

# Approximating the Mathematical Structure of Psychodynamics

Bryce Allen Bagley<sup>1,2,3</sup> and Navin Khoshnani<sup>1,2</sup>

<sup>1</sup>Mathematical Medicine Group, Department of Neurosurgery, Stanford University, Stanford, CA, USA

<sup>2</sup>Petritsch Lab, Department of Neurosurgery, Stanford University

<sup>3</sup>Physician-Scientist Training Program, Stanford University

October 31, 2025

## Abstract

The complexity of human cognition has meant that psychology makes more use of theory and conceptual models than perhaps any other biomedical field. To enable precise quantitative study of the full breadth of phenomena in psychological and psychiatric medicine as well as cognitive aspects of AI safety, there is a need for a mathematical formulation which is both mathematically precise and equally accessible to experts from numerous fields. In this paper we formalize human psychodynamics via the diagrammatic framework of process theory, describe its key properties, and explain the links between a diagrammatic representation and central concepts in analysis of cognitive processes in contexts such as psychotherapy, neurotechnology, AI alignment, AI agent representation of individuals in autonomous negotiations, developing human-like AI systems, and other aspects of AI safety safety.

## 1 Introduction

Across the breadth of psychological and psychiatric conditions where therapy has been shown to be an effective intervention, modalities within the category of psychodynamic psychotherapy (PDP) have proven broadly applicable and effective. However, a key factor adding difficulty to research on PDP is the difficulty of quantifying what PDP is *doing* in a process sense. Forms of therapy such as cognitive behavioral therapy (CBT) can have discrete, comparatively easily quantifiable units, such a number of exposures to a distressing stimulus, but PDP in some sense involves a great deal more information. A range of psychometrics have been applied to successful studies on PDP, and the insights produced are of great value. Yet endpoint and outcomes data are limited in that they describe a snapshot of mental states. The ability to quantify the dynamics underlying PDP could open the door to a range of new directions in research on optimizing and modifying PDP.

Many of the notions from PDP have the potential for substantial relevance to research questions related to human cognition in other, emerging fields. Cognition represents a critical attack surface for AI-powered attacks on neurotechnology,[1, 2, 3, 4, 5] and also one of the attack surfaces which current methods are least equipped to protect.[6] Likewise, in other realms of AI safety there is concern for the abilities of LLMs and other models to manipulate human cognition,[7, 8, 9, 10] and analysis and quantification of static states represents only part of the picture of human cognition. Recent work in this second area has provided timely insights into risks associated with human cognition as an attack surface for AI,[7, 8, 9, 10] but been largely heuristic in nature, and when modeling human cognition quantitatively relied on classical probability, which is well-established to yield incorrect results.

In recent decades an extensive body of literature in cognitive science and mathematical psychology has demonstrated a superficially particularly odd feature of human cognition. When one attempts to explain many features of cognition using classical probability, paradoxical outcomes are found in experimental studies. Yet by switching from classical probability to quantum probability, these paradoxes resolve in an elegant and empirically validated manner.[11, 12, 13, 14, 15, 16, 17, 18, 19, 20] It cannot be overly emphasized that quantum cognition has absolutely no connection whatsoever with unsupported hypotheses or pseudoscientific proposals asserting consciousness to be a quantum phenomenon. The sole link between quantum cognition and quantum mechanics is purely mathematical, in that they both use an alternative form of probability and statistics which is based on complex projective geometry (a different area of mathematics than the standard form of probability).[21]

In a recent paper, Bagley & Petritsch described how to take existing modeling approaches from the quantum cognition literature and extend it to the more general, high-dimensional case.[6] They termed this *holographic cognition* on account of its connections with hyperdimensional computing (centered on what are termed *holographic representations* in the hyperdimensional computing literature), area of computing that also centers on high-dimensional data.[22, 23] The structure of holographic cognition is very well-suited to offering a means of quantifying PDP-and psychodynamics more broadly—at a level of complexity able to capture rich psychological features. However, the presentation of holographic cognition in that paper is in a form typical for papers on mathematics and its applications. We recognized that this approach would greatly impair collaboration between researchers who would take a more mathematical approach and those from less quantitative or non-quantitative backgrounds, in addition to essentially crippling much of its potential for broader utility.

What is needed is a new way of describing concepts in holographic cognition and standard quantum cognition. This means of description requires two core properties. First, it must be capable of expressing mathematical statements in a precise and rigorous manner, so that those studying the approach from a quantitative perspective can guarantee their results are likewise precise and rigorous. Second, it must be presented in such a way that it is *equally* approachable for experts from many fields—and ideally even reasonably approachable for laypeople. If achieved effectively, the combination of these two features would provide a means for scientists, mathematicians, policy experts, and more to apply ideas and results from holographic cognition while all "speaking the same language", in some sense.

To accomplish this task we make use of a special case of the mathematical field of category theory.[24] Specifically process theory, an approach to the study of systems via the framework of a Symmetric Monoidal Category (SMC). There already exists a reformulation of quantum mechanics in terms of process theory, which conveniently provides some of the mathematical apparatus required for our objective.[25, 26, 27, 28, 29, 30, 31, 32, 33] A great many mathematicians and physicists have contributed to the development of that process theory.[26, 27, 28, 29, 30, 31, 32, 33, 34, 35] and to demonstrating that it is mathematically rigorous.[36, 37, 38, 39, 40] The results of their efforts are invaluable here. However, while their work provides those elements of a process theory required for the use of quantum probability, it is necessary to construct a new process theory describing psychodynamics due to key differences between cognition and quantum mechanics. While both operate based on complex projective probability, many phenomena are not shared. Development of this psychodynamics process theory is the central task of this paper.

## 2 Methods

### 2.1 Necessary Properties of a Process theory

Process theories shift the focus of mathematical models away from the states occupied by systems at varying times, instead emphasizing the processes which act to evolve states over time. In this way, they are quite naturally suited to how one reasons clinically about human psychology and psychopathology. There are three features required of any process theory, as summarized by Coeke and Kissinger.[24] In summary, processes act on types, and one can link together networks of processes so long as the correct types serve as inputs and outputs to those processes.

1. A collection T of *types* represented by wires.
2. A collection of P of *processes* represented by boxes, where for each process the input and output types are from the collection T.
3. A means of linking processes, which is an operation that interprets a diagram of wires and processes in the collection P such that the entire system represents a process in P.

A crucial consequence of this is that for a diagram comprised of collections of processes, that diagram as a whole will always be a process itself. In mathematical terms, it is closed under compositions of processes. In clinical terms, a sequence or grouping of psychodynamic and therapeutic processes for a given individual are collectively a psychodynamic process for that same individual.

### 2.2 Constructing a Process Theory Description of Psychodynamics

Concepts from quantum cognition translate quite directly to the central ideas of psychodynamics and psychotherapy. *Cogits*, proposed by Bagley and Petritsch as a term for the elementary units of a mental state, serve as a shorthand for the variety of things which can be represented in quantum cognition. Cogits can represent any of preferences, beliefs, memories, subconscious components of a mental state, and more. Complex instances of such facets of cognition can be represented by collections of cogits, for example by using some number n cogits to describe a more nuanced or complex belief. Because all can be described with the same mathematics, a common name frequently helps clean up discussion of the ideas. Taking these ideas and the requirements for a process theory as delineated by Coeke and Kissinger,[24] we develop what follows.

There are at least two ways one could define the elements of the set T. Perhaps the most immediately intuitive option is to have T include vectors of cogits,<sup>1</sup> vectors of stimuli coming from the outside world, and vectors describing the various outputs of cognition such as speech, movement, other actions, and so on. However, we argue that when conceptualized as processes which alter a person's mental state, the symmetry between internal dynamics of cognition and dynamic changes when exposed to external stimuli is such that stimuli are better represented as processes in the set P within a psychodynamics process theory.

As such, we instead use the following elements for a psychodynamics process theory.

- The types in T are:
  - Cogit state vectors representing relevant information about an individual's mental state at some time t. One need not describe the entirety of a person's cognition, and indeed this would be completely infeasible, so the size and content of these vectors is limited to what is sufficiently relevant to a given case. In cases where one needs to distinguish between cogit vectors, these vectors will be denoted by c with subscripts.<sup>2</sup>
  - Outputs of cognition, such as speech, actions, and so on. These will be denoted as a with subscripts in the same fashion as cogit vectors.

<sup>1</sup>For clinical or other non-quantitative readers: A vector is an ordered list of values. For example, if one listed out their favorites in each of various categories of foods, you would describe this information using a vector of some number of cogits.

<sup>2</sup>For clinical readers: When there is a set of the same sort of mathematical object, the convention is to distinguish them with subscripts like  $x_1, x_2, \dots, x_{19}$ , and so on. When you wish to distinguish between multiple items without specifying an exact item of that type, you would instead use a variable for the subscript, such as  $C_i, C_j, C_k$  to denote three separate but yet-unspecified cogit state vectors.

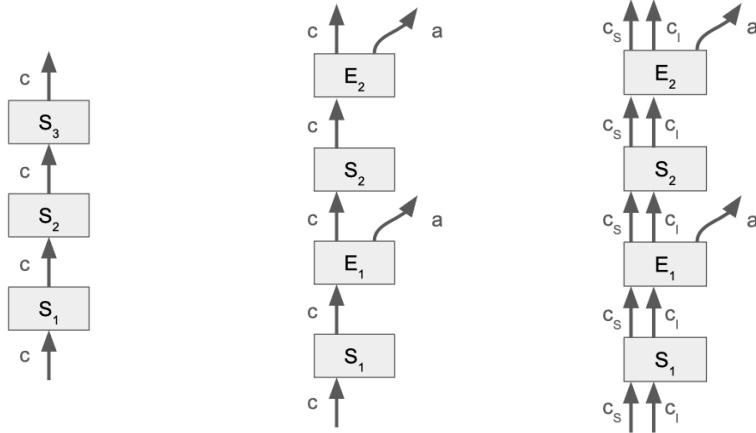
- The processes in P fall into a few key categories:

- Internal measurement processes, where one is consciously aware of their own thoughts—i.e. taking internal measurements of some limited portion of their cogit state vector—and the changes occurring in their mental state. These will be denoted by K, sometimes with an additional subscript.
- External measurement processes not related to communication from another entity, where one's cognition is altered by some external stimuli in a fashion which may or may not generate a response a. These processes will be denoted by E, with or without subscripts.
- External measurement processes related to communication from another entity, which prompt a response. We denote these as O.
- Cognitive dynamics separate from K. These will be denoted by S, with or without subscripts.

In many cases there will be overlap between these types of processes or ambiguity in how to represent them, so the use of notation is flexible as we will see in the below diagrams.

We now turn to showing the graphical notation for these same ideas. Let us begin with perhaps the simplest case, with the evolution of a mental state over time via subconscious processes. The diagram on the left below shows a cogit state vector  $c$  evolving over time via subconscious processes  $S_t$  at times  $t = 1, 2, 3$ .

The scale of time over which these processes take place can be chosen on the basis of the use case. In tests of instinctive reactions, they might be separated by seconds. In the context of psychotherapy looking at how a person's behavior, nonverbal communication, and such progress over time, they could instead represent time spans of weeks or even months. Of note, the process *itself* occurs over some time span. So  $S_1$  could represent the impact of subconscious processes over seconds, weeks, months, or whatever time scale is most relevant. This idea is perhaps clarified by representing stimuli as an external measurement process, as in the middle diagram.

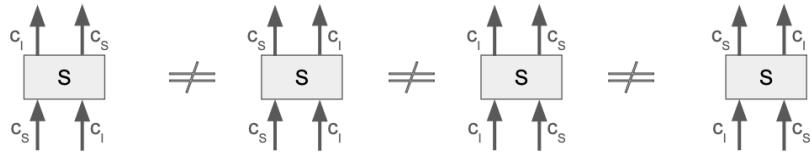


One has the option to use subscripts for further clarification, such as  $c_0, c_1, c_2$ , and  $c_3$  to denote the initial state and those resulting from each of the subconscious processes. Likewise,  $a_1, a_2$ , and  $a_3$  for the outputs from the external measurement operators. Concretely, the patient's words and behavior in response to three stimuli. These stimuli could represent part of an experiment in cognitive science, the effects of entire therapy sessions, or even a sequence of individual statements by a therapist during a single session. As previously mentioned, the timescale can be adjusted according to whatever is more useful for a given application.

Suppose we wanted to distinguish between the portion of an individuals cogs representing conscious aspects of their mind and the portion representing unconscious aspects. We would then use parallel wires, and we selected  $c_s$  and  $c_i$  to respectively denote the subconscious portion of their mental state and the portion of which a person has conscious internal awareness. This can be seen in the diagram on the right above. However it is not so simple to delineate between the conscious and subconscious portions of a mental state, as the very same information could have a conscious effect on actions and thoughts in one setting and a subconscious impact in another. Because of this, we do not separate the two except for the purpose of illustrating the above idea. In the formulation of quantum probability via a SMC, this separation would require the conscious and subconscious cogs be *bipartite*, meaning that there is no entanglement between the two subsets. It is a basic fact of psychology that this is incorrect, which is why we do not distinguish between the two and can in fact abandon the use of the  $c$  label for cogit vectors in general. Likewise, we will later see that we can equally drop the  $a$  notation for outputs.

One could add subscripts indicating time-points for each of the  $c$  and  $a$  wires, but this is not needed. One of the helpful facets of these diagrammatic representations is that the distinction between different  $c$  and  $a$  (as well as different  $S$ ,  $E$ , and  $I$ ) is made clear through the diagram itself. By having the lower and upper  $a$  be separate wires, we are inherently stating them to be different things, albeit of the same type. Likewise, we could remove the subscripts from the  $S$  and  $E$  processes seen here without losing any of what the diagram conveys.

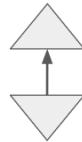
However, it should be noted that the mathematical properties of these diagrams are such that the horizontal and/or vertical ordering of wires' entries and exits from processes are not interchangeable. This property is displayed in the examples below.



While thus far we have used arrows to indicate the direction of time, we have the option of removing them and simply using the standard formation of time progressing upwards along a diagram. For the sake of clarity, however, we will continue to use arrows for the time being.

Finally, we describe the diagram features used to denote probability. In the process theory for quantum probability, it is standard to use triangles to denote states, with downward-pointing triangles reflecting the cogit vector(s) state(s) at the start of some process being considered, and upward-facing triangles reflecting probabilities of measuring various options from the set of possible states.

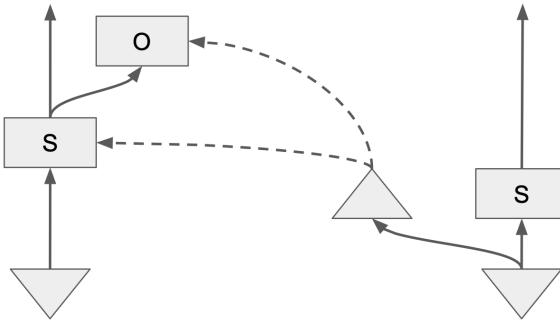
For example, suppose a case where we simply wish to describe the probability of measuring a cogit vectors as being in some specific state, absent the influence of whatever means of measurement was used as a readout. This is of course not possible in reality, and is shown here for the sake of simplicity.



### 2.3 Multiparty Processes

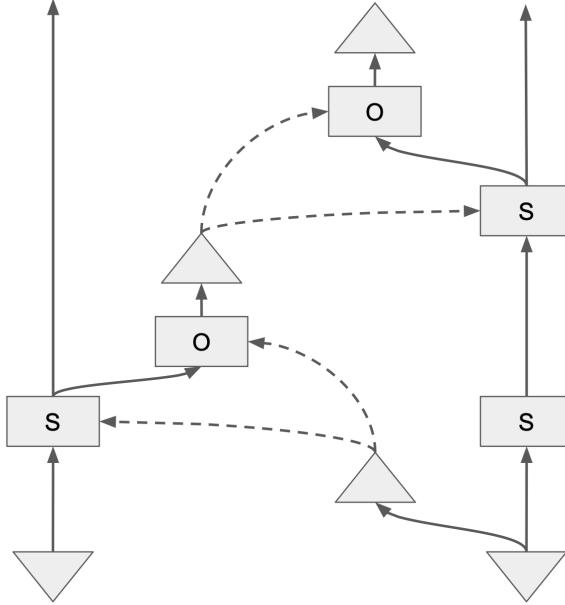
We now turn to the case which is more interesting for our purposes: that of diagrams representing interactions between individuals and/or entities of other sorts.

Take, for example, the following diagram:



At the start, each party's initial state is respectively represented by the pair inverted triangles. Both have some generic processes  $S$  by which their cogit vectors evolve, and for all  $S$  in this diagram we denote generic processes rather than specific and identical ones. Some output determined probabilistically by the initial state of the right-side party (RSP) initial state produces an effect on the left-side party (LSP). While the right-side party's cogit vector continues to evolve on its own, the probabilistic output of their initial state then applies two operators to the LSP's cogit vector. First, some operator by which the output from the RSP influences the evolution of the LSP's cogit vector. Second, some operator  $O$  which reflects a measurement. The dotted lines reflect the causal links between the measured output values—or more generally the distribution over possible measured outputs—and the  $\{S, O\}$  pair which any given output applies to the other party's cogit vector. The fact the outputs reflect a distribution over possible cogit vector states means that the paired  $S$  and  $O$  induced by an output are themselves operator-valued distributions (OVDs), which will be discussed at greater length later.

In the following diagram, we will see a reciprocal process induced by these output  $\{S, O\}$  pairs. To help visually distinguish between the evolution of states over time vs the application of operators due to outputs, we use dashed lines to link the latter.



Here we see a process of repeated, reciprocal interactions. The stimulus from the prior diagram now produces some output from the LSP which then applies an  $\{O, S\}$  pair to the RSP, which respectively both evolves the RSP's cogit vector and produces an output which again applies some general  $\{S, O\}$  pair to the LSP's cogit vector.

All  $S$  and  $O$  in this diagram, and indeed all processes in all diagrams within the psychodynamics process theory, are in fact operator-valued distributions (OVDs). While a full analysis of the algorithmic problem of inferring properties of these OVDs is beyond the scope of this paper, we intend to give such an analysis in a follow-up paper. In brief, algorithms from the field of quantum control theory can be applied. While traditionally they have been used to design optimal operators for accomplishing some desired transform to a quantum state, they can just as easily be used to take a specific state transformation and infer the structure of the operator which accomplished the transformation. By applying this over some set of sufficiently similar experiments or natural observations, one can approximate the OVD(s) corresponding to a given process.

It is important to note that one can define cogs in terms of the observed outputs in response to various  $O$  operators. Because of this, the  $O$  are actually the starting point for describing psychodynamic processes. This is one of the most helpful features of the psychodynamic processes, because it grounds all systems in terms of measurable properties. That said, it is not possible to have various  $O$  correspond to random unitary matrices, and some set  $\{O_i\}_{i=1,\dots,n}$  must be defined in a fashion which are with high probability internally consistent. For example, "Are you hungry?" and "Do you feel hungry?" cannot be defined such that they give opposite answers with high probability. As such, one would need to begin with some specific operator  $O_0$  and then progressively add additional operators such that the probabilities match empirical values. However, any given  $O_i$  need not be so simple as "Are you hungry?" As a large-scale example, a battery of numerous questions could be approximated as a single operator, though a more fine-grained analysis would reveal that the ordering of the questions would to at least some degree influence the outcomes.[13, 16, 19, 20] However, if the battery of questions were universally given in a consistent order, then it would be perfectly valid to approximate the entire battery as a single operator when considering processes with time-steps of durations longer than the period over which such a battery of questions were given. This hearkens back to the earlier discussion of the temporally scale-free nature of the mathematical framework of holographic cognition, as discussed at greater length in Bagley & Petritsch.[6]

### 3 W-L Language Games

Armed with this process theory, we turn to central example which further illustrates its use. A very general problem would be the form of "language game" described by Wittgenstein.[41] Two parties each have a "private language" which both reflects, is the product of, and generates (at least some aspects of) their understandings of the world, internal lives, and experiences. There is a public language, which is the means by which the two parties communicate, and the goal of one or both parties is to gain insight into the private language of the other. In honor of Wittgenstein's use of a lion as an example of entities which he supposed would possess a private language which humans could never understand, we will term these two parties Lion and Wittgenstein, denoted  $L$  and  $W$  in the diagrams.

$W$  would like to infer information about both the cogit vector state as well as the OVD(s) corresponding to the evolution of  $L$ 's cogit vector. The public language of the  $O$  operators is a shared ground-truth, as is the basis which it fixes for both  $W$ 's and  $L$ 's respective cogit vectors. However, at initialization of the W-L game  $W$  has no information about  $L$ 's cogit state  $\mathcal{L}_0$ . Some subset of  $W$ 's cogit state  $\mathcal{W}_0$  corresponds to their Bayesian priors for both  $\mathcal{L}_0$  and the OVD(s) of the  $S(\mathcal{L})_t$  which would be applied by the presentation by  $W$  at time  $t$  of measurement operators  $O_i$  from some set of options  $i = \{1, \dots, N\}$ .

For the sake of simplicity, let us assume a case where  $W$ 's initial hypothesis  $\mathcal{W}_0$ , a subset of  $W$ 's cogit vector, is a completely uniform prior over all possible values of the cogit vector  $\mathcal{L}_0$ . Additionally, we assume that  $L$  either cannot deceive  $W$ , or is sufficiently disincentivized against lying that it chooses to provide truthful responses.<sup>3</sup> For the sake of realism let us assume that only a subset of  $\mathcal{L}_t$  can actually be read out by any given  $O_i$ , and that the set cogs  $\mathcal{L}$  which could even possibly be measured by  $\cup_i O_i$  is a strict subset of some larger cogit collection describing the whole of  $L$ 's cognition. That is,  $\mathcal{L} \subset L$ . Let us additionally

<sup>3</sup>An analysis of cases with deception is beyond the scope of this paper, and will be assessed in future work.

assume that all  $O_i$  have the same upper bound on the Bures distance  $B(\tilde{\mathcal{L}}_0, \mathcal{L}_0)$ .<sup>4</sup> In this case, the choice of  $O_i \rightarrow O_{t=1}$  can be random across the set of available  $O$  while remaining optimal. For  $W$ , the hypothesis

Because the basis of the cogit vectors for  $W$  and  $L$  are the same—fixed by their relationship with the public language of the  $O_i$ —the true objective of the inference problem is to reduce the entropy of the OVD over  $S_L(O_i)_t$ .  $W$  thereby reduces uncertainty with respect to their hypothesis of the private language of  $L$ .

There is thus an output sampled from the distribution  $T_1 = O_i(\mathcal{L}_0)_{t=1}$ , which applies a dynamics operator  $S_{W,1}$  to  $\mathcal{W}_1$  to produce  $\mathcal{W}_2$ .

This leads us to a more formal mathematical description of the inference problem underlying W-L games.

### 3.1 Formulation of W-L Games as Hierarchical Bayesian Inference

Here we formalize the Wittgenstein Lion (W-L) Language game as an entropy reduction problem within the holographic cognition framework. We use hierarchical Bayesian modeling to show how Wittgenstein ( $W$ ) infers the cognitive dynamics or private language of an individual  $L$  through the use of properties of a population. As  $W$  interacts with the  $L$ , we recursively update a joint posterior belief over both the  $L$ 's hidden cognitive state and their unique OVD parameters. We define the objective function for an algorithm aimed at deriving an individual's OVD based on expected information gain (EIG), which allows us to actively select the next question that maximally reduces our uncertainty (entropy) about the individual  $L$ 's specific OVDs, refining the population prior to characterize the individual's private language.

#### 3.1.1 Cognitive States and Dynamics

The state space of the cognitive system of the  $L$  is modeled within a  $D$ -dimensional complex Hilbert space equipped with an inner product, denoted  $\mathcal{H}_D \cong \mathbb{C}^D$ . Let  $\mathcal{D}_D$  denote the space of density matrices on  $\mathcal{H}_D$ .

**Definition 1** (Cognitive State). *The mental state of the  $L$  at time  $t$  is represented by a density matrix  $\rho_t \in \mathcal{D}_D$ . This is a positive semi-definite operator on  $\mathcal{H}_D$  ( $\rho_t \succeq 0$ ) with unit trace ( $\text{Tr}(\rho_t) = 1$ ).*

This provides us a complete description of the  $L$ 's cognitive state at a given time, handling both the definite beliefs and statistical uncertainty of mixed states.

The observer ( $W$ ) interacts with  $L$  using a finite set of public language operators,  $\mathcal{O} = \{O_1, \dots, O_N\}$ .

**Definition 2** (Public Language Operators). *Each operator  $O_i \in \mathcal{O}$  is formalized as a Positive Operator-Valued Measure (POVM). A POVM is a set of elements  $\{E_{i,a}\}_a$ , where  $a$  indexes the possible outcomes (Answers,  $A$ ). These elements satisfy  $E_{i,a} \succeq 0$  and the completeness relation  $\sum_a E_{i,a} = \mathbb{1}$ .*

We model cognitive evolution, which is the change in mental state due to internal processes  $S$  or external stimuli, through the application of unitary operators. These operators belong to the Unitary Group  $U(D)$ , which means that the evolution preserves the probabilistic structure, and the state remains normalized. It should be noted that for any given set of empirically measurable cogs there will always be unmeasured or unmeasurable cogs—at least in any real-world setting. This means that operators describing cognitive dynamics would not truly be unitary. However, one can add “junk” cogs to the state vector as an approximate way of in some sense absorbing the non-unitarity. This is a traditional approximation from quantum mechanics when dealing with a mixed state—i.e. one in which the qubits are entangled not only within the system but also with the environment surrounding the system. In holographic cognition, the analogy to the “environment” is all portions of an individual's cognition not captured by a given set of cogs.

The probability of observing outcome  $a$  when applying a public operator  $O_i$  to state  $\rho_t$  is determined by the Born rule:

$$P(a|O_i, \rho_t) = \text{Tr}(E_{i,a} \rho_t). \quad (1)$$

This is the probability that the  $L$  will give a specific answer when  $W$  asks a question, given  $L$ 's current mental state.

#### 3.1.2 Operator-Valued Distributions

Applying a stimulus  $O_i$  induces a dynamic evolution of the cognitive state, giving us  $\rho'$ . We model this as a stochastic evolution, where we draw from a distribution of unitaries:

$$\rho' = U \rho U^\dagger \quad (2)$$

where  $U$  is a unitary matrix drawn from the Unitary group  $U(D)$ , and  $U^\dagger$  is its conjugate transpose. The specific transformation  $U$  is stochastic and governed by an OVD associated with  $O_i$ . This describes how  $L$ 's mental state changes. When  $L$  processes a stimulus, their internal state  $\rho$  is transformed into a new state  $\rho'$  by a cognitive process  $U$ .

**Definition 3** (Operator-Valued Distribution). *An OVD  $\mathcal{P}_i$  is a probability distribution over  $U(D)$ , characterizing the stochastic cognitive dynamics induced by the operator  $O_i$ .*

We assume the OVDs belong to a parameterized family of distributions  $\mathcal{P}_i(U|\theta_i)$ . The parameters  $\theta_i$  belong to some parameter space  $\Omega_{\theta_i}$ .

$$U \sim \mathcal{P}_i(U|\theta_i). \quad (3)$$

This is the specific cognitive process  $U$  that occurs when  $W$  asks question  $O_i$ . It is drawn from a distribution  $\mathcal{P}_i$  that characterizes  $L$ 's private language defined by  $\theta_i$ .

The  $L$ 's private language is characterized by the complete set of these parameters  $\Theta = \{\theta_1, \dots, \theta_N\}$ , which exist in the joint parameter space  $\Omega_\Theta = \Omega_{\theta_1} \times \dots \times \Omega_{\theta_N}$ . The primary objective is the inference of  $\Theta$ .

---

<sup>4</sup>See Bagley & Petritsch 2024 for a discussion of why the Bures distance is used in this context.[6]

## 3.2 The Hierarchical Bayesian Framework

We use a hierarchical Bayesian strategy to manage individual variability and the complexity of hidden state dynamics. We start with analysis on the population, which gives us a prior for analysis at the individual level.

1. **Population Level:** W first establishes a population-level prior  $P_{\text{pop}}(\Theta)$  by analyzing interactions across a population of Ls. This estimates the general distribution of OVD parameters, marginalizing over individual variations and unknown initial states.
2. **Individual Level:** W then engages with a specific individual. The population distribution serves as the initial prior belief:  $P(\Theta|H_0) = P_{\text{pop}}(\Theta)$ . Here,  $H_0$  represents the initial condition which is the state of knowledge before any interactions have occurred with a specific individual L. We also assume a prior over the initial cognitive state  $P(\rho_0)$  (e.g., the maximally mixed state  $\mathbb{1}/D$ ).

It is important to note that the maximally mixed state prior  $\mathbb{1}/D$  is the prior over the initial cognitive state, rather than using information from the population level. The hierarchical model informs the dynamics  $\Theta$  which describe how the mind changes. It does not necessarily inform the initial state  $\rho_0$  of a specific individual at the start of the interaction. The cogit states of each individual are unknown variables.

The dialogue proceeds sequentially. Let  $H_t = \{(O_1, A_1), \dots, (O_t, A_t)\}$  denote the history of interactions up to time t. At step  $t+1$ :

1. W selects  $O_{t+1} \in \mathcal{O}$ .
2. A hidden unitary  $U_{t+1} \sim \mathcal{P}_{O_{t+1}}(U|\theta_{O_{t+1}})$  is realized.
3. L's state evolves:  $\rho_{t+1} = U_{t+1}\rho_t U_{t+1}^\dagger$ .
4. L generates a response  $A_{t+1}$  based on  $\rho_{t+1}$  via the Born rule (Equation (1)).

The inference requires the simultaneous estimation of the static parameters  $\Theta$  and the dynamic hidden state  $\rho_t$ . W must maintain and update a joint posterior belief distribution  $P(\Theta, \rho_t | H_t)$ . This distribution represents the updated belief about both the static parameters  $\Theta$  and the dynamic cognitive state  $\rho_t$ , conditioned on the observed history  $H_t$ . The state and parameters are strongly interdependent. The observations  $A_t$  depend directly on the current state  $\rho_t$  via the Born rule, and the evolution of  $\rho_t$  to  $\rho_{t+1}$  depends directly on the parameters  $\Theta$  via the OVD. The joint distribution allows us to capture the correlations and shared uncertainty between these variables for more accurate inference.

### 3.2.1 Recursive Bayesian Update

We derive the recursive update for the joint belief state when W applies  $O_{t+1}$  and observes  $A_{t+1}$ . First, we calculate the predictive distribution  $P(\Theta, \rho_{t+1} | O_{t+1}, H_t)$  using the Chapman-Kolmogorov equation, marginalizing over the previous state  $\rho_t$ :

$$P(\Theta, \rho_{t+1} | O_{t+1}, H_t) = \int_{\mathcal{D}_D} P(\rho_{t+1} | \Theta, \rho_t, O_{t+1}) P(\Theta, \rho_t | H_t) d\rho_t. \quad (4)$$

Before W hears the answer to the next question  $O_{t+1}$ , W predicts the new mental state  $\rho_{t+1}$  and the private language parameters  $\Theta$ . This involves considering all possible current mental states  $\rho_t$  and how they would evolve according to the hypothesized dynamics.

The state transition probability  $P(\rho_{t+1} | \Theta, \rho_t, O_{t+1})$  encapsulates the dynamics of the stochastic evolution. It is calculated by integrating over all possible unitary realizations  $S \in U(D)$ :

$$P(\rho_{t+1} | \Theta, \rho_t, O_{t+1}) = \int_{U(D)} \delta(\rho_{t+1} - U\rho_t U^\dagger) \mathcal{P}_{O_{t+1}}(U|\theta_{O_{t+1}}) dU, \quad (5)$$

where  $\delta(\cdot)$  is the Dirac delta function generalized for the space of density matrices. For reference, The standard Dirac delta  $\delta(x - a)$  is a generalized function that is zero everywhere except at  $x = a$ , and its integral over the domain is 1. In the above equation, the Dirac delta is generalized to the space of density matrices:

$$\delta(\rho_{t+1} - U\rho_t U^\dagger)$$

This generalized function enforces the constraint of unitary evolution. When integrated, it yields a non-zero value only if  $\rho_{t+1}$  is exactly equal to the result of applying the specific unitary  $U$  to  $\rho_t$ . It acts as a selector, enforcing the deterministic relationship between  $\rho_t$ ,  $U$ , and  $\rho_{t+1}$  for a specific realization of  $U$ .

The next step is the correction and measurement update. Upon observing  $A_{t+1}$ , the belief is updated using Bayes' theorem. The new history is  $H_{t+1} = (H_t, O_{t+1}, A_{t+1})$ .

$$P(\Theta, \rho_{t+1} | H_{t+1}) = \frac{P(A_{t+1} | \Theta, \rho_{t+1}, O_{t+1}, H_t) P(\Theta, \rho_{t+1} | O_{t+1}, H_t)}{P(A_{t+1} | O_{t+1}, H_t)}. \quad (6)$$

Here, W observes L's answer  $A_{t+1}$  and updates their belief. Hypotheses that predicted the observed answer well become more probable, while those that did not become less probable.

The likelihood of the observation exhibits the Markov property and is given by the Born rule:

$$P(A_{t+1} | \Theta, \rho_{t+1}, O_{t+1}, H_t) = P(A_{t+1} | O_{t+1}, \rho_{t+1}) = \text{Tr}(E_{O_{t+1}, A_{t+1}} \rho_{t+1}). \quad (7)$$

It is important to clarify what the Markov property means here. This assumption states that the future is conditionally independent of the past, given its current state. This appears in Equation (4), Equation (5), and Equation (7). It means that if we know the current cognitive state  $\rho_{t+1}$ , the history  $H_t$  provides no additional information about the observation  $A_{t+1}$ . The density matrix  $\rho_t$  is defined as the complete representation of the cognitive state, and in the holographic cognition framework,  $\rho_t$  encapsulates all relevant information, including memories, beliefs, traumas, and subconscious elements, as described in section 2.2 of

this paper. This assumption does not say that the L forgets its past, but rather the L's entire relevant past is summarized within its current mental state  $\rho_t$ .

The denominator is the marginal likelihood (evidence) of the observation:

$$P(A_{t+1}|O_{t+1}, H_t) = \int_{\Omega_\Theta} \int_{\mathcal{D}_D} P(A_{t+1}|O_{t+1}, \rho_{t+1}) P(\Theta, \rho_{t+1}|O_{t+1}, H_t) d\rho_{t+1} d\Theta. \quad (8)$$

This is the overall probability of observing the answer  $A_{t+1}$ , averaged across all of W's hypotheses about L's mental state and private language.

### 3.3 Optimal Experiment Design for Entropy Reduction

W aims to select the sequence of operators  $\{O_t\}$  that maximizes the efficiency of learning  $\Theta$ . This is formalized as a Bayesian optimal experiment design (OED) problem focused on uncertainty reduction.

The primary objective is the inference of the OVD parameters. We define the marginal parameter posterior  $P(\Theta|H_t)$  by integrating out the cognitive state:

$$P(\Theta|H_t) = \int_{\mathcal{D}_D} P(\Theta, \rho_t|H_t) d\rho_t. \quad (9)$$

To focus on the private language  $\Theta$  rather than the momentary mental state  $\rho_t$ , W calculates the belief about  $\Theta$  by averaging over all possibilities for the hidden mental state  $\rho_t$ .

Uncertainty is quantified by the differential entropy  $H$  of this marginal posterior:

$$H(\Theta_t) \equiv H(\Theta|H_t) = - \int_{\Omega_\Theta} P(\Theta|H_t) \log P(\Theta|H_t) d\Theta. \quad (10)$$

W employs an active learning strategy by selecting the operator  $O_{t+1}$  that maximizes the expected information gain (EIG), defined as the expected reduction in entropy:

$$EIG(O_{t+1}|H_t) = H(\Theta_t) - \mathbb{E}_{A_{t+1}|O_{t+1}, H_t}[H(\Theta_{t+1})]. \quad (11)$$

This calculates how much W expects their uncertainty to decrease, on average, by asking a specific question  $O_{t+1}$ .

The expectation is taken over the potential outcomes  $A_{t+1}$ , weighted by their predictive probability (Equation (8)). The optimal action policy (in a greedy sense) is:

$$O_{t+1}^* = \arg \max_{O_i \in \mathcal{O}} EIG(O_i|H_t). \quad (12)$$

This objective function selects the operator  $O_{t+1}$  that is expected to yield the most information about the parameters  $\Theta$ , thereby minimizing the entropy of the OVDs. We employ a greedy approach here due to computational necessity. The truly optimal policy would require anticipating all possible future sequences of questions and answers, which would not be feasible for a high-dimensional continuous state.

The EIG is equivalent to the Mutual Information  $I(\Theta; A_{t+1}|O_{t+1}, H_t)$ . This can be reformulated in terms of the Kullback-Leibler (KL) divergence:

$$EIG(O_{t+1}|H_t) = \int_{\Omega_\Theta} P(\Theta|H_t) \left[ \sum_{A_{t+1}} P(A_{t+1}|\Theta, O_{t+1}, H_t) \log \frac{P(A_{t+1}|\Theta, O_{t+1}, H_t)}{P(A_{t+1}|O_{t+1}, H_t)} \right] d\Theta. \quad (13)$$

This leads to the interpretation of EIG as the expected KL divergence between the likelihood predicted by a specific model realization  $\Theta$  and the marginalized likelihood:

$$EIG(O_{t+1}|H_t) = \mathbb{E}_{\Theta|H_t} [D_{KL}(P(A_{t+1}|\Theta, O_{t+1}, H_t) || P(A_{t+1}|O_{t+1}, H_t))]. \quad (14)$$

This formulation shows that expected information gain is maximized when the predictions of different plausible models  $\Theta$  are highly varied compared to the average prediction. W seeks the question that best discriminates between competing hypotheses.

This criterion favors operators for which the predictions of different plausible models  $\Theta$  are most divergent, thereby maximizing the ability to discriminate between hypotheses.

Evaluating the EIG requires computing the model-specific likelihood terms  $P(A_{t+1}|\Theta, O_{t+1}, H_t)$ . This necessitates marginalization over the hidden cognitive states:

$$P(A_{t+1}|\Theta, O_{t+1}, H_t) = \int_{\mathcal{D}_D} P(A_{t+1}|O_{t+1}, \rho_{t+1}) P(\rho_{t+1}|\Theta, O_{t+1}, H_t) d\rho_{t+1}. \quad (15)$$

The term  $P(\rho_{t+1}|\Theta, O_{t+1}, H_t)$  is the state prediction conditioned on a specific model  $\Theta$ :

$$P(\rho_{t+1}|\Theta, O_{t+1}, H_t) = \int_{\mathcal{D}_D} P(\rho_{t+1}|\Theta, \rho_t, O_{t+1}) P(\rho_t|\Theta, H_t) d\rho_t. \quad (16)$$

This calculates the probability distribution over the next mental state  $\rho_{t+1}$ , assuming a specific private language  $\Theta$  is true. It propagates the uncertainty about the current state  $\rho_t$  given  $\Theta$  through the dynamics defined by  $\Theta$ .

## 4 Connections to AI Safety

In this section we lay out a series of potential applications for the methods above to problems in AI safety. Some are immediately feasible, others highly plausible, and in the case of the final example more speculative.

## 4.1 Analysis of AI-driven Attacks on Cognition

Suppose  $W$  were instead a misaligned or maliciously designed agent, and its goal was not only to infer the structure of  $L$ 's private language but then also use it to select an optimal sequence of  $O_t$  in order to selectively steer  $L$ 's cogit vector towards some goal state which serves the objectives of the agent. It should be stressed that this is not some fantastical notion of brainwashing or indoctrination. Examples of such manipulation could extend to extremes of that sort, but are much more likely to be along the lines of more subtle effects like political manipulation, instilling desire for products, alteration of preferences, or gradual shifts in beliefs. All of these can be seen around us constantly in the forms of advertisements, media programs, and more. All such inputs to our cognition are public language operators,, so an AI agent would merely need to learn how to engineer content (e.g. via generative AI models) to optimally steer a target individual's or population's cognition by leveraging its model(s) of their private language(s).

Out of a desire to avoid introducing any nomenclature which might unjustly associate Ludwig Wittgenstein with misaligned and malicious agents, we will instead select a more fitting acronym. Given a radically different Ludwig–Ludwig Mueller–played a central role in efforts to rewrite German religious beliefs into Nazi propaganda, his surname appears fitting.

In such an M-L game, the W-L game is extended with an additional quantum optimal control problem. Via inference over  $L$ 's private language and cogs using the joint belief state analysis described above,  $M$  then attempts to leverage existing  $O$  (in the simplified discrete case above) or design novel  $O$  by solving an optimal quantum control problem over the Unitary Group  $U(d)$  and translating them into whatever form of stimuli are available as vectors for applying the  $O$  designed in this fashion. This phrasing is quite general for a reason. In the context of an agent which interacts via language, speech or text would be the stimuli. If an agent producing imagery or other sounds, the corresponding senses. Yet this generalizes further, as in the contexts of brain-computer interfaces and other neurotechnology, the stimuli can be even more varied and far more direct. In each case, the general mathematical structure should be unchanged from what we describe here, and indeed that generality is a deliberate part of the approach we have described.

Let us now take this a step further in two directions.

First, suppose a group of  $M$ , or a single  $M$  agent acting via multiple channels, aimed to collectively steer an individual's cogit vector towards some goal state. If framed in terms of stimuli transmitted to  $L$ , the single- $M$  case may be easier to detect given it will need to be a single, stronger signal. By contrast, numerous less-impactful signals may be easier to disguise as noise by hiding them amidst other channels and signals—which may themselves even be transmitting information which has been inferred by  $M$  to optimize the tradeoff between obscuring the interference and hampering that same interference.

One could mitigate this difficulty by instead attempting to monitor the cogit dynamics of  $L$ . Yet this presents its own quandry, as in the absence of a method for doing so *without* violating  $L$ 's privacy, one opens the door to further abuse—if not simply making the objectives of  $M$  agents easier to achieve.

Second, suppose an  $M$  agent or group of  $M$  were to target the cognition of a group of  $L$ . Given a sufficiently accurate model of the dynamics by which interactions between  $L$ s shape the cognition of others within the group,  $M$  gain(s) the option of further obscuring its influence by leveraging network effects within the group's social dynamics.<sup>5</sup>

Full analysis of these problems is beyond the scope of this paper, but we introduce the notions here to help clarify the relevance of this paper to multiple problem spaces.

## 4.2 Application to AI Alignment

A central problem in AI alignment is the optimization of some function reflecting the discrepancy between human values and behavioral properties of AI systems. While this problem is immensely complex and multi-faceted, one aspect of this complexity comes from the supposed "irrationality" of human beliefs, preferences, values, decisions, and so on. However, these properties are precisely those which have been empirically shown to be effectively modeled by cogs, and they are irrational only through the lens of classical probability. While by no means a panacea or magic bullet, the use of holographic cognition—and particularly the W-L game—provides a way to infer a mathematical model of cogs and psychodynamics at scale. Most importantly, such a model reflects the complex projective probability underlying human cognition, in contrast with the approach of modeling cognition via classical probability which has historically been the norm in AI alignment.

## 4.3 Representation Agents

There has been growing interest in the concept of AI agents capable of serving as representatives of human users in settings such as planning, negotiating sales or contracts, information exchange, and more. Some of these tasks can rely on purely logical, rule-based protocols, making them amenable to existing methods (so long as "bullshit" by the agents in question is sufficiently minimized and/or mitigated).<sup>6</sup>

By contrast, tasks like contract negotiation are more amorphous, and effectively carrying them out is more contingent on an agent having an effective model of the cogit state(s) and private language(s) of the human(s) whom it aims to represent. Inference of such a mathematical model via a W-L game can reflect the empirically correct probability structure, though it remains to be seen what applied methods would be required in order to accomplish this with sufficiently high accuracy and comprehensiveness. It will also be

<sup>5</sup>See Appendix A for a note on an open question regarding a relevant property of holographic cognition in social contexts.

<sup>6</sup>We use the term "bullshit" (as described in the philosophy literature after being introduced by Frankfurt to mean speech without any concern for truth value) rather than the term "hallucination" which is more popular within the AI literature.[42] This is because "hallucination" assigns to AI anthropomorphic properties which are not based in evidence, inaccurately conflating a pathological internal process with expressed assertions about reality, and attempts to falsely transfer responsibility for malfunctions and failures of products from their creators to the products themselves. This obfuscation is highly counterproductive, as it distracts from the underlying source of such failure modalities being the design and engineering of models themselves.[43, 44, 45, 46, 47, 48]

important to determine how much an individual’s private language changes over time, and at what rate, in order to estimate how effective a model inferred via W-L game can be. Additionally, it will likely be important to time-stamp any such private language model, so that both the agent using the model and other agents with which it might interact can all have information as to the recency—and thus viability—of the private language model in question.

If such applied problems can be effectively resolved, we anticipate two branches for using W-L games to train agents to effectively represent individuals.

#### 4.3.1 Cognitive Dossier

In the first, one imagines an agent using the private language model obtained via W-L game as a guidebook of sorts. Much as one might use a dossier on an individual to gain insight into their cogits, and could leverage this to represent that individual, an AI agent could reference this model to estimate how effectively different courses of action align with the cogits of the individual whom the agent is tasked to represent.

Another effective metaphor is that of a will—a document meant to reflect and describe the intent and preferences of an individual in their absence.

#### 4.3.2 Gogol

Mathematically similar yet functionally quite different is the notion of a gogol. A concept introduced by the physicist and author Hannu Ranjaniemi, a gogol is an approximate duplicate of a human mind. Such entities operate semi-autonomously or fully so, carrying out a variety of tasks.[49, 50, 51]

It is interesting that Ranjaniemi describes gogols as existing on quantum computers, as there does not appear to be any indication he was aware of the use of complex projective probability to model cognition, and the mathematics necessary for constructing a gogol were completely absent at the time—indeed being first derived in Bagley & Petritsch and this paper.[6]

While most of the applications of gogols proposed by Ranjaniemi are highly speculative, there is at least one very concrete class of applications which are much more imminently feasible—assuming quantum computers with sufficiently high logical qubit counts. Though quantum cognition—and thus holographic cognition—shares the same form of probability as quantum mechanics, because there is no connection to quantum physics itself there is no analog of the fundamentally physical properties which enable certain quantum protocols. For such cases, creating a gogol and translating it to a quantum computer could enable necessarily-quantum algorithms and protocols to operate on models of human cognition. Examples could include quantum teleportation, quantum games, or secure multi-party computations such as in a quantum Trusted Execution Environment. If the error tolerance of a desired application and the minimum required cogit count are sufficiently non-demanding, it may be possible to implement such methods in the very near future.

## 5 Discussion

In this paper we have provided a mathematically precise formulation of psychodynamics and general cognitive processes via a modification of the Symmetric Monoidal Category used to describe quantum probability. Thanks to the advantageous properties of the diagrammatic form of category theory and especially process theories, this formulation and its notation will hopefully be easily interpretable to experts from a range of fields. The fact it is mathematically precise and rigorous despite this approachability is the greatest advantage of the psychodynamic process theory we present here, as such mathematical rigor will be essential for many of the most urgent and critical applications that we foresee.

Second, we formulated the W-L game, a model of the problem of inferring another party’s cognitive states and dynamics which is adapted from the language games of Ordinary Language Philosophy. Third, we analyzed the Bayesian optimization problem which reflects the core of a W-L game from a quantitative perspective. While a full discussion of algorithmically solving this problem are beyond the scope of this paper, we intend to explore them in follow-up work. This connects to the potential applications discussed in Section 4, and given the correct experimental environment, along with algorithms which are sufficiently efficient for the scale of data to be studied, this provides a protocol for solving inference problems with the potential for significant impact on multiple problems in AI safety. Further application to clinical psychology and psychiatry is also highly plausible, and we are in the process of developing these connections with collaborators from psychiatry.

Having expanded upon the framework of holographic cognition, here we have presented a more concrete exploration of its properties and use across multiple domains relevant to human health and AI safety, areas where it holds great potential for application.

### **Acknowledgements**

We would like to thank Nicholas Godfrey, PhD, for helpful discussions in the planning and ideation stages of this paper. We would also like to thank Dany Lamothe, MD, and Daniel Fishman, MD for discussions of the application of holographic models of cognition to psychodynamic psychotherapy, as well as feedback on the relationship between psychotherapy and W-L games. Finally, we thank Claudia Petritsch, PhD, for her mentorship and support of the Mathematical Medicine Group.

**Author Contributions** B.A. Bagley: Project ideation, development of the process theory and notation, analysis of the Bayesian optimization problem, writing of the paper; N. Khoshnan: Analysis of the Bayesian optimization problem, writing of the paper.

## References

- [1] Tamara Denning, Yoky Matsuoka, and Tadayoshi Kohno. “Neurosecurity: security and privacy for neural devices”. In: *Neurosurgical Focus* 27.1 (July 2009), E7. ISSN: 1092-0684. DOI: 10.3171/2009.4.focus0985. URL: <http://dx.doi.org/10.3171/2009.4.FOCUS0985>.
- [2] Pieter R. Roelfsema, Damiaan Denys, and P. Christiaan Klink. “Mind Reading and Writing: The Future of Neurotechnology”. In: *Trends in Cognitive Sciences* 22.7 (July 2018), pp. 598–610. ISSN: 1364-6613. DOI: 10.1016/j.tics.2018.04.001. URL: <http://dx.doi.org/10.1016/j.tics.2018.04.001>.
- [3] Sergio López Bernal et al. “Cyberattacks on Miniature Brain Implants to Disrupt Spontaneous Neural Signaling”. In: *IEEE Access* 8 (2020), pp. 152204–152222. DOI: 10.1109/ACCESS.2020.3017394.
- [4] Sergio López Bernal et al. “Security in Brain-Computer Interfaces: State-of-the-Art, Opportunities, and Future Challenges”. In: *ACM Comput. Surv.* 54.1 (Jan. 2021). ISSN: 0360-0300. DOI: 10.1145/3427376. URL: <https://doi.org/10.1145/3427376>.
- [5] Sergio López Bernal, Alberto Huertas Celdrán, and Gregorio Martínez Pérez. “Eight Reasons to Prioritize Brain-Computer Interface Cybersecurity”. In: *Communications of the ACM* 66.4 (Mar. 2023), pp. 68–78. ISSN: 1557-7317. DOI: 10.1145/3535509. URL: <http://dx.doi.org/10.1145/3535509>.
- [6] Bryce Allen Bagley and Claudia Katherina Petritsch. “A Mathematical Framework for the Problem of Security for Cognition in Neurotechnology and AI”. In: *ArXiv* (2004).
- [7] Yizhou Zhang et al. “Toward Mitigating Misinformation and Social Media Manipulation in LLM Era”. In: *Companion Proceedings of the ACM Web Conference 2024*. WWW ’24. Singapore, Singapore: Association for Computing Machinery, 2024, pp. 1302–1305. ISBN: 9798400701726. DOI: 10.1145/3589335.3641256. URL: <https://doi.org/10.1145/3589335.3641256>.
- [8] Alexander S. Choi et al. *The LLM Effect: Are Humans Truly Using LLMs, or Are They Being Influenced By Them Instead?* 2024. arXiv: 2410.04699 [cs.CL]. URL: <https://arxiv.org/abs/2410.04699>.
- [9] Yuan Sun and Ting Wang. *Be Friendly, Not Friends: How LLM Sycophancy Shapes User Trust*. 2025. arXiv: 2502.10844 [cs.HC]. URL: <https://arxiv.org/abs/2502.10844>.
- [10] Marcus Williams et al. *On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback*. 2025. arXiv: 2411.02306 [cs.LG]. URL: <https://arxiv.org/abs/2411.02306>.
- [11] Emmanuel M. Pothos and Jerome R. Busemeyer. “A quantum probability explanation for violations of ‘rational’ decision theory”. In: *Proceedings of the Royal Society B: Biological Sciences* 276.1665 (Mar. 2009), pp. 2171–2178. ISSN: 1471-2954. DOI: 10.1098/rspb.2009.0121. URL: <http://dx.doi.org/10.1098/rspb.2009.0121>.
- [12] Jennifer S. Trueblood and Jerome R. Busemeyer. “A Quantum Probability Model of Causal Reasoning”. In: *Frontiers in Psychology* 3 (2012). ISSN: 1664-1078. DOI: 10.3389/fpsyg.2012.00138. URL: <http://dx.doi.org/10.3389/fpsyg.2012.00138>.
- [13] Zheng Wang and Jerome R. Busemeyer. “A Quantum Question Order Model Supported by Empirical Tests of an A Priori and Precise Prediction”. In: *Topics in Cognitive Science* 5.4 (Sept. 2013), pp. 689–710. ISSN: 1756-8765. DOI: 10.1111/tops.12040. URL: <http://dx.doi.org/10.1111/tops.12040>.
- [14] Zheng Wang et al. “The Potential of Using Quantum Theory to Build Models of Cognition”. In: *Topics in Cognitive Science* 5.4 (Sept. 2013), pp. 672–688. ISSN: 1756-8765. DOI: 10.1111/tops.12043. URL: <http://dx.doi.org/10.1111/tops.12043>.
- [15] Zheng Wang et al. “Context effects produced by question orders reveal quantum nature of human judgments”. In: *Proceedings of the National Academy of Sciences* 111.26 (June 2014), pp. 9431–9436. ISSN: 1091-6490. DOI: 10.1073/pnas.1407756111. URL: <http://dx.doi.org/10.1073/pnas.1407756111>.
- [16] Peter D. Bruza, Zheng Wang, and Jerome R. Busemeyer. “Quantum cognition: a new theoretical approach to psychology”. In: *Trends in Cognitive Sciences* 19.7 (July 2015), pp. 383–393. ISSN: 1364-6613. DOI: 10.1016/j.tics.2015.05.001. URL: <http://dx.doi.org/10.1016/j.tics.2015.05.001>.
- [17] Jerome R. Busemeyer and Zheng Wang. “What Is Quantum Cognition, and How Is It Applied to Psychology?” In: *Current Directions in Psychological Science* 24.3 (June 2015), pp. 163–169. ISSN: 1467-8721. DOI: 10.1177/0963721414568663. URL: <http://dx.doi.org/10.1177/0963721414568663>.
- [18] James M. Yearsley and Jerome R. Busemeyer. “Quantum cognition and decision theories: A tutorial”. In: *Journal of Mathematical Psychology* 74 (Oct. 2016), pp. 99–116. ISSN: 0022-2496. DOI: 10.1016/j.jmp.2015.11.005. URL: <http://dx.doi.org/10.1016/j.jmp.2015.11.005>.

- [19] Emmanuel M. Pothos and Jerome R. Busemeyer. “Quantum Cognition”. In: *Annual Review of Psychology* 73.1 (Jan. 2022), pp. 749–778. ISSN: 1545-2085. DOI: 10.1146/annurev-psych-033020-123501. URL: <http://dx.doi.org/10.1146/annurev-psych-033020-123501>.
- [20] Jerome R. Busemeyer and Emmanuel M. Pothos. “Quantum Models of Cognition”. In: *The Cambridge Handbook of Computational Cognitive Sciences*. Cambridge University Press, May 2023, pp. 242–274. ISBN: 9781108755610. DOI: 10.1017/9781108755610.010. URL: <http://dx.doi.org/10.1017/9781108755610.010>.
- [21] John von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Berlin: Springer-Verlag; English translation: Mathematical Foundations of Quantum Mechanics, Princeton: Princeton University Press, 1955.
- [22] Denis Kleyko et al. “A Survey on Hyperdimensional Computing aka Vector Symbolic Architectures, Part I: Models and Data Transformations”. In: *ACM Comput. Surv.* 55.6 (Dec. 2022). ISSN: 0360-0300. DOI: 10.1145/3538531. URL: <https://doi.org/10.1145/3538531>.
- [23] Denis Kleyko et al. “A Survey on Hyperdimensional Computing aka Vector Symbolic Architectures, Part II: Applications, Cognitive Models, and Challenges”. In: *ACM Comput. Surv.* 55.9 (Jan. 2023). ISSN: 0360-0300. DOI: 10.1145/3558000. URL: <https://doi.org/10.1145/3558000>.
- [24] Bob Coeke and Aleks Kissinger. *Picturing Quantum Processes*. Cambridge University Press, 2017.
- [25] Bob Coeke. *The logic of entanglement. An invitation*. Tech. rep. Department of Computer Science, Oxford University, 2003.
- [26] Samson Abramsky and Bob Coeke. “A categorical semantics of quantum protocols”. In: *Proceedings of the 19th Annual IEEE Symposium on Logic in Computer Science* (2004).
- [27] Bob Coeke. “Quantum Theory: Reconsiderations of the Foundations III”. In: AIP Press, 2005. Chap. Kindergarten quantum mechanics.
- [28] Samson Abramsky and Bob Coeke. “Abstract physical traces”. In: *Theory and Applications of Categories* (2005).
- [29] John Carlos Baez. “The Structural Foundations of Quantum Gravity”. In: Oxford University Press, 2006. Chap. Quantum quandries: a category-theoretic perspective, pp. 240–266.
- [30] Peter Selinger. “Dagger compact closed categories and completely positive maps”. In: *Electronic Notes in Theoretical Computer Science* (2007).
- [31] John Carlos Baez and Aaron Lauda. “Deep Beauty: Understanding the Quantum World through Mathematical Innovation”. In: Cambridge University Press, 2011. Chap. A prehistory of n-categorical physics.
- [32] John Carlos Baez and Mike Stay. “New Structures for Physics”. In: Springer, 2011. Chap. Physics, Topology, logic and computation: a Rosetta Stone.
- [33] Chris Heunan, Mehrnoosh Sadrzadeh, and Edward Grefenstette. *Quantum Physics and Linguistics: A Compositional, Diagrammatic Discourse*. Oxford University Press, 2012.
- [34] Roger Penrose. “Applications of negative dimensional tensors”. In: *Combinatorial Mathematics and Its Applications* (Academic Press).
- [35] David Bohm and Francis David Peat. *Science, Order, and Creativity*. Routledge, 1987.
- [36] Andre Joyal and Ross Street. “The geometry of tensor calculus I”. In: *Advances in Mathematics* (1991).
- [37] Andre Joyal, Ross Street, and Dominic Verity. “Traced monoidal categories”. In: *Cambridge University Press* (Mathematical Proceedings of the Cambridge Philosophical Society).
- [38] Masahito Hasegawa, Martin Hofmann, and Gordon Plotkin. “Pillars of Computer Science”. In: Springer, 2008. Chap. Finite dimensional vector spaces are complete for traced symmetric monoidal categories.
- [39] Peter Selinger. “New Structures for Physics”. In: Springer-Verlag, 2011. Chap. A survey of graphical languages for monoidal categories.
- [40] Aleks Kissinger. “Pictures of processes: automated graph rewriting for monoidal categories”. PhD thesis. University of Oxford, 2012.
- [41] Ludwig Wittgenstein. *Philosophical Investigations*. Blackwell Publishing, 1968. ISBN: 9780631103202.
- [42] Harry G Frankfurt. “On bullshit”. In: (2009).
- [43] Sarah A. Fisher. “Large language models and their big bullshit potential”. In: *Ethics and Information Technology* 26.4 (2024). DOI: 10.1007/s10676-024-09802-5.
- [44] Abeba Birhane and Marek McGann. “Large models of what? Mistaking engineering achievements for human linguistic agency”. In: *Language Sciences* 106 (2024).
- [45] Michael Townsen Hicks, James Humphries, and Joe Slater. “ChatGPT is bullshit”. In: *Ethics and Information Technology* 26.2 (2024). DOI: 10.1007/s10676-024-09775-5.

- [46] Daniel W. Tigard. “On bullshit, large language models, and the need to curb your enthusiasm”. In: *AI and Ethics* 5.5 (2025).
- [47] Robert Sparrow and Gene Flenady. “Bullshit universities: the future of automated education”. In: *AI & SOCIETY* (2025).
- [48] Mark Coeckelbergh. “LLMs, Truth, and Democracy: An Overview of Risks”. In: *Science and Engineering Ethics* 31.1 (2025).
- [49] Hannu Rajaniemi. *The Quantum Thief*. London, England: Gollancz, Sept. 2010.
- [50] Hannu Rajaniemi. *The Fractal Prince*. London, England: Gollancz, Sept. 2012.
- [51] Hannu Rajaniemi. *The Causal Angel*. Jean Le Flambeur. Tor Books, July 2014.

## A Entanglement Structures in Individual and Group Cogit Vectors

There is an interesting question which arises from the structure of holographic cognition, and which is made particularly clear in the category-theoretic formulation of the same. As discussed in [6], holographic cognition is scale-free, or perhaps more precisely scale-flexible, with respect to both time and the space of possible cogit hypervectors. One could use operators to reflect changes across any time range of interest, and can likewise construct a cogit hypervector so that it is a concatenation of the cogits from multiple individuals, scaling to however many individuals one might wish to study.

However, is such a concatenation actually the correct description of a group's collective state? Phrased differently, we know that the entanglement structure for a single individual scales with dimension in such a way that there is a rapid decline in the probability of being able to partition the cogits within their hypervector into fully non-entangled subsets. At a group level, the null hypothesis is that such partitionings of a group cogit hypervector would be along the same lines as reversing a simple concatenation. As such, interactions between the cogit hypervectors of different individuals would in general reflect graphs over which classical information processes are occurring.

But is that true in the general case? An entanglement structure represents some collection of properties in the informational content and structure of an individual's nervous system, and it would be perfectly reasonable—indeed more plausible—for there to be no useful notion of entanglement structures bridging different individuals. Yet there are known phenomena which suggest it may be possible for the same mathematical structure to persist between individuals. It has been shown in the experimental neuroscience literature that certain patterns of neural activity become more similar in individuals with certain sorts of shared traits as those individuals interact. If this or other phenomena result in a group cogit hypervector no longer being partitionable into hypervectors corresponding to the individuals, there would be an entanglement structure representing the group.

There is no reason a priori to expect that this latter option is the case. While heuristic analogies of societies as intelligent superorganisms have existed for some time, there is no reason to think they would reflect underlying mathematical and information properties in a way directly equivalent to a rescaling of an individual's mind. However, if group cogit hypervectors are in general not partitionable, this would demonstrate such heuristic analogies have instead stumbled upon something quite substantive from a computational and phenomenological perspective. In mentioning this, we are not hypothesizing that group cogit hypervectors cannot be partitioned. Rather, that this is a property which should be quite amenable to experimental study, and could have significant and fascinating implications for quantifying phenomena at the interfaces of psychology and sociology.