





# Sources of richness and ineffability for phenomenally conscious states

Xu Ji <sup>1,2,\*</sup>, Eric Elmoznino <sup>1,2</sup>, George Deane <sup>3</sup>, Axel Constant<sup>4</sup>, Guillaume Dumas <sup>1,5</sup>, Guillaume Lajoie<sup>1,6</sup>, Jonathan Simon<sup>3</sup>, Yoshua Bengio<sup>1,2,7</sup>

<sup>1</sup>Mila - Quebec AI Institute, Montreal, Quebec H2S 3H1, Canada

<sup>2</sup>Department of Computer Science and Operations Research, University of Montreal, Pavillon André-Aisenstadt 2920, chemin de la Tour, Montreal, Quebec H3T 1J4, Canada

<sup>3</sup>Department of Philosophy, University of Montreal, Pavillon 2910, boul. Édouard-Montpetit, Montreal, Quebec H3C 3J7, Canada

<sup>4</sup>School of Engineering and Informatics, University of Sussex, Sussex House, Falmer, East Sussex BN1 9RH, United Kingdom

<sup>5</sup>Department of Psychiatry and Addiction, University of Montreal, Pavillon Roger-Gaudry 2900, boul. Édouard-Montpetit, Montreal, Quebec H3T 1J4, Canada

<sup>6</sup>Department of Mathematics and Statistics, University of Montreal, Pavillon André-Aisenstadt (AA-5190) 2920, chemin de la Tour, Montreal, Quebec H3T 1J4, Canada

<sup>7</sup>CIFAR - Canadian Institute for Advanced Research, MaRS Centre, West Tower 661 University Ave., Suite 505, Toronto, Ontario M5G 1M1, Canada

\*Correspondence address. Mila - Quebec Artificial Intelligence Institute, Artificial Intelligence, Montreal, Canada. E-mail: [xu.ji@mila.quebec](mailto:xu.ji@mila.quebec)

## Abstract

Conscious states—state that there is something it is like to be in—seem both rich or full of detail and ineffable or hard to fully describe or recall. The problem of ineffability, in particular, is a longstanding issue in philosophy that partly motivates the explanatory gap: the belief that consciousness cannot be reduced to underlying physical processes. Here, we provide an information theoretic dynamical systems perspective on the richness and ineffability of consciousness. In our framework, the richness of conscious experience corresponds to the amount of information in a conscious state and ineffability corresponds to the amount of information lost at different stages of processing. We describe how attractor dynamics in working memory would induce impoverished recollections of our original experiences, how the discrete symbolic nature of language is insufficient for describing the rich and high-dimensional structure of experiences, and how similarity in the cognitive function of two individuals relates to improved communicability of their experiences to each other. While our model may not settle all questions relating to the explanatory gap, it makes progress toward a fully physicalist explanation of the richness and ineffability of conscious experience—two important aspects that seem to be part of what makes qualitative character so puzzling.

## Introduction

Conscious states—state that there is something it is like to be in (Nagel, 1974)—present many apparent contradictions. On the one hand, every time we have a thought, look out at the world, or feel an emotion, we have a rich experience that seems impossible to fully describe. At the same time, conscious experiences are conceptualizable, with similar properties across individuals, and can often be communicated with a degree of fidelity.

This paper provides an information theoretic dynamical systems perspective on how and why consciousness may appear to us the way it does, namely, as both “rich” or full of detail, and “ineffable” or hard to fully describe or recall—in other words, why it seems that an experience is “worth a thousand words.” Our key contention is that these aspects of consciousness are implicated by a dynamical systems model of neural processing, in particular by “attractors”: patterns of joint neural activity that remain relatively stable over short timescales and yield a discrete partition over neural states. Importantly, interpreting cognitive processing

through the lenses of dynamical systems and information theory will give us the ability to reason about richness, ineffability, and communicability in general terms, without relying on implementation details of the neural processes that may give rise to consciousness. Broadly, the suggestion is that the rather abstract level of explanation afforded by information theory is the commensurate level of explanation for some key questions about richness and ineffability.

By “consciousness,” we mean phenomenal consciousness, i.e. the felt or subjective quality of experience. A state is phenomenally conscious when, in the words of Nagel (1974), there is “something it is like” to be in that state. Phenomenal consciousness is the form of consciousness that gives rise to what Joseph Levine calls the “explanatory gap” (Levine, 1993) and what David Chalmers calls the “hard problem of consciousness” (Chalmers, 1996): the problem of showing that phenomenal consciousness can be explained in terms of, or reduced to, underlying physical processes. The explanatory gap is one of the central problems in

Received 1 May 2023; revised 3 January 2024; accepted 23 January 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

the philosophy of mind, and it relies heavily on the intuition that “physicalist theories leave out [phenomenal consciousness] in the epistemological sense, because they reveal our inability to explain qualitative character in terms of the physical properties of sensory states” (Levine, 1993).

Here, we address one aspect of this problem by developing a structural/mechanistic explanation of the richness and ineffability of conscious experience, one that is given entirely in terms of information processing in a dynamical system such as the brain. Our model assumes that conscious experiences are derived from neural processes according to known physical laws and can therefore be understood using the standard methods of cognitive neuroscience. While our model may not settle all questions related to the explanatory gap, it will make progress toward a fully physicalist explanation of the richness and ineffability of conscious experience—two important aspects that seem to be part of what makes qualitative character so puzzling. The aim of this paper is to propose and justify a formal description of how neural dynamics could give rise to the ordinary sense of richness and ineffability in the brain. Our key contributions are summarized as follows:

- (i) We relate the philosophical notions of richness and ineffability to the computational notion of information. Assuming that brain dynamics are cast as information processing functions, we contend that the richness of conscious experience can be interpreted as the amount of information in conscious state, and ineffability as the amount of information lost in processing.
- (ii) Attractor dynamics are empirically ubiquitous in neural activity across cortical regions and have been proposed as a computational model for working memory (Khona and Fiete, 2022; Rolls, 2010), while prominent models of consciousness argue that conscious experience is a projection of working memory states (Baars, 2005; Dehaene and Naccache, 2001). We connect these theories by contending that significant information loss induced by attractor dynamics offers an account for the significant ineffability of conscious experience.
- (iii) By considering information at multiple stages during interpersonal communication, we show how different point-to-point pathways of information loss arise during cognitive processing, going beyond the specific case of ineffability of conscious experience at verbal report.
- (iv) Using Kolmogorov information theory (Kolmogorov, 1965), we prove a formal result that connects cognitive dissimilarity between individuals with increasing ineffability of conscious experience. This highlights the difference between cognitive dissimilarity and knowledge inadequacy, shedding light on the philosophical conundrum of what color scientist Mary learns when leaving her black and white room (Jackson, 1986).
- (v) Since information loss is a function of neural states, it can be approximately computed by cognitive processing, providing a mechanistic justification for the report of ineffability or the contention that consciously inaccessible rich representations exist (Sperling, 1960).

Instead of defining “access” as triggering correct behavior on a per-experience basis (Colagrosso and Mozer, 2004), we contend that there is a natural correspondence between access and preservation of information, which allows for quantification

using mutual information and analysis by applying information theoretic reasoning to the abstract computation graph. Casting ineffability as information loss allows us to reason about the ineffability of conscious experience from the computation graph without depending on the exact definition of conscious experience.

The paper is structured as follows. We present our dynamical systems model of conscious experience in the “An information theoretical dynamical systems perspective on conscious experience” section, beginning with the “Motivating attractor dynamics as a model for conscious experience” section, which motivates the use of attractor dynamics for modeling conscious processing using prior arguments from the literature that are independent of our own, including evidence for the Global Workspace Theory (Baars, 1993; Baars, 2005; Dehaene et al., 1998). The “Richness and ineffability” section formalizes the notions of richness and ineffability using both Shannon information theory (Shannon, 1948) and Kolmogorov complexity (Kolmogorov, 1965), which play a central role in making our later arguments precise. Core contributions are presented in “Intrapersonal ineffability” and “Interpersonal ineffability” sections, which discuss various sources of ineffability in conscious experience and explain the conditions under which these experiences can be partially communicated to others. We then briefly discuss the implications of our model on the debate surrounding “phenomenal” vs. “access” consciousness (Block, 1995), before concluding with a high-level discussion in the “Conclusion” section. A background summary of related work, concepts in philosophy, and computational neural dynamics is given in Appendices “Illusionism and overflow”, “Computation through neural dynamics” and “Related Work”.

## Results

### Motivating attractor dynamics as a model for conscious experience

To contextualize our argument, we begin by drawing on existing work on working memory to highlight several connections between attractor dynamics and conscious experience.

The contents of working memory are typically considered to be the attended contents of short-term memory: a function of short-term representations held in the brain and context from task information or other executive functioning objectives (Engle, 2002; Cowan, 2008). A central claim in several leading theories of consciousness is that what we are consciously aware of is the contents of working memory. For example, the Global Workspace Theory (Baars, 1993; Baars, 2005) and its neuronal extension (Dehaene et al., 1998) state that information becomes conscious by gaining entry into a limited workspace that serves as a bottleneck for the distributed activity present across the brain. Pairs of brain regions are largely isolated from each other, and arbitrary point-to-point communication is only possible via the workspace, which itself can both receive and broadcast information globally. The workspace, then, serves as a hub capable of coordinating brain-wide activity for centralized control and decision-making. It is easy to see the connection between the concepts of a global workspace and working memory (attentional selectivity, influence on executive decision-making, availability to verbal and behavioral reporting processes, limited capacity, and arbitrary modalities), and there is little distinction between them in the Global Workspace Theory (Dehaene and Naccache, 2001). Similarly, the notion of “access consciousness” introduced in Block (1995) can

be framed through the lens of a working memory whose contents are globally accessible across the brain.

The link between working memory and attractor dynamics, in turn, is well established. Empirical studies have demonstrated that attractor dynamics are ubiquitous in the brain, both across species and levels in the brain's hierarchy (Rolls, 2010; Khona and Fiete, 2022). The attractor model for working memory postulates that working memory emerges from recurrently connected cortical neural networks that allow representations to be maintained in the short term (on the order of seconds) by self-generated positive feedback (Durstewitz et al., 2000; Curtis and D'Esposito, 2003; Deco and Rolls, 2003; Barak and Tsodyks, 2014; Seeholzer et al., 2019). Attractor dynamics can support both "suppression" of inputs, for example, in decision-making where the brain state flows rapidly toward a discrete attractor and subsequent inputs or perturbations are discounted, as well as "integration" over inputs, where the incremental response to inputs causes reversible flow along continuous attractor manifolds (Redish et al., 1996; Wang, 2008; Khona and Fiete, 2022). Neural winner-take-all models implement hybrid analog-discrete computation (Wang, 2008; Wong and Wang, 2006). Robustness, discreteness, and temporal integration of information are all traits apparent in working memory (Khona and Fiete, 2022).

## Richness and ineffability

What is meant by the richness of experience? Intuitively, while we find it easy to communicate certain aspects of our mental state, we struggle to convey their full content or meaning. One can consider color as an example. We are tempted to think of color space as a simple 3-dimensional surface, on the basis of perceptual similarity judgments that people tend to make. However, there is a far richer and higher dimensional structure to experiencing color. For instance, most people would describe the color "red" as warm and aggressive. There are myriad associations that we make with various colors that are not functions of their nominal definitions, and all these associations as a whole contribute to the richness of the experience (Chalmers, 2010).

Broadly, richness means having a lot, the condition of being "well supplied or endowed" (Merriam Webster Dictionary, 2023). In the context of mental state attribution, richness gauges the amount of specificity—detail, texture, nuance, or informational content—contained by a mental state. It is a common principle in aesthetics that experience is rich (a picture speaks a thousand words), and many philosophers acknowledge that conscious states at least appear to be highly detailed, nuanced, and contentful (Tye, 2006; Chuard, 2007; Block, 1995), although some takes this appearance to be ultimately illusory (Dennett, 1993; Cohen et al., 2016).

This conception of richness corresponds well to the mathematical notion formalized by Shannon (1948), where richness of a random variable  $X$  is given by its entropy  $H(X)$ . Here, a random variable represents a state type, e.g. experience of some face or other. To say that such a variable is high in entropy is to say that the number of values it could take (the number of possible states the system could be in, e.g. the different experiences of faces one could possibly have) is relatively large and the probability distribution over these is relatively flat, and thus, the state is unpredictable. Specifically, Shannon entropy  $H(X)$  quantifies the average number of bits (answers to yes-or-no questions) required to specify which state  $X$  takes as a measure of informational content.

The notion of ineffability is closely related. In popular usage, ineffable can be defined as "too great for words" (Oxford English Dictionary, 2023). The concept is often used in theological contexts, but it has been applied to descriptions of qualitative experience since at least Dennett (1993). Given the term's theological associations, the claim that experience is ineffable might sound like a profession of dualism: consciousness is something magic that no physicalist theory can account for. However, strictly speaking, to claim that experience is ineffable is simply to claim that its informational content exceeds what we can remember or report. Much hinges on what exactly we mean by "can remember or report." Of course, one can say a thousand words, so the fact that a picture speaks that many words do not necessarily make a picture ineffable. Later, we will develop tools to allow us to precisely refine the senses of ineffability at issue, and we will see that experience is ineffable in multiple senses (although none of them need involve magic or anything anathema to physicalist theories).

We propose that ineffability corresponds to the mathematical notion of information loss when trying to express a conscious state in words. Given a function that processes an input variable  $X$  and produces an output variable  $Y$ , information loss of the input incurred by the output is measurable by conditional entropy  $H(X|Y)$  or entropy of the input variable given the output variable. Intuitively, conditional entropy  $H(X|Y)$  measures how well  $Y$  describes  $X$ : how much uncertainty remains about the value of  $X$ , once the value of  $Y$  is given. Conditional entropy  $H(X|Y)$  is mathematically equivalent to the entropy of the input  $X$  minus the mutual information between input and output,  $H(X|Y) = H(X) - I(X; Y)$ , where the latter is a measure of information shared between them; the amount of information about the state of one variable obtained by observing the state of the other. Note the difference between conditional entropy and mutual information: mutual information is how much uncertainty one random variable removes from another, while conditional entropy describes how much uncertainty remains in the first variable after the value of the second is given.

Usefully, quantifying ineffability in this manner allows us to offer a precise definition of ineffability as the negation of ineffability. Where ineffability is given by  $H(X|Y)$ , negating ineffability gives effability:  $-H(X|Y) = I(X; Y) - H(X)$ . Recalling that entropy is a measure of uncertainty or spread in a probability distribution, the smaller the  $H(X|Y)$  is, the less uncertain the  $X$  is given  $Y$ , the less information is lost, and the more effable or communicable  $X$  is via  $Y$ .

In the foregoing, we draw on the framework of Shannon information, but there are advantages, for our purposes, to using Kolmogorov information (Kolmogorov, 1965) as an alternative way to characterize richness and ineffability. In the Kolmogorov formalism, richness of a state  $x$  corresponds to its complexity  $K(x)$ , which is the length in bits of the shortest program written in a general programming language that outputs  $x$  and halts. Ineffability then corresponds to conditional Kolmogorov complexity of an input  $x$  given an output  $y$ ,  $K(x|y)$ , the length of the shortest program needed to produce  $x$  if  $y$  is given, or intuitively the complexity of  $x$  minus the number of bits that can be saved from knowing  $y$ , which is the Kolmogorov analog of Shannon information loss as conditional entropy. Note that since Kolmogorov complexity is defined on strings of bits, we restrict the domain of our functions to discrete variables and assume that floating point representation is used to encode real values (Box 1). Floating point representations are discrete in the sense that they form

a countable finite size space. They are, however, an approximation of the reals, with an approximation error that decreases as the precision/computer memory allowed for the representation increases. Construction of the computational model thus incurs a separate form of information loss resulting from discretization of real-valued continuous-time observations of neural states.

### Box 1. Notation

Let  $x$  denote an instance of random variable  $X$ ,  $\mathcal{X}$  denote the set of possible states for  $X$  with probability distribution  $P(X)$ ,  $\sum_{x \in \mathcal{X}} P(X=x) = 1$ ,  $p(x)$  denote  $P(X=x)$ , expectation  $\mathbb{E}_{p(x)}[f(x)]$  denote  $\sum_{x \in \mathcal{X}} p(x)f(x)$ , and likewise for other variables. We restrict function domains to discrete variables including floating point representation of reals.  $[n]$  denotes the list of natural numbers  $1, \dots, n$ .

Shannon entropy and Kolmogorov complexity are closely related metrics of richness and are described in more detail in [Box 2](#) and [Fig. 1](#). If the probability distribution over states is given, taking an expectation over the distribution on Kolmogorov complexity of its states allows Shannon entropy to be approximately recovered ([Grünwald and Vitányi, 2004](#)). Under either framework, richness is characterizable as information measured in bits, ineffability is characterizable as information loss or richness reduction, and communicability and ineffability are neither separate nor Boolean traits, but direct opposites of each other and varying on a scale.

A major difference between Shannon entropy and Kolmogorov complexity is that the former requires knowledge of the probability distribution over variable states, whereas the latter is defined on individual states without assuming a given probability distribution. The distribution may be undefined or highly privileged information in itself (that is, the meta-distribution over the distribution's parameters is rich). Consider, for example, measuring the amount of information in a book by considering the set of all possible books and the distribution over them ([Grünwald and Vitányi, 2003](#)) or the information in a temporal snapshot of a high-dimensional brain state by considering the distribution over all possible states. In these cases, we want a way to measure informational content that does not require knowledge of a hard-to-specify distribution. This is especially salient for us where interpersonal ineffability is concerned. Even if we assume that a brain's parameters fully determine the distribution over its own states (and so in some sense, individuals have direct access to their own distributions), still individuals cannot have this level of knowledge of the distributions of their interlocutors' brains. Explicitly allowing the communicator's distributional parameters to be unknown is therefore convenient for characterizing interpersonal ineffability from the perspective of the listener.

A second drawback of Shannon's framework is that entropy is a measure of statistical determinability of states; information is fully determined by the probability distribution on states and unrelated to the meaning, structure, or content of individual states ([Grünwald and Vitányi, 2003](#)). For example, consider again a case where we want to measure interpersonal ineffability, as a relationship between a communicator's experience and a listener's. It might just happen to turn out that whenever Alice thinks and talks about tennis, Bob almost always thinks about Beethoven. In this case, conditional entropy will be low, but conditional Kolmogorov complexity will be high and therefore

### Box 2. Metrics for richness and ineffability

Shannon entropy is given by

$$H(Z) = \mathbb{E}_{p(z)} [-\log p(z)]. \quad (1)$$

If variable  $Y$  is produced by processing  $Z$ ,  $y = f(z)$ , with joint distribution denoted by  $p(z, y)$  and  $f$  stochastic in the general case, then information loss from  $Z$  to  $Y$  is given by conditional entropy  $H(Z|Y) = H(Z) - I(Z; Y)$ , where  $I(Z; Y)$  denotes Shannon mutual information between variables,

$$I(Z; Y) = \mathbb{E}_{p(z, y)} \left[ \log \frac{p(z, y)}{p(z)p(y)} \right], \quad (2)$$

and  $H(Z|Y)$  is given by

$$H(Z|Y) = \mathbb{E}_{p(z, y)} [-\log p(z|y)]. \quad (3)$$

The Kolmogorov complexity of a state  $z$ ,  $K(z)$ , is the length  $l(r)$  in bits of the shortest binary program  $r$  that prints  $z$  and halts. Specifically, let  $U$  be a reference prefix universal machine. The prefix Kolmogorov complexity of  $z$  is

$$K(z) = \min_r \{l(r) : U(r) = z, r \in \{0, 1\}^*\}. \quad (4)$$

Conditional Kolmogorov complexity is the length of the shortest program that takes  $y$  as an input, prints  $z$ , and halts. It is given by

$$K(z) - I(z : y) = K(z|y) \stackrel{+}{=} K(z|y, K(y)) \stackrel{\log}{=} K(z|y) \quad (5)$$

where  $I(z : y)$  denotes Kolmogorov mutual information between states,  $y^*$  denotes the shortest program that produces  $y$  and halts, standard notation  $\stackrel{+}{=}$  and  $\stackrel{\log}{=}$  are used to denote equality up to constant and logarithmic factors, respectively ([Grünwald and Vitányi, 2004](#); [Li et al., 2008](#)). As [Eq. \(5\)](#) shows,  $K(z|y^*)$  and  $K(z|y)$  are comparable and either may be used to characterize information loss; in subsequent sections, we will generally refer to  $K(z|y)$ .

Shannon entropy and Kolmogorov complexity are related by the following constraints ([Grünwald and Vitányi, 2004](#)):

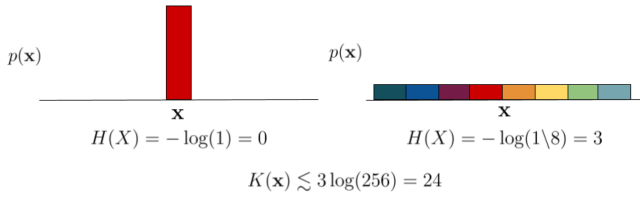
$$0 \leq (\mathbb{E}_{p(z)}[K(z)]) - H(Z) \leq K(p), \quad (6)$$

$$I(Z; Y) \stackrel{+}{=} \mathbb{E}_{p(z, y)} [I(z : y|p)], \quad (7)$$

$$I(Z; Y) - K(p) \stackrel{+}{<} \mathbb{E}_{p(z, y)} [I(z : y)] \stackrel{+}{4} I(Z; Y) + 2K(p), \quad (8)$$

which conveys how Kolmogorov complexity pays a penalty for not assuming knowledge of the distribution, since it must be encoded within the program.

suites to capture the absolute difference between their experiences. For these reasons, we argue that particularly in the case of interpersonal communication, Kolmogorov complexity should be used to characterize richness and ineffability of experiences.

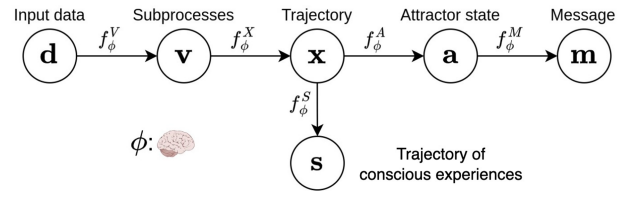


**Figure 1.** Illustrating Shannon entropy and Kolmogorov complexity for discrete color distributions. Entropy (Eq. (1)) involves an expectation over the states of stochastic variable  $X$ , whereas Kolmogorov complexity (Eq. (4)) is defined for an instance of state,  $x$ . Assume a universal red-green-blue (RGB) representation for colors where each RGB component ranges between 1 and 256. Without assumptions on the distribution over colors, the Kolmogorov complexity of each state is not greater than 24 (excluding program overheads) since color can be represented with 38-bit binary sequences, but may be lowered for smaller RGB values that do not require 8 bits if an optimized number encoding scheme is used (Grünwald and Vitányi, 2003). While entropy is the same for the same probability distribution over “any” states, Kolmogorov complexity would increase for states whose values are algorithmically more difficult to construct.

However, Shannon entropy is functionally equivalent if the distribution is given, and we will refer to both frameworks.

## Intrapersonal ineffability

In this section, we will develop our model of intrapersonal ineffability, that is, ineffability between stages of processing within a single experienter. We will be concerned with the following variables: Let  $X$  (with value  $\mathbf{x}$ ) be a trajectory of neural activities that determine working memory content and conscious experience, and let it consist of a sequence of transient states  $X_t$  for  $t \in [T]$ , where length  $T$  is fixed and sufficiently large such that all trajectories terminate near an attractor state. By “value,” we mean that  $X$ , for example, is a random variable ranging over neural trajectories and  $x$  stands for a value of that variable, i.e. a specific trajectory. Let  $A$  (with value  $\mathbf{a}$ ) denote the terminating attractor,  $S$  (with value  $\mathbf{s}$ ) denote a trajectory of conscious experience consisting of transient states  $S_t$  for  $t \in [T]$ ,  $D$  (with value  $\mathbf{d}$ ) denote external input datum,  $V$  (with value  $\mathbf{v}$ ) denote a list of  $N$  subprocess states  $V_n$  for  $n \in [N]$  and fixed  $N$  that comprise computation affecting working memory trajectory  $X$ , and  $M$  (with value  $\mathbf{m}$ ) denote the verbal report or output message of the individual. In addition, let  $\phi$  denote the brain’s synaptic weights that parametrize its dynamics. These variables are connected by a computation graph of functions (Fig. 2), given by  $\mathbf{v} = f_\phi^V(\mathbf{d})$ ,  $\mathbf{x} = f_\phi^X(\mathbf{v})$ ,  $\mathbf{a} = f_\phi^A(\mathbf{x})$ ,  $\mathbf{s} = f_\phi^S(\mathbf{x})$ , and  $\mathbf{m} = f_\phi^M(\mathbf{a})$ . The functions  $f_\phi^A$  (returns final attractor state) and  $f_\phi^S$  (outputs conscious experience that is fully determined by  $\mathbf{x}$ ) are deterministic, while  $f_\phi^M$ ,  $f_\phi^V$ , and  $f_\phi^X$  are generally stochastic, meaning that outputs may be dependent on hidden stochastic variables within the function that encodes historical states or neural processing noise. We assume that the correspondence between neural trajectory and conscious trajectory is a function, which is plausible given type identity and functionalist theories of consciousness (Smart, 2022) that assume a function between instantaneous neural and conscious state. Not speaking is encoded by a state of  $V$  corresponding to “no verbal report.” Subscripting with  $\phi$  denotes that function behavior is determined by cognitive parameters  $\phi$ . The computation graph defines a joint probability  $p_\phi(\mathbf{d}, \mathbf{x}, \mathbf{s}, \mathbf{a}, \mathbf{m})$ , from which conditional and marginal probability distributions on individual variables may be obtained. Entropy  $H_\phi$  is also parameterized since it depends on  $p_\phi$ . Finally, denote the transient state by  $\bar{X}$ , where  $p_\phi(\bar{X}) = \frac{1}{T} \sum_{t \in [T]} P_\phi(X_t = \bar{X})$



**Figure 2.** A model of intrapersonal ineffability. Information is channeled through the stages of input ( $\mathbf{d}$ ), subprocess state ( $\mathbf{v}$ ), working memory ( $\mathbf{x}$ ,  $\mathbf{a}$ ), trajectory of conscious experiences ( $\mathbf{s}$ ), and verbal report ( $\mathbf{m}$ ). A trajectory  $\mathbf{x}$  in the state space of working memory follows attractor dynamics, converging near an attractor  $\mathbf{a}$ . Each step transforming one variable to another is executed by the dynamics of the individual’s brain, which is determined by parameters  $\phi$ . A trajectory of conscious experiences  $\mathbf{s}$  is a function of the subject’s cognitive parameters  $\phi$  and working memory trajectories  $\mathbf{x}$  and encodes the experiences’ meanings.

is the probability that any transient state takes the value  $\bar{x}$ .

Our dynamical systems model of working memory distinguishes between two kinds of working memory state, attractor states and transient states, where the latter includes all time-varying states occupied by the system and the former corresponds to system output or the accessible contents of working memory (Khona and Fiete, 2022). Our model remains neutral about whether conscious states correspond exactly to transient states or attractor states but allows conscious state to be more generally a deterministic function of these states, thus conveying part of their information. Specifically, since  $\mathbf{s} = f_\phi^S(\mathbf{x})$ , conscious experience is not restricted to be identical to transient working memory states or attractor states, but is the output of a deterministic time-varying function of the trajectory through working memory states, where the function depends on cognitive parameters  $\phi$ . While we will not focus on the implementation details of how conscious experiences might relate to neural processes, intuitively,  $\mathbf{s}_t$  is some mathematical object (e.g. a vector of real numbers) representing one state in an abstract space of possible experiences (for background on what it means to formalize  $\mathbf{s}_t$  as a mathematical object and current approaches, see Kleiner (2020); Kleiner and Ludwig (2023)). Subsequently, information theory gives us the ability to reason about the relative richness and ineffability of conscious experience based on the computation graph, without needing implementation details of the functions.

## Information loss from attractor dynamics

The relation of trajectories  $\mathbf{x}$  to a smaller subset  $\mathbf{a}$  of attractor states is a defining characteristic of attractor dynamics, whether the subset consists of a discrete number of fixed points or a set of states that trace out a complex shape such as a curved manifold. In this section, we argue that the presence of attractor dynamics decreases the richness of working memory states and conscious experience. We will identify two related effects. First, at the level of comparison between systems, the presence of attractors concentrates the probability mass of transient states onto a smaller subspace, reducing the richness of transient states. Second, we show that at the level of comparison between states, since attractor states are less rich than transient states in general and the former constitute outputs of the system, the richness of attractor states limits the richness of downstream variables. Since dynamics are characterized by the flow of transient states toward an attractor in  $\mathcal{A}$  followed by persistent membership in  $\mathcal{A}$  and attractors  $\mathcal{A}$  typically constitute a significantly smaller subset

**Box 3.** Implications of reducing transient state richness

Reducing the richness of transient states  $H_\phi(\bar{X})$  also reduces a ceiling on the richness of full trajectories  $H_\phi(X)$ , since  $H_\phi(X) = H_\phi(X_1 \dots X_T) \leq \sum_{t \in [T]} H_\phi(X_t) \leq T(H_\phi(\bar{X}) + C)$  by the addition rule of entropy, where constant  $C = \max_{t \in [T]} (H_\phi(X_t) - H_\phi(\bar{X}))$  limits the maximum deviation of entropy between individual timesteps and the temporal average. This in turn reduces a ceiling on the richness of conscious experience as  $H_\phi(S) \leq H_\phi(X)$ . The latter can be shown as follows: the joint entropy  $H_\phi(S, X) = H_\phi(X) + H_\phi(S|X) = H_\phi(X)$  since  $f_\phi^S$  is deterministic, i.e.  $H_\phi(S|X) = 0$ .  $H_\phi(X) = H_\phi(S, X) = H_\phi(S) + H_\phi(X|S)$ , Shannon entropy is non-negative, and thus,  $H_\phi(S) \leq H_\phi(X)$ .

**Box 4.** Richness of attractors strictly less than richness of trajectories

As the full trajectory determines the attractor it terminates in,  $f_\phi^A$  is a deterministic function. It follows that  $H_\phi(A|X) = 0$ . We also know that  $H_\phi(X|A) > 0$  since multiple possible trajectories terminate in the same attractor state. Our result follows from this asymmetry. By the general relationship between joint and conditional entropies, we have  $H_\phi(X, A) = H_\phi(X) + H_\phi(A|X)$ . Since  $H_\phi(A|X) = 0$ , we have  $H_\phi(X, A) = H_\phi(X)$ . Reapplying the relation between joint and conditional probabilities, we also have  $H_\phi(X, A) = H_\phi(A) + H_\phi(X|A)$ . From these observations together, we know that  $H_\phi(A) + H_\phi(X|A) = H_\phi(X)$ . Since  $H_\phi(X|A) > 0$ , this yields  $H_\phi(A) < H_\phi(X)$ .

of all possible transient states  $\bar{X}$  (Khona and Fiete, 2022), the presence of attractors decreases the richness of transient states  $H_\phi(\bar{X})$ . Since entropy is a measure of distributional spread, dynamics with larger nonattractor transient state sets  $\bar{X} \setminus \mathcal{A}$ , implying more time spent in nonattractor states, yield richer distributions over transient states  $P_\phi(\bar{X})$ ; conversely, faster convergence to attractors and more time spent at attractors yield lower  $H_\phi(\bar{X})$ . (Note that  $\bar{X}$  is the set of all possible transient states, distinct from  $\bar{X}$  which is the variable for a transient state.) In turn, reducing the richness of transient states limits the richness of full trajectories and conscious experience (Box 3).

The same reasoning applies under Kolmogorov's formalism if the probability distribution is known, because expected Kolmogorov complexity  $\mathbb{E}_{p_\phi(\bar{\mathbf{x}})} K(\bar{\mathbf{x}}|p_\phi)$  is equivalent to entropy  $H_\phi(\bar{X})$  up to an additive constant if the distribution  $p_\phi$  is given (Eq. (6), since the program merely needs to access  $p_\phi$ , print it, and halt). Intuitively, this is because knowing the distribution gives the encoder a short-cut; the shortest lossless descriptor of  $\bar{\mathbf{x}}$ , given knowledge of the distribution  $P_\phi(\bar{X})$  and thus support  $\bar{X}$ , has length  $-\log p(\bar{\mathbf{x}})$  under Shannon's noiseless coding theorem (Grünwald and Vitányi, 2004). Given knowledge of  $P_\phi(\bar{X})$ ,  $-\log p(\bar{\mathbf{x}})$  bits are all that is additionally needed to determine the state using a descriptively simple (but not necessarily computationally short) computer program.

Thus far, we have described how the presence of attractors can decrease the richness of transient states overall, i.e. as a matter of comparing between systems (e.g. two brains). We now turn to a second way in which attractors reduce richness, as a matter of comparison between states in a given system.

Global Workspace Theory postulates that the access of representations from working memory by diverse processes across the brain depends on the representations being "amplified and maintained over a sufficient duration," for instance, for a minimum of approximately 100 ms (Dehaene and Naccache, 2001). In the language of the attractor framework, this amounts to the claim that the variable released to downstream processes such as verbal-behavioral reporting and long-term memory is  $A$ , not  $X$ . Crucially, attractor states are strictly less rich than trajectory states  $H_\phi(A) < H_\phi(X)$ , as explained in Box 4. Thus, selective release of attractor working memory states to downstream processing functions such as  $f_\phi^M$  implements an information bottleneck that limits the richness of downstream inputs. This constitutes an important source of ineffability, where our in-the-moment experiences  $S$  are richer than our later recollections, since richness of experience is upper bound by the richness of trajectories (i.e.  $H_\phi(X) \geq H_\phi(S)$ , Box 3),

so the higher the richness of trajectories, the higher the ceiling on information loss from conscious experience to the attractor state and downstream variables. This will be relevant to our discussion of phenomenal overflow (Block, 2007) later. In practice, one would expect the magnitude of information loss from trajectory  $X$  to working memory output  $A$  to be significantly large, since trajectories are sequences of brain states specifying the activity of billions of neurons, whereas working memory appears to be limited to representing a handful of items (Sperling, 1960), which gives us a clue to the magnitude of the bottleneck.

Researchers do not agree on the definition of phenomenal conscious experience  $S$ , and to allow for this, our results do not depend on the exact definition of  $f_\phi^S(X)$ , which does not need to be known to reason about bounds on richness and ineffability. For example, we do not assume an equality between  $S$  and  $A$  (the contents of working memory or access consciousness), which would amount to equating phenomenal consciousness with access consciousness; however, this is a special case that the model supports. Note that without prior assumption (i.e. with uniform expectation) on the degree of information loss incurred by unknown  $f_\phi^S$ , the expected richness of  $S$  is more rich than  $A$ , because the mean of  $\text{Uniform}(0, H_\phi(X))$  is  $H_\phi(X)/2$ , which is not as reductive as  $H_\phi(A) \ll H_\phi(X)$ . More statements on richness (and not its bound) can be made by considering special cases of  $f_\phi^S$ . One intuitive special case is  $S = X$  (conscious experience corresponds to the trajectory through neural state space) where the ineffability or relative richness of  $S$  with respect to  $A$  and  $M$  is maximized out of all choices of  $f_\phi^S$ . Another special case is  $S = [A, \dots, A]$  (no material difference between phenomenal and access consciousness) in which case  $S$  has the same richness as  $A$  but is still at least as rich as  $M$ . Importantly, as will be discussed in the "Existence and report of phenomenal experience" section, the report of ineffable richness at the point of  $A$  is justifiable in both cases.

**Information loss at verbal report**

The ineffability of an experience is perhaps most obvious when we attempt to put it into words, due to the highly compressed nature of language (Kirby et al., 2015). From the computation graph, we can say that ineffability or information loss from conscious experience to verbal report is at least as great as information loss from conscious experience to the working memory attractor (Box 5). Additionally, it would be reasonable to assume that information losses  $H_\phi(A|M)$  and  $H_\phi(S|M)$  are strictly positive (i.e.  $H_\phi(A|M) > 0$  and  $H_\phi(S|M) > 0$ ) if message  $M$  is a low-dimensional symbolic variable

**Box 5.** Ineffability of conscious experience to verbal report

From the computation graph,  $S-X-A-M$  forms a Markov chain even though  $S$  is computed from  $X$  ( $S$  is conditionally independent of  $A$  if given  $X$ ), and thus,  $S-A-M$  is also a Markov chain ( $S$  is conditionally independent of  $M$  if given  $A$ ). Thus,  $I_\phi(S;A) \geq I_\phi(S;M)$  from the data processing inequality theorem, implying  $H_\phi(S) - H_\phi(S|A) \geq H_\phi(S) - H_\phi(S|M)$  and  $H_\phi(S|M) \geq H_\phi(S|A)$ .

(such as a few words), whereas  $A$  and  $S$  are snapshots of working memory and conscious experience, since conditional entropy is strictly positive if every mapping is either one-to-one or one-to-many and there is at least one case of the latter. While it might appear that language is rich, note that  $n$  characters with an alphabet of 256 possible characters require not more than  $8n$  bits to represent, whereas the neural state is determined by the activity of up to approximately 100 billion neurons (Herculano-Houzel, 2009).

Information loss from attractor  $A$  or conscious experience  $S$  to verbal message  $M$  means that the latter do not fully identify the former and instead divide the space of attractors and conscious experiences more coarsely. For instance, saying that one “saw a fat cat” leaves out significant details about the specific attractor that generated the message, which would be difficult to communicate fully (e.g. the cat’s color, size, pose, the surrounding environment, etc.). Positive information loss  $H_\phi(S|M)$  implies that it is generally impossible to recover the conscious experience from the verbal message with certainty. Note that as long as  $H_\phi(A|M)$  is strictly positive, this means that conscious experience is somewhat ineffable to verbal report even if we identify conscious experience with the working memory attractor state.

An additional source of ineffability is that attractors can have more complex and high-dimensional structures than simple fixed points, which is common in high-dimensional systems. Such a system would exhibit increased richness of attractor state  $H_\phi(A)$  and increased ineffability, as the same richness of messages  $H_\phi(M)$  and an increase in joint entropy  $H_\phi(A,M)$  imply an increase in information loss  $H_\phi(A|M)$ , since  $H_\phi(A,M) = H_\phi(M) + H_\phi(A|M)$ .

### Hierarchical attractor dynamics

The brain is hierarchical in nature with many levels of spatial and temporal organization that can be studied, ranging from molecular and synaptic activities to local networks and large-scale networks (Changeux and Dehaene, 1989). Attractor dynamics appear to be ubiquitous across organizational levels and cortical regions of the brain, with processing in the neocortex hypothesized to support many attractor networks each concerned with a different type of processing (executive function and working memory, general short-term memory, long-term memory, etc.) (Rolls, 2007; Rolls, 2010; Khona and Fiete, 2022). The presence of multiple weakly coupled neocortical attractor networks yields benefits including specialization and increased memory capacity and in addition has ramifications for understanding conscious experience.

Anatomically, the inferior temporal cortex is an example of a sensory processing area that responds discriminatively to novel stimuli, whereas the prefrontal cortex is implicated in maintaining attention-modulated projections of such representations in

working memory (Rolls, 2007; Miller et al., 1993; Renart et al., 1999). Neural activity in both regions maintains persistence over time and exhibits attractor dynamics, but the content of sensory memory is akin to the state of a worker subprocess, whereas the content of working memory corresponds to the state of executive control; working memory representations exhibit increased temporal stability, persisting for longer durations of up to several seconds, and provide top-down feedback to diverse regions of the brain, including the inferior temporal cortex (Rolls, 2010; Chelazzi, 1999; Bushnell et al., 1981). The ability of the prefrontal attractor to stabilize in its high firing rate attractor state is attributable to positive feedback from strong internal recurrent connections that suppress incoming stimuli (Renart et al., 1999). The need to maintain information in working memory during periods where new stimuli may be perceived exemplifies why working memory and subprocess memory necessitate distinct attractor networks (Rolls, 2010; Rolls, 2007). The limits imposed on the richness of working memory state by subprocess memory states may be illustrated in an information theoretic manner by considering that the latter is an input to the former (Box 6).

**Box 6.** Richness of subprocess states constrains richness of conscious experience

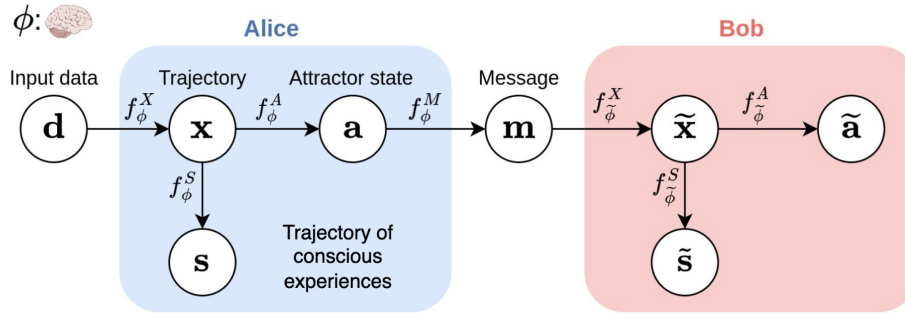
Extracting the stochasticity in  $f_\phi^X$  into an input variable  $\omega$ , meaning assuming that computation of  $X$  is cast as  $X = \hat{f}_\phi^X(V, \omega)$  where  $\hat{f}_\phi^X$  is deterministic, the richness of  $X$  is bound as  $H_\phi(X) \leq H_\phi(V_1, \dots, V_N, \omega) \leq \sum_{n \in [N]} H_\phi(V_n) + H_\phi(\omega)$  due to deterministic data processing and addition rule of entropy. That is, given a limit on the richness of noise  $H_\phi(\omega)$ , a ceiling on the richness of working memory trajectories  $H_\phi(X)$  scales with the richness of the subprocess states that constitute its inputs. In turn, this restricts ceilings on the richness of downstream variables such as conscious experience and working memory attractors (Box 3).

### Interpersonal ineffability

Communication channels are not limited to personal sensory processes and verbal or behavioral reporting processes but extend to channels between individuals. In this section, we will consider communication between two individuals using the model summarized in Fig. 3 in which a speaker, Alice, wishes to communicate her experience to a listener, Bob. We use the same variables as as in section Intrapersonal ineffability, but denote Bob’s variables using “~” (e.g.  $\tilde{\mathbf{s}}$  denotes Bob’s conscious experience). Again, we assume a computational chain of states  $\mathbf{x} \rightarrow \mathbf{a} \rightarrow \mathbf{m} \rightarrow \tilde{\mathbf{x}} \rightarrow \tilde{\mathbf{a}}$  that elicit an experience  $\tilde{\mathbf{s}} = f_\phi(\tilde{\mathbf{x}})$  in Bob. In section Intrapersonal ineffability, we have already considered sources of ineffability up to  $H_\phi(\mathbf{s}|\mathbf{m})$  and  $K(\mathbf{s}|\mathbf{m}, p_\phi)$  in this chain. What remains is to identify additional sources of ineffability after the message is transmitted. In this section, we use the Kolmogorov formalism, since we assume that the parameters  $\phi$  of Alice’s brain are not available to Bob.

#### A blank slate listener

Before considering the case in which Bob is a typical human listener, we begin with a discussion of ineffability when Bob is a blank slate (setting  $\tilde{\phi} = \emptyset$ ,  $\tilde{\mathbf{x}} = \emptyset$ ,  $\tilde{\mathbf{s}} = \emptyset$ ,  $\tilde{\mathbf{a}} = \emptyset$ , where  $\emptyset$  denotes the null value). In this case, the chain of communication ends



**Figure 3.** A model of interpersonal ineffability. We model the communication pipeline between a speaker Alice and a listener Bob. A trajectory  $\mathbf{x}$  in Alice's state space of working memory follows attractor dynamics, converging near an attractor  $\mathbf{a}$ . Alice then attempts to communicate the experiences with a message  $\mathbf{m}$ . On Bob's end, the message is decoded and influences his working memory trajectory  $\tilde{\mathbf{x}}$ , which in turn converges near an attractor  $\tilde{\mathbf{a}}$ . Each step transforming one variable to another is executed by the dynamics of the subject's brain, denoted by  $\phi$  for Alice and  $\tilde{\phi}$  for Bob. Trajectories of conscious experiences  $\mathbf{s}$  and  $\tilde{\mathbf{s}}$  are functions of the subjects' cognitive parameters  $\phi$  and  $\tilde{\phi}$  and working memory trajectories  $\mathbf{x}$  and  $\tilde{\mathbf{x}}$ , respectively, and encode the experiences' meanings. We are interested in the ineffability  $K(\mathbf{s}|\tilde{\mathbf{s}}, p_{\tilde{\phi}})$  of Alice's conscious experiences  $\mathbf{s}$  given the experiences  $\tilde{\mathbf{s}}$  elicited in Bob.

at  $\mathbf{m}$ , and thus, a quantity of interest is the ineffability  $K(\mathbf{s}|\mathbf{m})$  (without assuming access to Alice's cognitive parameters  $\phi$ , as we did in the "Information loss at verbal report" section). Intuitively, what this quantity refers to is the "intrinsic" ineffability of an experience given its message, without conditioning on extra information such as cognitive parameters  $\phi$  or  $\tilde{\phi}$ . Taking an expectation to express average ineffability of conscious experience  $\mathbf{s}$ , we have  $\mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m}) \geq \mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m}, p_{\phi})$  trivially since conditioning on more information cannot increase the length of the shortest program that outputs  $\mathbf{s}$ , but it is important to note that one would additionally expect the reduction to be significant, i.e.  $\mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m}) \gg \mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m}, p_{\phi})$ . This is because under Shannon's noiseless coding theorem, knowledge of Alice's state distribution  $p_{\phi}$  reduces the problem of describing  $\mathbf{s}$  in the general space of high-dimensional vectors to the problem of describing its index among the set of all possible conscious experiences associated with  $\mathbf{m}$  for a brain parameterized by  $\phi$ .

The inequality  $\mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m}) \gg \mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m}, p_{\phi})$  relates to an observation at the core of the philosophical debate on ineffability: our descriptions of our experiences never seem to come close to capturing their full richness. The gap is so significant that it has at times led some philosophers, scientists, and laypersons to the dualistic conclusion that conscious experiences are intrinsically indescribable, such that there is something more to their content than physically embodied information encoded in neural activity. Using our model, we argue that these intuitions do not necessarily imply a nonphysical basis for conscious experience but may be explained by physically grounded and significant information loss that is a natural consequence of computational processing between the cognitive states underlying our experiences and the linguistic messages that we use to express them. While the representation of  $\mathbf{m}$  is shared among individuals who speak the same language, the representation of  $\mathbf{a}$  is unique to communicator Alice. Therefore, under the Kolmogorov formalism, there is complexity or information content in  $\mathbf{a}$  that requires adopting Alice's representation space to reconstruct.

The significant magnitude of  $\mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m})$  captures the blank slate or tabula rasa case of the problem of ineffability: without assuming knowledge of the parameters of Alice's brain, experiences are highly ineffable using low-dimensional descriptions such as typical verbal messages. Nonetheless,  $K(\mathbf{s}|\mathbf{m}) \leq K(\mathbf{s}) < \infty$ ; our experiences are describable "in principle," even to a blank slate observer where no additional information is assumed. Using a numerical scale to quantify ineffability allows us to convey the

dual sense in which our experiences are, to varying degrees, both communicable and ineffable.

### A typical listener Cognitive similarity and effability

In a realistic communication scenario, the cognitive parameters of listener Bob  $\tilde{\phi}$  are given by a high-dimensional vector that provides information about Alice's parameters  $\phi$  within the generic space of high-dimensional vectors, due to the shared physical environment (including cultural experience) and shared evolutionary background, and thus may be used to reduce the description length of  $p_{\phi}$ . Trivially, we have that the expected ineffability of Alice's conscious experience can only improve by conditioning on Bob's parameters  $\mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m}) \geq \mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}K(\mathbf{s}|\mathbf{m}, p_{\tilde{\phi}})$ . However, we also obtain that a ceiling on the disadvantage of using Bob's parameters compared to Alice's parameters scales with the difference between them (Box 7).

#### Box 7. Cognitive dissimilarity and ineffability

From Grünwald and Vitányi (2004, Theorem 2.10) we obtain for given  $\mathbf{m}, p_{\phi}, p_{\tilde{\phi}}$  that  $0 \leq \mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}[K(\mathbf{s}|\mathbf{m}, p_{\tilde{\phi}})] - H_{\phi}(\mathbf{S}|\mathbf{m}) \leq K(p_{\phi}(\cdot|\mathbf{m})|p_{\tilde{\phi}}, \mathbf{m}) + c \leq K(p_{\phi}|p_{\tilde{\phi}}) + c$ , where  $c$  is a constant, and  $0 \leq \mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}[K(\mathbf{s}|\mathbf{m}, p_{\phi})] - H_{\phi}(\mathbf{S}|\mathbf{m}) \leq K(p_{\phi}(\cdot|\mathbf{m})|p_{\phi}, \mathbf{m}) + c = \epsilon + c$ , where  $\epsilon$  is the negligible descriptive complexity of  $p_{\phi}(\cdot|\mathbf{m})$  given  $p_{\phi}$ . Note  $H_{\phi}(\mathbf{S}|\mathbf{m}) \geq H_{\phi}(\mathbf{S}|\mathbf{m}, p_{\tilde{\phi}})$  where the underlying joint distribution includes the meta-distribution over  $p_{\tilde{\phi}}$ , and likewise  $H_{\phi}(\mathbf{S}|\mathbf{m}) \geq H_{\phi}(\mathbf{S}|\mathbf{m}, p_{\phi})$ . Then,  $\mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}[K(\mathbf{s}|\mathbf{m}, p_{\tilde{\phi}})] \leq H_{\phi}(\mathbf{S}|\mathbf{m}) + K(p_{\phi}|p_{\tilde{\phi}}) + c$  and  $\mathbb{E}_{p_{\phi}(\mathbf{s}|\mathbf{m})}[K(\mathbf{s}|\mathbf{m}, p_{\phi})] \leq H_{\phi}(\mathbf{S}|\mathbf{m}) + \epsilon + c$ . The difference between upper bounds on ineffability is  $K(p_{\phi}|p_{\tilde{\phi}}) - \epsilon$ .

The mismatch between Alice and Bob's parameters, which is formalized by  $K(p_{\phi}|p_{\tilde{\phi}})$  or the minimum number of bits required to encode a program that produces Alice's parameters from Bob's, loosely corresponds to the difference between Bob and Alice's cognitive function (Box 8), which depends on the extent to which they differ in genetic biases and lived experiences. This result supports the common intuition that our experiences are more effable or communicable to people who are similar to ourselves. It also resonates with the empirical observation of greater interbrain syn-



**Box 8.** Difference in functionality and difference in parameters

For a scalar valued function  $h$  with bound gradient magnitude, we have  $h(\mathbf{x}, \tilde{\theta}) = h(\mathbf{x}, \theta) + (\tilde{\theta} - \theta)^\top \nabla_\theta h(\mathbf{x}, \theta) + \mathcal{O}(\|\tilde{\theta} - \theta\|^2) \leq h(\mathbf{x}, \theta) + \|\tilde{\theta} - \theta\| \|\nabla_\theta h(\mathbf{x}, \theta)\| + \mathcal{O}(\|\tilde{\theta} - \theta\|^2)$  by the Taylor expansion. Assuming that first-order gradients are bound by positive constant  $C$ , then we have  $|h(\mathbf{x}, \tilde{\theta}) - h(\mathbf{x}, \theta)| \leq C\|\tilde{\theta} - \theta\| + \mathcal{O}(\|\tilde{\theta} - \theta\|^2)$ , i.e. an upper bound on the mismatch in functional output given parameterization  $\theta$  and  $\tilde{\theta}$  scales with the Euclidian distance between them.

chronization in related individuals (Goldstein et al., 2018) and how the brain's anatomical structure (i.e.  $\phi$  and  $\tilde{\phi}$ ) affects the propensity to communicate at the interpersonal level (Dumas et al., 2012).

Consider a prototypical example of inter-personal ineffability, in which Bob has been blind from birth and Alice is attempting to convey her experience of seeing the color red. In this case, Bob's brain might be so different from Alice's that the distance between their cognitive parameters  $K(p_\phi | p_{\tilde{\phi}})$  is sufficiently high that the benefit of conditioning on his own parameters is negligible. In other words, since  $\mathbb{E}_{p_\phi(\mathbf{s}|\mathbf{m})} K(\mathbf{s}|\mathbf{m}, p_{\tilde{\phi}}) \leq K(p_\phi | p_{\tilde{\phi}}) + c + H_\phi(\mathbf{S}|\mathbf{m})$  (Box 7), if  $K(p_\phi | p_{\tilde{\phi}})$  is large, then the ceiling on  $\mathbb{E}_{p_\phi(\mathbf{s}|\mathbf{m})} K(\mathbf{s}|\mathbf{m}, p_{\tilde{\phi}})$ , the ineffability of Alice's conscious experience given the message from Bob's perspective, is also large. Intuitively, when  $K(p_\phi | p_{\tilde{\phi}})$  is small, the information required to communicate the functions  $f_\phi^M$ ,  $f_\phi^A$ , and  $f_\phi^S$  in order to reconstruct  $\mathbf{s}$  from  $\mathbf{m}$  is offloaded to  $p_{\tilde{\phi}}$ , which is given, thus reducing a ceiling on expected program length  $\mathbb{E}_{p_\phi(\mathbf{s}|\mathbf{m})} K(\mathbf{s}|\mathbf{m}, p_{\tilde{\phi}})$ .

The cognitive dissimilarity factor  $K(p_\phi | p_{\tilde{\phi}})$  is also implicated in Frank Jackson's famous thought experiment, color scientist Mary who has lived her whole life in an entirely black and white room and has learned exhaustive knowledge about the process of color perception, but nonetheless possesses a brain that is incapable of understanding the experience of color (i.e. she does not know what it is like to see red) (Jackson, 1986; Alter and Walter, 2006). Since her knowledge is exhaustive, she knows everything that anyone could possibly tell her about the experience of seeing something red. Jackson argues that when she finally sees something red, she nevertheless learns something new ("what it is like to see red"). It has been argued that since she already knew all the physical facts, what she learned must have been a nonphysical fact (Jackson, 1986; Chalmers, 2010). Many philosophers have responded to this argument, developing different conceptions of how what Mary learns might be physical after all (Alter and Walter, 2006). Our model can be understood as offering support to the physicalist account. It highlights how the ineffability  $\mathbb{E}_{p_\phi(\mathbf{s}|\mathbf{m})} K(\mathbf{s}|\mathbf{m}, p_{\tilde{\phi}})$  of Alice describing her experience of color to Mary (who is playing the role of Bob) may be explained in part by the difference in their cognitive function. In other words, the ability to empathize with another person from a verbal report of their experience is aided by cognitive similarity or ease of reconstructing their cognitive function based on knowledge of one's own cognitive function, but simply memorizing a description of how the brain behaves in response to color does not imply that one's brain is capable of responding in that manner upon being exposed to it or its reference (i.e. hearing the word "red"), and it is similarity in cognitive behavior that is implicated in  $K(p_\phi | p_{\tilde{\phi}})$ .

The result in Box 7 states that high ineffability of Alice's experience of color to Mary implies high cognitive dissimilarity between Alice and Mary. Cognitive dissimilarity is not equivalent

to knowledge inadequacy; knowing how brain should respond does not imply being able to execute such a response. The view that Mary learns different cognitive behavior upon exposure to the color red is closest to the interpretation that she acquires a new ability (Lewis, 1990), as opposed to a new mode of presentation (Loar, 1990), a new relation of acquaintance (Conee, 1985), or a reminder of something that in principle she must have had access to all along (Dennett, 2006; Rabin, 2011).

### Theory of mind

Evolution has optimized human beings to be skilled at inferring the thoughts of others, an ability termed "Theory of Mind" (Premack and Woodruff, 1978; Graziano and Kastner, 2011; Graziano and Webb, 2015; Kelly et al., 2014). In our model, there is a link between theory of mind and ineffability. If cognitive functions  $f_\phi^X$  and  $f_\phi^S$  that produce Bob's conscious experience  $\tilde{\mathbf{s}}$  are optimized for decoding  $\mathbf{m}$  into Alice's conscious experience  $\mathbf{s}$ , then ineffability is reduced compared to reconstructing Alice's conscious experience from the raw message,  $K(\mathbf{s}|\mathbf{m}, \tilde{\phi}) \geq K(\mathbf{s}|\tilde{\mathbf{s}}, \tilde{\phi})$ , because part of the computation of reconstructing  $\mathbf{s}$  is executed during inference of  $\tilde{\mathbf{s}}$ , meaning that the smallest program from  $\tilde{\mathbf{s}}$  and  $\tilde{\phi}$  to  $\mathbf{s}$  would make use of  $\tilde{\mathbf{s}}$  to reduce its residual work, shortening the descriptive length of the program. In the extreme case, if  $K(\mathbf{s}|\tilde{\mathbf{s}}, \tilde{\phi}) \doteq 0$ , then by definition, Bob's cognitive function is optimal for inferring Alice's conscious experience, since no material additional information is required to determine  $\mathbf{s}$ .

In turn, if Alice's parameters  $\phi$  contain information about Bob's cognitive function or parameters  $\tilde{\phi}$ , she is capable of producing her message  $\mathbf{m}$  in a way that maximizes effability and minimizes  $K(\mathbf{s}|\tilde{\mathbf{s}}, \tilde{\phi})$ , since her cognitive functionality, including verbal reporting function  $f_\phi^M$ , depends on  $\phi$ .

### Phenomenal and access consciousness

Having provided an information theoretic dynamical systems perspective on richness and ineffability, we now turn explicitly to the question of whether rich phenomenal experience exists and why we self-report that it does. We first highlight ambiguities in the meaning of access before contrasting two hypotheses for explaining the report of phenomenal experience.

#### Effability, accessibility, and reportability

Notions such as "accessible," "reportable," and perhaps "effable" are somewhat ambiguous. A benefit of our framework is that it allows us to distinguish between (at least) three distinct notions in the vicinity.

First, as we have presented it earlier, the notion of "effability" refers to the ability to accurately describe one variable by another, which implies that it can be formalized using mutual information ("Richness and ineffability" section).

Second, "access" is interpretable in two different ways. Direct access is the notion of a variable  $X$  being a direct input to a function or process  $g$ , meaning that  $g$  is defined on variable  $X$ , whereas informational access is the notion of  $g$ 's input variable  $A$  sharing mutual information with  $X$ ,  $I(A; X) > 0$ , corresponding to  $X$  being effable with respect to  $A$ . A process that has no direct access to  $X$  may still have access to its information via inputs; if  $M = g(A)$ , process  $g$  has access to information about  $X$  if  $I(A; X) > 0$ . Thus, a variable may be effable with respect to the input and output variables of a process without being directly accessible to the process.

Third, while a reporting process is in general a process or transformation that outputs to another process, we stipulate that

“reporting process” may be understood to refer specifically to those that output to processes outside the cortex, such as cortical processes that encode speech or motor movements. We may then say that a variable is directly (or informationally) reportable if it is directly (or informationally) accessible by a reporting process, where the report corresponds to the output of the reporting process.

### Existence and report of phenomenal experience

According to the Global Workspace Theory, information from diverse brain regions corresponding to a variety of perceptual or cognitive processes is selected for inclusion in the contents of a centralized processing workspace associated with working memory that coordinates and communicates with multiple subsystems (Baars, 1993; Baars, 2005; Dehaene et al., 1998).

The features of this global workspace system make it suitable as a framework for an analysis of consciousness (i.e. phenomenal consciousness), even if we do not assume that only items in workspace are conscious. The features of the global workspace system also make it a suitable target for modeling in terms of attractor dynamics, since by their nature, states amplified and sustained in a central processing workspace are attractors. Thus, our model allows for the refinement of these concerning the relationship of consciousness to the global workspace.

Global workspace models of consciousness (Dehaene and Naccache, 2001) generally divide representations into three classes:

- 1 Those not computed by working memory processes (unconscious).
- 2 Those mobilized in the workspace via amplification and made accessible to downstream processing (conscious).
- 3 Those computed by working memory processes but not sufficiently amplified or attended to be released by the workspace.

The latter includes nonattractor transient states in an attractor model of working memory and being rich and unreportable, which are clear candidates for the basis of phenomenal experience (Dehaene and Naccache, 2001). It is a point of debate between adherents of the global workspace framework, whether items from the third class are indeed conscious. Some say no (Naccache, 2018; Cohen et al., 2016), and others say yes (Prinz, 2012). By allowing  $f_{\phi}^S$  to be abstract, our model is compatible with both views. We account for the report of ineffable phenomenal conscious experience for either case later.

- (i) Intermediate states of neural trajectory included in conscious experience  $S$ . If one assumes that attractor states are included in the content of consciousness and that the physical basis of transient states and attractor states in working memory is the same (i.e. they are differentiated by duration of attentional amplification, not location of neural circuitry), it would be reasonable to believe that transient states are also included in conscious awareness. If this is the case, then transient states are rich states that are consciously experienced but not directly accessible or reportable by downstream processes, while being partially verbally effable because of shared information with attractors which are directly reportable. In this paradigm, the fleeting nature of transient states impacts their direct reportability but not their inclusion in conscious experience.

- (ii) Metacognitive representation in working memory and access consciousness  $A$ . Regardless of whether transient states are included in the contents of phenomenal consciousness  $S$ , the attractor model for working memory suggests a second explanation for the self-report of phenomenal experience: an attractor state may encode information about its basin of attraction and thus information loss. For example, point attractor states may include dimensions whose values estimate the size of its local basin, which is a measure of the information loss when going from transient states in trajectories within that basin to the attractor state itself (Appendix “Metacognitive Representation of Ineffability”). This posits that rich experience exists, whether inside or outside the delimitation of consciousness, and its properties—such as richness—would be reportable, even if the transient states that support them are not. It is plausible that conscious awareness of abstract attributes of transient states such as richness would be advantageous, for instance, when reasoning about one’s uncertainty, including for the purpose of anticipating the listener’s uncertainty when engaging in theory of mind to minimize ineffability (“A typical listener” section). Note that the existence of metacognitive representation in  $A$  does not conflict with the first case (metacognitive representations may exist in  $A$  even if  $S = X$ ).

Our model supports an interpretation for Sperling’s experiments (Sperling, 1960), where subjects briefly exposed to a grid of characters were generally able to report character identities for “any” prompted row (containing  $\sim 4$  characters) but subsequently not other rows, in addition to being able to report that they experienced observing more characters. An account for this behavior is that upon receiving the prompt to report a specific row, working memory contents represented by attractor state  $\mathbf{a}$  contained the identities of characters in the prompted row, a summary over the grid (e.g. the number of characters and their arrangement) and an estimate of the information lost by the summary, while information sufficient to discriminate all characters existed in the processing pipeline but in upstream sensory state  $\mathbf{v}$ , from which  $\mathbf{x}$  and  $\mathbf{a}$  were computed. Subsequently, as attractor state  $\mathbf{a}$  is directly accessible to verbal reporting process  $f_{\phi}^M$ , the characters in the prompted row, grid details at the summary level, and the presence of information loss were directly reportable, and full grid details (identities of all characters) were not. The latter holds irrespective of where the distinction between conscious and unconscious is drawn, i.e. whether  $\mathbf{x}$ , which might have contained sufficient information from  $\mathbf{v}$  to discriminate all characters, is considered conscious.

These arguments suggest that Block’s distinction between phenomenal and access consciousness (Block, 1995) can be attributed to a difference in the representational stage of information processing (Dehaene and Naccache, 2001) and that the existence of a rich phenomenological experience that exceeds our reporting abilities (Sperling, 1960) is both justifiable and veridically reportable. Unpacking the implications of the model is an important task for future work.

## Conclusion

This paper characterizes the rich and ineffable nature of conscious experience from an information theoretic perspective. It connects the ordinary notion of ineffability with mathematical formalisms of information loss, describing how the latter arises as a result of computation in cognitive processing, how it is implemented by an

attractor model for working memory, and how it may be increased by the compressed nature of language as well as differences in the cognitive processing functions of individuals.

Attractor dynamics may be considered an attentional process: out of many, one or a few states are selected. This connects our work not only to Global Workspace Theory but more broadly to research in machine learning on attention mechanisms. We generally observe that attention, e.g. as introduced in deep learning by Bahdanau et al. (2015), may be used to name any function that incurs significant information loss and is present in both artificial and biological cognitive systems, where it is—at present—commonly modeled by the family of attention-based and transformer architectures (Bahdanau et al., 2015; Devlin et al., 2019; Khan et al., 2022; Chorowski et al., 2015) and dynamical systems (Khona and Fiete, 2022; Rolls, 2007) respectively.

In this work, we use a simple model to reason about emitter-receptor communication, where the past is conditioned on implicitly via parameters  $\phi$  and stochasticity in dynamics. An alternative would be to model more complex communication patterns explicitly. We have also not considered learning objectives for function parameters. Doing so would enable a discussion on the generalization benefits of the inductive bias (Goyal and Bengio, 2022) giving rise to this information loss: intuitively, how simpler representations support robustness (Mathis and Mozer, 1994) and the successful extrapolation of behavior beyond previously seen inputs. Information bottlenecks are a popular training regularizer in machine learning (Tishby et al., 2000; Alemi et al., 2017; Kawaguchi et al., 2023), but are understudied in the context of biologically plausible models, despite generalization ability being a key difference between humans and current artificial learning systems. Considering the benefits of information loss may allow us to understand ineffability more deeply, not just how it arises, but also why.

## Funding

The authors thank the following institutions for sources of funding: the Canadian Institute for Advanced Research (CIFAR) AI Chair Program, the Canada Research Chair Program, the Good Ventures Foundation, Unifying Neuroscience and Artificial Intelligence in Quebec (UNIQUE), IVADO, the Natural Sciences and Engineering Research Council of Canada (NSERC), Samsung, the Social Sciences and Humanities Research Council (SSHRC), and the Quebec government.

## Data availability

No data were gathered or produced in this study.

## Conflicts of interest

The authors have no conflict of interest to declare.

## Appendix 1

### Illusionism and overflow

Richness and ineffability figure in several important live debates about consciousness in the philosophical literature. Here, we summarize two: the illusionism debate and the overflow debate.

Illusionists argue that consciousness is an illusion, while realists deny this (Frankish, 2016). Illusionists generally argue that our expectations for consciousness are too high: that the job

of describing a conscious experience is too demanding for any physical process to fulfill and that (rather than rejecting physicalism) we should conclude that there is no such thing as consciousness (or at least, make do with a diminished conception of it) (Dennett, 1993; Graziano et al., 2020; Humphrey, 2020). Daniel Dennett famously lists ineffability as one of the hard-to-fulfill conditions that should lead one to illusionism: the prospect that conscious contents somehow escape our attempts to fully describe them is, for Dennett, a sign that consciousness is chimerical (Dennett, 1993). Notably, illusionists acknowledge that something gives rise to the relevant illusions: there must be an explanation of why it seems plausible to us, on introspection, that we are the subjects of (ineffable) conscious states. Qualia realists, in contrast, see conscious experience as the subjective viewpoint from which all else is observed or known and therefore consider it to be an explanandum that cannot be discarded (Tononi and Edelman, 1998; Chalmers, 2010; Descartes, 1986).

The overflow debate is between those who hold that consciousness is indeed rich and ineffable and those who deny it (while still maintaining that consciousness exists). Richness is a relative term and one contender for a reference object that justifies the characterization of consciousness as rich is the accessible content of working memory. Empirically, there appears to be a clear bandwidth limitation on the latter (Sperling, 1960; Miller and Buschman, 2015; Cohen et al., 2016), which is what makes it difficult, for example, to remember all the names of the people you meet at a party or all the digits of a phone number. Proponents of overflow say that consciousness is considerably richer than this sort of working memory and includes ineffable content unavailable for report (Block, 2007; Bronfman et al., 2019; Lamme, 2007; Vandenbroucke et al., 2012), while the staunchest opponents of overflow will maintain that consciousness is no richer than the bandwidth-restricted content of working memory, generally because they take consciousness to just “be” working memory or a supporting system for it (Ward, 2018; Phillips, 2016; Naccache, 2018; Cohen and Dennett, 2011).

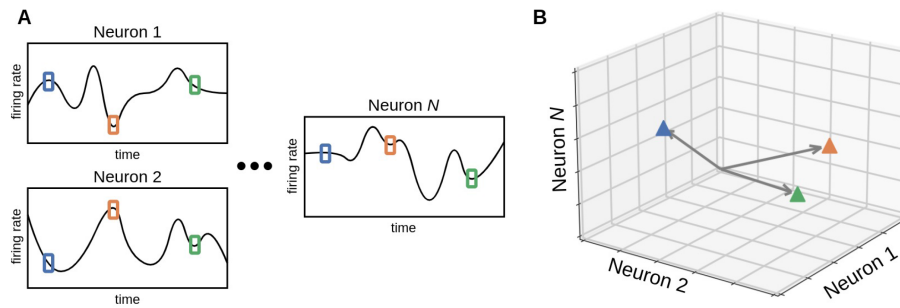
We thus have two important debates where those on both sides may benefit from a formal model of ineffability: illusionists and realists who deny overflow may benefit from a general model of why it seems to us that we are the subjects of rich and ineffable experiences, while realists who accept overflow may benefit from a characterization of how it emerges.

## Appendix 2

### Computation through neural dynamics

In the “An information theoretical dynamical systems perspective on conscious experience” section, we argue that we can account for the richness and ineffability of experience by modeling conscious states as neural trajectories in a high-dimensional dynamical system with attractors. In this section, we provide a brief overview of the essential concepts needed to understand the model.

First, we will introduce the notion of a neural activation space, in which temporally evolving states of neural activity follow trajectories governed by recurrent dynamics in the brain. Next, we will explain how state attractors, which are emergent properties of dynamical systems, can allow neural networks to solve computational problems that require some form of persistent memory.



**Figure A1.** Visualization of neural state space. (A) The activity trace for multiple neurons, where activity can be quantified in several different ways (e.g. firing or not, firing rate over some time window, membrane voltage, etc.). Boxes denote joint activity patterns across all neurons at specific timepoints. (B) At any particular timepoint, the joint activity pattern across  $N$  different neurons can be expressed as a vector in an  $N$ -dimensional state space.

Along the way, we will highlight key examples from the computational neuroscience literature where this dynamical systems framework was used to explain how populations of neurons solve perceptual and cognitive tasks.

### Neural activation state space

At any given moment, every neuron in the brain has some level of activity, and this activity can be numerically quantified in several different ways (e.g. firing or not, firing rate over some time window, membrane voltage, etc.), which we illustrate in Fig. 1a. Together, this instantaneous pattern of activity defines the brain's current "state", which may be compactly represented as a vector in an  $N$ -dimensional state space, where  $N$  is the number of neurons in the brain (or in the subpopulation of interest). In such a representation, each index in this vector identifies a particular neuron, and the value of a particular index corresponds to that neuron's current level of activity (Fig. 1b). We reason at the level of neuronal activity for clarity, but strictly, our framework makes no assumptions about the appropriate level of granularity: where these make direct contributions to cognitive information processing, other cells such as astrocytes or cell components such as dendrites may be state space parameters in their own right (Godfrey-Smith, 2016).

A benefit of describing neural activity in this manner is that it allows us to draw on the mathematical framework of dynamical systems theory to reason about mental states. For example, we can now talk about what a pattern of neural activity "represents" by projecting the state onto lower-dimensional subspaces that encode some meaningful feature. To explain this with example, it might be the case that when perceiving an object, certain dimensions of the elicited state represent its color, others represent its shape, yet others represent its function, etc. In addition, given a probabilistic transition model for states that account for noise in neural activity and other sources of uncertainty, we can measure quantities such as the likelihood and information content of a state. We can also quantify the similarities between states according to some distance metric between their vectors.

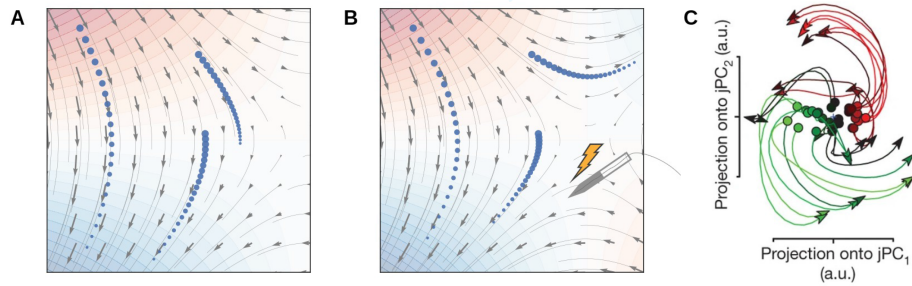
### Neural dynamics

While neural states can be used to represent an instantaneous pattern of activity, the brain is a complex dynamical system and must ultimately be understood in terms of how neural activity unfolds in time. The temporal evolution of neural activity—and any other dynamical system—is governed by two factors.

First, neurons in the brain have a large number of synapses that form recurrent loops. Recurrency means that even in the absence of any sensory input, brain states will evolve dynamically; the activity of one neuron at a particular time will influence the future activity of surrounding neurons, which may in turn influence the original neuron's activity at a later time in a causal loop. The dynamics governing these neural state trajectories are defined by the joint synaptic connectivity profile between all neurons in the brain. Any given connectivity profile results in a set of rules for how each state transitions to the next. This can be visually illustrated for the entire system using a "vector field" as shown in Fig. 2a: each vector indicates how a state at that location would evolve in the next instant in the absence of noise and where the size of the vector denotes the speed of the change. Intuitively, one can understand the dynamics of the system by starting off at an initial point in neural state space and tracing a trajectory that follows the vector field at each point in time. A different connectivity profile would yield different transition dynamics (i.e. a different vector field), and therefore, the same initial neural state would follow a different trajectory.

Another factor that governs neural dynamics is the input to the system, which may itself evolve over time. The dynamics of a subpopulation of neurons (e.g. a particular brain region) are modulated extrinsically by signals from surrounding neurons that synapse onto the population, including information from the stream of sensory signals entering the brain. Illustrated visually in Fig. 2b, this means that inputs warp the vector field that defines transitions from the current state to the next, ultimately resulting in potentially very different trajectories from those that would have occurred given other inputs.

Much of the field of computational neuroscience is concerned with understanding neural population coding through the lens of dynamical systems, thanks to their rich theoretical underpinnings and the mechanistic models they provide (Favela, 2021). Historically, this approach has been particularly fruitful in two systems: sensory integration (Burak, 2014; Zhang, 1996) and motor control (Churchland et al., 2012; Michaels et al., 2016; Shenoy et al., 2013). For example, Churchland et al. (2012) recorded from a population of neurons in the primate motor cortex and found that they exhibited rotational dynamics during a simple reaching task (Fig. 2c). While this was initially surprising because the movement itself was not rhythmic, the authors proposed a theory that muscle activity is constructed from an oscillatory basis, which was later supported by additional experiments. The neural dynamics, then, can be understood as pattern generators that generate



**Figure A2.** Neural dynamics and trajectories in activation space. (A) A dynamical system whose behavior is depicted using vector fields and example trajectories. (B) External inputs can modulate the behavior of a dynamical system (compare vector fields and trajectories with those in (A)). (C) An example of neural dynamics empirically observed in the primate motor cortex. As the neural dynamics are high-dimensional,  $j$ -Principal Component Analysis ( $j$ PCA) was used to reduce their dimensionality for visualization. The figure was reproduced with permissions from Churchland et al. (2012).

sequences of muscle activity optimized for producing natural movements.

Despite the success of this framework in sensory and motor domains, much less is understood about the dynamical underpinnings of higher-level cognition, although such dynamical systems are also implemented with neural substrates and would presumably share similar mechanisms. A contribution of our work is the application of dynamical systems to high-level conscious cognition and analysis of the implications for explaining the richness and ineffability of experience.

### State attractors

When neural dynamics are used to solve computational tasks, it is often the case that the solutions require some form of persistent memory, meaning that at least some projections of the neural activity must be self-sustaining. A dynamical system can implement this behavior by forming regions in its state space where states are drawn toward steady states (Fig. 3a). These regions are called “basins of attraction” because any state trajectory that enters them would progress toward the steady state in the absence of noise or changes in external inputs and dynamics. By steady states, we mean regions within the basins that deterministic trajectories eventually converge to. More generally, these sets of states are called “attractors” because neural activity trajectories that have reached the basin progress toward attractor states and remain there—approximately, in the presence of intrinsic noise in neural activity or changes in external input—until sufficient noise or external input activity nudges the state to escape the attractor basin. In general, dynamical systems can produce attractors that have a complex and high-dimensional structure within the basin (e.g. manifold, fractal structure) and can exhibit their own internal dynamics, as in the case of chaotic attractors (also called “strange”). Other common attractors contain fewer points, such as stable periodic orbits, or stable fixed points—single state points that do not change in time. In this section, we will focus on fixed point attractors for simplicity, but arguments throughout the paper apply to the general case of attractor subspaces. The important aspect of attractors for our purposes is that they are distinct and have nonoverlapping basins of attraction.

Since trajectories that have converged to attractors have a tendency to remain there in the absence of strong external inputs, attractors can endow a dynamical system with a form of self-sustaining memory over short timescales that are useful for performing many computations essential to real-world tasks. Attractor dynamics can also be used for efficient long-term memory, without the brain having to directly store the high-dimensional

vectors of the attractors in state space. As we will explain in the “Attractors are mutually exclusive: contractive dynamics discretize the state” section, attractors are mutually exclusive and thus have a discrete structure; they can be identified with symbols (e.g. words) that label “which” attractor the system is in without describing the attractor’s location in state space. The system could thus store a concise symbol in long-term memory rather than a high-dimensional vector. Afterward, the memory could be retrieved by using the symbol as an input “key” that drives the state to any location in the basin of the attractor, at which point the dynamics of the system will cause the trajectory to converge to the attractor. For example, to memorize an image of a face (represented by a high-dimensional vector) and associate it with a discrete entity like the name of a person, a learning process could update the parameters of the dynamical system, so that the image vector is an attractor state and the system enters its basin of attraction when the name (or rather a neural code for it) is provided as an input.

It is important to emphasize that the existence of these attractors and the particular properties they have (e.g. cardinality, location, and shape) are purely functions of the internal dynamics of the system. Neural networks are therefore particularly well-suited for implementing diverse computations through dynamical systems since they are composed of simple units whose connectivity can be flexibly tuned to achieve many possible complex attractor configurations, with the capacity for universal function approximation in the limit of large networks (Schäfer and Zimmermann, 2007).

A dynamical system can be modulated by external inputs, and therefore, the nature of its attractors can also be driven by contextual signals. In the human brain, for example, this context could include both external sensory input and the content of short- and long-term memory. In particular, the previous content of working memory (which is a part of short-term memory) might have a strong influence, so that our thoughts form coherent sequences and we can alternate between mutually exclusive interpretations of the world that are compatible with the context (e.g. flipping between different interpretations of the Necker cube—an ambiguous 2D line drawing of a cube that can be in one of two possible 3D orientations).

As was summarized in review articles by Rolls (2010) and Khona and Fiete (2022), the framework of attractor dynamics has been used to mechanistically explain the neural computations underlying decision-making (Wang, 2002; Wong and Wang, 2006; Wang, 2008), long-term memory (Hopfield, 1982; Chaudhuri and Fiete, 2019; Ramsauer et al., 2021), working memory (Durstewitz et al., 2000; Curtis and D’Esposito, 2003; Deco and Rolls, 2003;

Barak and Tsodyks, 2014; Seeholzer et al., 2019), and the performance of simple cognitive tasks (Driscoll et al., 2022). Attractors have also been observed empirically across several experiments investigating decision-making (Kurt et al., 2008; Lin et al., 2014; Stevens, 2015) and working memory (Gnadt and Andersen, 1988; Constantinidis et al., 2001; Curtis and D'Esposito, 2003).

### Attractors are mutually exclusive: contractive dynamics discretize the state

An important property of attractors is that they are mutually exclusive: each attractor  $\mathbf{a}$  is associated with a basin of attraction  $B(\mathbf{a})$ , which is the region in state space such that any state  $\mathbf{x}$  in  $B(\mathbf{a})$  will necessarily converge through the dynamics into  $\mathbf{a}$ , in the absence of noise or external perturbations. This division into mutually exclusive basins of attraction thus creates a partition of the state space: one can associate with any state  $\mathbf{x}$  the attractor  $\mathbf{a}$  corresponding to the basin of attraction  $B(\mathbf{a})$  in which  $\mathbf{x}$  falls.

As a consequence of this mutual exclusivity, any attractor  $\mathbf{a}$  has a dual discrete and continuous nature (Jaeger, 1999): the symbol or composition of symbols  $i(\mathbf{a})$  that identify  $\mathbf{a}$  among all the other possible attractors in the current dynamics is discrete, while a fixed point  $\mathbf{a}$  is associated with a real-valued vector [also called embedding (Bengio et al., 2000; Roweis and Saul, 2000; Morin and Bengio, 2005) in the deep learning literature] corresponding to the state of the system at that fixed point. If the dynamics are not attractive over all dimensions, the same statement can be made for the subspace that is attractive, which means that this discretization effect need not cover every possible dimension and nondiscretized dimensions may represent values in a continuous space.

Note that introducing randomness in the dynamics makes it possible to sample one of the attractors that may be reachable from the current state when that noise is taken into consideration. For example, if the state  $\mathbf{x}$  is close to the boundary between basins of attraction of attractors  $A$  and  $B$ , a small amount of additive noise would suffice to stochastically sample one destination or the other, with probabilities that would vary depending on how far  $\mathbf{x}$  is from the boundary and the specific dynamics in its area (for instance, basin depth or slope).

An example of discrete attractor dynamics in the brain can be found in the auditory cortex. Bathellier et al. (2012) studied firing rate patterns in local neural populations in mice and found abrupt shifts between small (1–3) numbers of distinct response modes, with response mode identity across multiple local populations providing a discrete code that allowed sound prompts to be identified with 86.2% accuracy by a linear classifier (compared to 87.3% using nondiscretized activity). Abrupt transitions between neural steady states have also been observed in the rat hippocampus and the zebrafish olfactory bulb (Niessing and Friedrich, 2010; Wills et al., 2005).

### Emergent attractors in task-optimized networks

To demonstrate how attractors naturally emerge as solutions to cognitive tasks, we briefly summarize relevant results from Sussillo and Barak (2013), where an artificial recurrent neural network (RNN) was trained to solve a simple memory task. An RNN is a network of artificial neurons, which can be connected through recurrent feedback loops. Neurons can also form connections to special input and output units, which allow the network to interface with a task. The connection strength between each directed pair of neurons is parameterized using a scalar weight that modulates the

degree to which activity in the first neuron drives future activity in the second, and these weights are optimized in order to minimize error on the task. Like the brain, RNNs have recurrent connections between neurons that define a dynamical system optimized to perform some computation and are therefore useful models for studying emergent neural dynamics.

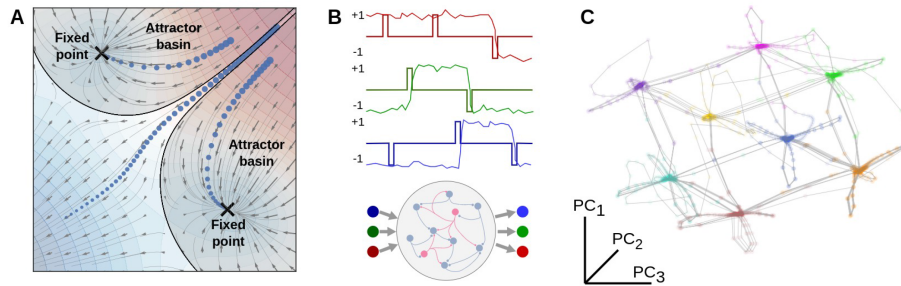
Sussillo and Barak (2013) train an RNN on the 3-bit flip-flop task (Fig. 3b), in which the network must learn to continuously output the sign (+1 or -1) of the last binary spike across three input channels (which we can call the “red,” “green,” and “blue” channels). For instance, following the input sequence [red=+1, green=-1, blue=+1], the correct output should be the vector {red=+1, green=-1, blue=+1}. If the next input spike was red=-1, the new output would change to {red=-1, green=-1, blue=+1}. Importantly, while each input spike only has a short duration, the network must continuously output the value of each channel's last spike, which imposes a memory demand.

When Sussillo and Barak (2013) inspected the learned dynamics of the RNN, they found that it solved the task through the use of fixed point attractors. Since the number of possible outputs is  $2^3 = 8$ , the model represented each of these using an attractor. Due to their stability, the model was then able to continuously read out from whichever attractor the trajectory had most recently converged to. Whenever a new spike appeared in one of the input channels (with a value different from that channel's previous spike), the state escaped the current basin of attraction and followed transient dynamics toward the attractor for the new output. This simple task demonstrated how attractor dynamics can naturally emerge in neural networks and implement non-trivial computations, such as those involving transitions between discrete memory states.

### Stability and robustness of conscious states

In addition to attractor dynamics models of working memory (“Motivating attractor dynamics as a model for conscious experience” section), the qualitative nature of self-reported conscious state also suggests a connection between attractor dynamics and consciousness. As a model of conscious processing, discrete attractor dynamics predict that our experience consists of a sequence of relatively stable states that transition swiftly from one to another. Such types of sequential dynamics have been hypothesized to be a key component of conscious thought and perception (James, 1892; Varela, 1999; Rabinovich et al., 2008; Tsuda, 2015). Empirically, one of the characteristics that distinguishes conscious vs. unconscious neural representations in psychophysics tasks is that they are significantly more stable in the “aware” condition (Schurger et al., 2015).

Qualitatively, subjects commonly report on the emergence of stable discrete “choices” within conscious perception. For instance, when looking at the Necker cube, subjects only perceive one single interpretation of its structure and orientation rather than a mixture of both possibilities. Occasionally, this interpretation will change to the alternative one, but the change will happen rapidly as an abrupt transition. Similarly, in the case of binocular rivalry, only a single image presented to one of the eyes will be consciously perceived rather than a mixture of the two, and which image is consciously perceived will abruptly change at random times. Such cases are characterizable by attractor dynamics that converge to one attractor and remain stable until sufficient input change or noise results in a rapid transition to another attractor.



**Figure A3.** Attractor dynamics in neural networks. (A) Attractors in a 2D state space. When a trajectory enters an attractor's basin, it begins to converge to the attractor and remains there until sufficient external input or intrinsic noise allows it to escape. (B) Sussillo and Barak (2013) train an artificial RNN to solve the 3-bit flip-flop memory task. In this task, the model must continuously output the sign of the most recent binary spike on three separate input channels. (C) Fixed point attractors emerge in the learned dynamics of a recurrent neural network (RNN) as a solution to the task. Each of the attractors corresponds to one of the  $8 (2^3)$  possible bit configurations, providing a stable memory state from which the output can be continuously read out. The plot shows a trajectory in the RNN's state space for changing inputs, where points along the trajectory are colored according to the correct output. The dimensionality of the state space was reduced using principal component analysis for visualization.

Input change or noise may also result in basin transitions that occur without complete convergence to attractors. This is familiar in the cases of thought and speech. One common example is thought-disruptive external stimuli, in which external stimuli distract or interrupt one's chain of thought. A less well-known but equally important example is the role of internal time-saving mechanisms. These are active in cases where one does not need to spell something out in full detail. For example, in speech production, phonemes are often not fully articulated: this may be understood by noting that once one has arrived at an attractor basin, it is disambiguated which point one converges toward (Roessig et al., 2019). A similar mechanism may explain the utility of verbal or symbolic thought, where the key may serve as synecdoche for the value.

Schurger et al. (2010) suggested that conscious states were associated with increased robustness to noise in psychophysics experiments. A signature of neural representations in the "conscious" condition was that they were highly reproducible; given the same stimulus presentation across different trials, patterns of neural activity were similar, as long as the subjects reported awareness of the stimulus. In contrast, patterns of activity during the "nonconscious" condition in which subjects were unaware of the stimulus exhibited greater variability. Both robustness to noise and reproducibility of states, in turn, are core properties accommodated by attractor dynamics.

### Appendix 3 Concrete Bounds for Neural Architectures

Adding more assumptions about the underlying architecture of the brain can establish concrete losses in richness (i.e. ineffability) as opposed to upper bounds, which was discussed in the "Information loss from attractor dynamics" section. In addition, learning parameters  $\phi$  of a computational model for neural processing from empirical Magnetic Resonance Imaging data (and verbal output) would allow  $H_\phi(X|A)$  (and  $H_\phi(A|M)$ ) to be computed and thus establish, in numerical terms, concrete and anatomically relevant values for ineffability, under the assumption that  $S = X$  or  $S = [A, \dots, A]$  ("Information loss from attractor dynamics" section).

### Appendix 4 Related Work

Several existing works argue that attractor dynamics have the right functional characteristics to serve as a computational model

for consciousness (Colagrosso and Mozer, 2004; Mozer, 2009; Mathis and Mozer, 1994; Mathis and Mozer, 2019; Grossberg, 1999; Rumelhart et al., 1986) but do not examine how information loss arising from such dynamics relate to the rich and ineffable aspects of conscious experience. Metzinger (2009) suggests that ineffability relates to nonidentifiability of conscious experience, but does not formalize this or make the connection to information theory.

In the context of Predictive Processing and the Free Energy Principle (Friston and Kiebel, 2009), the view offered in this paper broadly coheres with recent efforts at expressing consciousness-related phenomena, in information theoretic terms, as stemming from limitations in cognitive processing, such as limitations related to the ability to form accurate mental representation of sensory causes. Under the view of predictive processing and the free energy principle as applied to active inference (e.g. Friston et al. (2024)), cognitive systems able to access and communicate mental representations of the causes of their sensations do so by learning sparse prior beliefs. According to the view developed in Friston et al. (2024), sparse and therefore communicable representations have low entropy, which aligns with our notion of richness, and with the more general claim according to which effability comes at the cost of richness; the richer the mental state  $Z$  (i.e. the higher the entropy of the distribution over the random variable  $Z$ ), the higher the information loss ( $H(Z|Y)$ ).

### Appendix 5 Metacognitive Representation of Ineffability

In the "Existence and report of phenomenal experience" section, we discussed the attractor state encoding information about the basin of attraction as a mechanism supporting the report of ineffability. This information would correspond to a measure of entropy such as basin width across dimensions [the entropy of a uniformly distributed i.e. maximum entropy discrete variable scales with the size of the distribution's support, and the entropy of a multivariate Gaussian scales with covariance (Ahmed and Gokhale, 1989)]. There are multiple potential mechanisms that would allow the inclusion of such information in the attractor state and thus working memory. In the simple case, for a given attractor model, state could be extended with an extra dimension containing a width metric that is constant within each attractor basin. This has the property that true entropy within the basin of attraction does not change, as the value is constant within each basin. Alternatively, the attractor model for working memory could be composed of multiple attractor modules executing in parallel ("Hierarchical

attractor dynamics” section), where one module increments its state value until attractor dynamics in other modules converge, thereby providing a measure of basin size that is included in the composite attractor state for working memory and access consciousness. In general, such values would be determined by the parameters (weights) encoding neural state dynamics, which are typically learned from training data in computational and biological contexts. Note that information on information loss would be encoded by the parameters, since the parameters determine the basins of attraction, and the mechanisms described can be viewed as ways of enabling this information to be extracted from parameters into neural activation state and conscious awareness, given initial state or context.

Information loss or reductiveness in the thought process constitutes useful information for communication and task solving more generally, so one would expect it to be computed in cognitive processes. Nevertheless, we are not claiming that this theory on its own is sufficient for explaining how such information enters into consciousness, rather it offers a plausible account that may be further studied in the context of metacognition more generally (Yeung and Summerfield, 2012; Fleming, 2024). Empirically determining the presence of such mechanisms would require instantiating an attractor model, either learned from neuroimaging or artificially optimized for a task, and analyzing correlations between attractor state values and basin size.

## References

- Ahmed N A, Gokhale D. Entropy expressions and their estimators for multivariate distributions, *IEEE Transactions on Information Theory* 1989;**35**:688–92.
- Alemi AA, Fischer I, Dillon JV, Murphy K, Deep variational information bottleneck. In: *International Conference on Learning Representations*, Toulon, France, April 24–26, Vol. 5, 2017.
- Alter T, Walter S. *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford, United Kingdom: Oxford University Press, 2006.
- Baars B J. *A Cognitive Theory of consciousness*. Cambridge, United Kingdom: Cambridge University Press, 1993.
- Baars B J. Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. In: Laureys S, (ed.), *The Boundaries of Consciousness: Neurobiology and Neuropathology*. *Progress in Brain Research*, Vol. 150, Amsterdam, Netherlands: Elsevier, 2005, 45–53.
- Bahdanau D, Cho K, Bengio Y, Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations*, San Diego, CA, USA, May 7–9, Vol. 3, 2015.
- Barak O, Tsodyks M. Working models of working memory, *Current Opinion in neurobiology* 2014;**25**:20–4.
- Bathellier B, Ushakova L, Rumpel S. Discrete neocortical dynamics predict behavioral categorization of sounds, *Neuron* 2012;**76**:435–49.
- Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model. In: *Advances in Neural Processing Systems*, Denver, CO, USA, December, Vol. 13, 2000, 932–8.
- Block N. On a confusion about a function of consciousness, *Behavioral and Brain Sciences* 1995;**18**:227–47.
- Block N. Overflow, access, and attention, *Behavioral and Brain Sciences* 2007;**30**:530–48.
- Bronfman Z Z, Jacobson H, Usher M. Impoverished or rich consciousness outside attentional focus: Recent data tip the balance for overflow, *Mind & Language* 2019;**34**:423–44.
- Burak Y. Spatial coding and attractor dynamics of grid cells in the entorhinal cortex, *Current Opinion in neurobiology* 2014;**25**:169–75.
- Bushnell M C, Goldberg M E, Robinson D L. Behavioral enhancement of visual responses in monkey cerebral cortex. i. Modulation in posterior parietal cortex related to selective visual attention, *Journal of Neurophysiology* 1981;**46**:755–72.
- Chalmers D J. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford, United Kingdom: Oxford University Press, 1996.
- Chalmers D J. *The Character of consciousness*. Oxford, United Kingdom: Oxford University Press, 2010.
- Changeux J.-P., Dehaene S. Neuronal models of cognitive functions, *Cognition* 1989;**33**:63–109.
- Chaudhuri R, Fiete I. Bipartite expander hopfield networks as self-decoding high-capacity error correcting codes. In: *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, December 8–14, Vol. 32, 2019, 7686–97.
- Chelazzi L. Serial attention mechanisms in visual search: a critical look at the evidence, *Psychological research* 1999;**62**:195–219.
- Chorowski J K, Bahdanau D, Serdyuk D, Cho K, Bengio Y. Attention-based models for speech recognition. In: *Advances in Neural Information Processing Systems*, Montreal, QC, Canada, December 7–12, Vol. 28, 2015, 577–85.
- Chuard P. The riches of experience, *Journal of Consciousness Studies* 2007;**14**:20–42.
- Churchland M M, Cunningham J P, Kaufman M T, Foster J D, Nuyujukian P, Ryu S.I., Shenoy K V. Neural population dynamics during reaching, *Nature* 2012;**487**:51–6.
- Cohen M A, Dennett D C. Consciousness cannot be separated from function, *Trends in Cognitive sciences* 2011;**15**:358–64.
- Cohen M A, Dennett D C, Kanwisher N. What is the bandwidth of perceptual experience? *Trends in Cognitive Sciences* 2016;**20**:324–35.
- Colagrosso M, Mozer M C. Theories of access consciousness. In: *Advances in Neural Information Processing Systems*, Vancouver, BC, Canada, December 13–18, Vol. 17, 2004;289–96.
- Conee E. Physicalism and phenomenal properties, *Philosophical Quarterly* 1985;**35**:296–302.
- Constantinidis C, Franowicz M N, Goldman-Rakic P S. Coding specificity in cortical microcircuits: a multiple-electrode analysis of primate prefrontal cortex, *Journal of Neuroscience* 2001;**21**:3646–55.
- Cowan N. What are the differences between long-term, short-term, and working memory?, *Progress in Brain research* 2008;**169**:323–38.
- Curtis C E, D’Esposito M. Persistent activity in the prefrontal cortex during working memory, *Trends in Cognitive sciences* 2003;**7**:415–23.
- Deco G, Rolls E T. Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex, *European Journal of Neuroscience* 2003;**18**:2374–90.
- Dehaene S, Kerszberg M, Changeux J.-P. A neuronal model of a global workspace in effortful cognitive tasks, *Proceedings of the National Academy of Sciences* 1998;**95**:14529–34.
- Dehaene S, Naccache L. Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework, *Cognition* 2001;**79**:1–37.
- Dennett D C. *Consciousness explained*. London, England: Penguin UK, 1993.
- Dennett D. What robomary knows. In: Alter T and Walter S, (eds.), *Phenomenal Concepts and Phenomenal Knowledge: New Essays on Consciousness and Physicalism*. Oxford, United Kingdom: Oxford University Press, 2006.
- Descartes René. *Discourse on method*. New York, NY: Macmillan; London: Collier Macmillan (1986). Originally published: Indianapolis: Bobbs-Merrill(1950).; Translation of Discours de la méthode.; Bibliography, 1986, p. xxii.



- Devlin J, Chang M.-W., Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. In: *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN, USA, June 2–7, Vol. 1, 2019, 4171–4186.
- Driscoll L, Shenoy K, Sussillo D. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs, *bioRxiv* 2022.
- Dumas G, Chavez M, Nadel J, Martinerie J. Anatomical connectivity influences both intra- and inter-brain synchronizations, *PLOS ONE* 2012;**7**:e36414.
- Durstewitz D, Seamans J K, Sejnowski T J. Neurocomputational models of working memory, *Nature neuroscience* 2000;**3**:1184–91.
- Engle R W. Working memory capacity as executive attention, *Current Directions in Psychological Science* 2002;**11**:19–23.
- Favela L H. The dynamical renaissance in neuroscience, *Synthese* 2021;**199**:2103–27.
- Fleming S M. Metacognition and confidence: a review and synthesis, *Annual Review of Psychology* 2024;**75**:241–68.
- Frankish K. Illusionism as a theory of consciousness, *Journal of Consciousness Studies* 2016;**23**:11–39.
- Friston K, Kiebel S. Predictive coding under the free-energy principle, *Philosophical Transactions of the Royal Society B: Biological Sciences* 2009;**364**:1211–21.
- Friston K J, Ramstead M J, Kiefer A B et al. Designing ecosystems of intelligence from first principles *Collective Intelligence* 2024;**3**:1.
- Gnadt J W, Andersen R A. Memory related motor planning activity in posterior parietal cortex of macaque, *Experimental Brain Research* 1988;**70**:216–20.
- Godfrey-Smith P. Mind, matter, and metabolism, *Journal of Philosophy* 2016;**113**:481–506.
- Goldstein P, Weissman-Fogel I, Dumas G, Shamay-Tsoory S G. Brain-to-brain coupling during handholding is associated with pain reduction, *Proceedings of the National Academy of Sciences* 2018;**115**:E2528–37.
- Goyal A, Bengio Y. Inductive biases for deep learning of higher-level cognition, *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 2022;**478**:20210068.
- Graziano M S, Guterstam A, Bio B J, Wilterson A.I. Toward a standard model of consciousness: reconciling the attention schema, global workspace, higher-order thought, and illusionist theories, *Cognitive Neuropsychology* 2020;**37**:155–72.
- Graziano M S, Kastner S. Human consciousness and its relationship to social neuroscience: a novel hypothesis, *Cognitive Neuroscience* 2011;**2**:98–113.
- Graziano M S, Webb T W. The attention schema theory: a mechanistic account of subjective awareness, *Frontiers in Psychology* 2015;**6**:500.
- Grossberg S. The link between brain learning, attention, and consciousness, *Consciousness and Cognition* 1999;**8**:1–44.
- Grünwald P D, Vitényi P. Kolmogorov complexity and information theory. with an interpretation in terms of questions and answers, *Journal of Logic, Language and Information* 2003;**12**:497–529.
- Grünwald P, Vitényi P. M. B., *Shannon information and kolmogorov complexity*, CoRR, 2004, cs.IT/0410002.
- Herculano-Houzel S. The human brain in numbers: a linearly scaled-up primate brain, *Frontiers in Human Neuroscience* 2009;**3**.
- Hopfield J J. Neural networks and physical systems with emergent collective computational abilities, *Proceedings of the National Academy of Sciences* 1982;**79**:2554–8.
- Humphrey N. The invention of consciousness, *Topoi* 2020;**39**:13–21.
- Jackson F. What Mary didn't know, *The Journal of Philosophy* 1986;**83**:291–5.
- Jaeger H. From continuous dynamics to symbols. In: Tschacher W and Dauwalder J-P (eds.), *Dynamics, synergetics, autonomous agents: Nonlinear systems approaches to cognitive psychology and cognitive science*. Singapore: World Scientific, 1999, 29–48.
- James W. The stream of consciousness. In: James W (ed.), *Psychology*. Cambridge, Massachusetts, United States: MIT Press, 1892, 71–82.
- Kawaguchi K, Deng Z, Ji X, Huang J. How does information bottleneck help deep learning?. In: *International Conference on Machine Learning*, Honolulu, HI, USA, July 23–29, Vol. 202, 2023, 16049–96.
- Kelly Y T, Webb T W, Meier J D, Arcaro M J, Graziano M S. Attributing awareness to oneself and to others, *Proceedings of the National Academy of Sciences* 2014;**111**:5012–17.
- Khan S, Naseer M, Hayat M, Zamir S W, Khan F S, Shah M. Transformers in vision: a survey, *ACM Computing Surveys (CSUR)* 2022;**54**:1–41.
- Khona M, Fiete I R. Attractor and integrator networks in the brain, *Nature Reviews Neuroscience* 2022;1–23.
- Kirby S, Tamariz M, Cornish H, Smith K. Compression and communication in the cultural evolution of linguistic structure, *Cognition* 2015;**141**:87–102.
- Kleiner J. Mathematical models of consciousness, *Entropy* 2020;**22**:609.
- Kleiner J, Ludwig T. What is a mathematical structure of conscious experience?, 2023; preprint, arXiv:2301.11812.
- Kolmogorov A N. Three approaches to the quantitative definition of information', *Problems of Information Transmission* 1965;**1**:1–7.
- Kurt S, Deutscher A, Crook J M, Ohl F W, Budinger E, Moeller C K, Scheich H, Schulze H. Auditory cortical contrast enhancing by global winner-take-all inhibitory interactions, *PLOS ONE* 2008;**3**:e1735.
- Lamme V A. Sue ned block!: Making a better case for p-consciousness, *Behavioral and Brain Sciences* 2007;**30**:511–12.
- Levine J. On Leaving Out What It's Like. In: Davies M and Humphreys G W (eds.), *Consciousness: Psychological and Philosophical Essays*. Cambridge, MA, United States: MIT Press, 1993, 543–57.
- Lewis D K (1990). What experience teaches. In: Lycan W G (ed.), *Mind and Cognition*. Hoboken, NJ, United States: Blackwell, 29–57.
- Lin A C, Bygrave A M, De Calignon A, Lee T, Miesenböck G. Sparse, decorrelated odor coding in the mushroom body enhances learned odor discrimination, *Nature neuroscience* 2014;**17**:559–68.
- Li M, Vitényi P et al. *An Introduction to Kolmogorov Complexity and its applications*. Vol. 3, Berlin, Germany: Springer, 2008.
- Loar B. Phenomenal states, *Philosophical Perspectives* 1990;**4**:81–108.
- Mathis D, Mozer M C. On the computational utility of consciousness. In: *Advances in Neural Information Processing Systems*, Denver, CO, USA, Vol. 7, 1994, 11–18.
- Mathis D W, Mozer M C. Conscious and unconscious perception: A computational theory, In: *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society*. Routledge, 2019, 324–8.
- Merriam Webster Dictionary. (2023). Rich. <https://www.merriam-webster.com/dictionary/rich> (30 January 2023, date last accessed).
- Metzinger T. *The Ego Tunnel: The Science of The Mind and The Myth of The Self*. New York City, NY, United States: Basic Books (AZ), 2009.
- Michaels J A, Dann B, Scherberger H. Neural population dynamics during reaching are better explained by a dynamical system than representational tuning, *PLOS Computational Biology* 2016;**12**:e1005175.
- Miller E, Buschman T. Working memory capacity: limits on the bandwidth of cognition, *Daedalus* 2015;**144**:112–22.
- Miller E K, Li L, Desimone R. Activity of neurons in anterior inferior temporal cortex during a short-term memory task, *Journal of Neuroscience* 1993;**13**:1460–78.

- Morin F and Bengio Y (2005). Hierarchical probabilistic neural network language model. In: *International workshop on artificial intelligence and statistics*, Westminster, United Kingdom: PMLR, 246–52.
- Mozer M C. Attractor networks, *Oxford Companion to Consciousness* 2009; **1**:88–9.
- Naccache L. Why and how access consciousness can account for phenomenal consciousness, *Philosophical Transactions of the Royal Society B: Biological Sciences* 2018; **373**:20170357.
- Nagel T. What is it like to be a bat? *The Philosophical review* 1974; **83**:435–50.
- Niessing J, Friedrich R W. Olfactory pattern classification by discrete neuronal network states, *Nature* 2010; **465**:47–52.
- Oxford English Dictionary. ineffable. <https://www.oed.com/view/Entry/94904?redirectedFrom=ineffable#eid>. (30 January 2023, date last accessed).
- Phillips I. No watershed for overflow: recent work on the richness of consciousness, *Philosophical Psychology* 2016; **29**:236–49.
- Premack D, Woodruff G. Does the chimpanzee have a theory of mind?, *Behavioral and Brain sciences* 1978; **1**:515–26.
- Prinz J. *The Conscious Brain: How Attention Engenders Experience*. USA: OUP, 2012.
- Rabin G. Conceptual mastery and the knowledge argument, *Philosophical Studies* 2011; **154**:125–47.
- Rabinovich M.I., Huerta R, Varona P, Afraimovich V S. Transient cognitive dynamics, metastability, and decision making. *PLOS Computational Biology* 2008; **4**:e1000072.
- Ramsauer H, Schäfl B, Lehner J et al. Hopfield networks is all you need. In: *International Conference on Learning Representations*, Virtual Event, Austria, May 3–7, Vol. 9, 2021.
- Redish A D, Elga A N, Touretzky D S. A coupled attractor model of the rodent head direction system, *Network: Computation in Neural systems* 1996; **7**:671–85.
- Renart A, Parga N, Rolls E. A recurrent model of the interaction between prefrontal and inferotemporal cortex in delay tasks. In: *Advances in Neural Information Processing Systems*, Denver, CO, USA, November 29 – December 4, Vol. 12, 1999, 171–7.
- Roessig S, Mücke D, Grice M. The dynamics of intonation: Categorical and continuous variation in an attractor-based model, *PLOS ONE* 2019; **14**:5.
- Rolls E T. An attractor network in the hippocampus: theory and neurophysiology, *Learning & Memory* 2007a; **14**:714–31.
- Rolls E T. *Memory, Attention, and Decision-Making: A Unifying Computational Neuroscience Approach*. Oxford, United Kingdom: Oxford University Press, 2007b.
- Rolls E T. Attractor networks, *Wiley Interdisciplinary Reviews: Cognitive Science* 2010; **1**:119–34.
- Roweis S T, Saul L K. Nonlinear dimensionality reduction by locally linear embedding, *science* 2000; **290**:2323–6.
- Rumelhart D E, Smolensky P, McClelland J L, Hinton G. Sequential thought processes in pdp models, *Parallel Distributed processing: Explorations in the Microstructures of cognition* 1986; **2**:3–57.
- Schäfer A M, Zimmermann H -G. Recurrent neural networks are universal approximators, *International Journal of Neural systems* 2007; **17**:253–63.
- Schurger A, Pereira F, Treisman A, Cohen J D. Reproducibility distinguishes conscious from nonconscious neural representations, *Science* 2010; **327**:97–9.
- Schurger A, Sarigiannidis I, Naccache L, Sitt J D, Dehaene S. Cortical activity is more stable when sensory stimuli are consciously perceived, *Proceedings of the National Academy of Sciences* 2015; **112**:E2083–92.
- Seeholzer A, Deger M, Gerstner W. Stability of working memory in continuous attractor networks under the control of short-term plasticity, *PLOS Computational biology* 2019; **15**:e1006928.
- Shannon C E. A mathematical theory of communication, *The Bell System Technical journal* 1948; **27**:379–423.
- Shenoy K V, Sahani M, Churchland M M et al. Cortical control of arm movements: a dynamical systems perspective, *Annu Rev Neurosci* 2013; **36**:337–59.
- Smart J. J. C. (2022). The Mind/Brain Identity Theory. In: Zalta E N and Nodelman U, (ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA, United States: Metaphysics Research Lab, Stanford University, Winter 2022 edn.
- Sperling G. The information available in brief visual presentations, *Psychological monographs: General and applied* 1960; **74**:1–29.
- Stevens C F. What the fly's nose tells the fly's brain, *Proceedings of the National Academy of Sciences* 2015; **112**:9460–5.
- Sussillo D, Barak O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks, *Neural computation* 2013; **25**:626–49.
- Tishby N, Pereira F C, Bialek W, *The information bottleneck method*, 2000, preprint, arXiv:physics/0004057.
- Tononi G, Edelman G M. Consciousness and complexity, *Science* 1998; **282**:1846–51.
- Tsuda I. Chaotic itinerancy and its roles in cognitive neurodynamics, *Current Opinion in Neurobiology* 2015; **31**:67–71.
- Tye M (2006). Nonconceptual content, richness, and fineness of grain. In: Gendler T S and Hawthorne J, (eds.), *Perceptual Experience*. Oxford, United Kingdom: Oxford University Press, 504–30.
- Vandenbroucke A R, Sligte I G, Fahrenfort J J, Ambroziak K B, Lamme V A. Non-attended representations are perceptual rather than unconscious in nature, *PLOS ONE* 2012; **7**:e50042.
- Varela F (1999). The specious present: A neurophenomenology of time consciousness. In: Petitot J, Varela F J, Pacoud B, and Roy J, (eds.), *Naturalizing Phenomenology*. Stanford, CA, United States: Stanford University Press, 266–314.
- Wang X.-J. Probabilistic decision making by slow reverberation in cortical circuits, *Neuron* 2002; **36**:955–68.
- Wang X.-J. Decision making in recurrent neuronal circuits, *Neuron* 2008; **60**:215–34.
- Ward E J. Downgraded phenomenology: how conscious overflow lost its richness, *Philosophical Transactions of the Royal Society B: Biological Sciences* 2018; **373**:20170355.
- Wills T J, Lever C, Cacucci F, Burgess N, O'Keefe J. Attractor dynamics in the hippocampal representation of the local environment, *Science* 2005; **308**:873–6.
- Wong K.-F., Wang X.-J.. A recurrent network mechanism of time integration in perceptual decisions, *Journal of Neuroscience* 2006; **26**:1314–28.
- Yeung N, Summerfield C. Metacognition in human decision-making: confidence and error monitoring, *Philosophical Transactions of the Royal Society B: Biological Sciences* 2012; **367**:1310–21.
- Zhang K. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory, *Journal of Neuroscience* 1996; **16**:2112–26.

---

*Neuroscience of Consciousness*, 2024, **2024(1)**, 1–18

DOI: <https://doi.org/10.1093/nc/niae001>

**Research article**

Received 1 May 2023; revised 3 January 2024; accepted 23 January 2024

© The Author(s) 2024. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)