Abstract

In the literature on explanation, philosophers have proposed different conceptions of structural explanations. In this paper, I explore how several seemingly disparate accounts of structural explanation can be tied together with a central notion of abstraction borrowed from the philosophy of mathematics. Some explanations involve abstracting from a subject as an individual to seeing that individual as a node in a network of explanatory relations. I will then tie this account of structural explanation by abstraction to Wilkenfeld's (2019) account of understanding to show how this class of structural explanations by abstraction is understanding-conducive.

Understanding Abstracting Structural Explanations

§1 On the Very Idea of an Abstracting Structural Explanation

In the literature on explanation, philosophers have proposed different conceptions of structural explanations. Specifically, explanations that have involved higher order causes, systematic constraints, and competition that restricts the possible combinations of individuals have all been dubbed "structural" at some point or other. In this paper, I explore how several seemingly disparate accounts of structural explanation can be tied together with a central notion of abstraction borrowed from the philosophy of mathematics. Some explanations involve abstracting from a subject as an individual to seeing that individual as a node in a network of explanatory relations. This way of looking at structural explanations is not wholly novel, but rather combines the literature on structural explanation with relevant insights from the philosophy of mathematics. I will then tie this account of structural explanation by abstraction to Wilkenfeld's (2019) account of understanding to show how this class of structural explanations by abstraction is understanding-conducive.

The general idea of an abstracting explanation is that the very same object can be viewed either as an individual token, subject to causal and explanatory forces at one level of description, or as a node in a higher-order structure, subject to different causal and explanatory forces.

This notion of abstracting explanation has been articulated and defended clearly by Christopher Pincock (2015), who argues that appeal to objects of higher degrees of abstraction than the explanandum can account for the explanatory nature of certain examples in mathematics. (We borrow Pincock's (2015, p. 865) characterization of the relevant notion of "abstraction" in terms of types that have instances.)[1]

[1] "A canonical case [of abstraction] to keep in mind is the relationship between a type and its tokens. A given type may have many tokens as instances. Exactly what makes a token an instance of this or that type is subject to debate. For our purposes, we need only assume that the instantiation relation in asymmetric. A type has a token as

This paper has four main contentions. First, in §2, I suggest that at least one important type of explanation by abstraction concerns abstracting from a token's role as an office-holder in one explanatory framework to its role as an office in another explanatory structure. Second, in §3, I argue that the process of abstracting to a higher order node and then articulating the explanatory relations that that node enters into can account for much of what has been discussed under the header of "structural explanation", particularly in the social sciences. In §4, I explore the possibility that the explanatory network need not be causal,[2] but rather can incorporate a much broader class of theories of explanation. In §5 I discuss how the class of explanations described here can be conducive to understanding.

§2 Offices, Office-Holders, and Abstracting to a Higher Structure

In this section, I introduce one particular form of explanation by abstraction—abstracting structural explanations (ASEs). These explanations have two parts. First, we view the subject of the explanation as being characterizable as a mere occupant of a node in a higher order structure. Second, we place that higher order node in an explanatory framework, saying first that it explains the explanandum, but also what in turn explains it.

To borrow Stewart Shapiro's (1997) terminology from the philosophy of mathematics, abstracting explanations are those that relate individuals in their role as offices (such as being the President) as opposed to individuals as particular officeholders.[3] Importantly, what is an office at one level of description could be seen as an office holder from the perspective of a higher level of description, and so abstracting explanation can be iterated to ever higher order kinds.

The point can be illustrated by an example (adapted from Haslanger 2016). Jill and Jack have a baby, Jeremy, and we try to explain why Jill stayed home from work to care for Jeremy. We could provide a singular causal explanation, focusing on Jill's individual motivations. Jill stayed home because she decided it was the best thing for her and her family. If "why did Jill stay home?" is the question, "because she decided to" is a perfectly good answer. However, we can then ask "Why did Jill decide to stay home?" This sort of question can be answered at the same level of description as the former, treating Jill as an individual. Maybe her friends said staying home would be fun. Maybe she hated her job. But we can also look at the role Jill played in a broader system, and say that she stayed home because she filled the office of mother. That she is a mother makes Jill susceptible to different kinds of forces. For example, mothers get paid less than fathers for doing the same work (Connley 2021). One might also think that mothers are expected to stay home with their children, if their partners continue to work and they do not jointly have enough money for paid childcare (this particular causal relation will be marked out

---

an instance, but not vice versa. Many mathematical structures have concrete systems as instances." (Pincock 2015, p. 865)

[2] Or even, as in Pincock, based on genuine dependence relations.

[3] One famous source of confusion is that the very same utterances can be used to articulate relations between offices and between office-holders—saying "Of course the President believes in bombing Iran" has starkly different practical implications depending on whether "the President" is being used to pick out an office or a particular individual.

as a special case—competitive abstracting explanation—discussed below). These forces would affect whomever happens to occupy the office of "mother", and so abstract away from Jill's individual Jill-ness.

As noted, the process of abstraction and explaining can be iterated, going to ever higher levels of description. Just as we can ask why Jill decided to stay home, we can ask why mothers decide to stay. Again, the explanation can be what we might think of as horizontal—cast at the same level of description as the subject of the explanandum. Mothers stay home because women are paid less for the same work. But we could go another level higher, switching the role of "mother" from being an office to being an office-holder. Specifically, we could argue that the reason mothers get paid less is that "mother" occupies the "minority that did not have historical control of capital" role, which has nothing to do with mothers' gender per se. If other minorities without historical control of capital would have been likewise discriminated against by the market, the new abstracting explanation has removed motherhood from the equation in exactly the same way the original abstracting explanation removed Jill. The same office can also be described in ways that treat the same office-holders as offices in other higher order structures. It could be (as implied by Haslanger 2000) that the role of "mother" is necessarily part of a network of subordination, which would lead to other explanations in terms of subordinating power dynamics.

If ASEs explain phenomena characterized at one level of description by explaining the same phenomenon in terms of explanans at a higher level of description, two natural questions are whether there can be any absolute metric for when an explanation is an ASE—doesn't whether or not we're abstracting depend on where we started?—and whether one particular level of explanation is to be privileged as better. The answer is the same in both cases, which is that it depends on context—albeit in slightly different ways. Whether an explanation is abstracting depends on one's starting point, but that is no more surprising than (and is in fact a complement to) the fact that whether an explanation is reductive depends on one's starting point as well.[4] In §5 I argue that which ASEs are better will also depend on context, specifically insofar as there is a contextual parameter in accounts of understanding.

Not just any abstraction is the foundation of an ASE. Only an abstraction to a node whose occupation explains the explanandum by whatever theory of "normal" explanation one has (see §4) will qualify. For example, abstracting away from Jill to someone who occupies the astrological "Aries" node does not figure in an ASE, because nothing about occupying that node explains the decision (on any plausible account of genuine explanation). In fact, even nodes that correlate with the explanandum but do not explain it—say only people born in September happened to stay home with their children this week—would not figure into an ASE. Conversely, being a mother does figure into at least one standard explanation.

[4] The point that abstraction and reduction are symmetrical is borrowed from [acknowledgement omitted].

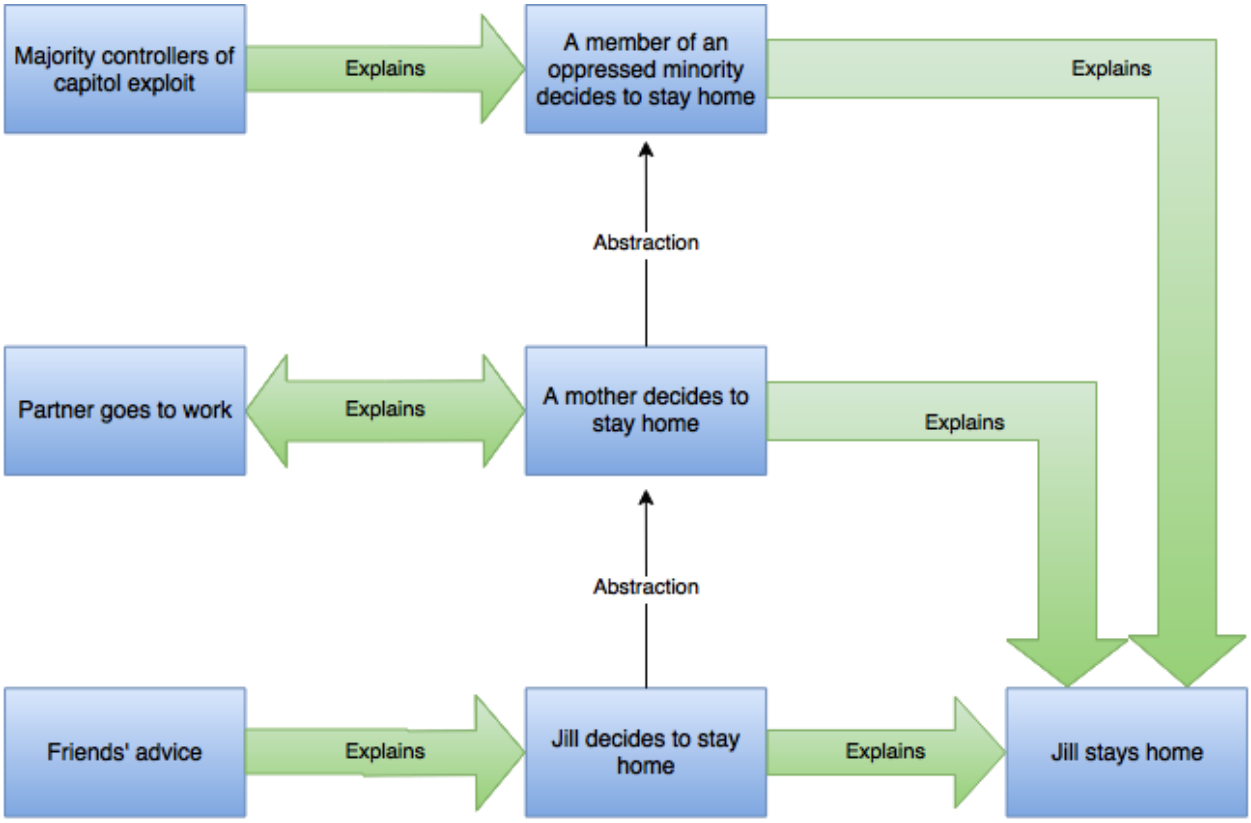

Figure 1: The general (simplified) form of abstracting structural explanations. In real ASEs there would likely be more horizontal links on the left-hand side.

§3 Unifying Different Strands of Structural Explanation

One can see this notion of abstracting structural explanation as what drives Sally Haslanger's (2016) account of structural explanation, in that objects are being considered as constrained by other forces in a complex structure (e.g., social hierarchies). Haslanger seems to see herself as offering the social equivalent of what Shapiro gives us for mathematics (e.g., 2016, p. 118).[5] But it can also make sense of much more mundane explanatory endeavors. Suppose we want to know why John broke his leg, and are told that he broke his leg because he fell down the

[5] However, see Barnes (2017) for a distinction between the projects regarding the question of whether individual existence depends on the structure.

stairs. We could then ask why he fell down the stairs, and be told that he fell because his strength gave out—this is the singular causal explanation. We could also ask why he fell down the stairs, and abstract away from his individual characteristics and be told that he fell down the stairs because he suffered from Chronic Fatigue Syndrome and our buildings are not sufficiently accessible. This latter explanation is what Dretske (1991) calls a structuring cause, but it is really just another form of seeing an element of the original explanation as a node in a higher order structure. By focusing only on one particular set of John's properties, the explanation treats John as a mere occupant of a particular position that's subject to higher order forces.

Somewhat similarly, Jackson & Pettit (1992) argue that structural explanations are those that make it extremely probable that some cause or other would have brought about the effect. To take their example, a raise in the temperature of the water in a flask will explain the flask's breaking—at a finer level of description, what really caused the break was some particular molecule striking the inside of the flask at a fortuitous spot. Nevertheless, Jackson & Pettit maintain that the story about boiling water is still explanatorily "interesting" (107) and "useful" (117), because it "makes it probable" (118) that some water molecule or other would impact as it did. At first, the talk of "making" something probable might seem causal, but on their account it is not. They say the explanation comes from the fact that "the rise in temperature means nothing more or less than that the rate of motion of the water molecules will increase, and if the rate of motion increases then it is more than likely that some molecule will have the effect explained." (ibid.) In their (1990), they are clear that the notion of "A meaning nothing more or less than B" is that A is a characterization of the same thing as B, but at a higher level of abstraction. In this case, the abstraction is that statements of temperatures rising are really abstractions from statements about the momenta of a group of molecules. The explanation of this particular water molecule's effects are a token of a type of water-molecules-in-a-heated-liquid. This explains a particular phenomena by casting it as an instance of an abstract kind.

But we can go farther. Consider the notion of structural explanation expounded by Garfinkel (1980). Garfinkel argues that we have a structural explanation when there is a truncation of the possibility space to disallow certain otherwise potentially possible combinations of states of individual members of a system. In one of his main illustrations, he asks us to consider a teacher who assigns grades on a hard curve, according to which his 50 person class will have 1 A, 24 Bs, and 25 Cs (Garfinkel 1980, p. 41). If we ask why Mary got an A, one could attempt to give a singular causal explanation, but many/most such efforts would fail. In this circumstance it is insufficient to say that she got an A because she wrote a thoughtful and well-researched essay, as the best B recipient might have met those criteria as well. Rather, the structure of the system puts Mary in direct competition with her peers, so we can only fully explain her performance by exploring the relation of that performance to everyone else's.[6] This

[6] Garfinkel's account is best explicated by its relation to individualistic models of explanation. According to individualistic models, the total distribution of outcomes in a system is simply the aggregate of all the outcomes of the individuals that system comprises. This can, however, create a false picture of the possibility-space in scenarios where there are systemic constraints that eliminate certain outcome distributions from obtaining. This is the case with the hard curve, where the range of possible outcomes is not the Cartesian product of all the individual grade

sort of competition between individuals in a system is not solely a feature of idiosyncratic grading curves—the motion of molecules of a gas that cannot take up a position already taken and the distribution of a zero-sum resources like salaries in a firm with a fixed budget exhibit a similar structure.

Another way to describe this restriction on possible phase space is by noting that there is a connection between the occupant of the best essay writer node getting an A and the occupant of the second best essay writer getting an A—in this case, a deterministic inhibitory connection. If the best essay got an A, that precludes the second best essay from getting an A as well. Alternately, there is a bidirectional explanatory relation between the best paper getting an A and the second best paper *not* getting an A (this is how we depict this sort of relationship in Figure 1). We can thus see competitive structural explanation as a particular kind of abstracting structural explanation, where among the causal and explanatory forces operating on the node are inhibitory links constraining the possible distribution of other nodes along some particular dimension. One might wonder what advantage is to be gained by recasting Garfinkel-style competitive explanations in this way; we see three such advantages. First, and perhaps most important, theoretical elegance recommends unifying the different threads of structural and structuring explanations, if possible; the current theory does so. Second, modeling the competition in question as inhibitory links between abstract nodes is easily extended to non-deterministic cases, where the best essay getting an A just makes it less probable that the second essay will get one as well (suppose the teacher does not have a hard curve, but really hates giving out As). Finally and relatedly, the recasting of competitive explanation in terms of abstraction and explanatory inhibition invites the tantalizing possibility that more distant explanatory connections—seeing how causes or other explanatorily prior states earlier in a system affect much later states—can be explored using the formal tools of neural networks.

assignments ($3^{50}\approx10^{23}$ possibilities), but only those that meet the teacher's pre-established criteria ($\approx10^{15}$ possibilities).

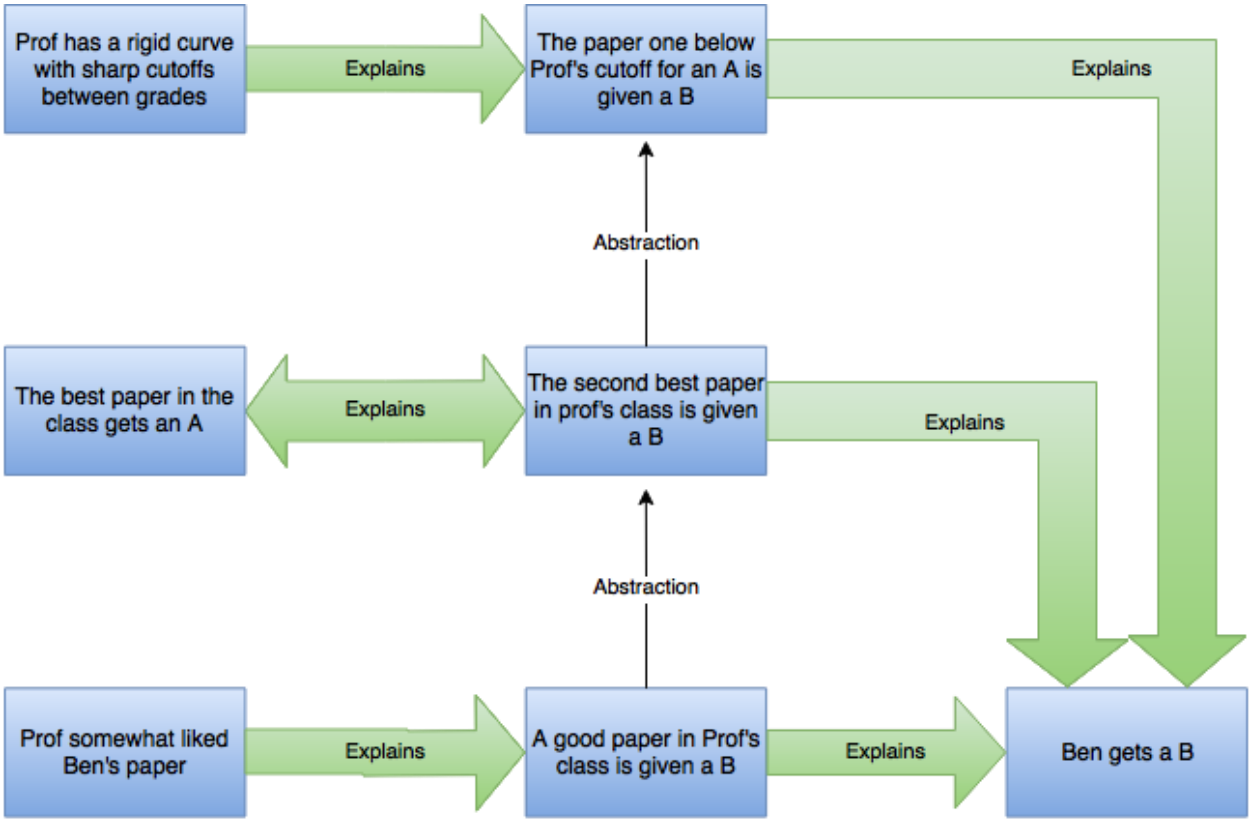

Figure 2: The link in the middle level is inhibitory. Interestingly, the nature of this link encodes the information from the highest level (that there is a hard curve) without actually stating it, which furthers the parallel with neural nets.

This framework can be utilized to see the connection between the factors influencing Jill's decision to stay home. There is a one-directional causal link going from women's only getting paid 75% of what men get paid and her decision not to go to work, but a two-direction inhibitory link between her going to work and her husband going to work. The basic model of explaining by relating a higher order description of the original phenomenon to other nodes via same-level explanatory links is the same in both cases though.

In summation, we can use some of the machinery developed for explanation in mathematics to account for what the various species of structural explanation have in common.

§4 ASEs and Causal Explanation

So far, we have defined abstracting structural explanations in terms of a preexisting notion of intra-level "normal" explanation. This maintains an intentional neutrality regarding the general form of explanations. In this section we discuss three accounts of explanation, and argue that ASEs form an interesting complement to each of them.

First, we address the dominant model of explanation, which is that explanations function by picking out causes. On an intuitive level, ASEs of the sort described in Figure 1 illustrate the way causal descriptions can be offered at different grains of analysis. Eschewing the intuitive for a more specific account of causation, we can see how abstracting explanations relate to Woodwardian causal explanations. Woodward (2003) argues that causal explanations put forth relationships between variables in a preexisting variable set, such that one explains when one relates how an "intervention"[7] on an earlier variable can affect the value taken by later variables. At first, it might not be clear how to make room for ASEs on Woodward's picture—how can a variable at a lower level of description be an alternative to a variable at a higher level of description, when the two are logically (rather than causally) related? I argue that the best way to make sense of ASEs is as conscious shifts to a different variable set, with more abstract variables that (relative to the new set) still cause the explanandum. What an ASE does is suggest that we consciously shift from a lower order phenomenon to a higher order one, and explain the phenomenon in terms of the role the higher order phenomenon plays in a causal network described at its own level. The higher level variable (e.g., being a member of an oppressed minority) is in effect an office occupied by the lower level variable (e.g., being a mother).

A second model of explanation generally that nicely complements our account of ASEs in particular is the mechanistic accounts of philosophers such as Machamer, Darden, and Craver (2000) and Bechtel (2008). These authors suggest that we explain by decomposing a system into hierarchically arranged parts and operations. At first blush, mechanistic accounts might seem the reverse of ASEs, as they reduce systems to component parts. However, one can see that the two forms of explanation are intimately linked, as mechanistic explanations also provide the resources for explaining lower-level phenomena in terms of their roles in higher-order systems. (Thus the otherwise inexplicable regularity that blood keeps reentering the heart can be explained by its role in the broader system as blood-pumper.)[8] Some mechanistic explanations—those that characterize the functioning of a system in terms of its role (read: office) in a bigger system—are a species of ASE.[9] The conception of ASE, however, is broader than this, as it applies to systems without hierarchically arranged parts, and even to systems without obvious parts at all (return to the decision to stay home discussed in §2).

---

[7] "Intervention" is for Woodward a term of art, but one close to its natural meaning

[8] This is related to Craver and Bechtel's (2006) contention that lower order components can be affected by higher order causes by "going along for the ride."

[9] This assumes that going up between "levels", as defined in Craver & Bechtel (2006), necessarily involves a degree of abstraction from a characterization of a component at one level to merely serving a particular role at a higher level.

Yet another model of explanation that can account for the legitimacy and power of ASEs is Strevens (2004) account of causal explanations as ways to put forward the causal difference makers in what would otherwise be an intractable causal story. Strevens' account gives us the materials to say why ASEs can be a particularly powerful kind of explanation in some instances and be wholly inane in others. In at least most of the examples cited above, abstracting to a higher level of explanation can allow us to jettison whole layers of details from our causal stories. If Jill's decision was primarily a product of the fact that as a woman she got paid less than her male coparent, then the ASE lets us delete from our explanation all those idiosyncratic details of her particular case that wouldn't have made a difference anyway. This can also show why some ASEs might be relatively bad—if there *was* something idiosyncratic about Jill's decision, then looking at her in her role as mother would not be particularly helpful and could even be overtly misleading. To take an extreme example, suppose Jill was planning to go back to work precisely because she was outraged by the way women are forced out of the workforce by lower pay, but then was unable to return to work after childbirth due to the last second development of a medical condition such as peripartum cardiomyopathy. The explanation of both good and bad ASEs is all for the best—a theory of a form of explanation should demonstrate both why that form can be powerful and why putative explanations of that form can sometimes fail.

§5 ASEs and Understanding

One role of explanations is, presumably, to provide understanding. There are a myriad of ways to parse the explanatory connections between explanations and understanding. To name just a few possibilities, understanding could just be the product of a good explanation (e.g., Hempel 1965), it could aspire to the ideal of having a good explanation (Khalifa 2017), it could help mark out good explainers (Hannon 2018), or it could even be what defines whether something is an explanation in the first place (Wilkenfeld 2014). But no one seems to deny that the two are intimately linked. (Note here that we are talking about genuine understanding, and not the sense of understanding derided in Trout 2002.) Notice though that on all of these accounts of the connection we can draw some conclusions about the quality of the explanation based on the quality of the understanding produced, even if there is disagreement about which is responsible for which.

What then can we say about the quality of understanding produced by an ASE? Accounts of understanding defined largely (Khalifa 2017) or entirely (Hannon 2018) in terms of explanation will not give us much independent footing here—the value of the explanation comes first, and only then can we reach conclusions about the quality of the understanding generated. However, accounts of understanding that are designed independently of (and perhaps explanatorily prior to) accounts of explanation provide the opportunity to independently assess the quality of ASEs. In this section we look at Wilkenfeld's (2019) account of "Understanding as Compression".

Wilkenfeld's account (itself compressed for the PSA word limit…) is that we understand more to the extent that we can get more information useful in a given context from a more

compressed representation/process pair. While an ASE is unlikely to change the process by which representations are decoded, it is potentially conducive to understanding along both other dimensions—useful output increased and minimization of representation (as measured by, e.g., description length as in Grunwald 2005). Knowing that Jill made her choice because she is a mother potentially allows us to delete information about her particular circumstances, thus reducing our description length. Moreover, it will potentially allow us to draw conclusions about other mothers, thus increasing our informational output in those contexts where we might be wondering about people other than just Jill. By contrast, a completely failed ASE (say one in terms of Jill as an Aries) will not allow us to reach accurate conclusions about either the actual causal structure of Jill's world or other mothers'.

As noted, one advantage of seeing the link between ASEs and an independent account of understanding is that it provides us with the tools to assess whether a particular ASE is a good one. It is important to note the critical role of context here. If we are concerned with a pattern of behavior for people in similar circumstances to Jill, the ASE will likely aid compression, thus generating understanding, and thus counting as a better explanation. By contrast, if we're in a context where the details of Jill's particular causal history are crucial—say we're her therapist trying to assess why she's anxious about her particular decision process—then the ASE will not provide much understanding, and so not be a good explanation in that context.

§6 Conclusion

The goal of this paper has been to show how using a concept primarily deployed in the philosophy of mathematics—abstracting structural explanations—can unify various strands of other explanations we see in daily life. When we then connect those explanations to an account of understanding, we can also then see both why they might be particularly valuable in some contexts and why they might not be good explanations in others. I would be remiss not to note that the goal of this paper could be construed as abstracting away from different notions of structural explanation to one overlying type, which I hope will allow the reader to compress their information about structural explanations, gain better understanding, and so consider this paper as providing a contextually valuable explanation of the existence of a variety of explanations generally.

References

Barnes, Elizabeth. 2017. "Realism and Social Structure." *Philosophical Studies* 174 (10): 2417–33.

Bechtel, William. 2008. *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. Taylor & Francis.

Connley, Courtney. 2021. "U.S. Moms Working Full-Time Are Paid $0.75 for Every Dollar Paid to Fathers, Leading to Devastating Economic Losses." *CNBC*, May 5, 2021. https://www.cnbc.com/2021/05/05/full-time-working-moms-are-paid-0point75-for-every-dollar-paid-to-fathers.html.

Craver, Carl, and William Bechtel. 2006. "Mechanism."

Dretske, Fred. 1991. *Explaining Behavior: Reasons in a World of Causes*. MIT press.
Garfinkel, Alan. 1981. *Forms of Explanation*. Yale University Press New Haven.
Grünwald, Peter. 2005. "A Tutorial Introduction to the Minimum Description Length Principle." http://sites.google.com/site/mkrishnarhul/IntroMDL.pdf.
Hannon, Michael. 2018. *What's the Point of Knowledge?: A Function-First Epistemology*. Oxford University Press.
Haslanger, Sally. 2000. "Gender and Race:(What) Are They?(What) Do We Want Them to Be?" *Noûs* 34 (1): 31–55.
———. 2016. "What Is a (Social) Structural Explanation?" *Philosophical Studies* 173 (1): 113–30.
Hempel, Carl. 1965. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press.
Jackson, Frank, and Philip Pettit. 1990. "Program Explanation: A General Perspective." *Analysis* 50 (2): 107–17.
———. 1992. "Structural Explanation in Social Theory."
Khalifa, Kareem. 2017. *Understanding, Explanation, and Scientific Knowledge*. New York: Cambridge University Press.
Machamer, Peter K., Lindley Darden, and Carl F. Craver. 2000. "Thinking About Mechanisms." *Philosophy of Science* 67 (1): 1–25.
Pincock, Christopher. 2015. "Abstract Explanations in Science." *The British Journal for the Philosophy of Science* 66 (4): 857–82.
Strevens, Michael. 2004. "The Causal and Unification Approaches to Explanation Unified—Causally." *Noûs* 38 (1): 154–76.
Trout, J. D. 2002. "Scientific Explanation and the Sense of Understanding." *Philosophy of Science* 69 (2): 212–33.
Wilkenfeld, Daniel A. 2014. "Functional Explaining: A New Approach to the Philosophy of Explanation." *Synthese* 191 (14): 3367–91.
———. 2019. "Understanding as Compression." *Philosophical Studies* 176: 2807–31.
Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.