RICE UNIVERSITY HONORS PROJECT IN COGNITIVE SCIENCE

X-Phi X-Planation:
Investigating researchers' concepts of explanation as a lens on the integration challenge in cognitive science

Kira Wegner-Clemens

Advised by: Alexander Morgan, Department of Philosophy

INTRODUCTION

Cognitive science is not a discipline in the traditional sense. Rather than provide a method of answering questions about the world, it instead poses a question: what is mind? Cognitive scientists are welcome to address that question from multiple avenues. It has even been proposed that cognitive science must be multidisciplinary, or at least relevant to multiple disciplines (Schunn, 1998). This interdisciplianry orientation theoretically allows cognitive science to incorporate the strengths of each component discipline and build off one another. If we zero in on the question from every possible direction, we knock away distractions and discover "truth." In practice, it ends up being much messier because findings produced by any component discipline needs to fit with findings from each of the others. If we take the premise that psychology, philosophy, computer science, and biology produce truths about the way mind works, we also must assume those truths combine to characterize mind more wholly.

The fitting together of different fields has been referred to as the "integration problem" and has received moderate attention within the field. However, no solution has emerged. Terms and definitions tangle together and quickly create roadblocks to cross-discipline communication. Intelligence or memory does not necessarily mean the same thing for a computer scientist and for a psychologist, even if both want to use the terms to answer questions about mind and consider themselves cognitive scientists. Each cognitive scientist is trained in a specific field and operates with slightly different sets of "baked in" assumptions about mind and focuses on different aspects. Few researchers are explicitly trained in cognitive science and a set of core methodology and assumptions does not exist beyond an orientation to and celebration of interdisciplinary endeavor. The act of fitting these explanations together

raises additional questions - what is the relation between distinct fields studying mind? How can each be combined? Should they be combined? These questions remain an open debate in the field.

Here, we investigate two questions:

1. How can different models of explanation inform the integration challenge?
2. How do researchers use the concept “explanation” in their work?

Due to the complex nature of these questions and limitations with experimental philosophy methodology, we cannot make firm claims. However, this work provides a rough mapping of the ways in which cognitive scientists use and understand explanation in their work. Through a literature review and survey of cognitive scientists, we characterize the range of attitudes towards explanation and inter-field integration.

BACKGROUND

Science is the work of creating explanations about why and how the world is. However, no strict definition of what exactly constitutes an explanation exists. Throughout the literature and in every day life, "to explain" is used to refer to a number of different concepts. Philosophers of science have extensively debated what of this should be considered a full explanation, identifying numerous possible modes of explanation and necessary features.

I. What constitutes an explanation?

In order to understand this debate, we need to understand the debate about explanation more generally. An extensive debate exists in philosophy attempting to establish what constitutes an explanation, particularly a scientific explanation. Accumulation of evidence or even a cohesive set of findings does not necessarily explain anything. To take an example from classic biology, taxonomies of different species do not seem to explain speciation, however Charles Darwin's theory of natural selection does seem to. Similar distinctions have been made throughout the history of science and in nearly every field. However, exactly why one model is an explanation and another is not proves tricky to define.

To explain may be to predict. It is intuitive to think of explanation as prediction or association, using past events to determine future events. Matteo Carindini notes that when a layperson reads a popular science article announcing a behavior was linked to a given brain region, they think neuroscientists have explained the behavior (Carindini, 2012). However, typically these links are based on simple correlation or prediction based on past events. If my model can take

information about the clouds and tell you if it will rain tomorrow, that is unambiguously a successful model – now you know to take your umbrella to work. However, it is less clear whether this model successfully explains rain or whether an extensive simulation would explain the brain. Determining future events based on past events is often regarded as determining the cause of the event. However, it also could easily be a mistaken correlation. David Barrett argues for a disambiguation of prediction and explanation with the example of the Copernican model of the solar system (Barret, 2013). This model could accurately predict positions of stars, but did not refer the actual organization of the solar system. In order to actually determine if the weather model described is an explanation, it might help to know the kind of information you know about the clouds.

To explain may be to identify mechanism. Mechanists typical argue that a model requires details about the structure underlying the process. Carl Craver refers to models like Copernicus' as "how-possibly" models rather than "how-actually models (Craver 2006). A description is how a system possibly might work, while an explanation is how that system actually works. It includes not just the input and the outputs, but actual information about how the inputs change into the outputs. To take our weather example from earlier, it is not sufficient to know that when it is cloudy, it will rain. An explanatory model would also need to explain how raindrops form in clouds and what causes them to fall. To Craver, mechanistic details like these are required if we want to distinguish how-actually and how-possibly models. He additionally argues for a continuum of explanations, ranging from a "sketch" to a "complete description." A sketch might identify interactions between components, but does not actually provide

information about how that interaction occurs. Craver further elucidates these differences using the Hodgkin-Huxley model of the action potential (Craver, 2006). Hodgkin and Huxley claimed, and Craver agrees, that their model is not explanatory because it simply represents relationships between different electrophysiological elements, such as voltage and membrane permeability, mathematically. He argues that the model only becomes explanatory when taken with additional information about the molecular nature of the channels involved. He additionally underlines the importance of temporal and spatial considerations of a mechanism, rather than simply the physical, allowing for psychological theory to be mechanistic.

To explain may be to identify function. Functional analysis involves decomposing a process into smaller processes and identifying goals. Rather than identify physical, structural, spatial or temporal information, these analyses identify sub-capacities and goals. Functional decomposition outlines the problem a system needs to solve and generalized components that do so. Mechanists argue this is not a sufficient explanation, but rather a useful sketch. Functionalists argue that even without a full understanding of mechanistic detail, a model can provide an explanation. Under this view, functional analysis provides a form of explanation that is both distinct and autonomous from mechanistic explanation (Piccinini & Craver). Few functionalists argue that information about mechanisms do not inform explanation as well, but instead argue that functional decomposition provides a different kind of explanation or answer different questions. In particular, Daniel Wieskopf argues that some systems must primarily be understood functionally and cannot be explained mechanistically (Wieskopf, 2011). People additionally are more likely to generate functional explanations of fictional processes and more

likely to generalize from a functional explanation, especially for biological information (Lombrozo 2014). However, a psychology preferecne does not necessarily mean it best captures explanation. People are easily tricked by logical fallacies and other errors in reasoning. Functional information may be compelling, but still fail to fully explain behavior.

To explain may be many distinct but related concepts. The two views outlined already are one of many that have been proposed in a long-running debate in philosophy of science. Aristotle outlined four different types of explanation, arguing explanation includes not only mechanism and function, but also necessary properties or material (Lombrozo, 2006). Even if philosophy of science debate came to an end, competing views of explanation can be found throughout scientific discourse. For this reason, some argue explanation cannot be thought of as a unitary concept or process. Ingo Brigandt highlights the wide variation of explanation, noting that even the same research group may use different types of explanation on different projects (Brigandt, 2013). Matteo Colombo similarly has argued that we have a multitude of competing models of explanation because explanation actually encompasses many processes with distinct purposes. Explanation may be thought of as a group of interrelated concepts and mental representations useful to understanding our world (Colombo, 2016). These pluralists argue that the different models of explanation, only a few of which are outlined here, each represent some aspect of explanation and cannot be collapsed with one another. Some hold that multiple forms of explanation are possible for the same phenomenon, rather than certain types of information being descriptive or sketches of explanation (Brigandt, 2013). Psychology further supports this concept, representing explanation as a series of interrelated cognitive features. Experiments

have shown that both functional and mechanistic reasoning presented simultaneously can cause subjects to generalize a given feature (Lombrozo, 2014). As mentioned above, psychology theory does not necessarily provide us an answer the philosophical question of what scientific explanation should be and should represent. However, it is worth considering, even if only as a useful concept to conceive the diversity of ways "explanation" is used.

II. How can we use these models of explanation as a lens on the integration challenge?

Understanding how scientists use the concept of explanation can be powerful for scientists as they do their own work and read other's work. However, we contend that for cognitive science, it can also inform understandings of relations between fields. In this work, we do not intend to define explanation or advocate for a specific view regarding integration. We instead intend simply to characterize how different understandings of explanation in different fields contribute to the integration challenge. Two major theories of integration are discussed here.

"Higher" fields may reduce to "lower" fields. Reductionism claims with a sufficiently advanced level of knowledge, all investigations of the natural world will reduce and become equivalent to a field below it (Brigandt, 2013). Ultimately, the world can be described in terms of physics. This theory certainly has seen a degree of success: biological processes rely on chemical processes rely on physics. However, it has been more controversial with regards to studies of the mind. Psychological theory under this framework may be useful, but fundamentally incomplete. The debate primarily deals with psychology and neurobiology, however it is easy to fit linguistics, computer science, and philosophy in at a similarly or more abstract level than psychology.

Reductionists often imagine the natural world as split into a grand hierarchy, each a higher level of abstraction that will be subsumed and entirely included in the lower level. Under this framework, every natural phenomenon will be explicable purely by physics once we have reached sufficient understanding of physics (Sober, 1999). If we had a complete understanding of the biology of the brain, psychology would not provide any additional information above and beyond that. This theory dovetails well with a strict mechanist interpretation. Mechanists argue that functional decomposition, often used in psychological theory can be thought of as a "sketch of a mechanism" to be filled in with further details later (Piccinini & Craver, 2011). Explanatory power only comes with information about mechanism, which is often treated as equivalent to molecular or biophysical details although it is not necessarily so. The connectomics community is built around this mechanistic focus, with a goal to understand brain by recreating all neurons and connections (Jonas & Kording, 2017). However, it has been considerably criticized on the basis of multiple realizability (Sober, 1999). A behavior cannot be reduced if it can be produced by distinct neural differences, or a non-neural system such as artificial intelligence. Despite this criticism, reductionism remains strongly held within contemporary neurobiology, especially as new electrophysiology and imaging techniques emerge (Bickle, 2014). A recent article even called for neuroscience to "correct the reductionist bias" and focus again on behavior (Krakauer, et al 2017)

Different fields may investigate different levels of a phenomenon. The alternatives to reductionism largely argue that different types of investigation provide different types of information that can equally contribute to explanation. A number of anti-reductionist theories have emerged. Most famously, David Marr and Tomaso Poggio proposed a tri-level hypothesis of analysis (Marr, 1982). Marr argued that the visual system could be understood at three levels of analysis: computational, algorithmic, and implementational. These levels can similar as the goal or purpose of the process (computational), the rules and representations (algorithmic), and the implementation (physical). Each of these levels provides distinct, autonomous information about the system of interest. These levels can be thought of in terms of explanation types, with functional analysis largely being computational or algorithmic and mechanistic information largely concerning implementation. Aside from Marr's theory, there are few well-known characterizations of what the fields of explanation is. Much of the discussion instead is anti-reductionist in nature, seeking to establish psychology as autonomous and distinct from neurobiology (Kaplan, 1984). One strategy for doing this, which has further been extended to computational theories, is been to claim this can be derives from differences in explanation types. Anti-reductionists also criticize the more-details-the-better view, saying that at times additional information about mechanism does not actually improve the explanation (Kaplan, 1984). All in all, a variety of arguments have been used to criticize reductionism, largely with the idea that different types of evidence can provide distinct explanations.

METHODOLOGY

I. How do we investigate researchers use and understanding of "explanation"?

Identifying an individual's attitude towards explanation, or even how they use it in their work, in a rigorous quantitative way proves difficult. Although researchers frequently use terminology such as "this explains" in their published work and talks, it rarely is entirely clear what conception of explanation is intended. We intend to use a survey to broadly characterize whether members of different fields tend to a specific type of explanation.

This investigation will largely be preliminary because no prior investigations on explanation among cognitive scientists have been conducted. However, the emerging field of experimental philosophy provides a useful template for our methodology. This field intends to seek insight for philosophical questions from non-philosophers in a rigorous way, borrowing techniques from social science (Sosa, 2007; Woolfolk, 2013). Ideally, experimental philosophy, or xphi, allows for a quantified understanding of how a particular group understands a concept and whether there are differences between groups. For example, it is easy to imagine a child answering a question differently at different developmental stages, or that someone's cultural background or upbringings plays a role in their answer to ethic questions. Experimental philosophy hopes to investigate these types of questions. Experimental philosophy of science investigates how non-philosophers, often scientists, understand questions relevant to philosophy of science. Experimental philosophy has been heavily criticized, largely due the lack of rigor expected for these techniques. Notably, problems with sampling, reliability and validity of survey instruments, and the effectiveness of self reported questionnaires (Woolfolk, 2013). Keeping this in mind, we chose to undertake this as a pilot or exploratory study that would

better inform later, more rigorous studies. We also drew heavily on key experiments in experimental philosophy of science, as well as related work in psychology of explanation, when making our study.

A large project by Griffiths & Stotz investigated representations of "gene" among biologists (Stotz, et al. 2004). Genetics as a field has undergone rapid transformation over the past hundred years, largely due to new genomic sequences and editing techniques. Two representations now exist: either the classical "packet" of genetic material transferred to offspring or as strands of DNA, as informed by modern technology and understanding. This study tests both implicit and explicit attitudes. Griffiths and Stotz hoped to understand to what extent which biologists favored each representation and whether that depended on discipline membership. They studied both implicit and explicit attitudes. The researchers were asked to respond to a series of statements indicating various attitudes towards genes. Additionally, the philosophers asked researchers to perform a gene annotation task. In this survey, researchers were presented with genes spaced different distances apart and asked whether one or two genes were depicted, a task that closely mimics real life genetic annotation tasks.

Genes are a much more concrete and simpler concept than scientific explanation, so we also looked to at methodology techniques investigating more abstract terms such as innateness. In one study, non-biologists were given descriptions of birdsong and asked whether the trait was innate (Griffiths, et al 2009). Based on the information provided, the researchers were able to identify three features that explained 70% of the variance in the group's understanding of innateness. The authors are careful to note that that they do not intend to define innateness by necessary and sufficient conditions, but rather to understand features of

the cognitive structures when the term “innateness” is used. We have similar goals in that we do not intend to define explanation, but rather to better understand its usage.

II. *Conducting a survey*

We developed a three-part survey, similar in structure to the Representing Genes project described above. Each of our three parts (Usage, Stated View, and Scientific Profile) is intended to address a different aspect of the larger research question.

The first section (Usage) is intended to measure implicit usage of explanation. Essentially, we wanted to determine if researchers made decisions about whether a model is an explanation or not consistent with a specific view of explanation. We presented respondents with short descriptions of models in cognitive science and asked five questions about the explanatory power of each.

Each model was selected from discussions in relevant philosophical literature. Additionally, these were selected in order to provide a range of features that might be relevant. We hoped to represent both different levels of biological detail and the presence or absence of of mathematical equations. Well-established models were selected in order to avoid bias in more controversial and recent work or attitudes to specific researchers. Model C is a broad class, but has been subject to significant current debate in neuroscience (Love, 2011; Bowers, 2012; Colombo, 2012). Descriptions were intended to be partial and not reflective of the most recent scholarship. This choice allowed us to both distance the models from ongoing debate and to provide models where the features we asked about (networks, biomolecules, locations) were not elaborated. Additionally, due to concerns about the time to take the survey,

descriptions were kept to one paragraph with one supporting figure in order to keep the survey short and help the response rate. Although this does decrease the likelihood that researchers will view any of the models as a full description or explanation, it should be similar across each of the models.

MODEL	DESCRIPTION	Biological details?	Mathematical?	Cognitive/behavioral details?
A: Hodgkin-Huxley model of the action potential	Set of equations describing time course of action potential	YES	YES	NO
B: Baddeley-Hitch model of working memory	"Black box" model based on human reaction time and error rates	NO	NO	YES
C: Bayesian brain hypothesis	The brain uses a specific type of computation to use previous experience and sensory input to predict future events.	NO	YES	YES

The second section (Quotes) hopes is intended to measure explicit attitudes. Here, we used five different from cognitive scientists and asked respondents to indicate the extent to which they agreed or disagreed from 1 (fully disagree) to 5 (fully agree) on a Likert scale. Each of these was intended to represent attitudes towards the relation of fields within cognitive science or a mode of explanation.

QUOTE	VIEWPOINT REFLECTED
1: "Vis-à-vis explanations of behavior, neurological theories specify mechanisms and psychological theories do not." - Fodor (1965)	Functionalist
2: "The task of a psychologist trying to understand human cognition is analogous to that of a man trying to discover how a computer has been programmed." - Neisser (1967)	Mechanist
3: "The proper level of discourse for cognitive theory concerns information, not the medium used to carry it." – Palmer (1978)	Functionalist
4: "The mind can be studied independently from the brain. Psychology (the study of the programs) can be pursued independently from neurophysiology (the study of the machine and the machine code)." – Philip Johnson-Laird (1983)	Functionalist
5: "Trying to understand perception by studying neurons is like trying to understand bird flight by only study feathers: it cannot be done." – David Marr (1982)	Functionalist

The majority of the quotes represented functionalist attitudes because, as outlined earlier mechanistic explanations are often favored in contemporary neuroscience, meaning discussions on the topic tend to be arguments for functionalism. It is also worth noting that the quotes used are ambiguous in their meaning and in some cases, may be interpreted in multiple ways. Although quotes allow us to more clearly link it to a broader idea, it introduces an additional level of ambiguity because we could not control how the respondents interpreted the quote. In future work, direct statements may be better suited to this type of investigation.

II. Subject recruitment

Subjects were recruited from publicly available faculty emails at US institutions with cognitive science programs. We included institutions with both undergraduate and graduate programs in cognitive science, or similarly named programs (such as MIT's Department of Brain and Cognitive Sciences or Washington University St. Louis Philosophy-Neuroscience-Psychology program. In total, 874 researchers at 38 institutions were contacted. This population was selected due to the public availability of contact information.

Due to the nature of the field, this population represents only a small segment of cognitive scientists, as many are based in more traditional discipline departments. However, these traditional disciplines also include a number of researchers not working on questions related to cognition. For example, a traditional biology department may include neuroscientists, but may equally include plant biologists wholly uninterested in behavior. Rather than make subjective assessments of whether a researcher should be considered a cognitive scientist or not, we used department affiliation. The actual work conducted in and subfield composition of these departments varies considerably. We do not know how representative our contacted group or our group of respondents is for the population of cognitive scientists as a whole.

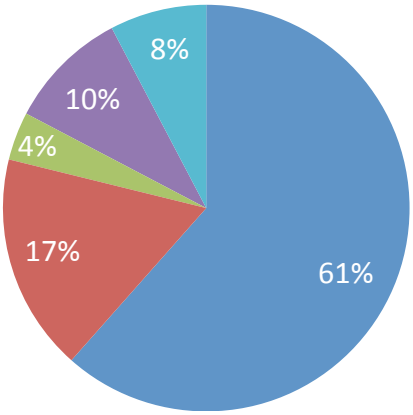
Potential subjects received a short email with a link to a Google Form.

RESULTS

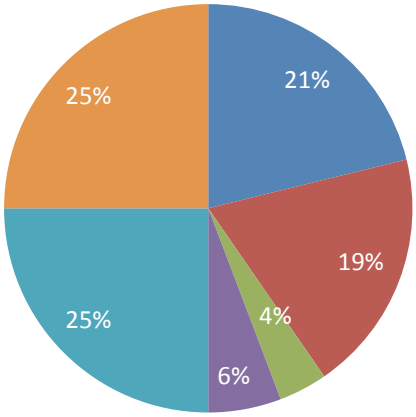
I. Respondents – Discipline Membership

Fifty-five researchers responded, for a response rate of 6.06%. This sample is admittedly small, but not unexpected because in most cases we had no prior contact with the respondents. However, due to this small sample size, we will not be able to compare responses between different groups. Additionally, our sample did not have groups of similar sizes. All but one researcher responded to the question about discipline membership. The majority (n = 32, 61.5%) selected psychology as their primary discipline. It is unknown whether this reflects a more general trend among cognitive science departments or a bias in our sample. The second largest group was neurobiology (n=9, 17.3%).

Primary Discipline



Secondary Discipline



Psychology Neurobiology Linguistics Philosophy Computer Science

No, I only work in the discipline identified above. (*Option provided only for second question)

In the free response question, 25 subjects (47.2%) included “cognitive” or “cognition” in their response. This result confirms that the group we targeted largely considers themselves cognitive scientists. Although roughly half self-identified as cognitive scientists , it was open ended and elicited a wide range of responses. Several responded with much more fine-grained descriptions of their specific topic of interest (visual attention; synaptic physiology; etc). One subject opted not to select a primary or secondary discipline.

II. Models – Responses

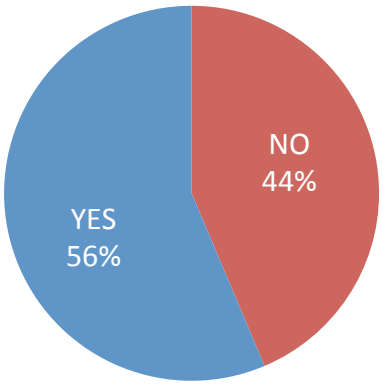
As mentioned earlier, the small sample size and uneven group sizes make it difficult to conduct rigorous statistical analyses. We instead discuss general trends.

1 – Is the model a description? Is it an explanation?

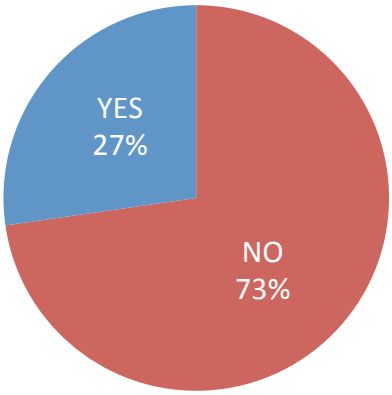
This set of questions was intended to gauge whether certain model features made it an explanation rather than a description. We had expected participants to be more likely to say each model was a description of the model than an explanation. However, we only saw this effect in model A (the Hodgkin-Huxley model of the action potential). For this model, slightly more than half (n=30, 56.6%) agreed that it sufficiently described the phenomenon, while only a quarter (n=13, 24.5%) said it explained the phenomenon.

MODEL A: HODGKIN-HUXLEY MODEL OF AN ACTION POTENTIAL

Description



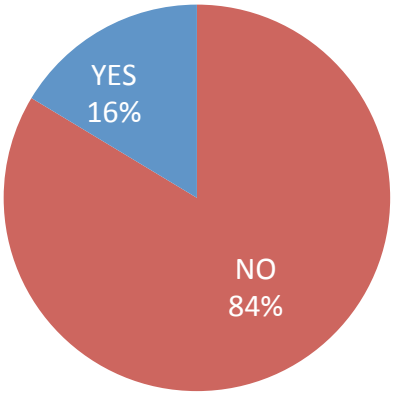
Explanation



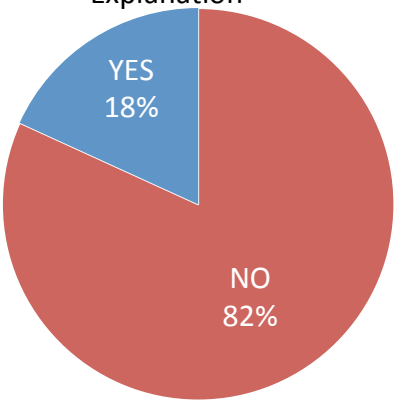
This distinction between explanation and description did not hold for model B (Baddeley-Hitch model of working memory) and model C (Bayesian brain hypothesis). In each case, the majority of respondents judged it neither a description or explanation.

MODEL B: BADDELEY-HITCH MODEL OF WORKING MEMORY

Description

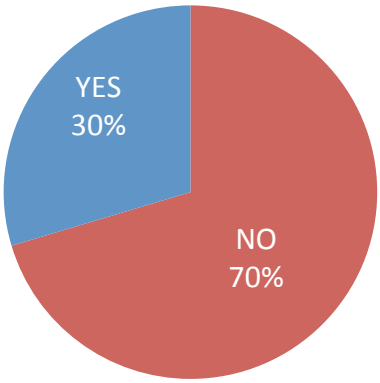


Explanation

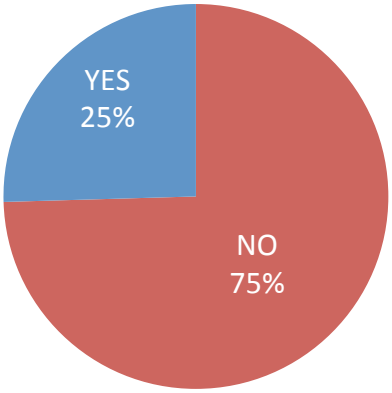


MODEL C: BAYESIAN BRAIN HYPOTHESIS

Description



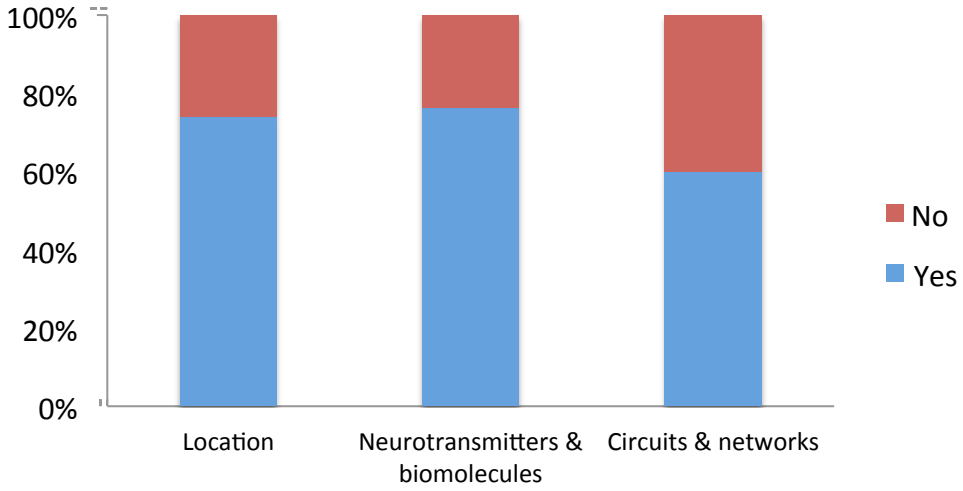
Explanation



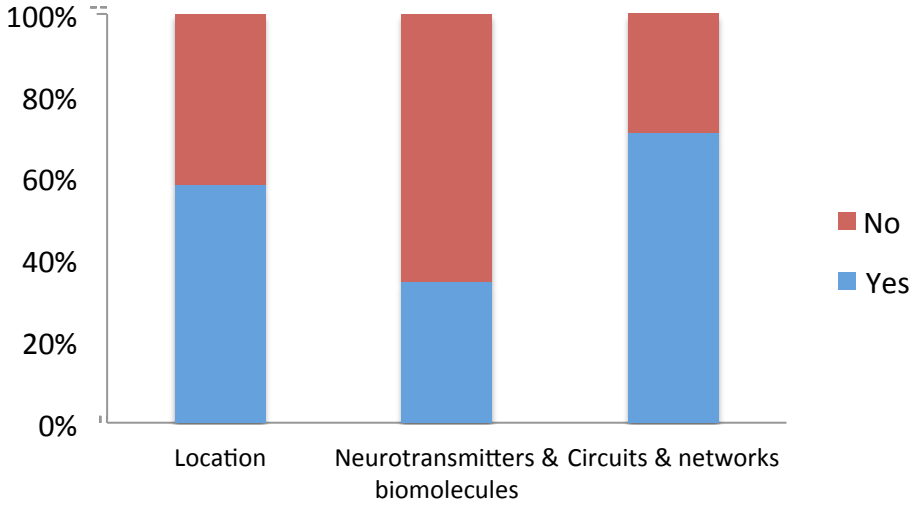
2 – What type of information strengthens an explanation?

Question 3, 4, and 5 were intended to test the more-details-better-view, as well as gauge what type of . Each question targeted a different scale: broad spatial information (where in the brain), appeal to "lower" level (biomolecules), and temporal/spatial combinations (networks). In nearly every case, the majority indicated more information would strengthen the information. We expected this result. Even someone holding a strong functionalist view might not view mechanistic evidence as necessary for explanation, but would likely still agree that it can provide explanatory power. Interestingly though, this pattern breaks for two questions. For both Model B (Baddeley-Hitch Model of Working Memoy) and Model C (Bayesian Brain Hypothesis), the majority of respondents did not view neurotransmitters and other biomolecules as important to the explanation of behavior. This finding reflects a functional viewpoint.

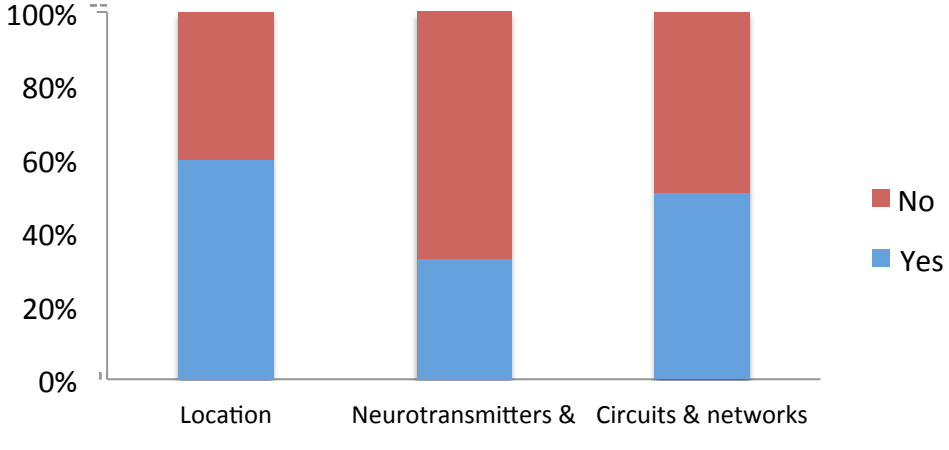
MODEL A: HODGKIN-HUXLEY MODEL OF AN ACTION POTENTIAL



MODEL B: BADDELEY-HITCH MODEL OF WORKING MEMORY



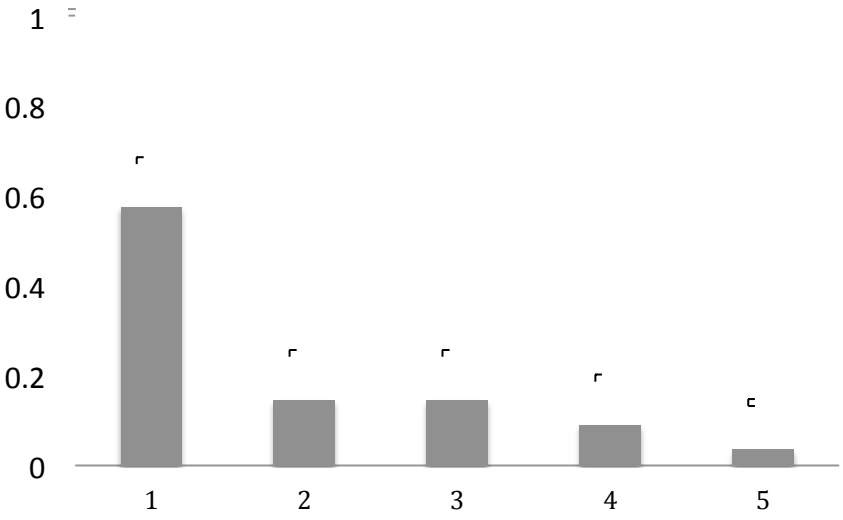
MODEL C: BAYESIAN BRAIN THEORY OF PERCEPTION



III. Statements – Responses

Here, we report the proportion of subjects who responded with a given value for each statement. One subject chose not to respond to statement 1, so those proportions were calculated out of 54 rather than 55. All subjects answered the other four statements. The median score for each statement is provided, as that is the typical way of reporting ordinal data from Likert scales (Sullivan, 2013). However, we additionally report mean to give a further understanding of variation in the responses. Subgroup means are listed for neurobiology, philosophy, and psychology. These were not calculated for linguistics and computer science because fewer than 5 respondents selected each. It is also worth noting that the psychology group is considerably larger than philosophy (n=7) or neurobiology (n=9). These group sizes are also too small to draw any conclusions. As mentioned earlier, one subject opted not to select a primary or secondary discipline and is included only in the full group averages.

Statement 1: *"Vis-à-vis explanations of behavior, neurological theories specify mechanisms and psychological theories do not." – Fodor (1965)*



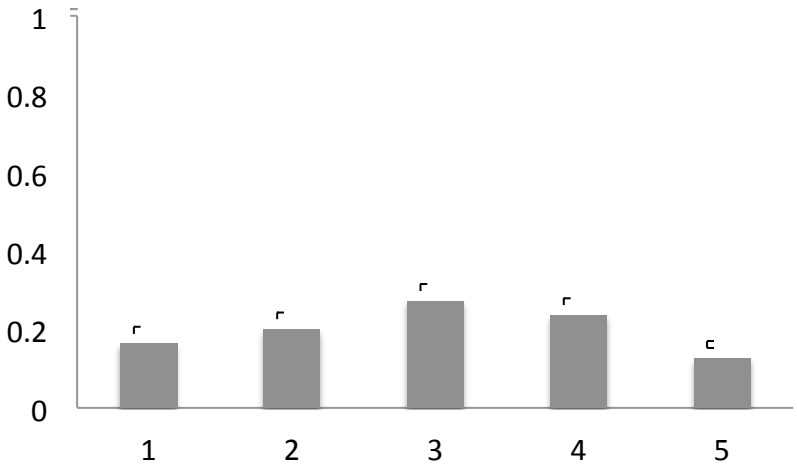
Median: 1

Mean 1.8

Mean by subgroup

- Psychology: 1.48
- Neurobiology: 3.1
- Philosophy: 1.83

Statement 2: *"The task of a psychologist trying to understand human cognition is analogous to that of a man trying to discover how a computer has been programmed." – Neisser (1967)*



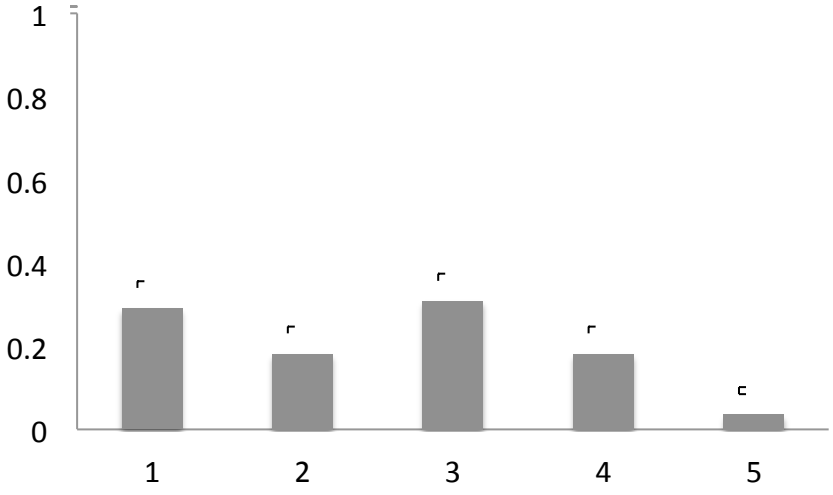
Median: 3

Mean: 2.96

Mean by subgroup:

- Psychology: 2.96
- Neurobiology: 2.89
- Philosophy: 3.42

Statement 3: *"The proper level of discourse for cognitive theory concerns information, not the medium used to carry it." – Palmer (1978)*



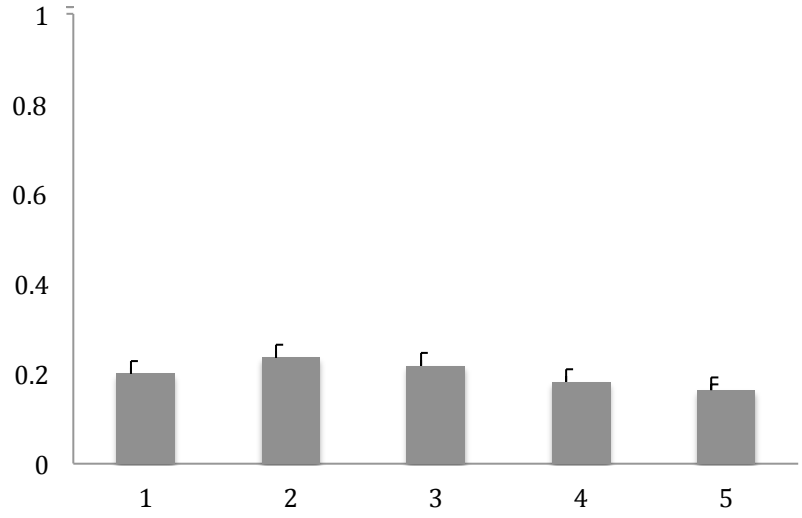
Median: 3

Mean: 2.49

Mean by subgroup:

- Psychology: 2.48
- Neurobiology: 2.11
- Philosophy: 3.71

Statement 4*: "The mind can be studied independently from the brain. Psychology (the study of the programs) can be pursued independently from neurophysiology (the study of the machine and the machine code)." – Johnson-Laird (1983)*



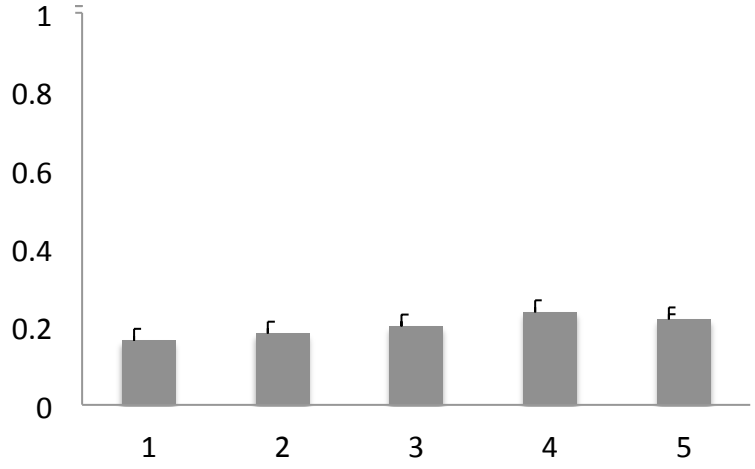
Median: 3

Mean: 2.87

Mean by subgroup:

- Psychology: 3.06
- Neurobiology: 2.33
- Philosophy: 3.14

Statement 5: *"Trying to understand perception by studying neurons is like trying to understand bird flight by only study feathers: it cannot be done." – Marr (1982)*



Median: 3

Mean: 3.16

Mean by subgroup:

- Psychology: 3.22
- Neurobiology: 2.55
- Philosophy: 3.71

DISCUSSION

As mentioned earlier, it is inherently difficult to make strong claims due to the complexity of the components. Our work sought to better understand how a complex process with numerous conceptions (scientific explanation) is understood by a diverse poorly defined group (cognitive scientists) in order to understand a complex problem (inter-field integration). Our small sample size, uneven groups, and problems with our survey instrument, discussed later in suggestions for future work, compounded these inherent difficulties. Our survey is best thought of as an exploratory analysis to begin the process of mapping understanding the concept space and range of attitudes held. With that caveat, we highlight several key findings.

Explanations about behavior differ from explanations about cells. Our respondents seem to draw a distinction between models B and C (Baddeley-Hitch model of working memory and Bayesian brain hypothesis of perception, respectively) and model A (the Hodgkin-Huxley model of the action potential). We see this distinction both in the first set of questions about the models, concerning the difference between explanation and description. The Hodgkin-Huxley model was the only model where a greater proportion of respondents thought it was a description than an explanation. It is generally understood in the philosophical literature that description requires less information than explanation, so we expected this distinction. However, we did not see this distinction in either of the behavioral models, suggesting description is not seen as a partial or incomplete explanation of behavior, but is for a neural mechanism. Of course, in model A and B, the majority did not agree it was an explanation or description, so we must be cautious not to over interpret these results. Additionally, in the

second set of questions, the majority of respondents answered that additional information about neurotransmitters or biomolecules would improve the explanation in model A, but not in model B or C. This finding could reflect the belief that neurotransmitters are not important to the explanation of behavior. We however must again be cautious, because some the models used are complex and well known, meaning some other feature of the Hodgkin-Huxley model or some other shared feature of the Baddeley-Hitch and Bayesian brain hypothesis could be a confound.

Cognitive scientists say neurological theories don't specify mechanism OR psychological theories do. A consensus only emerged for one of the statements from the statements section, with 72.2% either disagreeing or strongly disagreeing with statement 1 ("Vis-à-vis explanations of behavior, neurological theories specify mechanisms and psychological theories do not." – Fodor (1965). However, despite this strong consensus, we cannot determine the viewpoint held because of the structure of the quote. The respondents could have disagreed because they do not think neurological theories specify mechanisms or because they think psychological theories do. Given the composition of our group, predominantly psychologists, it seems likely they believe psychological theories do specify mechanisms, This explanation seems particularly likely when looking at the average response by subgroup. The average for neurobiologists was 3.1, while the average for psychologists was half that, at 1.48. Although we again must caution against a broad understanding given the small sample size, this suggests neurobiologists see their work as that of identifying mechanisms, but do not see psychology theories as specifying mechanisms. This attitude could reflect either a reductionist or layers stance toward inter-field

relations. The subjects could either see psychological theory as a level of explanation that does not require mechanism, or as useful but ultimately unimportant to mechanisms that will explain the phenomenon.

Cognitive theory is not about information OR is about medium. Statement 3 ("The proper level of discourse for cognitive theory concerns information, not the medium used to carry it." – Palmer (1978)) was also skewed towards disagreement, although the median was ultimately still 3. Only 3.6% of respondents fully agreed, while 30% selected 3 (neither agree nor disagree) and 47% selected 1 or 2 (fully or partially disagree). Again, this quote has two parts, making it difficult to interpret. It indicates respondents either believe that cognitive theory should concern medium, likely meaning mechanistic information, or that it should not concern information. The former seems more likely, given our background understanding of attitudes discussed in the literature. Again, this could reflect either a mechanistic/reductionist view where only the medium of information matters, or a levels view where both mechanism and information and function matter.

Taken, these findings do suggest reductionism is less established than typically though. Boone and Piccinini argue that with the rise of cognitive neuroscience, the autonomous claim of psychology has fallen away and aims of current work are mechanistic (Piccinini 2016). They argue that the question has been settled among practicing cognitive neuroscientists. Our findings suggest instead that the debate continues and that significant support remains for functional explanations. A more rigorous investigation into these questions is needed. In future

work, the surveys must be improved to allow for more easily interpreted results. Straightforward statements such as (“Cognitive theory concerns information”) with a similar Likert scale set up would more directly answer our questions of interest. Additionally, a more balanced sample must be obtained. It may be possible to achieve this by focusing on the two largest groups, neurobiologists and psychologists. If a similar size sample can be obtained, split equally between psychologists and biologists, we could draw stronger conclusions about how concepts differ between groups. This information is necessary for researchers to understand how their work relates to findings in related fields. An overall larger sample would also be helpful. Work must also be done to understand how well the distribution of fields in our sample reflects those who consider themselves cognitive scientists.

CONCLUSIONS

An astonishing diversity of opinions on what constitutes an explanation and how fields should be related exists. In this work, we identified major attitudes and conducted a survey of attitudes. Questions of philosophical explain seem rather remote from the actual work of cognitive science. However, even a broad understanding of concepts such as studied here is helpful to work in cognitive science. If multiple disciplines are to be combined, the attitudes and assumptions of each field must be understood. Even if we accept that explanation is pluralistic, the types of explanation particular to a given field or type of evidence must be understood for that work to inform understanding in a separate field or type of work.

The goal of cognitive science is to study mind, in any way possible. With the rise of neuroscience, especially cognitive neuroscience technique, mind has become synonymous with brain, neurons, and synaptic connections. Indeed, several opinion pieces published this year (Krakuer, et al. “Neuroscience Needs Behavior: Correcting a Reductionist Bias”, Jonas and Kording’s “Can a neuroscientist understand a microprocessor?”) call for further emphasis on behavior and function. As reflected in our findings, the debate about explanation as it relates to field studying the mind is far from over. Cognitive scientists need to develop a deeper understanding of their field and of the explanations they create.

REFERENCES

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Barrett, D. (2014). Functional analysis and mechanistic explanation. *Synthese*, *191*(12), 2695–2714. https://doi.org/10.1007/s11229-014-0410-9

Bateson, P., & Laland, K. N. (2013). Tinbergen's four questions: An appreciation and an update. *Trends in Ecology and Evolution*, *28*(12), 712–718. https://doi.org/10.1016/j.tree.2013.09.013

Bechtel, W. (2005). Reducing Psychology while Maintaining its Autonomy via Mechanistic Explanations. *Psychology*, 1–19. https://doi.org/10.1017/CBO9781107415324.004

Bechtel, W. (2015). The Non-Redundant Contributions of Marr ' s Three Levels of Analysis for Explaining Information-Processing Mechanisms, *7*, 312–322. https://doi.org/10.1111/tops.12141

Bickle, J. (2016). Revolutions in Neuroscience : Tool Development, *10*(March), 1–13. https://doi.org/10.3389/fnsys.2016.00024

Bickle, J. (2015). Marr and Reductionism, *7*, 299–311. https://doi.org/10.1111/tops.12134

Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese*, *193*(5), 1509–1534. https://doi.org/10.1007/s11229-015-0783-4

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological Bulletin*, *138*(3), 389–414. https://doi.org/10.1037/a0026450

Brigandt, I. (2013). Explanation in Biology: Reduction, Pluralism, and Explanatory Aims. *Science and Education*, *22*(1), 69–91. https://doi.org/10.1007/s11191-011-9350-7

Carandini, M. (2012). From circuits to behavior: a bridge too far? *Nature Neuroscience*, *15*(4), 507–9. https://doi.org/10.1038/nn.3043

Colombo, M. (2016). Experimental Philosophy of Explanation Rising : The Case for a Plurality of Concepts of Explanation. *Cognitive Science*, 1–15. https://doi.org/10.1111/cogs.12340

Craver, C. F. (2005). Beyond reduction: Mechanisms, multifield integration and the unity of neuroscience. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2 SPEC. ISS.), 373–395. https://doi.org/10.1016/j.shpsc.2005.03.008

Craver, C. F. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376. https://doi.org/10.1007/s11229-006-9097-x

Danks, D. (2013). Moving from Levels & Reduction to Dimensions & Constraints. *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, 2124–2129.

Friston, K. (2012). The history of the future of the Bayesian brain. *NeuroImage*, *62*(2), 1230–1233. https://doi.org/10.1016/j.neuroimage.2011.10.004

Glennan, S. (2002). Rethinking Mechanistic Explanation, *69*(September 2002).

Griffiths, Paul E, K. S. (2010). How the Mind Grows : A Developmental Perspective on the Biology of Cognition (Vol. 122, pp. 29–51).

Griffiths, P. E., & Stotz, K. (2008). Experimental Philosophy of Science. *Philosophy Compass*, *3*(3), 507–521. https://doi.org/10.1111/j.1747-9991.2008.00133.x

Griffiths, P., Machery, E., & Linquist, S. (2009). The Vernacular Concept of Innateness Paul Griffiths, Edouard Machery, Stefan Linquist . *Mind Language*, *24*(5), 1–40. https://doi.org/10.1111/j.1468-0017.2009.01376.x

Griffiths, T. L., & Tenenbaum, J. B. (2005). Structure and strength in causal induction. *Cognitive Psychology*, *51*(4), 334–384. https://doi.org/10.1016/j.cogpsych.2005.05.004

Hovarth, J., & Grundmann, T. (2007). Introduction: Experimental Philosophy and its Critics. *Experimental Philosophy and Its Critics*.

Itthipuripat, S., & Serences, J. T. (2015). Integrating Levels of Analysis in Systems and Cognitive Neurosciences: Selective Attention as a Case Study. *The Neuroscientist : A Review Journal Bringing Neurobiology, Neurology and Psychiatry*, 1073858415603312. https://doi.org/10.1177/1073858415603312

Jonas, E., & Kording, K. P. (2017). Could a Neuroscientist Understand a Microprocessor? *PLOS Computational Biology*, *13*(1), e1005268. https://doi.org/10.1371/journal.pcbi.1005268

Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *The Behavioral and Brain Sciences*, *34*(4), 169–188. https://doi.org/10.1017/S0140525X10003134

Kaplan, D. M., & Bechtel, W. (2011). Dynamical models: An alternative or complement to mechanistic explanations? *Topics in Cognitive Science*, *3*(2), 438–444. https://doi.org/10.1111/j.1756-8765.2011.01147.x

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*(3), 339–373. https://doi.org/10.1007/s11229-011-9970-0

Kaplan, D. (1984). Integrating Mind and Brain Science: A Field Guide. *Davidmichaelkaplan.Org*, 1–28. Retrieved from http://www.davidmichaelkaplan.org/uploads/6/4/2/9/64290681/kaplan-integrating-mind-brain-science-field-guide.pdf

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., Maciver, M. A., & Poeppel, D. (2017). Neuroscience Needs Behavior: Correcting a Reductionist Bias. *Neuron*, *93*(3), 480–490. https://doi.org/10.1016/j.neuron.2016.12.041

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332. https://doi.org/10.1016/j.cogpsych.2010.05.002

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, *8*(September), 700. https://doi.org/10.3389/fnhum.2014.00700

Machamer, P., Darden, L., Craver, C. F., Machamertt, P., & Craver, C. F. (2011). Thinking About Mechanisms, *67*(1), 1–25.

Marom, S., Meir, R., Braun, E., Gal, A., Kermany, E., & Eytan, D. (2009). On the precarious path of reverse neuro-engineering, *3*(May), 3–6. https://doi.org/10.3389/neuro.10.005.2009

Marr, D. (1982). Understanding Complex Information-processing Systems. *W. H. Freeman and Com- Pany, Used with Permission*, (1956), 69–?

Miller, G. A. (2003). The cognitive revolution : a historical perspective, *7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283–311. https://doi.org/10.1007/s11229-011-9898-4

Salmon, W. C. (1989). Four Decades of Scientific Explanation. *Minnesota Studies in the Philosophy of Science Volume XII Scientific Explanations*, *3*, 3–219. Retrieved from http://books.google.hu/books?id=u2ocAQAAIAAJ

Schunn, C. (1998). The growth of multidisciplinarity in the Cognitive Science Society. *Cognitive Science*, *22*(1), 107–130. https://doi.org/10.1016/S0364-0213(99)80036-6

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Sobert, E. (1999). The Multiple Realizability Argument against Reductionism. *Philosophy of Science*, *66*(4), 542–564.

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, *132*(1), 99–107. https://doi.org/10.1007/s11098-006-9050-3

Sternberg, S., & Sternberg, S. (2017). Modular processes in mind and brain Modular processes in mind and brain, *3294*(April). https://doi.org/10.1080/02643294.2011.557231

Stotz, K., & Griffiths, P. (2004). Genes : Philosophical analyses put to the test *. *History and Philosophy of the Life Sciences*, *26*(1), 5–28. https://doi.org/10.1080/03919710412331341621

Stotz, K., Griffiths, P. E., & Knight, R. (2004). How biologists conceptualize genes: An empirical study. *Studies in History and Philosophy of Science Part C :Studies in History and Philosophy of Biological and Biomedical Sciences*, *35*(4), 647–673. https://doi.org/10.1016/j.shpsc.2004.09.005

Sullivan, Gail M; Artino, A. (2013). Analyzing and Interpreting Data From Likert-Type Scales, (December), 541–542.

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, *183*(3), 313–338. https://doi.org/10.1007/s11229-011-9958-9

Woolfolk, R. L. (2013). Experimental philosophy: A methodological critique. *Metaphilosophy*, *44*(1–2), 79–87. https://doi.org/10.1111/meta.12016

APPENDIX: SURVEY MATERIALS

Below, a copy of the survey sent out is reproduced. It was formatted differently when presented to subjects, as Google Forms was used to provide an easily accessible web interface. However, the content is identical.

Welcome to the Explanation in Cognitive Science Survey

Thank you for agreeing to participate in our survey! This study is part of a senior honors project examining explanation in cognitive science. As part of the project, we would like to gauge what researchers view as an adequate explanation in cognitive science. There are three sections total and it will take approximately 10 minutes. Responses will be anonymous.

Section I (Model A)

Please read the model description and answer the subsequent questions.

An action potential is a rapid increase and decrease in the membrane potential of a neuron. Hodgkin and Huxley developed a model of how action potentials are initiated and propagated. This model is a set of differential equations that mathematically describe this capacity of neurons.

1. Does this model sufficiently describe the phenomenon? YES NO

2. Does this model explain the phenomenon? YES NO

3. Would additional information about where this phenomenon occurs strengthen the explanation? YES NO

4. Would additional information about specific neurotransmitters or other biomolecules involved increase this model's explanatory power? YES NO

5. Would detail about specific neural circuits or networks increase the model's explanatory power?

YES NO

Section I (Model B)

Working memory refers to the ability to store information for cognitive processing for short time periods. Baddeley and Hitch developed a three-component model of this ability by testing human subjects on a number of memory tasks. They distinguished component based on patterns of errors and reaction times on these tasks. The three components included are the central executive, visuospatial sketchpad, and phonological loop. The main system is the central executive, which is a general processing capability with limited capacity that directs attention to other subsystems. The other two subsystems deal with specific types of information.

1. Does this model sufficiently describe the phenomenon? YES NO

2. Does this model explain the phenomenon? YES NO

3. Would additional information about where this phenomenon occurs strengthen the explanation? YES NO

4. Would additional information about specific neurotransmitters or other biomolecules involved increase this model's explanatory power? YES NO

5. Would detail about specific neural circuits or networks increase the model's explanatory power?

YES NO

Section I (Model C)

Please read the model description and answer the subsequent questions. This class of models argues that the brain uses Bayes' rule, a statistical method, in order to perceive the world.

$P(world|data) \propto P(data|world) \times P(world)$

Posterior $\propto$ Likelihood x Prior

In words, to perceive what's happening in the world, the brain uses the likelihood (how likely this data would occur assuming a given state of the world) and the prior (how likely that state of the world is) to determine the posterior (state of the world).

1. Does this model sufficiently describe the phenomenon? YES NO

2. Does this model explain the phenomenon? YES NO

3. Would additional information about where this phenomenon occurs strengthen the explanation? YES NO

4. Would additional information about specific neurotransmitters or other biomolecules involved increase this model's explanatory power? YES NO

5. Would detail about specific neural circuits or networks increase the model's explanatory power?

YES NO

Section II

Please indicate your opinion towards the following quotes concerning explanations in cognitive science. Indicate 1 if you fully disagree and 5 if you fully agree with the statement.

1. Vis-à-vis explanations of behavior, neurological theories specify mechanisms and psychological theories do not.

1 2 3 4 5

2. The task of a psychologist trying to understand human cognition is analogous to that of a man trying to discover how a computer has been programmed.

1 2 3 4 5

3. The proper level of discourse for cognitive theory concerns information, not the medium used to carry it.

1 2 3 4 5

4. The mind can be studied independently from the brain. Psychology (the study of the programs) can be pursued independently from neurophysiology (the study of the machine and the machine code).

1 2 3 4 5

5. Trying to understand perception by studying neurons is like trying to understand bird flight by only study feathers: it cannot be done.

1 2 3 4 5

Section III

1. What discipline do most strongly identify as working within?

Psychology
Neurobiology
Linguistics
Philosophy
Computer Science

2. Do you work in a secondary discipline?

No, I only work in the discipline identified above
Psychology
Neurobiology
Linguistics
Philosophy
Computer Science

3. How would you describe your field?
(Free response)

Thank you!
Thank you for your time! Your response will allow us to better understand how researchers view explanation in cognitive science and how that relates to philosophical understandings of scientific explanation. Additionally, we are interested in how these understandings might different between the disciplines within cognitive science.

If you have any further questions about this survey, please contact Kira Wegner-Clemens at explain.cog.sci@gmail.com

Sources of figures & quotes, by section:

I. MODELS
Hodgkin-Huxley model of the action potential -
Hodgkin, Alan L., and Andrew F. Huxley. "Propagation of electrical signals along giant nerve fibres." Proceedings of the Royal Society of London. Series B, Biological Sciences (1952): 177-183.

Baddeley-Hitch model of working memory -
Baddeley, Alan. "Working memory: looking back and looking forward." Nature reviews neuroscience 4.10 (2003): 829-839.

Bayesian brain models of perception -
Vincent, Benjamin T. "A tutorial on Bayesian models of perception." Journal of Mathematical Psychology 66 (2015): 103-114.

II. QUOTES

1. "Vis-à-vis explanations of behavior, neurological theories specify mechanisms and psychological theories do not." - – Fodor (1965)

2. "The task of a psychologist trying to understand human cognition is analogous to that of a man trying to discover how a computer has been programmed." - Neisser (1967)

3. "The proper level of discourse for cognitive theory concerns information, not the medium used to carry it." – Palmer (1978)

4. "The mind can be studied independently from the brain. Psychology (the study of the programs) can be pursued independently from neurophysiology (the study of the machine and the machine code)" – Johnson-Laird (1983)

5. "Trying to understand perception by studying neurons is like trying to understand bird flight by only study feathers: it cannot be done." - David Marr (1982)