarrow \langle \alpha \notin X \rangle \in \mathrm{Tr})^1] \\
&X^2_{\mathrm{Con}} \in [(\langle \alpha \notin X \rangle \in \mathrm{Tr} \rightarrow \alpha \notin X)^2] \\
&X^3_{\mathrm{Con}} \in [(\alpha \notin X \rightarrow \Delta(X) \notin X)^3] \\
&X^4_{\mathrm{Con}} \in [(\Delta(X) \in X \rightarrow \Delta(X) \notin X)^4] \\
&X^5_{\mathrm{Con}} \in [(\Delta(X) \notin X \rightarrow \Delta(X) \notin X)^5] \\
&X^6_{\mathrm{Con}} \in [((\Delta(X) \in X) \vee (\Delta(X) \notin X) \rightarrow \Delta(X) \notin X)^6] \\
&X^7_{\mathrm{Con}} \in [((\Delta(X) \in X) \vee (\Delta(X) \notin X))^7] \\
&X^8_{\mathrm{Con}} \in [(\Delta(X) \notin X)^8] \\
&X^9_{\mathrm{Con}} \in [(\Delta(\mathrm{Tr}) \notin \mathrm{Tr})^9] \\
&X^{10}_{\mathrm{Con}} \in [(\Delta(\mathrm{Tr}) \in \mathrm{Tr})^{10}] \\
&X^{11}_{\mathrm{Con}} \in [((\Delta(\mathrm{Tr}) \in \mathrm{Tr})^9 \wedge (\Delta(\mathrm{Tr}) \notin \mathrm{Tr})^{10})^{11}]
\end{aligned}
$$

To prove a contradiction in time in a manner that could require a physically impossible brain state, one would need to show that $X^t_{\mathrm{Con}} \in [\phi^t]$ and $X^t_{\mathrm{Con}} \in [\neg\phi^t]$ for a single t. This does not happen. In this way, if we want to model deductive inferences as processes carried out by a physical systems such as the brain which evolves in time, we see that the contradictions appear not directly, but spread out in time and then recalled, and so they may be implemented by a machine such as the brain that operates in time (3). In particular, we do not encounter the fractal picture given in (Grim et al., 1993).

Moreover, because it is possible to have $X^t_{\mathrm{Con}} \in [\phi^t]$ and $X^{t'}_{\mathrm{Con}} \in [\neg\phi^{t'}]$ at different times $t \neq t'$, we see that the brain has on this model sufficient expressive power to treat truth values as dynamically changing quantities. This may be contrasted with Turing machines tasked with deciding truth values; the state of such a machine may evolve in time, but the truth value it aims to decide is static.

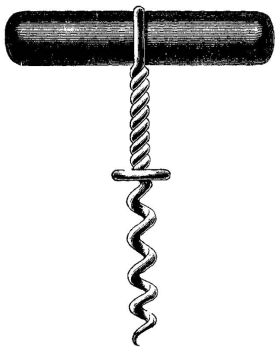

FIG. 3 Unfolding self-reference in time can be imagined as unfolding a circle many-times packed into a corkscrew, where the time dimension corresponds to the long dimension of the corkscrew. Equivalently, it can be imagined as the evolution of circularly polarised light.

B. Clarifying remarks

Let us make a few remarks on the conclusions reached so far. We are not denying the logical contradiction that appears in this above deduction. Indeed, what we have done here amounts to a temporal version of what (Priest, 2002) calls parameterisation; it is a standard approach to avoid paradoxes, and in general, any contradiction that is avoided by parameterisation will reappear at a higher level again when one analyzes the parameterised formalism. However, this is irrelevant to our aims: what we have shown is that an inference-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

making device that has a register that expresses its state of deduction in time (while some auxiliary system carries out further inference-making tasks leading to eventual update of the register) can effectively model contradictory scenarios without existing in a contradictory state itself. That is, there is never an instant where such a system need occupy two different physical states simultaneously.

Extending this to our model of the brain, the 'inference' column label could be replaced with 'the thought of which the conscious agent is aware' at each given time, while the 'rule' column label could just as well be interpreted as 'the cognitive process being carried out in the intermediate time window.' In this way, we have a rough picture for how the brain could physically model the contradictions that arise from self-reference paradoxes (noting that the above proof for the contradiction in Tarski's paradox is of the generic diagonalization form) without itself being in any strange superposition of disagreeing physical configurations.

What makes this temporal parameterization technique useful is that while in a purely logical setting, the relation between subsequent lines in a deduction is strictly a logical one (with no temporality and so forth), when represented on a physical system, is no longer an abstractly logic relation, but is instead a *causal* relation indicating an interaction between these two brain modules we have discussed. In particular, it is a causal relation which requires an intermediate physical process to commence and terminate. Hence, there is an intermittent time, and so the contradiction may be "stretched out" in time in the appropriate sense. (This is analogous to the Kantian view of time as a means for the thinking subject to experience contradictory perceptions without an actual contradiction obtaining (Kant, 1998, A32/B48).)

Integrated Information Theory

Integrated Information Theory (IIT; see (Oizumi et al., 2014; Tononi et al., 2016; Tononi and Koch, 2015)) is a framework that seeks to provide a constructive account of the origins of conscious experience by describing it as an emergent feature of causally integrated dynamical systems such as the brain. IIT begins by articulating those features of conscious experience that one might take to be constitutive, and then identifies features of the causal structure of a dynamical system that qualitatively realize these features (in a manner that can be made quantifiably precise via informational measures). A model of the IIT formalism is a dynamical system X as the SLM together with all of the probabilities of the form $p(x_i^{t+1}|X^t)$. From the causal probabilities defined in (5), the IIT formalism defines measures to quantify the extent to which a subsystem $S \subseteq X$ cannot be causally reduced, e.g. to a pair of subsystems G and H with $S = G \cup H$ and the extent to which every such S is causally integrated.

A state of a subsystem at some time is *irreducible* if and only if the probabilities that characterize its intrinsic causal structure cannot be exactly recovered by partitioning it into subsystems. Irreducibility is quantified using an informational measure; those subsystems of X that realize the maximum of this measure for the system X are called *maximally irreducible*. There are generally many different maximally irreducible subsystems.

According to IIT, only those subsystems that are maximally irreducible at a given moment contribute to consciousness at that moment, forming the instantaneous correlates of consciousness. The manner in which a maximally irreducible subsystem contributes to consciousness is dictated by its causal probabilities which populate points in a supposed space of qualia. The conscious experience realized by a physical substrate (a human brain or otherwise) is a byproduct of that substrate's maximally irreducible intrinsic causal structure.

Here we do not assess the plausibility of IIT as a theory of consciousness; rather, we note that our SLM can be recast within the IIT formalism straightforwardly.

C. Transforming logical inconsistency to incongruence

We now apply the IIT formalism (see Box 3) to the SLM, and show how the logical inconsistency of self-referential paradoxes is transformed to incongruence.

First observe that since the correlates of consciousness were taken to reside in X_{Con}, it is reasonable to suppose that for any subsystem of the brain $Z \subseteq B$, if Z is maximally irreducible while in some state Z^t, it must be the case that $Z \cap X_{\text{Con}} \neq \emptyset$. In most cases, Z will simply be a subsystem of X_{Con}. However from (7), there will be some irreducible subsystems that overlap with Y_{Cog} as well. In particular, $X_{\text{Con}} \cup Y_{\text{Cog}}$ is expected to be maximally irreducible.

Incongruence in IIT is defined as follows. For any system S, given a pair of subsystems $G, H \subseteq S$, G and H are incongruent if they make differing predictions about the past or future behaviour of some particular node $z \in S$ (see (Haun and Tononi, 2019) and (Albantakis and Tononi, 2019, p. 5)). This occurs, for instance, if $p(z^{t+1}|G^t) \neq p(z^{t+1}|H^t)$. When self-referential inferences are made, if we suppose ϕ is thought about at time t, $\neg\phi$ is thought about at $t+1$, and $\phi^t \wedge \neg\phi^{t+1}$ is thought about at time $t+2$, then it is because of the cognitive processes in Y_{Cog} implemented at time t and $t+1$ that this is the case. In particular, if we presently think some sentence is true, we expect that it will be true still at the next instant, so that

$$p(X_{\text{Con}}^{t+1} \in [\neg\phi^{t+1}] \mid X_{\text{Con}}^{t} \in [\phi^t], Y_{\text{Cog}}) \tag{10}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 November 2021 doi:10.20944/preprints202111.0524.v1

is large, while

$$p(X_{\text{Con}}^{t+1} \in [\neg\phi^{t+1}] \mid X_{\text{Con}}^{t} \in [\phi^{t}]) \tag{11}$$

is small. However, Y_{Cog} implements a rule of inference in this transition, which causes $X_{\text{Con}}^{t+1} \in [\neg\phi^{t+1}]$ to occur. During self-referential inferences, not only do two different subsystems disagree about the probabilities assigned to a particular node's future state (cf. (7)); rather, they assign essentially *opposite* probabilities to the future behaviour of the subsystem spanned by all maximally irreducible subsystems. Hence, incongruence arises in a strong way.

Put differently, causal incongruence in IIT offers a precise sense in which the parts of a system fail to describe the whole of the system, namely, taken separately, the parts may disagree with one another about the descriptions they provide. In the SLM framework, this is exploited as a *feature*: it is this disagreement that enables the brain to represent contradictions in the requisite manner needed to make sense of self-referential statements.

D. Avoiding unhalting cycles

We now argue that the cyclic behaviour of the SLM, as described in III.A, does not persist *indefinitely* (as it would for an unhalting Turing machine). When the thinking subject gets caught in a cognitive cycle of the same form, if their attention is drawn away from the cyclic inference at hand, the cycle will end. This is so because, as a thinking subject learns by repeating a task many times, they devote less and less attention and focus towards the task being learned (Kandel et al., 2013, Chapter 64). In the present context, this means that if the thinking subject cycles through the thought process associated with deriving disagreeing truth values for a self-referential statement, they will not get caught in a loop, but rather will pay less attention to the inference upon subsequent iterations. Since the brain actively monitors a large class of sensory stimuli and implements many cognitive processes in parallel, as this attention diminishes, the thinking subject is increasingly likely to refocus their attention elsewhere. In short, if attention is a resource, the architecture of the brain is such that the re-allocation of this resource inhibits the ensuing feedback and makes infinite inferential loops unstable.

This is analogous to binocular rivalry, where the subject's visual field is eventually changed, whence their visual sensations escape from flowing towards lock-in states (Clark, 2013; Hohwy et al., 2008), and to visual paradoxes, like the Necker cube (where two alternative possible attractors are present) or the recognition of ambiguous images (Inoue and Nakamoto, 1994; Kelso, 1995). This metastable behaviour due to self-reference can also be found in gene networks, where the causal feedback associated with cross-regulatory interactions can be spread in time or space leading to interesting phenomena (Isalan, 2009).

IV. CONCLUSIONS AND OUTLOOK

In this work, we have constructed a high-level discrete dynamical model of the brain, termed the Strange Loop Model (SLM; II), in order to describe inference-making, which uses causal feedback between conscious and cognitive processes. We have used the SLM to model self-reference and shown that logical inconsistencies unfold in time (III.A), and hence the contradictions dissolve, as one never encounters inconsistent truth values simultaneously. Rather, one deduces at different times that a sentence has different truth values and then remembers having carried out both such deductions. This flexibility enables the human brain to model self-reference in a manner that is inaccessible to usual computing devices by construction. We have also applied the SLM within the context of IIT and shown that logical inconsistencies are transformed into incongruences (III.C). Finally we have argued that, because the brain is receptive to a wide range of different stimuli, and because one devotes less attention to repetitive cognitive tasks as time passes, these cyclic inferences are unstable are thus terminated (III.D).

The SLM illustrates the extent to which the human mind is capable of understanding and making inferences. The interaction between X_{Con} and Y_{Cog} via the described causal feedback enables the human mind to be aware of the outcomes of cognitive inferences, and likewise further cognize about such an awareness. Put differently, the causal feedback here described enables the thinking subject to be aware of their own cognitive processes, and to then make inferences about their own cognition. This situation is reminiscent of universality encountered in Turing machines, spin models and neural networks (De las Cuevas, 2020).

Finally, we may compare the SLM with a Turing machine or any other standard computing machine. Unlike an algorithm running in a Turing machine, the processing carried out by the SLM is not a deciding process, because it need not reach a static truth value of a variable. Moreover, the only relevant features of a Turing machine are its input–output functionality (that is, the formal language it recognizes (Kozen, 1997)), whereas the intrinsic causal structure of the brain is crucial. In this way, we conclude that the process carried out by the brain and the computer is different.

Acknowledgments

We wish to thank Larissa Albantakis and Wolfram Hinzen for insightful discussions. PF is supported in part by funding from the Social Sciences and Humanities Research Council. RS thanks the Spanish Ministry of Economy and Competitiveness, grant PID2019-111680GB-I00, an AGAUR FI 2018 grant, and the Santa Fe Institute. GDLC acknowledges funding from the Austrian Science Fund (FWF) [START Prize Y-1261-N].

References

Albantakis, L. and Tononi, G. (2019). Causal Composition: Structural Differences among Dynamically Equivalent Systems. *Entropy* 21, 989. doi:10.3390/e21100989

Allman, J. M. (1999). *Evolving brains* (Scientific American Library)

Arbib, M. (2012). *Brains, machines, and mathematics* (New York: Springer,)

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36, 181–204. doi:10.1017/S0140525X12000477

Cobb, M. (2020). *The idea of the brain: The past and future of neuroscience* (Hachette UK)

De las Cuevas, G. (2020). Universality Everywhere implies Undecidability Everywhere

Descartes, R. (1993). *Meditations on First Philosophy* (Hackett Publishing)

Edelman, G. (1992). *Bright Air, Brilliant Fire: On the Matter of the Mind* (Basic Books)

Edelman, G. M. (2005). *Wider Than the Sky: The Phenomenal Gift of Consciousness* (Yale University Press)

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 360, 815–836. doi:10.1098/rstb.2005.1622

Friston, K. and Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1211–1221. doi:10.1098/rstb.2008.0300

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural networks* 1, 119–130

Grim, P. (1993). Self-Reference, Chaos, and Fuzzy Logic. In *Integration of Fuzzy Logic and Chaos Theory* (Springer). 317–359

Grim, P., Mar, G., Neiger, M., and St. Denis, P. (1993). Self-reference and paradox in two and three dimensions. *Comput. & Graphics* 17, 609–612

Haun, A. and Tononi, G. (2019). Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy* 21, 1160. doi:10.3390/e21121160

Hofstadter, D. (1979). *Gödel, Escher, Bach* (Basic Books)

Hofstadter, D. (2007). *I am a strange loop* (Basic Books)

Hohwy, J., Roepstorff, A., and Friston, K. (2008). Predictive coding explains binocular rivalry: an epistemological review. *Cognition* 108, 687–701. doi:10.1016/j.cognition.2008.05.010

Indiveri, G. and Liu, S. (2015). Memory and information processing in neuromorphic systems. *Proceedings of the IEEE* 103, 1379–1397

Inoue, M. and Nakamoto, K. (1994). Dynamics of cognitive interpretations of a necker cube in a chaos neural network. *Progress of Theoretical Physics* 92, 501–508

Isalan, M. (2009). Gene networks and liar paradoxes. *BioEssays* 31, 1110–1115

Jonas, E. and Kording, K. (2017). Could a neuroscientist understand a microprocessor? *PLoS computational biology* 13, e1005268

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A., and Hudspeth, A. J. (eds.) (2013). *Principles of Neural Science* (McGraw Hill Medical), 5 edn.

Kant, I. (1998). *Critique of Pure Reason*. The Cambridge Edition of the Works of Immanuel Kant (Cambridge: Cambridge University Press). doi:10.1017/CBO9780511804649

Kelso, J. S. (1995). *Dynamic patterns: The self-organization of brain and behavior* (MIT Press)

Kozen, D. C. (1997). *Automata and Computability* (Springer)

Krohn, S. and Ostwald, D. (2017). Computing integrated information. *Neuroscience of Consciousness* 2017. doi:10.1093/nc/nix017

Lawvere, F. W. (2006). Diagonal arguments and Cartesian Closed Categories. *Reprints in Theory and Applications of Categories* , 1–13

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444

Markovi, D., Mizrahi, A., Querlioz, D., and Grollier, J. (2020). Physics for neuromorphic computing. *Nature Reviews Physics* 2, 499–510

Martinez, P. and Sprecher, S. (2020). Of circuits and brains: The origin and diversification of neural architectures. *Frontiers in Ecology and Evolution* 8, 82

McCulloch, W. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* 5, 115–133

Oizumi, M., Albantakis, L., and Tononi, G. (2014). From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLOS Computational Biology* 10, e1003588. doi:10.1371/journal.pcbi.1003588

Park, H.-J. and Friston, K. (2013). Structural and Functional Brain Networks: From Connections to Cognition. *Science* 342. doi:10.1126/science.1238411

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (Cambridge University Press), 2 edn.

Priest, G. (2002). *Beyond the Limits of Thought* (Clarendon Press)

Priest, G. (2006). *In Contradiction* (Clarendon Press)

Prokopenko, M., Harre, M., Lizier, J., Boschetti, F., Peppas, P., and Kauffman, S. (2019). Self-referential basis of undecidable dynamics: from The Liar Paradox and The Halting Problem to The Edge of Chaos. *Physics of Life Reviews* 31, 134–156

Rashevsky, N. (1960). *Mathematical biophysics: physico-mathematical foundations of biology* (New York City, NY, USA: Dover P. Inc.), 3rd edn.

Roberts, D. M. (2021). Substructural fixed-point theorems and the diagonal argument: theme and variations. *arXiv:2110.00239*

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning representations by back-propagating errors. *Nature* 323, 533–536

Smee, A. (1850). *Instinct and reason: deduced from electro-biology* (Reeve and Benham)

Sporns, O., Tononi, G., and KÃ¶tter, R. (2005). The Human Connectome: A Structural Description of the Human Brain. *PLOS Computational Biology* 1, e42. doi:10.1371/journal.pcbi.0010042

Tegmark, M. (2000). Importance of quantum decoherence in brain processes. *Physical Review E* 61, 4194–4206. doi:10.1103/PhysRevE.61.4194

Tognoli, E. and Kelso, J. A. S. (2014). The Metastable Brain. *Neuron* 81, 35–48. doi:10.1016/j.neuron.2013.12.022

Tononi, G., Boly, M., Massimini, M., and Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience* 17, 450–461. doi:10.1038/nrn.2016.44

Tononi, G. and Edelman, G. M. (1998). Consciousness and Complexity. *Science* 282, 1846–1851. doi:10.1126/science.282.5395.1846

Tononi, G. and Koch, C. (2015). Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences* 370, 20140167. doi:10.1098/rstb.2014.0167

Wood, G. (2002). *Living dolls: a magical history of the quest for mechanical life* (Faber & Faber)

Yanofsky, N. S. (2003). A Universal Approach to Self-Referential Paradoxes, Incompleteness and Fixed Points. *Bulletin of Symbolic Logic* 9, 362–386. doi:10.2178/bsl/1058448677

Zylberberg, A., Dehaene, S., Roelfsema, P. R., and Sigman, M.

 doi:10.20944/preprints202111.0524.v1

(2011). The human Turing machine: a neural framework for mental programs. *Trends Cogn. Sci.* 15, 293–300