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

The"mind-brain supervenience" conjecture suggests that all mental properties are derived from the physical properties of the brain. To address the question of whether the mind supervenes on the brain, we frame a supervenience hypothesis in rigorous statistical terms. Specifically, we propose a modified version of supervenience (called ε-supervenience) that is amenable to experimental investigation and statistical analysis. To illustrate this approach, we perform 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 brain-graph or connectome). ε-supervenience allows us to determine whether a particular mental property can be inferred from one's connectome to within any given positive misclassification rate, regardless of the relationship between the two. This may provide motivation for cross-disciplinary research between neuroscientists and statisticians.

The mind-brain supervenience notion was canonized in 1970 with the following quote from Donald Davidson [1]:

> [mind-brain] 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.

We consider a special case of supervenience of considerable interest. Specifically, this work addresses a version of a local supervenience between mental properties and brain properties, with particular emphasize on brain-graphs (i.e., connectivity structure). The determination of supervenience (or lack thereof) of a mental property on a brain property has potentially important implications in a number of fields of inquiry. Neural network theory and artificial intelligence often implicitly take a generalized notion of local mind-brain supervenience as an assumption, which if falsified, might change modern approaches to learning [2, 3]. Cognitive neuroscience similarly seems to operate under such assumptions, which if falsified, might result in novel perspectives and theories [4, 5]. And the question of mind-brain supervenience continues to be debated amongst philosophers [6].

This work does not attempt to resolve any particular mind-brain supervenience debates. Rather, we propose a statistical approach for framing mind-brain supervenience questions. This approach depends on defining the space of mental and brain properties under investigation and a statistical model characterizing the possible distributions governing their relationship. Such definitions transform supervenience from a conjecture or an assumption into a hypothesis which can be tested.

Results

Statistical supervenience: a definition

Let $\mathcal{M} = \{m_1, m_2, \ldots\}$ be the space of all possible mental properties, including all possible thoughts, memories beliefs, etc. Similarly, let $\mathcal{B} = \{b_1, b_2, \ldots\}$ be the set of all possible brain properties, including the position and momentum of all subatomic particles within the skull. Given these definitions, Davidson's conjecture may be concisely and formally stated thusly: $m \neq m' \Rightarrow b \neq b'$, where $(m, b), (m', b') \in \mathcal{M} \times \mathcal{B}$ are mind-brain pairs. This mind-brain supervenience relation does not imply an injective relation, a causal relation, or an identity relation (see Appendix 1 for more details and some examples). To facilitate both statistical analysis and empirical investigation, we convert this supervenience relation from a logical to a probabilistic relation.

Let F_{MB} indicate a joint distribution of minds and brains. Statistical supervenience can then be defined as follows:

Definition 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{MB}$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $\mathbb{P}[m \neq m' | b = b'] = 0$, or equivalently $\mathbb{P}[m = m' | b = b'] = 1$.*

Statistical supervenience is therefore a probabilistic relation on sets (which could be considered a generalization of correlation; see Appendix 1 for details).

Statistical supervenience is equivalent to perfect classification accuracy

Given that minds statistically supervene on brains, $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, then if two minds differ, there must be some brain-based difference to account for the mental difference. This means that there must exist a deterministic function mapping each brain to its supervening mind, g. One could therefore, in principle, know this function. When the space of all possible minds is finite, that is, $|\mathcal{M}| < \infty$, any function $g : \mathcal{B} \to \mathcal{M}$ mapping from minds to brains is called a *classifier*. Define misclassification rate, the probability that g misclassifies b under distribution $F = F_{MB}$, as

$$L_F(g) = \mathbb{P}[g(B) \neq M] = \sum_{(m,b) \in \mathcal{M} \times \mathcal{B}} \mathbb{I}\{g(b) \neq m\} \mathbb{P}[B = b, M = m] \tag{1}$$

where $\mathbb{I}$ denotes the indicator function taking value unity whenever its argument is true, and zero otherwise. The Bayes optimal classifier g^* minimizes $L_F(g)$ over all classifiers, that is: $g^* = \operatorname{argmin}_g L_F(g)$. The *Bayes error*, or Bayes risk, $L^* = L_F(g^*)$, is the minimum possible misclassification rate.

The primary result of casting supervenience in a statistical framework is the following theorem that we state without proof:

Theorem 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{MB}$, denoted $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, if and only if $L^* = 0$.*

The above argument shows (for the first time to our knowledge) that statistical supervenience and zero Bayes error are equivalent. Statistical supervenience can therefore be thought of as a constraint on the possible distributions on minds and brains. Specifically, let $\mathcal{F}$ indicate the set of all possible joint distributions on minds and brains, and let $\mathcal{F}_s = \{F_{MB} \in \mathcal{F} : L^* = 0\}$ be subset of distributions for which supervenience holds. Theorem 1 implies that $\mathcal{F}_s \subsetneq \mathcal{F}$.

The existence of a consistent statistical test for supervenience

The above theorem implies that if we desire to know whether a particular mental property is supervenient on a particular brain property, we must check whether $L^* = 0$. Unfortunately, L^* is typically unknown. Fortunately, we can approximate L^* using training data.

Assume that training data $\mathcal{T}_n = \{(M_1, B_1), \ldots, (M_n, B_n)\}$ are each sampled identically and independently (iid) from the true (but unknown) joint distribution F_{MB}. Let g_n be a classifier induced by the training data, $g_n : \mathcal{B} \times (\mathcal{M} \times \mathcal{B})^n \mapsto \mathcal{M}$. The expected misclassification rate of such a classifier is given by:

$$\mathbb{E}[L_F(g_n)] = \sum_{(m,b) \in \mathcal{M} \times \mathcal{B}} \mathbb{I}\{g_n(b; \mathcal{T}_n) \neq m\} \mathbb{P}[B = b, M = m], \tag{2}$$

where the expectation is taken with respect to the training data $\mathcal{T}_n$. Calculating the expected misclassification rate is often intractable in practice because it requires a sum over all possible mind-brain property pairs. Instead, expected misclassification rate is approximated by "hold-out" error. Let $\mathcal{H}_{n'} = \{(M_{n+1}, B_{n+1}), \ldots, (M_{n+n'}, B_{n+n'})\}$ be a set of n' hold-out samples, each sampled iid from F_{MB}. The hold-out approximation to misclassification rate is given by:

$$\mathbb{E}[L_F(g_n)] \approx \widehat{L}_F^{n'}(g_n) = \sum_{(M_i, B_i) \in \mathcal{H}_{n'}} \mathbb{I}\{g_n(B_i; \mathcal{T}_n) \neq M_i\}. \tag{3}$$

which can be used as a surrogate for L^*. By definition of g^*, the expectation of $\widehat{L}_F^{n'}(g_n)$ (with respect to both $\mathcal{T}_n$ and $\mathcal{H}_{n'}$) is greater than or equal to g^* for any g_n and all n. Thus, we can construct a hypothesis test for L^* using the surrogate $\widehat{L}_F^{n'}(g_n)$.

A statistical test proceeds by calculating a test statistic and a critical value c_α. We reject if the test statistic is more extreme than the critical value, or equivalently, if the p-value is less than α, the allowable Type I error rate. The p-value, the probability of rejecting the least favorable true null hypothesis (the hypothesis within the potentially composite null closest to the boundary with the alternate hypothesis), is a functional of the distribution of the test statistic.

Ideally, we might consider the following hypothesis test: $H_0 : L^* > 0$ and $H_A : L^* = 0$. Unfortunately, because the null hypothesis is not closed, the alternate hypothesis lies on the boundary, so the p-value is always equal to unity [7]. To proceed, therefore, we introduce a relaxed notion of supervenience:

Definition 2. *Given $\varepsilon > 0$, $\mathcal{M}$ is said to ε-supervene on $\mathcal{B}$ for distribution $F = F_{MB}$, denoted $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$, if and only if $L_F(g^*) < \varepsilon$.*

Given this relaxation, consider the problem of testing for ε-supervenience:

$$H_0^\varepsilon : L^* \geq \varepsilon$$
$$H_A^\varepsilon : L^* < \varepsilon$$

Let $\widehat{n} = \widehat{L}_F^{n'}(g_n)$ be the test statistic. The distribution of $\widehat{n}$ is available under the least favorable null distribution to the alternate hypothesis. For the above hypothesis test, the p-value is given by the binomial cumulative

distribution function, $\mathbb{B}(\widehat{n}; n', \varepsilon) = \sum_{k \in [\widehat{n}]_0} \text{Binomial}(k; n'; \varepsilon)$, where $[\widehat{n}]_0 = \{0, 1, \ldots, \widehat{n}\}$. We therefore reject whenever this p-value is less than α; rejection implies that we are $100(1-\alpha)\%$ confident that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$.

A test is *consistent* whenever its power (the probability of a rejecting the null when it is indeed false) goes to unity as n increases. For any statistical test, if the p-value converges in distribution to δ_0 (point mass at zero), then whenever $\alpha > 0$, power goes to unity. For the above ε-supervenience statistical test, if $g_n \to g^*$ as $n \to \infty$, then $\widehat{L}_F^{n'}(g_n) \to L^*$ as $n, n' \to \infty$. Thus, if $L^* < \varepsilon$, the p-value converges: that is, $\mathbb{B}(\widehat{n}; n', \varepsilon) \to \delta_0$.

The definition of ε-supervenience therefore admits, for the first time to our knowledge, a viable statistical test of supervenience, given a specified ε and α. Moreover, this test is consistent whenever g_n converges to the Bayes classifier, g^*.

Unfortunately, the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ depends on the (unknown) distribution $F = F_{MB}$ [8]. Furthermore, arbitrarily slow convergence theorems regarding the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ demonstrate that there is no universal n, n' which will guarantee that the test has power greater than any specified target $\beta > \alpha$ [9]. 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 $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ holds, but we can never be confident in its negation; rather, it may be the case that the evidence in favor of $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ is insufficient for any number of reasons, including that we simply have not yet collected enough data. Thus, without restrictions on F_{MB}, arbitrarily slow convergence theorems imply that our theorem of ε-supervenience does not strictly satisfy Popper's *falsifiability* requirement [10].

The availability of a consistent statistical test for supervenience

The above considerations indicate the existence of a consistent test for supervenience whenever the classifier used is consistent. To actually implement such a test, one must be able to (i) measure mind/brain pairs and (ii) have a consistent classifier g_n. Unfortunately, we do not know how to measure the entirety of one's brain, much less one's mind. Moreover, even if we did, we do not know the joint distribution of mind/brain pairs, F_{MB}. We therefore must restrict our interest to a mental property/brain property pair. A mental property might be a person's intelligence, psychological state, current thought, gender identity, etc. A brain property might be the number of cells in a person's brain at some time t, or the collection of spike trains of all neurons in the brain during some time period t to t'. Regardless of the details of the specifications of the mental property and the brain property, given such specifications, one can assume a model, $\mathcal{F}$. We would like a classifier, g_n, that is guaranteed to be consistent, no matter which of the possible distributions $F_{MB} \in \mathcal{F}$ is the true distribution. A classifier with such a property is called an *universally consistent classifier*. Below, we show that under a very general mind-brain model $\mathcal{F}$, an universally consistent classifier is readily available.

***Gedankenexperiment* 1.** *Let the physical property under consideration be brain connectivity structure, so b is a brain-graph ("connectome") with vertices representing neurons (or collections thereof) and edges representing synapses (or collections thereof). Furthermore, let $\mathcal{B}$, the brain observation space, be the collection of all graphs on a given finite number of vertices, and $\mathcal{M}$ the mental property observation space, be finite. Now, imagine collecting very large amounts of very accurate identically and independently sampled brain-graph data and the associated mental property indicators from F_{MB}. A k_n-nearest neighbor classifier using a Frobenius norm is universally consistent (see Methods for details). Therefore, the existence of a universally consistent classifier guarantees that eventually (in n, n') we will be able to conclude $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$ for this mind-brain property pair, if indeed ε-supervenience holds. This logic holds for directed graphs or multigraphs or hypergraphs with discrete edge weights and vertex attributes, as well as unlabeled graphs (see [?] for details). Furthermore, the proof holds for other matrix norms (which might speed up convergence), and the regression scenario where $|\mathcal{M}|$ is infinite (again, see Methods for details).*

The feasibility of a consistent statistical test for supervenience

The above argument demonstrates the availability of a consistent test under certain restrictions. The arbitrary slow convergence theorem, however, demonstrates that convergence rates might be unbearably slow. We therefore provide a (hopefully illustrative) example of the feasibility of such a test on synthetic data.

Caenorhabditis elegans is a species whose nervous system is believed to consist of the same 279 labeled neurons for each organism [11]. Moreover, these animals exhibit a rich behavioral repertoire that seemingly depends on circuit properties [12]. These findings motivate the use of C. elegans for a synthetic data analysis [?].

Conducting such an experiment requires specifying a joint distribution F_{MB} over brain-graphs and behaviors. The joint distribution decomposes into the product of a class-conditional distribution (likelihood) and a prior, $F_{MB} = F_{B|M}F_M$. The prior specifies the probability of any particular organism exhibiting the behavior. The class-conditional distribution specifies the brain-graph distribution given that the organism does (or does not) exhibit the behavior.

Let A_{uv} be the number of chemical synapses between neuron u and neuron v according to [13]. Then, let $\mathcal{E}$ be the set of edges deemed responsible for odor-evoked behavior according to [14]. If odor-evoked behavior is supervenient on this subgraph $\mathcal{E}$, then the distribution of edges in $\mathcal{E}$ must differ between the two classes. Let each edge in each class be Poisson distributed, $F_{B_{uv}|M=m_j}$ = Poisson($\lambda_{uv|j}$). For class m_0, let $\lambda_{uv|0} = A_{uv} + \eta$, where $\eta = 0.05$ is a small noise parameter because it is believed that the C. elegans connectome is similar across organisms [11]. For class m_1, [illegible] where the signal parameter $\nu_{uv} = \eta$ for all edges not in $\mathcal{E}$, and ν_{uv} is uniformly sampled from $[-5, 5]$ for all edges within $\mathcal{E}$.

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 F_{MB} used here; however, we demand universal consistency.) Figure 1 shows that for this simulation, rejecting $\varepsilon = 0.1$ supervenience at $\alpha = 0.01$ only requires a few hundred training samples.

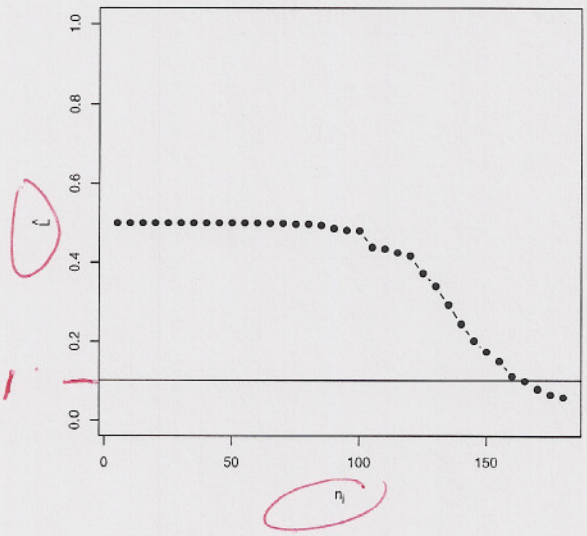

Figure 1: C. elegans graph classification simulation results. The estimated expected misclassification rate $\widehat{L}_F^{n'}(g_n)$ (with $n' = 1000$ testing samples) is plotted as a function of class-conditional training sample size $n_j = n/2$, suggesting that for $\varepsilon = 0.1$ we can determine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

Importantly, conducting this experiment *in actu* is not beyond current technological limitations. 3D superresolution imaging [15] combined with neurite tracing algorithms [16, 17, 18] allow the collection of a C. elegans 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 $M = m_1$ [12], and the class of each organism (m_0 vs. m_1) can also be determined automatically [19].

Discussion

This work makes the following contributions. First, we define statistical supervenience. This definition makes it apparent that supervenience implies the possibility of perfect classification. The determination of supervenience from data requires a statistical test. We therefore demonstrate the existence, availability, and feasibility of a consistent test of a relaxed notion of supervenience, ε-supervenience, given specified and measurable mental properties and brain properties. In other words, the proposed test is guaranteed to reject the null whenever the null is false, given sufficient data, for any possible distribution governing mental property/brain property pairs.

Alas, this is a one-sided test, so although power converges to unity, a failure to reject the null has many possible explanations. First, the amount of data might be insufficient given the particular distance implemented; collecting more data or utilizing a more informative distance might resolve this difficulty. Second, ε-supervenience is defined between a mental property and a brain property. Thus, a failure to reject ε-supervenience on a particular brain property does not entail non-ε-supervenience on any other brain property. Third, in general, neither mental properties nor brain properties are directly measurable; rather, one typically measures a function of such properties. For instance, instead of measuring intelligence, one may consider the results of an IQ test as a proxy for intelligence.

Thus, given measurements of mental and brain properties that we believe reflect the properties of interest, and a sufficient amount data satisfying the independent and identically sampled assumption, a rejection of ε-supervenience entails that we are $100(1-a)\%$ confident that the mental property under investigation does not ε-supervene on the brain property under investigation. Unfortunately, failure to reject is more ambiguous. $\varepsilon-$supervenience tests can therefore be thought of as constraining the space of possible brain properties upon which a mental property supervenes. Determining that a mental property does not supervene on *any* brain property is beyond the capacities of this formalism.

Interestingly, much of contemporary research in neuroscience and cognitive science could be cast as mind-brain supervenience investigations. Specifically, any investigation that aims to "explain" some behavioral or psychological phenomenon by reference to neural (or other brain) activity or structure could be thought of as mind-brain supervenience investigations. Thus, the proposed ε-supervenience framework is perhaps a useful unifying perspective for these fields. Perhaps this is especially true in light of the advent of "connectomics" [20, 21], a field devoted to estimating whole organism brain-graphs and relating them to function. Testing supervenience of various mental properties on these brain-graphs will perhaps therefore become increasingly compelling; the framework developed herein could be fundamental to these investigations. For example, questions about whether connectivity structure alone is sufficient to explain a particular mental property is one possible mind-brain ε-supervenience investigation. The above synthetic data analysis demonstrates the feasibility of ε-supervenience on small brain-graphs. Similar supervenience tests on larger animals (such as humans) will potentially benefit from either higher-throughput imaging modalities [22, 23], more coarse brain-graphs [24, 25], or both.

Methods

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, $d_{(1)} < d_{(2)} < \ldots < d_{(n)}$, and their corresponding mental properties, , $m_{(1)}, m_{(2)}, \ldots, m_{(n)}$, where parenthetical indices indicate rank order. The 1-NN algorithm predicts that the unobserved mind is of the same class as the closest brain's class: $\widehat{m} = m_{(1)}$. The k_n nearest neighbor is a straightforward generalization of this approach. It says that the test mind is in the same class as whichever class is the plurality class of the k_n nearest neighbors, $\widehat{m} = \text{argmax}_{m'} \mathbb{I}\{\sum_{i=1}^{k_n} m_{(i)} = m'\}$. 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.

Let g_n be the k_n nearest neighbor (k_nNN) classifier when there are n training samples. A collection of such classifiers $\{g_n\}$, with k_n increasing with n, is called a classifier sequence. A universally consistent classifier sequence is any classifier sequence that is guaranteed to converge to the Bayes optimal classifier regardless of the true distribution from which the data were sampled; that is, a universally consistent classifier sequence satisfies $g_n \to g^*$ as $n \to \infty$ for all F_{MB}.

The k_nNN classifier is consistent if (i) $k_n \to \infty$ as $n \to \infty$ and (ii) $k_n/n \to 0$ as $n \to \infty$ [26]. In Stone's

original proof [26], b was assumed to be a d-dimensional vector, and 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 [8]. When brain-graphs are represented by their adjacency matrices, one can stack the columns of the adjacency matrices, effectively embedding graphs into a vector space, in which case Stone's theorem applies. Stone's original proof also applied to the scenario when $|\mathcal{M}|$ was infinite, resulting in a universally consistent regression algorithm as well.

Note that the above extension of Stone's original theorem to the graph domain implicitly assumed that vertices were labeled, such that elements of the adjacency matrices could easily be compared across graphs. In theory, when vertices are unlabeled, one could first map each graph to a quotient space invariant to isomorphisms, and then proceed as before. Unfortunately, there is no known polynomial time complexity algorithm for graph isomorphism [27], so in practice, dealing with unlabeled vertices will likely be computationally challenging.

References

[1] Davidson, D. *Experience and Theory*, chapter Mental Eve. Duckworth (1970).

[2] Haykin, S. *Neural Networks and Learning Machines*. Prentice Hall, 3rd edition, (2008).

[3] Ripley, B. D. *Pattern Recognition and Neural Networks*. Cambridge University Press, (2008).

[4] Fodor, J. A. *Concepts: Where Cognitive Science Went Wrong (Oxford Cognitive Science)*. Oxford University Press, USA, (1998).

[5] Gazzaniga, M. S., Ivry, R. B., and Mangun, G. R. *Cognitive Neuroscience: The Biology of the Mind (Third Edition)*. W. W. Norton & Company, (2008).

[6] Kim, J. *Physicalism, or Something Near Enough (Princeton Monographs in Philosophy)*. Princeton University Press, (2007).

[7] Bickel, P. J. and Doksum, K. A. *Mathematical Statistics: Basic Ideas and Selected Topics, Vol I (2nd Edition)*. Prentice Hall, (2000).

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

[9] Devroye, L. *Utilitas Mathematica* **483**, 475–483 (1983).

[10] Popper, K. R. (1959). title

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

[12] de Bono, M. and Maricq, A. V. *Annu Rev Neurosci* **28**, 451–501 (2005).

[13] Varshney, L. R., Chen, B. L., Paniagua, E., Hall, D. H., Chklovskii, D. B., Spring, C., and Farm, J. *World Wide Web Internet And Web Information Systems* , 1–41.

[14] Chalasani, S. H., Chronis, N., Tsunozaki, M., Gray, J. M., Ramot, D., Goodman, M. B., and Bargmann, C. I. *Nature* **450**(7166), 63–70 November (2007).

[15] Vaziri, A., Tang, J., Shroff, H., and Shank, C. V. *Proceedings of the National Academy of Sciences of the United States of America* **105**(51), 20221–6 December (2008).

[16] Helmstaedter, M., Briggman, K. L., and Denk, W. *Current opinion in neurobiology* **18**(6), 633–41 December (2008).

[17] Mishchenko, Y. *J Neurosci Methods* **176**(2), 276–289 January (2009).

[18] Lu, J., Fiala, J. C., and Lichtman, J. W. *PLoS ONE* **4**(5), e5655 (2009).

[19] Buckingham, S. D. and Sattelle, D. B. *Invertebrate neuroscience : IN* **8**(3), 121–31 September (2008).

[20] Sporns, O., Tononi, G., and Kotter, R. *PLoS Computational Biology* **1**(4), e42 (2005).

[21] Hagmann, P. *From diffusion MRI to brain connectomics*. PhD thesis, Institut de traitement des signaux, (2005).

[22] Hayworth, K. J., Kasthuri, N., Schalek, R., Lichtman, J. W., Program, N., Angeles, L., and Biology, C. *World* **12**(Supp 2), 86–87 (2006).

[23] Bock, D. D., Lee, W.-C. A., Kerlin, A. M., Andermann, M. L., Hood, G., Wetzel, A. W., Yurgenson, S., Soucy, E. R., Kim, H. S., and Reid, R. C. *Nature* **471**(7337), 177–182 March (2011).

[24] Palm, C., Axer, M., Gräß el, D., Dammers, J., Lindemeyer, J., Zilles, K., Pietrzyk, U., and Amunts, K. *Frontiers in Human Neuroscience* **4** (2010).

[25] Johansen-Berg, H. and Behrens, T. E. *Diffusion MRI: From quantitative measurement to in-vivo neuroanatomy*. Academic Press, (2009).

[26] Stone, C. J. *The Annals of Statistics* **5**(4), 595–620 July (1977).

[27] Garey, M. R. and Johnson, D. S. *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).

Acknowledgments

The authors would like to acknowledge helpful discussions with J. Lande, B. Vogelstein, S. Seung, and two helpful referees.

Author Contributions

JTV, RJV, and CEP conceived of the manuscript. JTV and CEP wrote it. CEP ran the experiment.

Additional Information

The authors have no competing financial interests to declare.

Supplementary Materials for: 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

1 Relations between sets

In this appendix we aim to provide more intuition regarding supervenience by discussing the limitations and extent of its implications.

First, a supervenient relation does not imply an injective relation. An injective relation is any relation that preserves distinctness. Thus if minds are injective on brains, then $b \neq b' \implies m \neq m'$ (note that the directionality of the implication has been switched relative to supervenience). However, 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 (likely?) 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, supervenience does not imply causality. 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. The above example with the two coins demonstrates this, as the two coins are not the same thing, even if one has perfect information about one from the other.

What supervenience does imply, however, is the following. Imagine finding two different minds. If $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$, then the brains supervening 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.

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 $\mathcal{M} \overset{S}{\sim}_F \mathcal{B}$ and $\mathcal{B} \overset{S}{\sim}_F \mathcal{M}$. Thus, perfect correlation implies supervenience, but supervenience does not imply correlation. In fact, supervenience may be thought of as a generalization of correlation which can be applied to arbitrary valued random variables (such as mental or brain properties), unlike correlation which is only defined for scalar valued random variables.

References

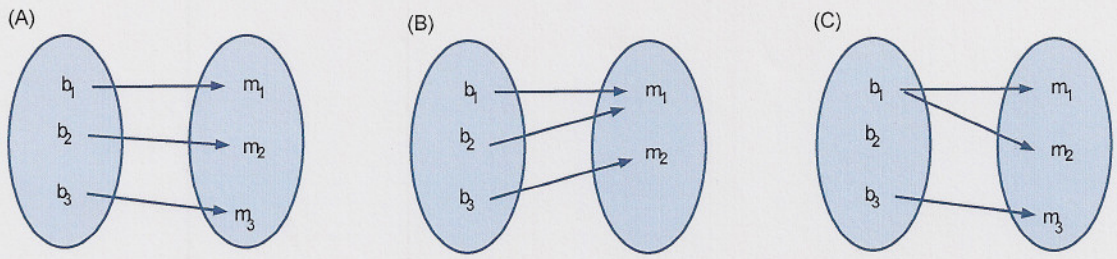

Figure 1: Possible relations between minds and brains. (A) Minds supervene on brains, and it so happens that there is a bijective relation from brains to minds. (B) Minds supervene on brains, and it so happens that there is a surjective (a.k.a., onto) relation from brains to minds. (C) Minds are *not*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

must be!

but NOT bijective