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 misclassification rate > 0, regardless of the relationship between the two. This may provide motivation for cross-disciplinary research between neuroscientists and statisticians.

Donald Davidson canonized the mind-brain supervenience relation 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.

While supervenience can be a very general claim about arbitrary sets of objects, we consider a special case of considerable interest. Specifically, this work addresses a novel version of a local supervenience between mental properties and brain properties. 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 drastically change modern approaches to learning [**?**]. Cognitive neuroscience similarly seems to operate under such assumptions, which if false, could cause a revolution [**?**, **?**]. And the question of mind-brain supervenience continues to be debated amongst philosophers [**?**].

This work does not attempt to resolve any 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

The $\mathcal{M} = \{m_1, m_2, \ldots\}$ be a set of possible mental properties. For example, m might therefore indicate a person's intelligence, psychological state, current thought, gender identity, etc. Similarly, let $\mathcal{B} = \{b_1, b_2, \ldots\}$ be a set of possible brain properties. For example, b might denote the number of cells in a person's brain at some time t, or the spike train of all neurons in the brain during some time period t to t'. Given these definitions, Davidson's conjecture may be concisely and formally stated thusly: $m \neq m' \implies b \neq b'$, 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 $F_{M,B}$ indicate a joint distribution of minds and brains. Statistical supervenience can thusly be defined:

Definition 1. *$\mathcal{M}$ is said to statistically supervene on $\mathcal{B}$ for distribution $F = F_{M,B}$, 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 (related to, but distinct from correlation; see Appendix 1 for details).

Statistical supervenience is equivalent to perfect classification accuracy

If 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 exists a deterministic function mapping each brain to its supervening mind, $g : \mathcal{B} \mapsto \mathcal{M}$; therefore, one could in principle construct this function. The optimal such classifier, g^*, has the smallest expected misclassification rate, $L_F(g^*)$, under distribuion F. The minimum misclassification rate is called Bayes error (see Methods for details). The primary result of casting supervenience as a statistical framework is the following theorem:

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

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. □

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 $\mathcal{F}_s$ be subset of distributions for which supervenience holds. Theorem 1 implies that $\mathcal{F}_s = \{F_{M,B} : L_F(g^*) = 0\} \subseteq \mathcal{F}$.

ε-supervenience admits hypothesis testing

Although the above theorem is of potential theoretical interest, the arguments rely on knowing the typically unknown $F_{M,B}$ and g^*, rendering them useless pragmatically. However, both $F_{M,B}$ and g^* could be estimated from data. 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 $F = F_{M,B}$, 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. Let the null hypothesis be H_0: $L_F(g_n) \geq \varepsilon$, and the alternative hypothesis be H_A: $L_F(g_n) < \varepsilon$. Hold-out error, $\widehat{L}_F^{n'}(g_{\widetilde{n}})$, as defined in Methods, is a test statistic for this hypothesis. We reject for values of the test statistic lower than the critical value, that is, we reject if and only if $\widehat{L}_F^{n'}(g_{\widetilde{n}}) < c_\alpha(n', \varepsilon)$. The critical value is available under the least favorable distribution $n'\widehat{L}_F^{n'}(g_{\widetilde{n}}) \sim$ Binomial(n', ε). Thus, rejection implies that we are $100(1-\alpha)$% confident that $\mathcal{M} \overset{\varepsilon}{\sim}_F \mathcal{B}$. 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.

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. 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 $\mathbb{E}[L_F(g_n)] \to L_F(g^*)$ as $n \to \infty$. Below, we show that under a very general mind-brain model, one can construct a consistent classifier whose power approaches unity with sufficient data.

Unfortunately, the rate of convergence of $L_F(g_n)$ to $L_F(g^*)$ depends on the (unknown) distribution $F = F_{M,B}$ [4]. 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 $\beta > \alpha$ [5]. 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 $F_{M,B}$, arbitrarily slow convergence theorems imply that our theorem of ε-supervenience does not strictly satisfy Popper's *falsifiability* requirement [6].

A *Gedankenexperiment* demonstrating consistency and unity power

To ensure consistency and therefore unity power, the classifier g_n must be able to converge to the truth, regardless of the true distribution, F. We therefore make explicit a model for brains, and show that under this very general model, universally consistent classifiers are available.

***Gedankenexperiment* 1.** *Let the physical property under consideration be brain connectivity structure ("connectome"), so b is a brain-graph (or, network) with vertices representing neurons (or neuroanatomical regions) and edges representing synapses (or white matter tracts). Further let $\mathcal{B}$, the brain observation space, be the collection of all graphs on a finite number of vertices, and $\mathcal{M}$, the mental property observation space, be finite. Now, imagine collecting very large amounts of very accurate independent and identically distributed brain-graph data and the associated mental property indicators. A k_n-nearest neighbor classifier using an isomorphism-matching Frobenius norm is universally consistent (see Appendix 2 for proof). 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. Furthermore, Appendix 2 also extends the proof to deal with other matrix norms (which might speed up convergence), and the regression scenario, where $|\mathcal{M}|$ is infinite.

Discussion

We have introduced the notion of ε-supervenience, which states that the Bayes optimal misclassification rate for any mind/brain property pair is less than ε. Furthermore, when we restrict the space of minds and brains to the setting of *Gedankenexperiment* 1, we have shown that k_n-NN classifiers 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 mind/brain pair properties. Alas, this is a one-sided test, so although power converges to unity, one can never determine whether (i) more data is necessary to get a lower p-value, or (ii) that the particular ε-supervenience does not hold.

A failure to reject the null therefore 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 supervenience on a particular brain property does not entail non-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 considers the results of an IQ test as a proxy for intelligence. Similarly, instead of measuring neural spike trains, one estimates spike trains from some measurable neural signal, like voltage. While one could conceivably ask questions such as: "is IQ score at some time supervenient on the the output of a voltage measuring device?", these are less elegant and general than their non-observable analogs, such as: "is intelligence supervenient on spike trains?"

Thus, given measurement of mental and brain properties that we believe reflect the properties of interest, and a sufficient amount data satisfying the exchangeability assumption, a rejection entails that we are $100(1-a)\%$ confident that the mental property under investigation does not 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.

A central (perhaps *the* central) quest in much of neuroscience, psychology, and cognitive science is to discover the brain properties that subvene under various mental properties, although questions are rarely cast within a supervenience formalism. Moreover, the particular brain properties that are often believed to subvene under these mental properties are neural circuits, or brain-subgraphs. To this end, many investigations in these fields include schematic diagrams showing a particular brain-subgraph subvening under a particular mental phenotype. This practice transcends the evolutionary hierarchy of neuroscientific research. For instance, in the invertebrate literature, vertices correspond to particular labeled neurons, and edges correspond to synapses [7]. In the vertebrate literature, vertices often correspond to types of neurons in particular regions, and edges correspond to tendencies of connections [8]. For primates [9] and humans [10] vertices frequently represent functionally distinct neuroanatomical regions, and edges represent regional interconnectivity. Furthermore, this practice also transcends analytical background, including anatomists [11], philosophers [12], statisticians [13], and physicists [14]. The near ubiquity of this practice suggests that a fundamental quest is to determine which brain-subgraphs subvene under which mental properties (although perhaps causality, not supervenience, is the true desideratum). Perhaps supervenience is therefore a framework that can fruitfully be applied to myriad and varied neurocognitive investigations.

In recent years, with the advent of the field of "connectomics" [15, 16], neuroimaging has driven an explosion of studies investigating the human connectome and relating connectomes to cognitive properties [17]. Many of these studies can be framed as ε-supervenience hypothesis tests. For instance, a recent study showed that using data from diffusion tensor imaging [18], one can nearly perfectly differentiate between schizophrenic individuals and normal (control) individuals [19]. As the resolution and signal-to-noise ratio of magnetic resonance imaging continue to improve, especially with more advanced techniques such as High Angular Diffusion Imaging

[20], Q-Ball Imaging [21], and diffusion spectrum imaging [22], similar results could be obtained with other, more subtle cognitive properties. Furthermore, the utilization of other imaging technologies, such as polarized light imaging [23] and high-throughput electron microscopy [24, 25], will continue to improve the effective resolution of these inferred connectomes from human brains. While determining from a brain scan whether a particular individual knows calculus might be quite distant, many other cognitive and psychological supervenience hypotheses have already been tested, and the gap between testing for calculus and testing for schizophrenia seems to be diminishing.

The *Gedankenexperiment* in Section 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 ever observe any activity of the brain, one must merely observe the model that determines the activity (in a potentially stochastic process). Thus, this approach to understanding the relationship between mind and brain is distinct from the standard systems neuroscience paradigm, in which the goal is typically to understand the neural activity "code." In contrast, if mind-brain supervenience holds, it motivates a search for the neural *connectivity* "code," a so-called "engram" for memories [30, 31, 32, 33, 34], or more generally a *mengram*, the neural signature of any mental property, be it cognitive, psychological, or otherwise (note that supervenience allows for the particular mengram of a mental property to vary both across individuals and time).

This *Gedankenexperiment*, together with (i) the formal definition of ε-supervenience as a constraint on distributions, (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. The above discussion suggests that many previously conducted investigations either assume supervenience, or test it. Further, new technologies facilitate testing supervenience of mental properties on brain-graphs more easily.

Methods

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 $L_F(g) = \mathbb{P}[g(B) \neq M]$ which denotes the probability that g misclassifies b. The Bayes optimal classifier g^* minimizes $L_F(g)$ over all classifiers, that is: $g^* = \operatorname{argmin}_g L_F(g)$. Thus, the *Bayes error*, or Bayes risk, $L_F(g^*)$ is the minimum possible misclassification rate.

Let $\mathcal{T}_n = \{(m_1, b_1), (m_2, b_2), \ldots, (m_n, b_n)\}$ be a set of random samples taking their values in $\mathcal{M} \times \mathcal{B}$, each independently and identically distributed according to $F_{M,B}$. 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 training data $\mathcal{T}_n$, and produces a classification: $g_n(b; \mathcal{T}_n) = \widehat{m}$. Misclassification rate for this classifier will be a random variable, because the training data $\mathcal{T}_n$ are random samples. The expected misclassification rate for this classifier is therefore approximated by "hold-out" error: $\widehat{L}_F^{n'}(g_{\widetilde{n}}) = F[g_{\widetilde{n}}(B) = M|\mathcal{T}_{\widetilde{n}}]$, where $\widetilde{n} = n - n'$, and $n' < n$ is the number of held-out training samples (samples not used to obtain $g_{\widetilde{n}}(\cdot)$). The approximate number of misclassified minds therefore has a binomial distribution: $n'\widehat{L}_F^{n'}(g_{\widetilde{n}}) \sim \text{Binomial}(n', L_F(g_{\widetilde{n}}))$.

References

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

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

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

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

[5] Devroye, L. *Probability Theory and Related Fields* **62**(4), 475–483 (1983).

[6] Popper, K. (1959).

[7] Geoffrey North, R. J. G., editor. *Invertebrate neurobiology*. CSHL Press, (2007).

[8] Shepherd, G. *The synaptic organization of the brain*. Oxford University Press New York, (2004).

[9] Felleman, D. and Van Essen, D. *Cerebral cortex* **1**(1), 1 (1991).

[10] S. Mori, S. Wakana, P. v. Z. L. N.-P. *MRI Atlas of Human White Matter*. Elsevier Science, (2005).

[11] Abeles, M. *Corticonics*. Cambridge University Press, (1991).

[12] Koch, C. and Davis, J. *Large-scale neuronal theories of the brain*. The MIT Press, (1994).

[13] Rao, R., Olshausen, B., and Lewicki, M. *Probabilistic models of the brain: Perception and neural function*. The MIT Press, (2002).

[14] Chow, C., Gutkin, B., Hansel, D., Meunier, C., and Dalibard, J., editors. *Methods and Models in Neurophysics*. Elsevier, (2003).

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

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

[17] Sporns, O. *Networks of the Brain*. MIT Press, (2010).

[18] Basser, P. J., Mattiello, J., and LeBihan, D. *Biophys J* **66**(1), 259–267 Jan (1994).

[19] Ardekani, B. A., Tabesh, A., Sevy, S., Robinson, D. G., Bilder, R. M., and Szeszko, P. R. *Hum Brain Mapp* Mar (2010).

[20] Tuch, D., Reese, T., Wiegell, M., Makris, N., Belliveau, J., and Wedeen, V. *Magnetic Resonance in Medicine* **48**(4), 577–582 (2002).

[21] Tuch, D. *Magnetic Resonance in Medicine* **52**(6), 1358–1372 (2004).

[22] Wedeen, V., Hagmann, P., Tseng, W., Reese, T., and Weisskoff, R. *Magnetic Resonance in Medicine* **54**(6), 1377–1386 (2005).

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

[24] W. Denk, W. and Horstmann, H. *PLOS Biol.* **2**, e329 (2004).

[25] Hayworth, K., Kasthuri, N., Schalek, R., and Lichtman, J. *Microscopy and Microanalysis* **12**(S02), 86–87 (2006).

[26] Penrose, R. and Gardner, M. *The emperor's new mind: concerning computers, minds, and the laws of physics*. Oxford University Press, USA, (1999).

[27] Satinover, J. *The quantum brain: the search for freedom and the next generation of man*. Wiley, (2002).

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

[29] Craver, C. F. (2009).

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

[31] Lashley, K. *Symposia of the society for experimental biology* **4**(454-482), 30 (1950).

[32] Zhang, W. and Linden, D. *Nature Reviews Neuroscience* **4**(11), 885–900 (2003).

[33] Shema, R., Sacktor, T., and Dudai, Y. *Science* **317**(5840), 951 (2007).

[34] Berry, J., Krause, W., and Davis, R. *Progress in brain research* **169**, 293–304 (2008).

Acknowledgments

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

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.