Qualifying Exam.

The final quests.

Causation & Explanation:

Is identifying the explanation for some event equivalent to identifying its cause(s)? How are the two similar? How might they differ? Support your response by (briefly) describing relevant accounts of explanation and causation from philosophy and psychology and empirical results from psychology.

Highest level answer:

Your definition of "event" and (consequently) the "cause(s)" and "explanation(s)" of that event, will change how one answers the question:

> Is identifying the explanation for some event equivalent to identifying its cause(s)? In the following essay consider two notions of an event, each of which lends itself to a particular interpretation of cause and explanation.

In the first case, an event is treated formally, as a realization of an abstract variable whose value is the result of its causes, which are also abstract variables. This is most closely related to statistical and law-covering notions of explanation (Hempel, cf. Salmon) and probabilistic accounts of causation such as that embodied in causal Bayes nets (Sloman & Lagnado, Steyvers et al., Griffiths and Tenenbaum) and their extensions (Kemp et al.). When looked at in this light, individual events are more obviously treated as instances of types rather than tokens, and there seem to be a number of theories in which causation and explanation are closely related if not equivalent (some exceptions notwithstanding).

In the second case, events are treated as physical occurrences that are caused by hidden mechanisms that carry causal forces (Shultz; Salmon; White). From this perspective, it is less clear what constitutes an explanation other than that

immediate cause that “passed a mark” to the event in question.

Finally, I briefly consider an additional framing of what an explanation is, in which the explanation of an event and the cause of an event will almost surely differ. Some have treated explanation as a purely linguistic phenomenon only understandable in the context of the pragmatics of answering a question (van Fraassen; Hilton & Slugoski; cf. Salmon). I know of no accounts of causation that treat causes as linguistic phenomena. In which case, the explanation and the cause of an event are guaranteed to differ.

What is an event?

While it is normally tacitly assumed, we need to know what counts as an event if we want to know anything about its causes or its explanations. Because they tend to be more comparable, I will consider two approaches to causal explanation. Under one perspective of an event is as a realization of an variable embedded in a formal system of such as a graph. Under another perspective, an event is an actual occurrence due to underlying causal powers being transferred. Which of these two perspectives one takes changes how you see the relation between the cause and the explanation.

Causes and Explanations as realized regularities

If an event is taken to be the realization/assignment of a (random) variable to a particular value, then causal attribution of that assignment and explanation will often end up under the same rubric. Or more accurately, the statistical properties that will lead to the assignment of “the cause(s)” will be those that lead to the assignment of “the explanation(s)”. In particular, I will consider this from the the perspective of early accounts of computational approaches to human causal induction, then the artificial intelligence community and their work on *abduction* in Bayesian networks and how that relates to results from the psychology of causation and explanation.

If we start with the assumption that we can directly sense no causal “connexions” as Hume had said, all we are going to have to rely on for our causes and our explanations are regularities of events. Mill proposed grounding this in terms of differences of proportions, which grew to become the psychological theory of identifying causal relations known as **Δ**-P (Novick & Cheng 1992). An alternative arose from social psychology with Kelley based on the Analysis of Variance (ANOVA) model in which he described the causes of people’s behavior as being either coming from the person(in general), the situation(in general) or the particular circumstances(see also, Malle 2011). While the ANOVA was posed as a way of explaining people’s behavior (by assuming people had knowledge of how common different events were, in general), **Δ**-P was emphasized more strongly as model of reasoning about types of events rather than explaining individual events. However, its motivating idea of “how much does this change the probability of an outcome” was later adapted to be used iteratively (i.e., at different time-steps in a sequence of events) as a model of how much causal responsibility should be attributed to a person and successfully modeled peoples judgments (Spellman).

This brings into sharp contrast the a similarity and difference between the cause and the explanation of some event (at least as it has been developed in this trend in the psychological literature). Causal inference as to whether a relationship

exists or the strength of that relationship can occur about types of events and not care about the particular cause of any individual event. On the other hand, while background rates are needed in order to make causal attributions/explanations (due to Humean assumptions), people are asked to provide judgments about explanation or "the cause" or causal responsibility of individual events. Even those that dissent from the pure ANOVA model, such as Hilton and Slugoski, will still make assumptions about what the background probabilities are in order to determine abnormality conditions for attributing cause or explaining events.

In the artificial intelligence community similar assumptions have been made, though the notion of causation and probabilistic dependence has been much more fleshed out. Beginning with Pearl's (1988) proposal that explanations in Bayesian networks are chosen based on assigning a value to every variable to maximize the posterior probability given the observations. This is known as the Most Probable Explanation. He labeled this the problem of *abduction* (loosely derived from Pierce's taxonomy of logics). This was later extended by Shimony to not require that every variable have a value in the final explanation, which was a shift of the problem from that of *abduction* to *partial abduction* and the solution was known as MAP (maximum a posteriori estimation), as he termed it. This is a much more difficult problem because it involves a combinatorial explosion of candidate explanations since any non-empty subset of unobserved variables could be included in these explanations. These approaches are very similar to the approaches that we had seen in psychology, but rather than thinking about differences in probabilities are considering absolute maximal probability values among structured variable distributions. In that these models were not explicitly causal (had no necessary notion of intervention) its hard to say that these explanations were "the causes of the events", but my guess is that given the similarity of the causal reasoning theorists in psychology and these approaches that had their analogs existed in psychology causation would have been invoked.

Doing a disservice to the literature is required due to temporal limitations, but two additional models worthy of consideration are the Most Relevant Explanation(MRE) (Yuan & Lu) and the Causal Explanation Tree(CET) (Nielsen, Pellet, & Elisseeff). The MRE decides explanations based the generalized Bayes factor (p(d|h)\p(d|¬h)), where the hypothesis that gives the highest ratio between the probability of the data under the hypothesis and the probability under the negation of the hypothesis is treated as the explanation. One interesting feature is that this is computationally related to both the Griffiths and Tenenbaum (2005) *causal support* model of causal structure induction and Griffiths and Tenenbaum (2007) measure of the coincidentality of an event under a causal hypothesis (h) versus a null hypothesis (¬h)(though in general they considered cases where ¬h was a simple known null hypothesis rather than the summation of all the hypotheses that were alternatives which is MRE's approach). Given that the identical measure has been claimed to be grounds for judging the explanation of an event and is involved in determining (in general) that it is causal and for understanding how coincidental a particular event is under a particular causal theory, it seems that the connection between choosing the cause and the explanation of an event has gotten yet closer. Interestingly Griffiths & Tenenbaum (2007) even mention the problem of explaining the coincidence, further bolstering this argument. The argument finds additional similarities in that CET model relies on iteratively choosing that node that carries the largest causal information flow (Polani and Ai), which is a measure of the causal influence of one variable on another. In this case, literally, the process of finding an

explanation is finding that event which is likely the biggest cause of the explanandum.

A final note is that unlike in the early psychological work on choosing among causes where the antecedent events were known, most of these artificial intelligence models assume that the states of the variables were *not* known. If they were known, by presumption they would need to be explained as part of "the event". This is not true for CET, which even allows observed events to be among "the explanations" for the event being explained, making it hold more closely with the original explanations.

So in summary looking at causation from this statistical regularity view, there are clear relationships between the two as evidenced in the psychological and artificial intelligence literatures, however the role of knowing versus not knowing the underlying values of the events (where you know in causal attribution and often did not know in the context of explanation) led them to differ. Additionally while explanation is generally concerned with individual events, causation seems to be involved in both individual events and the regularities of events. While those two tasks may be linked (as suggested by the similarities between MRE and *causal support*) that relationship is only implicit if it is there at all.

Mechanistic views of causation

If an event is a event is an actual occurrence due to underlying causal powers being transferred then we have a somewhat different tale to tell. Mechanistic theories of causation rely on the notion of transferring forces between objects where that transference is an event (Kant; Michotte; Shultz 1982; Ahn et al.; White). For example, you might consider a light shining on a wall or a fan blowing out a candle to be the kind information relevant to inferring a causal relationship. Shultz demonstrated that children across cultures and ages (incl. adults) made causal inferences based on these kinds of information in contradiction to (admittedly somewhat odd) simple counts of co-ocurrences of events as discussed in the previous section. Ahn et al., found that people would ask for mechanistic information rather than the kinds of general regularity information presupposed by the regularity models of causation. That is when people were asked for "the cause", they wanted to know particular information related to the particular occurrence and were only satisfied once they received information. This task is similar to the kinds of exploratory information tasks that are actually engaged in when searching for hidden causal mechanisms as in Kushnir et al; Cook, Goodman and Schulz; and Gweon and Schulz. Furthermore in Ahn et al., when regularity information was accompanied by mechanistic information, it was deemed more convincing.

From the philosophy of explanation, we get the notion of "mark passing" from one object to the next, e.g., a billiard ball colliding with another and sending it flying (Salmon; cf. Michotte). These marks are then what convey information about what the cause is (as there is only one cause, even if there are enabling conditions, see Hilton) and being that cause is explicitly the criterion needed to explain that event.

So from this perspective the explanation and the cause are thought to be the same. This largely stems from the fact that the meaningfulness of "the cause" is the end of the theory: it does not explicitly concern itself with types of events, but rather

what it is that individual events were caused by and how having that information could explain the events in question.

Pragmatic accounts of explanation: These are not the causes you are looking for

Briefly, van Fraassen poses explanation not as a question about identifying causes, but about different ways one might answer a question. It may be that there are causes and that they end up being explanations, but which variables you consider and which variables and events are even considered among explanations is not an easily answered question. This is related to the concerns about the pragmatics of explanation raised by Hilton, and the concern around abnormality raised by Hilton and Slugoski, though they are concerned with causation and not explanation. However, because van Fraassen's account is explicitly linguistic and causes, presumably are not linguistic events while there may be similarities in the end, explanations are fundamentally different kinds of things from causes and so the cause will not be the explanation, even if the explanation includes reference to the cause.

Epilogue: If there were more time. These both were parts of a far too ambitious introduction, but I thought they were interesting enough to merit including at the end.

This was to go within the mechanism section:
However, if we were to speculate based on work by Preston and Epley (2005), there may be cases in which chains of causes may alter peoples notion of what makes an important or valuable explanation which suggests that equivalence many not be an appropriate description, even if in most cases cause and explanation would refer to the same "difference maker"(in Strevens(2008) terms). Similar problems arise based on Lombrozo's (2010) work on causal pluralism, in which we see the the semantics of the situation (e.g., whether actions are intentional or accidental) will change whether one or another "cause" will be identified as "the" cause.

this was to follow:
I then return to the issue of whether it matters that we consider events as individual tokens that merit explaining per se, or only as instances of types that themselves are the real currency of causal attributions and explanations. In considering the types account, I examine attempts at unifying the two prior accounts such as that offered by Cheng (Novick & Cheng). I argue that while this is a promising approach for unifying mechanistic and statistical notions of causation, without further assumptions, its focus on types rather than tokens forces it to give neither an account of explanations or causes of *an* event in the sense intended by the mechanists (our second group). An attribution might be made, but then it will have to fit squarely inside the statistical regularity camp as described before.

This was to follow the van Fraasen section:
I conclude with a brief discussion about kinds of explanation that seem to not fit in a framework such as even to be compared with causation; e.g., the explanation of why something is the member of a particular category or why we cannot have analytic solutions for arbitrary quintic equations (i.e., mathematical facts).

Science :

Cognitive development and learning have been described as analogous, in some respects, to the process of scientific theory change. What are the key similarities prompting this analogy? What are some important dissimilarities and criticisms of the "children as scientists" position? In the end, does the connection between development and science improve our understanding of either human psychology or how science works? Be sure to support the similarities and differences that you identify by appeal to theoretical arguments and claims from your readings as well as relevant empirical results.

Highest level answer:

Cognitive development and scientific theory change are similar in that both have extensive reliance on social aspects including the cognitive division of labor which relies on both the explicit deference to experts and a clear discpline structure in the content such that there can be experts and the social support needed for exploration/play to occur.

Additionally within theories themselves there are parallels of between those theories that develop in their content and their dynamics. One hallmark of (most) science (to the logical positivists endless frustration) is the postulation of unobservable entities. Children too are known to postulation unobservable entities to explain the occurrence of observable events. Additionally, one of the common themes of developmental research are the basic postulation of ontological kinds, the claims and inferences that those kinds merit, and the changing of those kinds as greater data accumulates (e.g., plants and animals → living things).

In some ways the differences between the two are harder to pick out not due to the absence of differences but to their proliferation. Scientists have a full history (18+ years) of education, young children by definition, do not. Scientists often rely on formal reasoning to explicitly calculate what their results even mean, children, if they are using statistics (which the evidence suggests that they are), are doing so implicitly or at least not with the same numerical system used by their scientist-analogs.

More crucially, it seems that if theories regarding core cognition (Carey & Spelke) or the role of language in helping create complex categories are correct, young children's development looks very different from that of scientific theories, given that interdisciplinary or "bridge" science is some of the most rewarded on the most easily quantified currency used by scientists, citations (Shi et al.,). Which gives us a secondary crucial difference, if Kitcher, Hull and Strevens are correct in

supposing that one of the defining characteristics of science is not curiosity in and of itself, but the approbation/applause of their peers then the motivational consequences will be vast.

I conclude with a positive note on the prospects for a computational cognitive science of science as a way of unifying the study of these fields.

Similarities

Cognitive Division of Labor

Science (as we know it, at least) would be impossible were it not for the expertise and deference structure of science (Hull). That is the world is incredibly complicated, and in order to understand it with any great precision requires the cognitive division of labor (Keil, Kitcher, Weisberg & Muldoon, Strevens). That is, individuals specialize in particular subfields of information and are then trusted by other members of their community. How exactly that trust is established is yet undetermined, but with the appropriate checking measures in place (e.g., gossip and first-to-publish credit assignment(Kitcher; Strevens)) the structure of science will encourage self-correction and long-run accuracy. The social structure also plays a role in governing what it is that scientists discover, but that discussion will be reserved for exploration and play.

Children are faced with the same task of attempting to understand the world, and are in an even worse state ignorance than scientists about the complex patterns in both the social and physical worlds. Children similarly need to rely on the experts in order to gain from the accumulated "cultural capital" (Keil & Newman) in the surrounding culture. To do so requires that children can recognize not only which people are experts on what, but what sorts of experts there even could be (Keil et al., Keil, 2010). It also suggests that they would benefit from being able to track the reliability of testimony and understand pedagogical intent (Gopnik & Wellman).

Fortunately for children it seems there are experts, and fascinatingly these experts in science seem to cluster around the same domains that cognitive developmentalists have identified as being some of the domains specified in early childhood(to see the appropriate cognitive development domain merely append the word folk to the following list;(Gopnik; Carey; Keil; Gelman; others)) * biology * physics * psychology * cosmology * mathematics

A philosopher noted that Aristotle generally speaking happened to divide up his writings among the fields that we would have sorted them into today even if we would give them . I don't remember that philosopher's name, but the point is that it seems that there may be "joints" to be carved into in our world (as Gopnik put it), and those joints may be disciplinary boundaries that both children and adults perceive

Finally, because children will be getting conflicting information from their cultural sources and their own senses (e.g., the world is flat and they are told that experts think it is round) they will need to bring these facts to bear with one another. While some would argue that scientists merely walk there way past these

arguments by arguing that it is not in their area (Lakatos; Kuhn; Hacking) for theories to be able to be unified, some degree of reconciliation among competing parties need to occur (Hacking). We just hope that unlike children the state-of-art theories are not as malformed by our scientists as the notion of a flattened earth would be.

Exploration and play

One similarity noted by several writers (Hacking, Schulz, Gopnik) are the similarities between scientific exploration and play. Gopnik even argues that explanation is to thinking as orgasm is to sex in that the experience of an apparently successful explanation is motivation enough to encourage scientists and children alike to pursue learning (though Trout might argue given the cognitive biases that scientists and children alike face, those orgasms may be more along the lines of literal intellectual masturbation).

If Ullman, Goodman and Tenenbaum and Henderson et al. are correct though, the space that we have to search through in both cognitive development is very large (thus the need to rely on cultural capital in order to maximize whatever knowledge we do find). Because of the space though, and because of the problem of local maximums, we may not wish to optimize at every moment. Indeed, though children probably do not like it when their parents become angry that does not stop them from going through the "terrible twos", which Gopnik describes as searching through the space of acceptable social behavior. Similarly monumental scientific achievements often seem to require substantial departures from the expectations of the field, this is the distinction that Kuhn raises regarding paradigm shifts, and one of the motivations behind Feyerabend and Hacking's respective endorsement of scientific anarchism/dadaism and anarcho-rationalism. We need to have the freedom to fully explore the problem space without constraint in order to effectively be able truly novel solutions to our problems.

Differences

Superficial differences

Scientists tend to be adults and have years of formal education, extensive vocabularies throughout their career as scientists. Young children, do not; these skills are built up over time.

Scientists often rely on complicated statistical techniques (some of which they barely understand if they do at all) to explicitly be able to say what it is that they're assuming and what it is that their results mean. If children are using statistics they are using statistics (which the evidence suggests that they are), are doing so implicitly or at least not with the same numerical system used by their scientist-analogs.

In addition to statistical techniques scientists are often given extensive training on devices and apparatuses that act as extensions of themselves in ways children (however nice their toys) could not even begin to fathom. No one has predicted that a toddler will be the one to explain the Higgs Boson and the LHC's results to

the world.

Children's knowledge is whatever it is that they obtain, whereas in order for information to be passed through the scientific channels of communication, it needs to be vetted by their peers. Something children rarely have to do (though they may be reprimanded for what they say, not by their peers, but their parents).

Core differences

If Lakatos is correct that science can be correctly described by research programmes of (fairly) coherent sets of theories, then that would seem to be a difficult claim to hold if children have independent cognitive modules in which knowledge sits separated into neat bins. Some examples of those modules are those described pieces of "core cognition" of approximate and file-drawer number or the division of the biological into plants versus animals and never the twain shall meet.

Furthermore, scientists are rewarded for the unification of disparate theories (Shi et al.), with papers that bridge separated fields that are individaully densely clustered being given the highest citations, whereas seemingly arbitrary citations in the vein of core cognition being penalized heavily with very few citations.

Integration of knowledge

Core cognition? vs. research programmes

Common Reasons; Common ground

Scientists are humans endowed with human psychology. Additionally, they were all once children. So then the question to be interesting is not just how the scientists are affected by the cognitive processes that affect all humans but how scientific ideas are going to be paralleled by the progression of ideas in children, and vice versa.

In that sense, what I see as being the greatest potential benefit for science is the burgeoning of the computational work, particularly that of structured probabilistic models as embodied in Henderson et al for science and Ullman et al for cognitive development. The cognitive science of science, when made formal allows us to know precisely what our claims and assumptions are. This kind of formality while also having deep expressivitiy is lacking in the philosophy of science. On the other hand the detailed pictures attempting to integrate specific aspects of the development of individual ideas and theories (of the type studied by historians of science) are not often studied in cognitive development. We will study children on average but not children as particulars, though we argue that such prior knowledge is actively shaping the entirety of the learning process.

Epilogue

These would have been included if I'd had more time: in the similarities

Theories, hidden entities and mechanisms

undergoing fundamental shifts in theory (theory of mind (Gopnik; gopnik & wellman, intuitive biology (carey, 1985)) category shifts Discovery of hidden entities

Qualifying quests, 2: Rational Probabilistic Models of Higher Level Cognition

Learning causal theories, continuity and time and their role in the level at which we analyze human theories

Describe some of the recent work on rational models of theory learning about real-world phenomena (e.g., causal theories). How might this work be strengthened by incorporating temporal and continuous-dimensional concerns to these models of theories? How does adding this complexity affect our interpretation of the meaning of theories? More specifically, can time be understood at a computational level(as a part of the overall goal of theory-making), or is it relegated to the algorithmic level(where it merely affects the dynamics of theories, not their "meaning") and how does our answer to that question change our picture of theory-craft?

(note to readers: the order of events deviates slightly from that as predicted in the introduction. Unfortunately, such is the shifty nature of time you never know where its going to leave its effect :))

Introduction

One of the central problems for cognitive science that ranges across the disciplines that it spans, is understanding how we generate, learn and alter our theories. Because of the scope of such a question I will limit my analysis to the question of causal theories, which include information about ontology (including, categories and features), plausible relations among objects in those categories, and the form of those relations (including different ways to cache out causal influence). Though these descriptions are drawn from a particular proposal regarding causal theories, i.e. Griffiths & Tenenbaum (2009) *theory-based causal inference*, it is a useful framework to consider even if we move into higher-order causal theories based on Bayesian non-parametric priors of the sort considered by Kemp, Tenenbaum, Niyogi and Griffiths (2010a), Kemp, Goodman and Tenenbaum (2010b), or probabilistic, higher-order logical grammars such as Goodman, Ullman and Tenenbaum (2011), Bonawitz, Goodman, Tenenbaum and Gopnik(2012), and Ullman, Goodman and Tenenbaum (2012). I begin by briefly reviewing some of these frameworks and some of their computational and empirical results.

After discussing some of the basic pieces of theories of probabilistic theory learning, I will delve into some of the details of how the current systems can be extended to include other kinds of data, both continuous and continuous-time data. I focus then on ways in which temporal information changes how we might

understand theories. In particular, I look at how categories may shift in meaning over time (Navarro, Perfors and Vong), differences in plausible causal relations and their functional form, some of the problems that considering time solves, and the some of the new challenges that arise if we do let time into the semantics of our theories.

Finally, I return to the proposition in Ullman, Goodman and Tenenbaum(2012) in which they propose that we can understand theory learning as a form of stochastic search. They argue that their model is described at the algorithmic level (in terms of Marr's (1982) three levels) as opposed to the computational level at which these other theories have been proposed. I argue in line with Danks (2013) that the distinction between algorithmic and computational is less relevant than its assumptions about the underlying problem and its commitments to the realities of different entities. In particular, I take issue with the equivalence of the computational level theory with the use of an implicit universal hypothesis space (using terminology from Perfors, 2012), and suggest that, in light of considerations on the role of time in understanding causal relations, the relegation of time to the dynamics of theories sacrifices a good deal of the interesting problems yet to be solved in human cognition.

Ontology

When one considers the ontology of a theory, one is discussing the types of objects and the nature of their observable features. For example in the case of medicine we might have symptoms, pathogenic diseases, genetic diseases, pathogens and genotypes. These may have different observable features or may be not directly unobservable (as DNA was not until the 1950s). It is important to note that so long as they are systematically related to observable entities, we can perform inference about these unobservable entities, we just never directly observe them (Cheng; Griffiths & Tenenbaum, 2009).

In addition, the ontology will need to tell us what *kind* of data will be observable from these objects. This is true even if we are talking about the *same* data. For example, we may be able to see data about someone's temperature as

- binary data that can be present or absent
    - in Kemp's (2012) terminology, *additive features*
    - e.g., "whether or not someone has a fever",
- binary data that can take on various levels
    - in Kemp's 2012 terminology, *substitutive features*
    - e.g., "whether someone's temperature is high or low"
- continuous values
- e.g., "what someone's temperature is"
- time-series of these different data types
    - e.g., "the time-course of someone's temperature"
    - which can both be
    - discrete
        - e.g., "the time-course at each instance of measurement regardless of when the measurement was taken"

- or continuous
    - e.g., “the time-course at each instance of measurement indexed by the exact time that the temperature was measured”.

This is not an exhaustive list by any means. Additionally, I may not been as careful as Kemp (2012, Kemp & Jern, in press) would prefer in distinguishing between the observable data about the categories and the observable data about the features. Nor have I distinguished those data that are observable as features of and those data that are non-causal relations between objects(Kemp et al. 2010). However, my ambiguity is purposeful. Because we are putting no specifications on the semantics of the ontology and the categories and features within, there is no reason to limit ourselves to merely first-order relations as was done in the original *theory-based causal induction* model (Griffiths and Tenenbaum, 2009). We can in theory have higher-order relations among our relations and quantification over entities (as in Kemp, 2012). But we get ahead of ourselves.

Plausible relations and functional form

In addition to representing our entities, we will have to specify how they relate to one another. In many of the traditional settings these relationships are thought to be constantly present (a static feature of the objects), though if they are probabilistic they may be randomness introduced in terms of when the relation is apparent (e.g., all lottery tickets have a static chance of winning the lottery before the numbers are drawn, but only one will actually be realized). And if we take them as given, our plausible relations may give us grounds for solving some of the inferential problems that our representational flexibility forces upon us. For example, we might presume that genetic diseases only result from genes and pathogenic diseases only result from pathogens, but some symptoms are only caused by pathogens, some are caused only by genes, and some are caused by both. If you had a case where an gene-caused and a both-caused symptom appeared, the structure may suggest that the both-caused symptom is more likely to have occurred because of the gene rather than positing a gene and a pathogen (nb: this rests on some simplicity assumptions similar to those found in Lombrozo(2007)).

In addition to describing the plausible relations, we will need to specify the form of those relationships. In the binary, additive feature case (where cause and effect are additive features, and some background cause of the effect is always present) common functional forms are Noisy-OR (1-(1-w_b)*(1-w_c)^(|c|), where w_b is the background rate, |c| is whether or not the cause C occurred and w_c is the cause’s ability to produce the effect absent the background) which represents a generative relation where the cause is assumed to be independent of the background and if both “cause” the effect then it occurs, noisy-ANDNOT (w_b*(1-c)^|c|) which represents similar assumptions with a preventative cause, or a generic relation where the probability of the effect when the cause is present and absent are determined independently of one another (Cheng; Griffiths & Tenenbaum (2005), see also White).

One could imagine that if working with other kinds of data, one would have a preference for different functional forms, for example, raising something to the power of |c| makes sense in the additive binary context, but substitutive binary features and continuous dimensional objects do not carry the same semantics

associated with 0 or 1. It is likely then that we would need a different functional form depending on the kind of data we consider. But before considering different kinds of data we will explore some aspects of the higher-order frameworks for causal theories.

Higher-order causal theories

The ability to represent theories compositionally is one of the major advances offered by the probabilistic higher-order logical grammars, which allows blank predicates to be posited, composed and quantified over. Whereas *theory-based causal induction*could be seen as a way of representing causal Bayes nets (which had met previous success in accounting for causal inference), we can see the most general higher-order probabilistic grammar as representing generalizations of other generalizations to an arbitrary degree (at least, in theory). In doing so, it allows theories of theories, or *framework theories* as posed by Gopnik & Wellman (2012). This too could help explain some of the effects of domain differences as described in Griffiths & Tenenbaum, where different domains have different associated ontologies, plausible relations and functional forms and while learning to cross domains is possible, it takes much more data. But this ability to represent a variety of objects, features, domains, relations, functional forms, etc. introduces difficult inference problems, which unsurprisingly is a common critique of the structured probabilistic approach (see, McClelland et al. and Jones & Love).

One of the ways that this inference problem is made easier is by the use of *hierarchical Bayesian models*(HBMs) which allow arbitrary dependencies between variables that can have arbitrary values, including as grammars, programs or logics. This allows data at the bottom level to propagate their influence upward over complex structures. This seems to be of little note, but one feature of this is what is known as the *blessing of abstraction* – disparate data can combine their influence to help determine the distribution on *over-hypotheses*, that is, hypotheses about hypotheses (Goodman et al., Ullman et al., Gopnik & Wellman). Because of this, you may very rapidly learn abstract knowledge, even before you learn the concrete instances; this finding has developmental evidence supporting it with children learning basic-level categories (“dog”, “cat”) long before they learn more specific categories (“great dane”, “russian blue”)(Carey, 1985; Keil & Newman, 2008; Gopnik & Wellman 2012).

It is this kind of analysis that allows the simultaneous inference of categories for which meaning only exists in relation to one another (Kemp et al.) — which has been a long-standing problem in conceptual role semantics (Hall). It additionally is capable of simultaneously inferring the identity of objects from unknown categories related in unknown way and the rules for relating them, even with as little as one data point (with appropriate background knowledge)(Bonawitz et al. 2012). Using real world stimuli, it has been able to recover a good amount of held-out structural information from medical diagnosis and ontology databases (Kemp at al.). Indeed, the power of HBMs to represent non-linear “paradigm shifts” (Kuhn, 1962) has resulted in them being proposed as a model of not just human theory learning individually but the progress of scientific theories more generally (Henderson et al., 2010).

Additionally, the compositional structure of these frameworks gives them the ability to represent powerful assumptions while making those assumptions model explicit and transparent for analysis. Frameworks allow the structure to develop

complexity as needed in order to best accommodate the data. The models in these frameworks will often employ Bayesian non-parametric methods that postulate an infinite number of categories (thus non-parametric), but with only a finite amount of data only a finite number of categories will be created. If one thinks instead of categories of rules and relations between the categories it becomes apparent just how powerful such a open-ended "grammar" can be. Indeed, as bizarre as it may sound, from relatively limited data (with limited generality), Goodman, Ullman and Tenenbaum were able to extract a definition of causality itself (or at least some of the properties of the relation "intervention" as postulated by Pearl, 2000).

Levels of analysis

Most of these models are posed as a computational-level analysis (in which the model describes the goal of the cognitive system(Marr, 1982)) of human causal theory learning. However, Ullman, Goodman, and Tenenbaum pose a closely related model that uses Markov Chain Monte Carlo sampling in the theory space (or as they call it the *probabilistic language-of-thought*) to model the introduction of genuinely "new" theories and the evaluation and learning once those theories are proposed. To do so they assume that a cache of law-templates (e.g., reflexivity R(a,b)→R(b,a), transitivity R(a,b) + R(b,c)→R(a,c)) exist that learners can call upon to then fill in with the data they are obtaining from their senses.

Because they are proposing that people have a limited sample of theories at any one time and that sample changes gradually, they argue that they are proposing an algorithmic level theory that specifies "how" some cognitive system accomplishes the goal defined in the computational level. They argue that the computational-level theory effectively has a hypothesis space that literally covers the whole universe of coherent theories, because (absent any computational limitations) their model could effectively reach any potential theory (nb: including an infinitude of "correct theories").

I would argue rather that this is still answer the question of what the goal of the cognitive system is, but rather that it is a more realistic version of that goal. There has often been an equivalence made between "computational constraints" and "algorithmic level theories", but there need not be such a relation(Danks, 2013). Indeed, in the original proposal Marr suggests that one could have multiple computational level theories for the same phenomena and that they all would need to be constrained by lower level considerations. I see this as a problem of the way that temporal information is encoded in the model, namely as a way of exclusively capturing the dynamics of theory change rather than having the theories say anything about time. Once elements of time begin to be incorporated even at the ontological level (as in the exploding cans of Griffiths and Tenenbaum, 2009), there is little reason to not also consider the changing way in which meaning changes over time as well.

Then, if one is allowing the very meaning of the entities within the theory to change it seems even more untenable to consider the computational level to necessarily include the "implicit hypothesis space" as described by Perfors (2012). That is, the set of all hypotheses that could possibly ever be proposed by the theory. This is in contrast to the "explicit hypothesis space" that is actually being considered by the theorizer at any time. It seems as reasonable to say that the goal

of the theory making system is to make a theory not to find the "correct" theory for all time, but rather to produce one that gets the world as "right" as it can for now. That means that the theory itself, the entities within the theory, and the measures/data it has available will all be evolving in time. Thus, its seems reasonable that the potential hypothesis space will be growing contingent on those particular pieces of data and the theoretical positions that make different theories conceivable via the apparatuses that they render possible.

The idea that there was something to observe on the moon motivated Galileo to turn his telescope to the moon, data from which eventually lead to Newton's theory of gravitation, which in turn allowed us to guess Neptune's existence due to perturbations based Newton's theory. Each of these advances relies on the existence of the previous advance for it to even exist (there are no perturbations from Newton's theory without Newton's theory).

It seems that there is space for a computational level theory that needs to sample from the space of theories in precisely the way that *normal science* progresses(Kuhn, 1962) or research programmes (Lakatos, 1978) grow, by iteratively suggesting changes to the current theory until one runs into such a calamity that everyone scrambles to salvage as many of the phenomena as is possible in some other theoretical framework. There is no reason why finitude of sample need be a process claim, anymore than the decay of memory need be in the shape of an exponential distribution (Anderson, 1990) or the decay of a causal effect need be a dirac-delta function (Griffiths & Tenenbaum, 2005). Or perhaps the better way of putting it is as Danks (2013) did, that the issue is that these problems are not distinguished on the scale of computational vs. algorithmic; rather they merely postulate different assumptions about what things really underlie our theories.

A timely digression

Though out of order, before concluding, I want to attend to some of the central reasons why one attend to the inclusion of temporal information in the semantics of theories.

First and foremost temporal considerations affect people's causal cognition. Hagmayer and Waldmann point out that time helps people to parse events, to know which of several causes is likely to have resulted in an effect (see also Garcia effect and Buehner and May, 2003), and that different time schedules for experimental designs can suggest dramatically different effects. For example, if during a treatment something has an effect and then after treatment that effect does not last if you consider only "post-beginning-of-treatment" as your treatment effect if enough time is present (see also how temporal/interventional assumptions in experimental design and treatment effects can alter inferences in Freedman, 2006). Temporal information can sway people's judgments even if they are intervening and the contingency information suggests otherwise if that time does not reflect the accurate underlying causal structure (Lagnado & Sloman, 2005). Additionally, people represent temporal distribution of cause-effect delays in addition to some interpretations of covariation information, are dissuaded from causal beliefs if there is a substantial range of variation in those delays, and discount a causal effect even if that variation is induced by an observable "hastener"(Greville & Buehner 2007,2010; Lagnado & Speekenbrink).

Second it seems from some perspective it is the relative change in events rather than their absolute value that predicts which inferences people make. This has been established in both adults (Rottman and Keil, 2012) and children (Rottman, Kominsky and Keil, 2013) showing that they rely on alternations in the variables from trial to trial to infer causal relations rather than the absolute state values as would be suggested by a traditional causal Bayes net. Changes in the values of distribution too have been argued as a way of determining, beyond traditional covariation information, *which cause* was the *actual* cause of an event (Glymour et al., 2010).

Third, to speak in terms of our previous nomenclature, we have evidence that people are able to learn that the meaning of terms can change over time. Navarro, Perfors and Vong had people learn whether boxes should be categorized as "high" or "low" and the boundary between high and low went up as the experiment went on. They found people were able to do substantially better than chance at detecting the category membership which would have required updating their decision-boundary as time went on, suggesting that participants had no trouble learning that meanings could change over time (though they adjusted less quickly than would have been optimal, though that is consistent with a Bayesian updating formulation of the problem). Navarro, Perfors and Vong motivated their work from the idea that `cell phone` over time has had constantly meanings (even just in terms of getting smaller and lighter over time).

In continuous time we need to shift our ontology to represent both interval-like states (such as lights being on from [t_1,t_2]) and instantaneous events (such as turning the lights on at t_i). This necessitates developing systems for defining the functional forms and relations between these variables. As it stands, there are capable models for handling generic interval-interval relations (Nodelman, Shelton and Koller; Gopalratnam, Kautz and Weld), generative and preventative event-event relations (Simma; Simma & Jordan; Blundell, Heller, and Beck; Cho, Galstyan, Brantingham, Tita) and interval-event relations with conditionally-constant effects (Markov-modulated poisson processes). Event-interval relations are sufficiently more complicated since events occur, by necessity instantaneously, so it is less clear how they can affect intervals that exist extended over time. And these possibilities do not even begin to exhaust the list; the point is, that there's a lot of representational capacity there, and (at least in continuous-time) the information is there to make at least some of those inferences work.

And the rich nature of time (especially continuous-time) opens up an array of new *kinds* of inferential problems. For example Rottman and Ahn (2009) demonstrated that people will represent habituation and sensitization effects in individuals that are exposed to the same stimuli multiple times but will not do so between individuals even if those individuals are similarly sampled on later days. This may seem obvious that habituation and sensitization can only be diagnosed with repeated observations of the same individual, but the current models of causal induction that estimate fixed parameters will fail to capture this; this can be thought of as a causal analog to the changing categories in Navarro, Perfors and Vong study.

A fourth reason to incorporate time in our causal theories is simply that temporal information (especially continuous temporal information) is richer than non-temporal info. That can be seen in that there are kinds of relations such as cycles and even auto-loops such as self-prevention a neuron's refractory period after

having an action-potential. Additionally, if you have veridical continuous-time information and are representing the occurrence of events as Poisson processes, if you can identify any structure at all (cf. Blundell, Heller, and Beck), you should be able to identify otherwise Markov-equivalent graphs (such as a common-effect and a causal chain) merely from observations. This goes beyond what is possible with traditional contingency information.

Time for theories; a short-on-time conclusion.

Time is an incredibly important aspect of our causal theories in life, and understanding its role is integral to the very goal that cognition is attempting to deal with — predictive, explanatory and interventional inference in a world in flux. We will not have a full characterization of the mind without considering at least the phenomena described above, and I would argue that we will not understand theoretical meaning, generation or change unless we deeply interweave time into the goals of our formal frameworks. Ullman et al. should applaud themselves for making an excellent computational model (not algorithmic) of one way time says something about our theories (their dynamics). Now we just need to let theories have something to say about time and let the cycles continuous-time makes possible continue to spin.