Are mental properties supervenient on brain properties?

Joshua T. Vogelstein[1], R. Jacob Vogelstein[2], Carey E. Priebe[1]
[1]Department of Applied Mathematics & Statistics,
Johns Hopkins University, Baltimore, MD, 21218,
[2]National Security Technology Department,
Johns Hopkins University Applied Physics Laboratory, Laurel, MD 20723

October 14, 2010

Abstract

The "mind-brain supervenience" conjecture suggests that all mental properties (e.g. consciousness, intelligence, personality, etc.) are derived from the physical properties of the brain. The validity of this conjecture has been argued in philosophical terms for over 2,500 years. Alternative conjectures, including various non-physical causes of mental properties, seem rather implausible to many, but proving or disproving these alternatives has remained elusive.

To address the question of whether the mind supervenes on the brain through empirical means, here we frame a supervenience hypothesis in rigorous mathematical terms and propose a modified version of supervenience (called ε-supervenience) that is amenable to empirical investigations and statistical analysis. To elucidate this approach, we posit a thought experiment that illustrates how the probabilistic theory of pattern recognition can be used to make a one-sided determination of ε-supervenience. The physical property of the brain employed in this analysis is the graph describing brain connectivity (i.e., the *connectome*), and ε-supervenience allows us to determine whether a particular mental property can be inferred from one's connectome to within any given misclassification rate $\varepsilon > 0$, regardless of the relationship between the two. In addition to the theoretical results, we show via simulation that given reasonable assumptions about class conditional probabilities and the amount of data available, the thought experiment can actually be conducted on a simple organism, *Caenorhabditis elegans*, with currently available technology.

The potential significance of this work can be divided into distinct disciplines. To the philosopher, this work demonstrates that philosophical conjectures can be morphed into statistical hypotheses, amenable to experimental investigations, allowing the philosopher to add empirical support to their rational arguments. To the statistician, this work points out the limitations of hypothesis testing in a novel domain. To the neuroscientist, a theoretically possible experiment is proposed to garnish support for a hypothesis that is widely believed: that mental properties supervene on brain properties.

1 Introduction

Questioning the relationship between the mind (thoughts, beliefs, preferences, emotions, intelligence, etc.) and the brain (the physical structure inside our skulls) dates back at least as far as 400 BCE, when Plato wrote the dialogues, in which he posited immateriality of the soul [1]. Approximately two millennia passed before these ideas reached their canonical form through Descartes's discussion of mind-body dualism [2]. Then, in the 20th century, Donald Davidson stated and popularized the mind-brain supervenience conjecture, which claims that an agent cannot alter in some mental property without altering in some physical property [3]. Contemporary fields of neural network theory and neuroscientific inquiry often assume mind-brain supervenience, or an even stronger assumption about mind-brain causality, but no previously proposed notion of supervenience seems amenable to empirical investigation. Here we define new versions of supervenience that formulate the conjecture in rigorous mathematical terms and that can be experimentally tested as a hypothesis.

The primary contributions of this work are as follows. First, a notion of supervenience amenable to empirical investigation is formally introduced. This renders the mind-brain dualism debate a hypothesis, rather than an assumption. Second, in addition to expanding the space of questions amenable to hypothesis testing, we also demonstrate the limits of hypothesis testing. Third we posit a very general model of brains and their associated mental properties that facilitates analysis in a graph theoretical and statistical framework. Fourth, we prove that this formulation admits universally consistent classifiers that are guaranteed to find the relationship between minds and brains, if one exists. Fifth we demonstrate through simulation that the proposed universally consistent classifier has reasonable convergence properties on simulated brain-graph data.

2 Preliminaries

2.1 Supervenience

The intention in this work is to develop greater insight regarding the relationship between minds and brains, using statistical methods, with particular interest in notions of supervenience. Modern philosophy of science suggests that adequate scientific explanations must account for the objects of investigation and their relationships [4]. Therefore, here we define minds, brains, and supervenience.

Let b correspond to an agent's brain, which is a particular element from the set of all possible brains, $\mathcal{B}$. The set of possible brains $\mathcal{B}$ is completely unrestricted, meaning that $\mathcal{B}$ could be an infinite set, with arbitrarily complexity. In particular, b might represent the position, momentum, and type of each subatomic particle residing within the skull some agent. Thus, each different $b \in \mathcal{B}$ could correspond to some difference in the position, momentum, or type of at least one subatomic particle composing the brain.

Similarly, let m correspond to an agent's mind, which is a particular element from the set of all possible minds, $\mathcal{M}$. The set of possible minds $\mathcal{M}$ is also unrestricted. In particular, m might represent all an agent's thoughts, beliefs, and preferences. Thus, each different $m \in \mathcal{M}$ could correspond to a difference in at least one thought, belief, or preference.

The mind-brain supervenience conjecture is a relation between the set of mental states and the set of brain states. Donald Davidson canonized this conjecture in 1970 with the following quote: [3]

> supervenience might be taken to mean that there cannot be two events alike in all physical respects but differing in some mental respect, or that an object cannot alter in some mental respect without altering in some physical respect.

This conjecture may be concisely and formally stated: $m \neq m' \implies b \neq b' \, \forall (m, b), (m', b') \in \mathcal{M} \times \mathcal{B}$, where (m, b) is a mind-brain pair, and (m', b') is a different mind-brain pair. While mind-brain supervenience is a relatively strong claim, importantly, one can imagine far stronger relations.

For instance, it may be the case that minds supervene on brains, but one cannot form an *injective* relation from brains to minds. An injective relation is any relation that preserves distinctness. Thus if minds are injective on brains, then $b \neq b' \implies m \neq m' \, \forall (m, b), (m', b') \in \mathcal{M} \times \mathcal{B}$ (note that the directionality of the implication has been switched relative to supervenience). For instance, it might be the case that a brain could change without the mind changing. Consider the case that a single subatomic particle shifts its position by a Plank length, changing brain state from b to b'. It is possible that the mental state supervening on brain state b remains m, even after b changes to b'. In such a scenario, the mind might still supervene on the brain, but the relation from brains to minds is not injective. This argument also shows that supervenience is not necessarily a *symmetric* relation. Minds supervening on brains does not imply that brains supervene on minds.

Second, it may be the case that minds supervene on brains, but that brains do not cause minds. For instance, consider an analogy where M and B correspond to two coins being flipped, each possibly landing on heads or tails. Further assume that every time one lands on heads so does the other, and every time one lands on tails, so do the other. This implies that M supervenes on B, but assumes nothing about whether M causes B, or B causes M, or some exogenous force causes both.

Third, supervenience does not imply *identity*. Consider, for example, acceleration and velocity. Clearly, acceleration supervenes on velocity, as acceleration cannot change without velocity changing (assuming one does not consider gravity as acceleration). Similarly, velocity supervenes on position, as velocity cannot change without position changing. Therefore, acceleration supervenes on position, by the transitive property of supervenience, but it is not the case

that a change in acceleration is equal to a change in position. Rather, position can change with constant velocity, meaning without acceleration changing.

What supervenience does imply, however, is the following. Imagine finding two different minds. If $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$, then the brains subvening under those two minds must be different. In other words, there cannot be two different minds, either of which could supervene on a single brain. Figure 1 shows several possible relations between the sets of minds and brains.

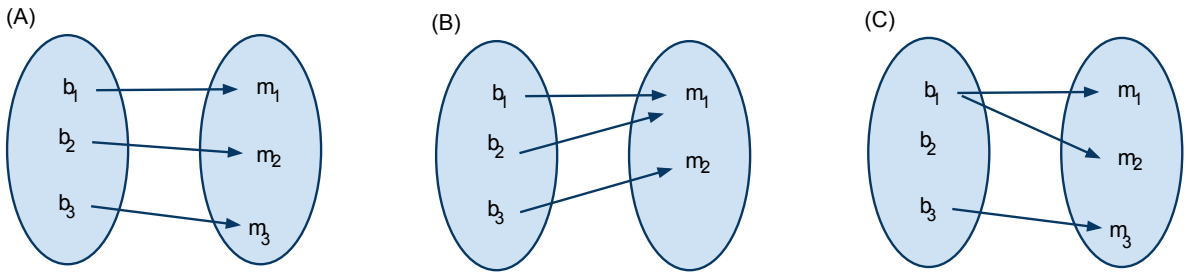

Figure 1: Possible relations between minds and brains. (A) Minds supervene on brains, and it so happens that there is a bijective relation from brains to minds. (B) Minds supervene on brains, and it so happens that there is a surjective (a.k.a., onto) relation from brains to minds. (C) Minds are *not* supervenient on brains, because two different minds supervene on the same brain.

All of the above relations are *logical* relations, not probabilistic relations. To facilitate both statistical analysis and empirical investigation, we project this supervenience notion into a statistical setting. To proceed, we first define a *model*, $\mathcal{P} = \{\mathbb{P}[M, B]\}$, the set of all possible joint distributions, $\mathbb{P}[M, B]$, that could describe the probabilistic relationship between minds and brains. Each distribution $\mathbb{P}[M, B]$ uniquely specifies the probability of any element (m, b) occurring from the space of all possible elements $\mathcal{M} \times \mathcal{B}$ (also called the *sample space*).Given a distribution, one can then calculate any marginal or conditional distributions. For instance, $\mathbb{P}[B] = \int_{m \in \mathcal{M}} \mathbb{P}[M, B] dm$, or $\mathbb{P}[M|B] = \mathbb{P}[M, B]/\mathbb{P}[B]$. Having specified a model, statistical supervenience can be defined as follows:

Definition 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $\mathbb{P} = \mathbb{P}[M, B]$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $\mathbb{P}[m \neq m'|b = b'] = 0 \, \forall (m, b), (m', b') \in \mathcal{M} \times \mathcal{B}$. Alternately, $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ can be written as $\mathbb{P}[m = m'|b = b'] = 1 \, \forall (m, b), (m', b') \in \mathcal{M} \times \mathcal{B}$.*

Statistical supervenience is therefore a probabilistic relation on sets. Note that statistical supervenience is distinct from statistical correlation. *Statistical correlation* between brain states and mental states is defined as $\rho_{MB} = \mathbb{E}[(B - \mu_B)(M - \mu_M)]/(\sigma_B \sigma_M)$, where μ_X and σ_X are the mean and variance of X, and $\mathbb{E}[X]$ is the expected value of X. If $\rho_{MB} = 1$, then both $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ and $\mathcal{B} \overset{\varepsilon}{\sim}_F \mathcal{M}$. Thus, perfect correlation implies supervenience, but supervenience does not imply correlation.

2.2 Brain-graphs

Given the above notion of statistical supervenience, one can then design a statistical test for supervenience by carefully considering an appropriate model. For mind-brain supervenience, a model of the brain is paramount, and has historically been a bit of a sticky wicket. It was not until 1891, that H. Waldeyer-Hartz first formally proposed the "neuron doctrine" [5], which states that the nervous system is a complex *network* of "neurons" (a term he invented), largely based on Ramon y Cajal's work using the Golgi stain [6]. This doctrine has been central to much of the development of neuroscience and artificial intelligence for over 100 years, including the development of neural network theory [7] and cognitive science [8]. Fundamental to nearly all models of the brain (or components therein) since the neuron doctrine's introduction is the existence of neural units, each connected to (some) other neural units. Therefore, it seems appropriate to assume that a reasonable model of a brain is graph, composed of a set of vertices (or nodes), each representing a neural unit, and a set of edges (arcs) connecting the nodes. Below, we formalize this concept to enable statistical testing in the graph domain.

3 Results

3.1 Theoretical results

If minds statistically supervene on brains, $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, then two different minds must supervene on two different brains. This means that there exists a unique mapping from each brain to a single mind. In other words, one can in principle construct a function $g(b) : \mathcal{B} \mapsto \mathcal{M}$, that is a deterministic mapping from brains to minds. It may be the case that subsets of brains form equivalence classes, such that any brain in that subset is mapped to the same mind (see, for example, b_1 and b_2 in Figure 1(A)). Assuming for the moment that the space of all possible minds is finite, that is $|\mathcal{M}| < \infty$, then we call any such function a *classifier* (this assumption will later be relaxed). Let $\widehat{m}$ denote the output of a classifier, $g(b) = \widehat{m}$. Define misclassification rate as:

$$L_{\mathbb{P}}(g) = \mathbb{P}[g(B) \neq M] = \frac{1}{|\mathcal{B}||\mathcal{M}|} \iint \mathbb{I}\{g(b) \neq m\} db dm \tag{1}$$

where $\mathbb{I}\{\cdot\}$ indicates the indicator function, taking unity value if its argument is true, and zero otherwise. $L_{\mathbb{P}}(g)$ therefore effectively counts the fraction of time g misclassifies b. The Bayes optimal classifier g^* minimizes $L_{\mathbb{P}}(g)$ over all classifiers, that is

$$g^* = \underset{g \in \mathcal{G}}{\operatorname{argmin}}\, L_{\mathbb{P}}(g) \tag{2}$$

where $\mathcal{G}$ is the set of all possible classifiers. Thus, the *Bayes error*, or Bayes risk, $L_{\mathbb{P}}(g^*)$ is the minimum possible misclassification rate. The primary result of casting supervenience is a statistical framework is the following theorem:

Theorem 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $\mathbb{P} = \mathbb{P}[M, B]$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $L_{\mathbb{P}}(g^*) = 0$.*

If minds supervene on brains, then, by the definition of supervenience, there exists a function that maps each brain deterministically to a particular mind. This means that one could draw a decision boundary between all equivalence classes of brains, each class corresponding to a different mind, and no mind will reside within two different equivalence classes. Thus, the optimal classifier would correctly find these decision boundaries, and therefore have no opportunity to err. □

This relationship between statistical supervenience and Bayes error can therefore be described concisely: $\mathcal{M} \overset{S}{\sim}_F \mathcal{B} \Leftrightarrow L_{\mathbb{P}}(g^*) = 0$. Thus, the above arguments shows (for the first time to our knowledge) that statistical supervenience and zero Bayes error are equivalent. Further, statistical supervenience can be thought of as a constraint on the possible models. Specifically, let $\mathcal{P}_s \subseteq \mathcal{P}$ be subset of models for which supervenience holds. Then, $\mathcal{P}_s = \{\mathbb{P}[M, B] : L_{\mathbb{P}}(g^*) = 0\} \subseteq \mathcal{P}$.

3.2 Hypothesis testing

While the above theorem is of potential theoretical interest, because the arguments rest on knowing $\mathbb{P}[M, B]$ and g^*, which are typically unknown, they are pragmatically useless. However, both $\mathbb{P}[M, B]$ and g^* could be estimated from data. Let $(m_1, b_1), (m_2, b_2), \ldots, (m_n, b_n)$ be random samples taking their values in $\mathcal{M} \times \mathcal{B}$, independently and identically distributed according to model $\mathbb{P}[M, B]$. Generalizing the concept of a classifier g to allow incorporation of training data, consider $g_n : \mathcal{B} \times (\mathcal{M} \times \mathcal{B})^n \mapsto \mathcal{M}$ which takes as input an observed brain b and n training pairs $\mathcal{T}_n = \{(m_1, b_1), \cdots, (m_n, b_n)\}$, and produces a classification $g_n(b; \mathcal{T}_n) = \widehat{m}$. Misclassification rate for this classifier will therefore be a random variable, because the training data $\mathcal{T}_n$ are random samples. Therefore, instead of calculating misclassification rate for g_n, we compute the expected misclassification rate:

$$\mathbb{E}[L_{\mathbb{P}}(g_n)] = \mathbb{E}[\mathbb{P}_F[g_n(B; \mathcal{T}_n) \neq M | \mathcal{T}_n]] = \int \mathbb{P}[g_n(B) = M | \mathcal{T}_n] \mathbb{P}[\mathcal{T}_n] d\mathcal{T}_n. \tag{3}$$

Unfortunately, in practice, computing $\mathbb{E}[L_{\mathbb{P}}(g_n)]$, requires integrating over all possible training data corpuses of size n, and by definition, we only have access to a single training data corpus. We therefore define "hold out" misclassification performance:

$$\mathbb{E}[L_{\mathbb{P}}(g_n)] \approx \widehat{L}_F^{n'}(g_n) = \sum_{\mathcal{T}_{n-n'}} \mathbb{P}[g_n(B) = M | \mathcal{T}_{n-n'}] \mathbb{P}[\mathcal{T}_{n-n'}], \tag{4}$$

where $n' < n$ is the number of held-out training samples, and the sum is taken over a sufficiently large number of subsets, $\mathcal{T}_{n-n'}$, such that $\widehat{L}_F^{n'}(g_n)$ converges. $n'\widehat{L}_F^{n'}(g_n)$ is the expected number of misclassified minds. Simplifying further by assuming just one hold-out set, then $n'\widehat{L}_F^{n'}(g_n)$ and has a binomial distribution, because for any of the n' held-out samples, the classifier could be either correct or incorrect, thus $n'\widehat{L}_F^{n'}(g_n) \sim \text{Binomial}(n', L_{\mathbb{P}}(g_n))$.

Before explicitly considering the problem of testing for statistical supervenience, we define a relaxed notion of supervenience:

Definition 2. *Given $\varepsilon > 0$, $\mathcal{M}$ is said to ε-supervene on $\mathcal{B}$ for distribution $\mathbb{P} = \mathbb{P}[M, B]$, denoted $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$, if and only if $L_{\mathbb{P}}(g^*) < \varepsilon$.*

Given this relaxation, consider the problem of testing for $\varepsilon-$supervenience. First, specify a significance level, α, such that if the p-value is less than α, then the null is rejected. Because we hope to reject the null, in favor of the alternative, let the null hypothesis be H_0: $L_{\mathbb{P}}(g_n) \geq \varepsilon$, and the alternative hypothesis be H_A: $L_{\mathbb{P}}(g_n) < \varepsilon$. We reject for low values of the test-statistic, $n'\widehat{L}_F^{n'}(g_n)$. Specifically, if $n'\widehat{L}_F^{n'}(g_n)$ is less than the critical value, $c_\alpha(n', \varepsilon)$, then we reject. The critical value is available under the least favorable distribution Binomial(n', ε). Thus, if $n'\widehat{L}_F^{n'}(g_n) < c_\alpha(n', \varepsilon)$, we can conclude with $100(1-\alpha)\%$ confidence that $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$. The definition of ε-supervenience therefore admits, for the first time to our knowledge, a statistical test of supervenience, given a specified ε and α. Similar to the above, one can define the set of ε-supervenience models as the set of models under which ε-supervenience holds, that is: $\mathcal{P}_\varepsilon = \{\mathbb{P} | L_{\mathbb{P}}(g^*) < \varepsilon\}$. One could then sort ε-supervenience subsets, $\mathcal{P}_s \subseteq \mathcal{P}_\varepsilon \subseteq \mathcal{P}_{\varepsilon'} \subseteq \mathcal{P}$, for any $\varepsilon < \varepsilon'$.

3.3 Power and consistency

Importantly, the utility of any statistical test depends both on the p-value, the probability of obtaining a test statistic at least as extreme as the observed value (under the assumed model), and its power, the probability that the test will reject a false null hypothesis (in other words, the probability that it will not make a Type II error). Ideally, the power of this test would go to unity, as $n, n' \to \infty$. A sufficient condition for power to approach unity is that g_n is a *consistent* classifier. A classifier is consistent if and only if its expected misclassification rate converges to the Bayes optimal limit with sufficient data, that is $\mathbb{E}[L_{\mathbb{P}}(g_n)] \to L_{\mathbb{P}}(g^*)$ as $n \to \infty$. As the notation suggests, consistency of a classifier is a function of the true model, F. Without any prior knowledge of what the model might be, one desires a *universally consistent* classifier, that is a classifier that is consistent for all $\mathbb{P} \in \mathcal{P}$.

Unfortunately, the rate of convergence of $L_{\mathbb{P}}(g_n)$ to $L_{\mathbb{P}}(g^*)$ depends on the (unknown) distribution $\mathbb{P} = \mathbb{P}[M, B]$ [9]. Furthermore, arbitrarily slow convergence theorems regarding the rate of convergence of $L_{\mathbb{P}}(g_n)$ to $L_{\mathbb{P}}(g^*)$ demonstrate that there is no universal n, n' which will guarantee that the test has power greater than any specified target $\beta > \alpha$ [10]. For this reason, the test outlined above can provide only a one-sided conclusion: if we reject we can be $100(1-\alpha)\%$ confident that $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ holds, but we can never be confident in its negation. This means that we can never be confident that $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ does *not* hold; rather, it may be the case that the evidence in favor of $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ is insufficient for any number of reasons, including that we simply have not yet collected enough data. Unfortunately, arbitrarily slow convergence theorems inform us that no matter how much data we collect, we cannot disambiguate between not yet having enough data, and $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ not holding. Thus, without restrictions on $\mathbb{P}[M, B]$, arbitrarily slow convergence theorems imply that our theorem of ε-supervenience does not strictly satisfy Popper's *falsifiability* requirement [11]. Given these limitations on even universal consistency, it is still the best one can hope for. Therefore, we hope to obtain a universally consistent classifier to test for ε-supervenience.

3.4 Universal consistency for brain-graphs

Formally, a brain-graph $G = (V, A)$ is characterized by a set of vertices (or nodes), $V = \{V_i\} = \{V_1, \ldots, V_n\}$, where n is the number of neural units, and arcs (or edges) $A = \{A_{ij}\}$, where A_{ij} represents the connectivity from neural unit V_j to V_i. For simplicity (to be generalized below), assume that brain-graphs are simple graphs, meaning that for all $i, j \in [n]$: (i) there are no self loops, $a_{ii} = 0$, (ii) all edges are binary, $a_{ij} \in \{0, 1\}$, and (iii) all edges are symmetric, $a_{ij} = a_{ji}$. Further assume that the vertices are *labeled*, meaning that there is a known one-to-one mapping from each vertex in any brain to a vertex in any other brain, and that each brain-graph has the same number of vertices. Thus, each brain-graph is characterized entirely by its adjacency matrix, $a \in \mathcal{A}$, where $\mathcal{A} \subseteq \{0, 1\}^{n \times n}$. Further, let the mental property under investigation be binary, so $m \in \mathcal{M} = \{0, 1\}$. The Bayes classifier is:

$$\widehat{m} = \underset{m}{\text{argmax}}\, \mathbb{P}[b, m] = \underset{m}{\text{argmax}}\, \mathbb{P}[b|m]\mathbb{P}[m] \tag{5}$$

where the likelihood term, $\mathbb{P}[B|M]$, specifies the probability of observing any particular adjacency matrix for any class, and the prior term, $\mathbb{P}[M]$, specifies the prior probability of either class.

Under the above model assumptions, the number of possible brain-graph/mental-property pairs is finite, that is, $|\mathcal{M} \times \mathcal{B}| = d < \infty$. Therefore, one can simply enumerate all possible brain-graphs for each class, such that each class is represented by a multinomial distribution. The maximum likelihood estimator (MLE) of the multinomial parameter for each class as guaranteed to converge to the true (but unknown) parameter. Formally, one can state $\widehat{\mathbb{P}}_{MLE}[B|M] \to \mathbb{P}[B|M]$ as $n \to \infty$, almost surely. Furthermore, the MLE estimator of the prior also converges to the true (but unknown) prior distribution, $\widehat{\mathbb{P}}_{MLE}[M] \to \mathbb{P}[M]$, as $n \to \infty$, almost surely.

The Bayes plugin classifier is defined by:

$$\widehat{m} = \operatorname*{argmax}_m \widehat{\mathbb{P}}[b|m]\widehat{\mathbb{P}}[m] \tag{6}$$

where $\widehat{\mathbb{P}}[B|M]$ and $\widehat{\mathbb{P}}[M]$ are plugin estimators. Since the MLE estimators defined above are plugin estimators, and consistent, one obtains a consistent classifier by plugging in the MLE parameters. This logic holds for any model in which the cardinality of the sample space is finite, that is $|\mathcal{M} \times \mathcal{B}| < \infty$. In particular, one can relax both the assumed symmetry and hollowness of the adjacency matrix, and this result still holds. Furthermore, one can allow for multi-graphs, in which there are a finite number of different kinds of edges (different colored edges), such that $a_{ij} \in \{0, 1, \ldots, K\}$, with $K < \infty$, and the set of possible brain-graphs remains finite. Finally, one can allow $\mathcal{M}$ to be any finite space, not just binary. Note that in all these cases, the process one uses to enumerate and sort the brain-graph collections is essentially irrelevant, assuming the same process is used for each class. Further note that more robust estimators, such as certain M-estimators and the *maximum a posteriori* estimators, also admit universally consistent classifiers, with perhaps faster convergence properties.

The above approach, however, is insufficient in several important cases. First, when the edges live in a continuous space, such as $a_{ij} \in \mathbb{R}$, corresponding to weighted graphs, the MLE plugin is not well defined. In such scenarios, the k_n nearest neighbor (k-NN) algorithm can be used (see Appendix 1 for a description of the k-NN algorithm) [12]. Although originally defined to operate on finite dimensional Euclidean space, the same algorithm can be applied here, once one has first embedded the brain-graphs into such a space. One possible embedding is to stack the columns of the adjacency matrix, yielding an n^2 element vector, instead of an $n \times n$ element matrix. Given such an embedding, the k-NN classifier may be used without modification, assuming one has defined a suitable distance operator (such as L_2 or L_1). This algorithm may also be used in the above cases where the support of the sample space is finite, and, convergence of the k-NN algorithm for those problems will be faster than the MLE plugins, whenever the distance used is "appropriate" (where appropriate is assessed by convergence rates under the true (but unknown) model). Furthermore, one can allow for multiple kinds of edges, as before, yielding adjacency tensors, with $a_{ij} \in \mathbb{R}^d$. Or, one can allow the mental property to be multidimensional as well, $m \in \mathbb{R}^l$. The k-NN algorithm converges in both these generalizations. Finally, one could allow the number of vertices to differ across brain samples. In such a scenario, one can "pad" the adjacency matrices of the smaller brain-graphs, and the result still holds.

The above universal consistency results hold for unlabeled graphs as well, with a minor modification. Assuming the brain-graphs are unlabeled, and we have a test brain-graph b, one could first align each training brain-graph with the test brain-graph, and then use the above algorithms. To align any pair of graphs, one can solve the graph isomorphism problem, $\widehat{Q}_i = \operatorname{argmin}_{Q \in \mathcal{Q}} \left\| Q b_i Q^\mathsf{T} - b \right\|$, where $\mathcal{Q}$ is the set of all permutation matrices (that is, zero matrices with a single unit value in each row and column). Given $\widehat{Q}_i$, one can compute $\widetilde{b}_i = \widehat{Q}_i b \widehat{Q}_i^\mathsf{T}$, and then proceed as one would normally. Unfortunately, the graph isomorphism problem is known to be NP-incomplete (that is, it is not known to believe in either P or NP-complete) [13], and therefore no polynomial time algorithm is available, making this approach quite time-consuming.

3.5 Main result

Returning to the question of supervenience, consider the following thought experiment. Let the physical property under consideration be brain-graphs, and the mental property under investigation be binary (such as, knows calculus or not). Now, imagine collecting very large amounts of very accurate exchangeable brain-graph data and the associated mental property values. A k_n-nearest neighbor algorithm using an isomorphism-matching Frobenius norm is universally consistent. Therefore, Theorem 1 applies and the existence of a universally consistent classifier guarantees that eventually (in n, n') one will be able to conclude $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ for this mental/brain property pair, if indeed ε-supervenience

holds. This logic holds for directed graphs or multigraphs or hypergraphs with edge weights, with or without labeled vertices.

3.6 Simulation

As an example of a feasible experiment, one may consider a species whose nervous system consists of the same (small) number of labeled neurons for each organism. *Caenorhabditis elegans* is believed to be such a species [14]. The hermaphroditic C. elegans' somatic nervous system consists of 279 interconnected neurons. While the graph with these neurons as vertices and edges defined by chemical synapses between neurons is not identical across individuals, it is reasonably consistent [14]. Furthermore, these animals exhibit a rich behavioral repertoire that depends on circuit properties [15]. Thus, one may design an experiment by describing the joint distribution $\mathbb{P}[M, B]$ via class-conditional distributions $\mathbb{P}[B|M = m_j]$ for the C. elegans brain-graph for two mental properties of interest, m_0 and m_1, along with the prior probability of class membership $\mathbb{P}[M = m_1]$. Here the mental property corresponds to the C. elegans exhibiting or not exhibiting a particular behavior (e.g., response to an odor).

Simulations suggest that one may build a classifier, practically and with a manageable training sample size n, that demonstrates ε-supervenience with reasonable choices for ε and α and a plausible joint distribution $\mathbb{P}[M, B]$ (Figure 2). To generate the data, we let the class-conditional random variable $E_{ij}|M = m_0$ be distributed Poisson$(A_{ij} + \eta)$, where A_{ij} is the number of chemical synapses between neuron i and neuron j according to [16], with noise parameter $0 < \eta \ll 1$. The class-conditional random variable $E_{ij}|M = m_1$ is distributed Poisson$(A_{ij} + z_{ij})$ for neurons $i, j \in \mathcal{D}$, where $\mathcal{D}$ is the set of edges deemed responsible for odor-evoked behavior according to [17], with signal parameter z_{ij} uniformly sampled from $[-5, 5]$. We consider k_n-nearest neighbor classification of labeled multigraphs (directed, with loops) on 279 vertices, under Frobenius norm. The k_n-nearest neighbor classifier used here satisfies $k_n \to \infty$ as $n \to \infty$ and $k_n/n \to 0$ as $n \to \infty$, ensuring universal consistency. (Better classifiers can be constructed for the joint distribution $\mathbb{P}[M, B]$ used here; however, we demand universal consistency.)

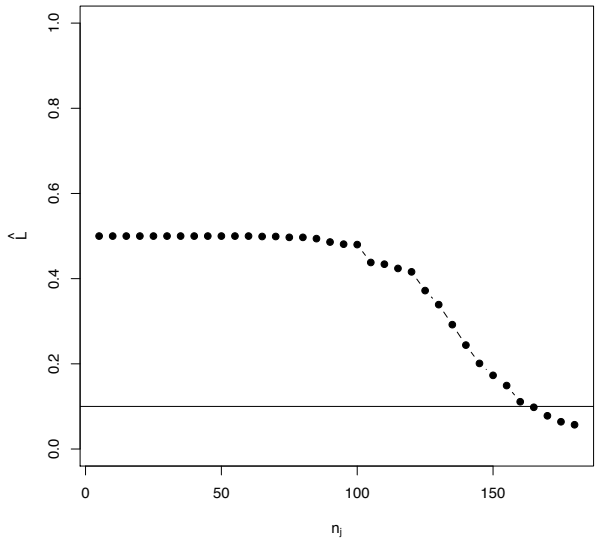

Figure 2: C. elegans graph classification simulation results. $\widehat{L}_{\mathbb{P}}^{1000}(g_n)$ is plotted as a function of class-conditional training sample size n_j, suggesting that for $\varepsilon = 0.1$ we can determine that $\mathcal{M} \overset{\varepsilon}{\sim}_{\mathbb{P}} \mathcal{B}$ holds with 99% confidence with just a few hundred training samples generated from $\mathbb{P}[M, B]$. Each dot depicts an estimate for $L_{\mathbb{P}}(g_n)$; standard errors are $(L_{\mathbb{P}}(g_n)(1 - L_{\mathbb{P}}(g_n))/1000)^{1/2}$; e.g., $n_j = 180$; $k_n = 53$; $\widehat{L}_F^{1000}(g_n) = 0.057$; standard error less than 0.01. We reject $H_0 : L_{\mathbb{P}}(g^*) \geq 0.10$ at $\alpha = 0.01$. $L_{\mathbb{P}}(g^*) \approx 0$ for this simulation.

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D superresolution imaging [18] combined with neurite tracing algorithms [19, 20, 21] allow the collection of a brain-graph within a day.

Genetic manipulations, laser ablations, and training paradigms can each be used to obtain a non-wild type population for use as $M = m_1$ [15], and the class of each organism (m_0 vs. m_1) can also be determined automatically [22].

4 Discussion

Summary We have introduced the notion of ε-supervenience, which states that the misclassification rate for any mental/brain property pair is less than ε. Furthermore, we have shown that certain algorithms are universally consistent, such that one can derive a hypothesis test, with confidence level α, that is guaranteed to converge to the Bayes optimal misclassification rate, given sufficient data, no matter the true (but unknown) distribution of mental/brain pair properties. Furthermore, this is a one-sided test, so although power converges to unity, one can never determine whether more data is necessary to get a lower p-value, or the particular ε supervenience does not hold. A simulation of a realistically performable experiment in C. elegans suggests that convergence rates are not so poor as to make these results completely impractical. Several points merit further discussion, in our humble opinion.

Practical issues A more informative and tractable distance on $\mathcal{B}$ may be desired, as the k_n-nearest neighbor classifier under an L_p norm may have a rate of convergence so slow and a computational demand so high as to be impractical. Improved distance measures may be considered by utilizing prior neurobiological knowledge. Further, collecting enough sufficiently accurate, exchangeable brain-graph data and the associated mental property values may be beyond current technological capabilities. Related experimental work includes collecting various types of brain graph data [23, 24, 25] and various approaches to inference on brain graphs [26, 20, 21], suggests feasibility of such an experiment in the future.

Quantum networks Recently, several authors have suggested the possibility that brains are better characterized as quantum networks, instead of classical networks. The above results hold regardless of whether computations in the brain are quantum or classical, as quantum networks merely speed up computation for certain classes of problems, they cannot, however, solve problems that classical computers cannot [27]. This means, if the above analysis failed to reject the null at level α, the interpretation does not change if one assumes quantum versus classical computations.

Stochastic Supervenience Possible explanations of how it might be the case that $\varepsilon > 0$ include stochastic supervenience [28], and supernatural causal effects. Thus, the above analysis could be considered an empirical test for whether we have souls, or, perhaps whether souls play a causal role in our mental properties over and above the physical role played by the brain, or whether the data we have suggests that the probability that our souls play a measurable causal role over and above the physical is less than ε.

Dynamics vs. statics The above *in silico* experiment did not require simulating any *dynamics*; rather, the dynamics are necessarily a function of the model parameters (statics). Similarly, for the question of mind-brain supervenience in humans, one need not every observe any activity of the brain, one must merely observe the model, which determines the activity (in a potentially stochastic process). Thus, this approach to understanding the relationship between mind and brain could be considered a rather drastic departure from most of systems neuroscience, in which the goal is typically to understanding the neural activity "code." In contrast, if mind-brain supervenience is true, this motivates a search for the neural connectivity "code," an *engram* for memories [29, 30, 31, 32, 33], or more generally a *mengram*, the neural signature of any mental property, be it cognitive, psychological, or otherwise.

Concluding thoughts This thought experiment, together with (i) the formal definition of ε-supervenience, (ii) the brain-graph model, and (iii) the universal consistency proof on graphs, is the first demonstration (to our knowledge) that empirically investigating supervenience is at least theoretically possible.

A k_n nearest neighbor algorithm

Assume that b is a real d-dimensional vector, $b \in \mathcal{B} \subseteq \mathbb{R}^d$, and m is a binary indicator, $m \in \mathcal{M} = \{0, 1\}$. Then, further assume that we have observed a collection of training data, $\mathcal{T}_n = \{(m_i, b_i)\}_{i=1}^n$, each sampled identically and

independently from some unknown joint distribution, $(m_i, b_i) \overset{iid}{\sim} \mathbb{P}[M, B]$. A new brain, b, called the "test brain", is then observed, and one desires to find the most likely class of the new brain, m. It is further assumed that the test mind/brain pair is sampled from the same distribution as the training data, $(m, b) \sim \mathbb{P}[M, B]$, and m is simply unobserved.

The 1-nearest neighbor (1-NN) classifier works as follows. Compute the distance between the test brain and all the training brains, $d_i = d(b, b_i)$ for all $i \in [n]$, where $[n] = 1, 2, \ldots, n$. Then, sort them, $d_{(1)} < d_{(2)} < \ldots < d_{(n)}$, where parenthetical indices, (i), indicate rank order. One can then also obtain a rank order for the training minds, $m_{(1)}, m_{(2)}, \ldots, m_{(n)}$, where $m_{(i)}$ is the class of the i^{th} closest training brain to b. The 1-NN algorithm predicts that the unobserved mind is of the same class as the closest brain's class: $m = m_{(1)}$. The k_n nearest neighbor is a straightforward generalization of this approach. It says that the test mind is in the same class as which ever class is the majority class of the k_n nearest neighbors, $m = \mathbb{I}\{\sum_{i=1}^{k_n} m_{(i)} > k_n/2\}$. Given a particular choice of k_n (the number of nearest neighbors to consider), and a choice of $d(\cdot, \cdot)$ (the distance metric used to compare the test datum and training data), one then has a relatively simple and intuitive algorithm.

Unfortunately, no such algorithm is universally consistent. Let g_n be the k_n nearest neighbor classifier when there are n training points. Then, a collection of such algorithms, $\{g_n\}$, with k_n increasing with n, can be universally consistent under certain constraints. In particular, as n increases, k_n must also increase, but not quite as quickly. Formally, k_n must satisfy: (i) $k_n \to \infty$ as $n \to \infty$ and (ii) $k_n/n \to 0$ as $n \to \infty$. In Stone' original proof, the L_2 norm ($d(b, b') = \sum_{j=1}^{d} (b_j - b'_j)^2$, where j indexes elements of the d-dimensional vector) was shown to satisfy the constraints on a distance metric for this collection of classifiers to be universally consistent. Later, others extended these results to apply to any L_p norm [9].

B Acknowledgments

The authors would like to acknowledge helpful discussions with J Lande and B Vogelstein.

References

[1] Plato, *Plato: complete works*. Hackett Pub Co, 1997.

[2] R. Descartes, *Meditationes de prima philosophia*. 1641.

[3] D. Davidson, *Experience and Theory*, ch. Mental Events. Duckworth, 1970.

[4] C. F. Craver, *Explaining the brain: mechanisms and the mosaic unity of neuroscience*. Clarendon Press, 2007.

[5] H. W. G. von Waldeyer-Hartz, "Ueber einige neuere forschungen im gebiete der anatomie des centralnervensystems," *Deutsche medicinische Wochenschrift*, 1891.

[6] S. Finger, *Origins of neuroscience: A history of explorations into brain function*. Oxford University Press, USA, 2001.

[7] C. Bishop, *Neural networks for pattern recognition*. Oxford University Press, USA, 1995.

[8] J. McClelland, D. Rumelhart, *et al.*, *Parallel distributed processing: Explorations in the microstructure of cognition*. MIT press Cambridge, MA, 1986.

[9] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[10] L. Devroye, "On arbitrarily slow rates of global convergence in density estimation," *Probability Theory and Related Fields*, vol. 62, no. 4, pp. 475–483, 1983.

[11] K. Popper, "The logic of scientific discovery," 1959.

[12] C. Stone, "Consistent nonparametric regression," *The annals of statistics*, vol. 5, no. 4, pp. 595–620, 1977.

[13] M. Garey and D. Johnson, *Computers and intractability. A guide to the theory of NP-completeness. A Series of Books in the Mathematical Sciences*. WH Freeman and Company, San Francisco, Calif, 1979.

[14] R. M. Durbin, *Studies on the Development and Organisation of the Nervous System of Caenorhabditis elegans*. PhD thesis, University of Cambridge, 1987.

[15] M. de Bono and A. V. Maricq, "Neuronal substrates of complex behaviors in c. elegans.," *Annu Rev Neurosci*, vol. 28, pp. 451–501, 2005.

[16] L. Varshney, B. Chen, E. Paniagua, D. Hall, and D. Chklovskii, "Structural Properties of the Caenorhabditis elegans Neuronal Network," *ArXiv*, 2009.

[17] S. H. Chalasani, N. Chronis, M. Tsunozaki, J. M. Gray, D. Ramot, M. B. Goodman, and C. I. Bargmann, "Dissecting a circuit for olfactory behaviour in caenorhabditis elegans.," *Nature*, vol. 450, pp. 63–70, Nov 2007.

[18] A. Vaziri, J. Tang, H. Shroff, and C. V. Shank, "Multilayer three-dimensional super resolution imaging of thick biological samples.," *Proc Natl Acad Sci U S A*, vol. 105, pp. 20221–20226, Dec 2008.

[19] M. Helmstaedter, K. L. Briggman, and W. Denk, "3d structural imaging of the brain with photons and electrons.," *Curr Opin Neurobiol*, vol. 18, pp. 633–641, Dec 2008.

[20] Y. Mishchenko, "Automation of 3d reconstruction of neural tissue from large volume of conventional serial section transmission electron micrographs.," *J Neurosci Methods*, vol. 176, pp. 276–289, Jan 2009.

[21] J. Lu, J. C. Fiala, and J. W. Lichtman, "Semi-automated reconstruction of neural processes from large numbers of fluorescence images," *PLoS ONE*, vol. 4, p. e5655, 05 2009.

[22] S. D. Buckingham and D. B. Sattelle, "Strategies for automated analysis of c. elegans locomotion.," *Invert Neurosci*, vol. 8, pp. 121–131, Sep 2008.

[23] J. White, E. Southgate, J. N. Thomson, and S. Brenner, "The structure of the nervous system of the nematode caenorhabditis elegans.," *Philosophical Transactions of Royal Society London. Series B, Biological Sciences*, vol. 314, no. 1165, pp. 1–340, 1986.

[24] W. W. Denk and H. Horstmann, "Serial block-face scanning electron microscopy to reconstruct three-dimensional tissue nanostructure," *PLOS Biol.*, vol. 2, p. e329, 2004.

[25] K. Briggman and W. Denk, "Towards neural circuit reconstruction with volume electron microscopy techniques," *Current opinion in neurobiology*, vol. 16, no. 5, pp. 562–570, 2006.

[26] J. H. Macke, N. Maack, R. Gupta, W. Denk, B. Schlkopf, and A. Borst, "Contour-propagation algorithms for semi-automated reconstruction of neural processes.," *J Neurosci Methods*, vol. 167, pp. 349–357, Jan 2008.

[27] M. A. Nielson and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[28] C. F. Craver, "Stochastic supervenience," 2009.

[29] R. W. Semon, *The Mneme*. G. Allen & Unwin ltd., 1921.

[30] K. Lashley, "In search of the engram," *Symposia of the society for experimental biology*, vol. 4, no. 454-482, p. 30, 1950.

[31] W. Zhang and D. Linden, "The other side of the engram: experience-driven changes in neuronal intrinsic excitability," *Nature Reviews Neuroscience*, vol. 4, no. 11, pp. 885–900, 2003.

[32] R. Shema, T. Sacktor, and Y. Dudai, "Rapid erasure of long-term memory associations in the cortex by an inhibitor of PKM {zeta}," *Science*, vol. 317, no. 5840, p. 951, 2007.

[33] J. Berry, W. Krause, and R. Davis, "Olfactory memory traces in Drosophila," *Progress in brain research*, vol. 169, pp. 293–304, 2008.