Causal Modeling and the Efficacy of Action

Holly Andersen

Simon Fraser University

handerse@sfu.ca

Forthcoming in *Mental Action*, Routledge, ed. Michael Brent.

I. Introduction

This paper brings together Thompson's naive action explanation with interventionist modeling of causal structure to show how they work together to produce causal models that go beyond current modeling capabilities. I will, in the process, show why the internal structure of action, where stages are unified by rationalizations into a coherent overarching action, cannot be causal. Actions, and action explanations, cannot be reduced or simplified to causation and mere causal explanation without genuine loss. Despite this, existing causal modeling techniques can be deployed to model action in some cases. By deploying well-justified assumptions about rationalization, we can strengthen existing causal modeling techniques' inferential power in cases where we take ourselves to be modeling causal systems that also involve actions. This capacity for naive action explanation to strengthen causal modeling inferences provides motivation to incorporate it into interventionist approaches to causation.

Action explanation and interventionism are, in many ways, an awkward fit. The former involves all the rich particularities of singular instances of action, rich with normative structure. The latter is built for general pre-specified variables with allowed values, lacking the rich normative structure that is distinctive of action. Unification might seem like a tempting motivation to accommodate (or more likely, offer a reduction of) action explanations within the ambit of causal explanation. But such a move would result in an under-description of genuine structure in the world. Action explanation cannot be reduced to or fully supplanted by causal explanation. And conversely, causal explanation can be better understood by contrasting it with the kind of structure Michael Thompson (2007) calls rationalization. Action explanations involve a modal strength connecting the relata that makes them much closer to what Lange (2012) has called distinctively mathematical explanations, rather than the modal strength by which causal relata are connected. Just as distinctively mathematical explanations cannot be reduced to any collection of causal explanations, no matter how exhaustive, neither can action explanations be adequately replaced by any collection of causal explanations.

Yet, we can rely on action theory to bring more inferential power to interventionist models, by treating rationalizations that unify stages of an action *as if* they were causal connections. It is important to emphasize that they are not *in fact* causal in character; the normativity of rationalizations cannot be adequately represented in interventionist modeling. Because rationalizations unify in a stronger way than mere causation, such a treatment *under*utilizes rationalization in terms of the inferences that could be justified on its basis. We can treat them as if they were causal, use these connections for making causal inferences, and thereby generate models that can make more predictions about what will happen in such systems.

I argue for this by laying out some key pieces of conceptual machinery that are required for using the approach to causal modeling variously referred to as interventionism, causal Bayes nets, or causal structural equation modeling. The Causal Markov and Causal Faithfulness assumptions are substantive, in that they make non-negligible claims about the underlying nature of the systems being modeled and can be empirically checked to ensure that they are warranted. By committing to these assumptions, we gain powerful techniques for inferring causal structure from probabilistic relationships among variables, and for predicting probabilistic relationships from causal structure. Similarly, the rationalization that explains an action by situating it as a means to another action, constitutes a form of constraint on the causal options available to genuinely rational agents. This constraint can be formalized into an assumption, the Rationalization condition, [1] that can be made about a system of causal variables, in a manner analogous to the Causal Faithfulness and Causal Markov conditions. Thus, Thompson's characterization of the internal unity of action can be incorporated into causal modeling once specific conditions are met.

Section II lays out a brief overview of naive action explanation and the relation of rationalization that holds between an action performed as the means and the action the performance of which is the end, highlighting the features that will turn out to useful in incorporating rationalization into causal modeling. Section III contrasts causal explanation with distinctively mathematical explanation in order to draw a distinction between two ways of applying model. It is a key part of the overall argument trajectory to show that naive action explanation behaves like the modally stronger distinctively mathematical explanations, because of the way it is 'applied' as a model, rather than with the comparatively weaker strength of causal explanation. Section IV introduces the role of conditions like Causal Markov and Faithfulness. Section V introduces Rationalization as a new condition for causal modeling. Section VI illustrates the use of the Rationalization condition with the example of driving. Section VII concludes.

II. Naive Action Explanation and Rationalization

This section lays out a brief overview of Thompson's naive action explanation, examining the character of rationalization as a relation that situates the action to be explained as a means towards or stage in another overarching or more encompassing action.

Thompson begins by identifying a characteristic pattern of explanation involved in actions. Following his lead, I will use the example of baking bread. Suppose someone walks into the kitchen, sees you reaching up into the cupboard, and asks why. We often explain the action of reaching up by situating it as a stage, means, or part of a more encompassing action, like getting down the flour. Getting the flour is itself a means or stage that can be explained with recourse to the more encompassing action of making bread. Naive action explanation thus explains by situating the explanandum as an action that is a smaller part of a larger structure that subsumes it and the other requisite action-stages *as* stages of the larger action. One is doing A as part of doing B; one is doing B, then C, then D, as part of doing X. There is a nested structure: getting down the flour is itself comprised of smaller actions, like reaching up, grasping, pulling, carrying. But getting down the flour is then a means to starting the dough, and starting the dough is itself given further naive action explanation as a means to the end of baking bread.

[1] It is key to distinguish between beliefs and action: the rationality of actions, given in their rationalization relations as they unfurl from beginning towards completion, is the specific target here, not an epistemological notion of rationality that applies primarily to beliefs.

Such explanation relies on the 'in order to' that connects the more concrete and limited action to the goal or overarching action into which it fits as a stage. In baking bread, the overall action is not one can *do* except by doing other actions. One bakes bread *by* getting the flour, adding the ingredients, kneading, letting it rise, and so on. There is no separate action of baking the bread that is additional to or separate from the instrumentally performed actions of kneading, rising, baking, and so on (a well-known point since Ryle 1949).

The relationship that bears the explanatory load in naive action explanation is that of rationalization. An action like getting down the flour has a special relation to the action of baking bread. It is not merely that both actions happen to be going on, nor is it that engaging in one causes one to engage in the other; it is rather that the first is done specifically because it is a stage in the second. The performance of the first action is in service to the performance of the second. It is only because of this relationship that explanatory illumination can be shed on the first action by situating it with respect to the second. This cannot be a straightforwardly causal relation: starting the dough by no means causes one to later knead the dough, or allow it to rise.

In explanation via rationalization, both relata are actions. They could not be otherwise, in order for it to be a relation of rationalization, rather than some other kind of relation. An explanation that involved an action as a relatum, either as explanans or explanandum, but involved merely a causal defined second relatum could not possibly be a naive action explanation. This is not to say such explanations cannot exist. It is to say that they would not qualify, by the very nature of naive action explanation, as an example of such explanations. Rationalization as an explanatory relation can only hold between two actions.

Rationalizations, on Thompson's view, can be given a non-final form: one action can be performed in service of another, without that further action being somehow an final end or overarching and self-complete end in itself. Thus, we can find that action A might serve as a stage in the unfurling of a larger action B, which is itself just a stage in some further action C. B can rationalize A, in providing a naive action explanation of it, without there by having to ground that in some final action. Action B may rationalize A; B may in its turn be rationalized by C (see chapter 5, section 2, in particular). B provides explanatory traction on A even though it may be incomplete considered as an explanation required to capture everything about action A. B need not be some final or end action, some not-itself-naively-explained action, to provide substantive explanatory work with respect to A. Rationalization of a means by an end action can explain without the end itself having to have some special quality of finality, or to be further judged in terms of its legitimacy to be undertaken. Even if we don't think someone should be baking bread right now, it is nevertheless the fact that they are baking bread that provides the explanation of their reaching for the flour.

This has the consequence of blocking calls for complete finality in allowable ends. The rationalization of kneading the dough as a means to the end of baking bread does not need to culminate with yet further naive explanation of how baking the bread then fits into some action of being healthy, or enjoying a hobby, or living a fulfilled life, and so forth. The end of having baked bread already rationalizes the stages, without further termination. We can simply explain one action by another, if it fits in the right way, and thereby have improved on our explanatory situation, even though the explanans action clearly itself could be a further explanandum. This feature will allow it to fit neatly into causal modelling, as we see in subsequent sections.

Naive action explanation cannot simply be a new type of causal explanation. There is nothing in starting bread dough that *causes* one to subsequently let dough rise or bake it. Yet knowing that someone has started bread dough does license one to infer to they will be letting it rise

and baking it later on. In such a case, it is not the rationalizing action of baking bread that is the direct subject of the inference. I might infer you are baking bread by noting that you are kneading dough, using naive action explanation; but it is not the same kind of relation that obtains when I note that you are kneading dough and infer that in an hour or two you will be baking it. Baking the dough is also a stage or means towards the end of baking bread, along with kneading the dough. This highlights how one can infer to future actions that are means of the same action: that two actions are rationalized by the same end action provides an inferential handle that connects them as means of the same end. This inferential connection between two actions that are means rationalized by the same action will, in section V, provide the foundation for using rationalization in causal modeling.

III. The model versus the system as primary target of inquiry: comparing distinctively mathematical explanations and naive action explanation

With this account in hand, this section turns to contrast rationalization and naive action explanation with causal connection and causal explanation. By the end of this section, I aim to have shown that action explanation is deployed in a manner closely analogous to distinctively mathematical explanations rather than causal explanations, in terms of how models and systems fit together. This in turn means that rationalization in naive action explanation offers a modally stronger degree of connection than does mere causal explanation.

Lange (2012) defends the claim that there are certain kinds of explanations, which he calls distinctively mathematical explanations, that have a distinctive degree of necessity and cannot be assimilated to causal explanation without loss. One example is that of a mother with 23 strawberries and 3 children. There is no way to evenly divide the strawberries among the children without cutting the fruit. The mother's failure to divide the strawberries evenly among the kids is, however, not merely some causal fact: it is not that she lacks a knife, or is counting incorrectly, or otherwise causally prevented from doing so. Lange points out that it is the mathematical fact that 23 is not evenly divisible by 3 that does the explanatory work. Even though it is something about the physical world being explained, rather than a purely mathematical fact, it is a mathematical explanation and not a physical one involving causation.

Andersen (2017) responds to Lange's claims in several ways. The key response that I want to redeploy here is to make a distinction between between two ways in which a model can be used. These reflect two different kinds of modeling tasks, with different orientations towards fitting a model to a system (Andersen 2017). In brief, one way to use a model for a system is such that the system being modeled has priority in determining what is 'wrong' when there is failure of model-system fit; and in the second kind of modeling tasks, the model itself has priority as an object of study, such that a system which fails to fit the model is rejected in search of systems that do fit the model. These are both legitimate modeling tasks - it is not that one should be endorsed over the other. Rather, it highlights how taking a different primary focus in terms of the object of study - the system being model or the model being used - leads to two different kinds of explanations of the system in question from the model in question. First I will illustrate this in a scientific case, and subsequently apply it to naive action explanation.

Consider the Lotka-Volterra toy model. The Lotka-Volterra (LV) equations give the population of a predator and prey population over time. The population size of either at a given time is a direct consequence of the birth rate and death rate at a previous time increment. For the prey population, the death rate is a function of the predator population at the relevant time. For the

predator population, the birth rate at a later time is a function of the earlier prey population. This model is a very useful example of a toy model: it is known as being a very simplified, idealized, and often numerically inaccurate model of actual predator and prey populations. Much of the failure to be numerically accurate stems from the fact that very few systems actually fit the model - it is hard to find genuinely isolated predator and prey populations that meet the conditions for these equations to fully apply. Nevertheless, they are extremely useful.

Sometimes, such well-developed toy model can be studied on their own, since many different scientists, with very different target systems, might use versions of it. The equations treated as a toy model can be used to derive the robust Volterra principle (Weisberg and Reisman 2008). This states that when a general biocide event (something that kills both predator and prey indiscriminately) occurs, then in the recovery period afterwards, the proportion of prey to predators goes steeply up. This turns out to be a mathematical result of the model: any simultaneous increase in the death rates can be shown to result in this change in proportion. It falls out as a purely mathematical consequence of the equations. It is useful and interesting to know of the LV equations that they have this feature, even if it turns out that no actual system ever follows those equations strictly.

This illustrates the distinction between two ways of applying a model: taking either the target system or the model itself as the primary focus of inquiry. In the first way of applying a model to a system, a particular system is being modeled, and if the assumptions do not fit that system, the model must be rejected. The system comes first, and the model must be tailored to fit that system. Many cases of modeling are like this. The wildlife biologists in charge of managing some specified conservation area will often have just this kind of focus. The ecosystem(s) are fixed, in that they are well specified as the target requiring a model for the purposes of, e.g., prediction of future population changes. If there is a general biocide of some kind and this change in proportion of prey to predators is not observed, one goes looking for another model other than LV. It doesn't disprove that the general result holds for LV; it demonstrates that the LV model does not fit the system.

In the second way of applying a model, the model itself is a focus for inquiry. The LV equations can themselves be studied, as clearly illustrated by the way in which Weisberg and Reisman derive the robust Volterra principle. In this kind of modeling task, one starts with the model and goes looking for a suitable system that it fits. It turns out that a case of chemicals dumped in the sea near Italy illustrates this general biocide result effectively; the ratio of prey fish to sharks shot up in the recovery period. If, however, the example from Italy ended up not fitting the model, then we could simply move on to look for some other system that better illustrates the effect. We would not, in this approach, reject the LV as not applying and continue modeling the chemical dump system. We would look for a better fit by taking the model with us and leaving that particular system behind.

What is extremely important in this contrast between model usages is that we already knew that the robust Volterra principle would obtain in any system of which the model held, before we ever even found a system of which the model held. It *had to* hold of any system of which the equations hold, because it is a straightforward mathematical consequence of the equations of that very model. This does not guarantee that we would ever find such a system of which the model holds. But it does ensure, with mathematical certainty, that *if* we find a system of which this model holds, *then* that system must also obey the robust Volterra principle. It holds with mathematical certainty, and nothing weaker, for the systems of which it does end up holding.

Causal explanations are generated when a model is applied in the first way. When we focus on the system in question first, the LV equations help us track the causal relationships governing changes in one population with respect to the influence from the other population. Causal explanations have some degree of strength of connection; they are not merely accidentally true generalizations, for instance. But since they are empirically dis/confirmable like this, they do not hold with mathematical necessity; mathematical necessity is stronger than causal connection.

Distinctively mathematical explanations are generated from the model applied in the second way. Metaphorically, it is like we are walking around with a bag, into which we only put a certain kind of of stone. We know that there will only be that kind of stone inside the bag, because we ensured that it would be so by using it as a selection criterion. We don't need to check each stone already in there to make sure the contents of the bag fit the criterion; we *enforced* the criterion in the first place. It might turn out that the bag is empty, because we have not come across any such stones yet. But we know with certainty that *if* there are ever any stones in the bag, they will be of that kind, because we will only put that kind in. In the second approach to modeling, we enforce the criterion that the system must fit the model that is the focus of inquiry, such that it must be the case that all systems that turn up success are already known to have certain features.

All of this is set-up to make the following point: in action explanation, especially in naive action explanation, naive action explanation is treated akin to the LV model applied in the second way. We can usefully explore the features of action as we use like a model that is a target of inquiry, and we can go looking for examples that fit the 'model' of action, or LV, by rejecting those that don't and looking until we find examples that do fit. If we discover that a particular example turns out to not be an action, for whatever reason, we have two options, mirroring the two approaches to modeling. We can distinguish psychological explanation as taking the first kind of approach, where we reject action explanation as providing sufficient traction on the example, but stick with the example and resort to merely psychological explanation instead of action explanation. Or, we take the second approach by sticking with action explanation and rejecting that candidate as not an action, and continue the search for some better example that is an action. Naive action explanation, by dint of holding between two actions, must pre-select for action; it cannot, by definition, end up holding of non-actions. This enforced pre-selection criterion ensures that anything that can be said of action explanation will hold, in systems of which it holds, with a strength like mathematical explanation, and not like mere causal explanation.

Action explanation enforces the selection criteria, like enforcing the criterion of only putting stones of a certain kind in the bag. As a consequence of this, it must be the case that whatever ends up in the action bag is already know to have certain features, which can be explored by taking action itself - in this case, naive action explanation - as the target of inquiry. The existence of behaviors that are not action are neither here nor there for that purpose; it merely means that we pass by those examples as we engage in naive action explanation. Thus, we can know things about any case of genuine action that we find in the world prior to ever finding it, by dint of the fact that we can draw inferences from the 'model' itself, studying action. This means that rationalization, as the relation that unifies actions performed as means as means to an action that is also an end, will yield explanations that are stronger than merely causal explanation.

IV. How Causal Markov and Faithfulness justify inferences in causal modeling

We turn now to see what makes the engine of interventionist or Causal Bayes Nets modeling work. These techniques are essentially a set of algorithms to make justified inferences between

probabilistic relationships in data and causal structure as represented in structural equations and directed acyclic diagrams (DAGs). The inferences work with the assistance of some background assumptions or conditions that provide the justificatory foundation for those inferences. In case these assumptions fail to hold, we would be unjustified in making inferences that require them. When those assumption fail, only limited versions of the algorithms can be used, resulting in weaker available inferences; the strongest inferences can be made when the full set of conditions hold. When modeling a system, we ought (in the epistemic sense of ought) use the strongest set of assumptions we are justified in believing obtains for the system in question.

One central assumption in causal modeling is the Causal Markov condition. This condition stipulates that causes are probabilistically independent of their non-effects, conditional on their parents (Spirtes, Glymour, and Scheines 2000; Pearl 2009; Woodward 2005; Hitchcock 2018). Put another way, after conditioning on the parents of a given variable, then the only remaining probabilistic dependencies are effects of that variable. This condition allows us make the inferences that are fundamental for causal modeling: using intervention to distinguish causal from correlational structure. Without conditionalizing on the parents of a target variable, then any other effect of those same parents will be probabilistically correlated with the target variable, even though they are not an effect of it. By conditionalizing on the parents of a cause, the dependencies with non-effects is 'broken' for common case structures. In a nutshell, then, the Causal Markov condition ensures that existing conditional dependencies are due to causal relationship(s) and not coincidence.

What would it look like if the Causal Markov Condition failed? How empirically substantive is this condition? This has been the focus of some back and forth (Hausman and Woodward 1999, 2004; Cartwright 1999, 2002). Part of what emerged from this disagreement is that Causal Markov is a genuine assumption about the world. It could fail, if we found that there were persistent probabilistic dependencies between variables that could not be accounted for by causal connections. The correlations would have to be both robust over time, and genuinely inexplicable with respect to causal connection. It would be pure 'spooky' correlation. Cartwright emphasized that this assumption is not trivial, a priori, or merely analytic in character. Hausman and Woodward emphasized that this is an assumption most of us are willing to commit to without much by way of metaphysical misgivings. An upshot is that the Causal Markov condition is empirical, in that we can genuinely consider what it would be like to find that it is violated somewhere, but also metaphysical, in that considering how it would be violated requires rejection of the principle of sufficient reason, for instance.

Another central assumption, the Causal Faithfulness condition, is also a key part of licensing inferences between causal structure and probabilistic relationships in data. It ensures that existing conditional probabilistic independencies reveal causally independent variables. Faithfulness is a feature that a directed acyclic diagram (DAG) may have to a set of probability relationships among the variables in that graph. A graph is faithful to the probability distribution when there are no causally connected variables in the graph that are independent in the distribution. Put another way, the causal faithfulness condition ensures that there are no 'hidden' causal dependencies that fail to show up in the probabilities in the data.

The Causal Faithfulness condition can be violated when there are precisely counterbalanced causal relationships that 'disappear' by being probabilistically independent in the data despite being causally connected in the true graph. Consider a cause C that has two pathways by which it is connected to effect E, one path that brings about C with a weight of .8, and another path where C causes D with weight 1 and D then suppresses C, with weight -.8. C and E are causally connected in the true graph, and if the weights of those pathways were anything other than precisely opposite,

they would be probabilistically dependent in the data. But because the two pathways have exactly opposing weights, so that C causes E with precisely the same strength that it suppresses E, it looks as if C has no influence on E.

If the parameter values for causal relationships in graphs were randomly distributed, then this violation would occur with measure 0 frequency (SGS 2000). But this is itself a substantive assumption about systems. Zhang and Spirtes (2008) argue that it may be violated in mechanisms like thermostats for maintaining a fixed room temperature. Andersen (2013) argues that it may be violated even more rampantly, since evolved systems use homeostatic mechanisms that are much more finely tuned than thermostats, and which are evolved precisely to maintain homeostasis. Modified versions of the condition, which would fail less often to hold of systems though also support somewhat weaker inferences, can be used (Zhang and Spirtes 2016; Forster et al 2017).

Taken together, the Causal Markov and Causal Faithfulness conditions ensure that the probabilistic dependencies and independences in data taken from a given system connect in reliable ways with the causal relationships in the true causal graph of that system. Without these, one cannot infer between data and *causal* relationships. Even though Causal Faithfulness and Causal Markov assumptions require substantive commitments about the systems in question, they return advantages in making genuine discoveries. Thus, making the strongest set of assumptions about Faithfulness and Markov that are warranted by the particular system being modeled allows us to use the strongest version of the inferential tools that are justified by those conditions.

V. Introducing the Rationalization Condition

The outcome of this section will be the introduction of a new condition, an addition to the two most commonly used ones in current practice. This assumption, the Rationalization condition, will ordinarily be violated: for the overwhelming majority of systems, it will fail to hold, and the default modeling apparatus is used. But there do exist systems in which the Rationalization condition holds. And in modeling such systems, use of this condition will strengthen the inferences we can make, in particular from the causal graph to predict probabilistic relationships in data. Insofar as we should use the strongest available set of inferences given the conditions that we are justified to by the conditions that obtain in the system(s) being modeled, this Rationalization condition is a useful way to add power to interventionist causal modeling.

My proposal, put very briefly, is that we add what I will call the Rationalization condition: variables representing distinct actions that are each rationalized as means of the same end have that shared rationalization treated as a causal arrow between them, where the causal order follows the temporal order of those actions. The rationalizing explanation, the naive action explanans, is not directly represented in the system with a variable. Only the two naive action explananda are represented. They are connected via the Rationalization relation with an arrow in the graph if and only if they are rationalized by the same (missing) action.

By treating rationalization relations that obtain between appropriately defined variables *as if* they are causal in the context of causal modeling, we can predict additional probabilistic relationships in the data. Since, as we saw in the previous section, rationalization is a stronger explanatory connection than causation, it can be weakened to mere causation without thereby overextending our justificatory base. Refraining from using the Rationality condition simply reverts to the same causal modeling techniques we currently use.

Begin with actions that can be naively explained by a further action. Distinct means to the same end may be used to defined variables such that those two means variables are treated as if causally connected. There is a two dimensional figure, that of the rationalization relation situating the first action as a means in an end action, and the second action as means to the same end action. This is projected onto a one dimensional arrow connecting the two means actions. The two means actions must be sufficiently distinct that one can occur without the other thereby occurring; the temporally earlier means action is treated as cause and the temporally later means action as effect. In this regard, the causal graph must simplify rationalization in a way that loses genuine structure. This both illustrates why causal relations could not be used to in general reduce rationalization relations, but also how such projection provides the requisite inferential basis to justify inferences about connections between those two action variables.

Key to using Rationalization is that two different stages of a single action can be explained with respect to that same action: if we ask why I am kneading dough, and ask why I am letting it rise, both are naively explained with respect to the same baking of the same bread. They are each distinct stages of that same single action, temporally differentiable means to the end of making bread.

In a system of variables that includes ones like Kneading Bread and Letting Dough Rise, these variables will be probabilistically connected: it is an empirical fact about the world that when we identify genuine instances of each of these two variables, using that action description, they will be consistently positively correlated. Yet it is also clear that they are not straightforwardly *causally* connected, in the way that dropping a glass and getting the floor wet are causally connected. Instead, these variables are connected via rationalization: an instance of each variable is given a naive action explanation rationalizing each with the same action. Kneading bread does not count as *causing* one to let it rise. Even the weaker sense of causal connection is lacking - nothing compels that connection, even weakly, except the aims of the agent performing them. But they go together with such consistency that it can be reliably used in prediction: when someone is Kneading Dough, it is quite likely that later they will let it rise. It is rationalization as uniting these as two means to a common end that provides the connection and the prediction, not causation.

Thus, we can add an arrow in the DAG between these two variables, just as if it were a regular causal relation, and makes inferences in terms of predicting probabilistic relationships in the data taken from such a system. We can thus predict, and explain, the systematic correlations we find between these two variables, by incorporating this additional arrow in the graph. Refraining from using the Rationalization condition in such cases won't lead to a model that makes inaccurate predictions. But it will lead to a model that generates weaker predictions and explanations than it could. We would be refraining from saying true things that we could say if we relied on the Rationalization condition.

What it takes to make this rationalization condition part of the formal apparatus is straightforward. The action variables should be defined so as to allow for fairly straightforward identification of instances of the variable. This is a generic feature of the craft aspect of modeling, and not particular to action. For these variables, a superscript R is added, indicating that a given variable is being used with respect to its rationalization relationships. It need not *only* be used for rationalization connections, but it *may* be so used. This superscript is then also added to the weight of the connection in the DAG. Thus, some variables, and some causal arrows in a DAG, will have an appended R superscript to remind us explicitly of the requirements for their use in the model.

Even when using the Rationalization condition, the independence of causal variables (e.g. Campbell 2010) must be ensured. If there is a single occurrence in the world, for any kind of system,

which ends up counting as an instance of two different variables in a system of variables, then those variables will appear causally connected even those they are not. To avoid double-counting, variables are defined to ensure that no single instance counts as an instance of both of them. This is implemented in systems involving Rationalization by leaving out any variable representing the shared end action, and only including the common means to that end as variables. Kneading Bread and Letting Dough Rise will be be independent variables in the appropriate way. Kneading Bread and Making Bread will not: some or most instances of Kneading Bread will also be instances of Making Bread. But for different actions that are stages or means in the same action for a naive action explanation, the conditions of variable independence will be met.[2]

The Rationalization condition will not be met in the vast majority of causal systems being modeled. When we think this condition is violated - when, for any number of reasons, we lack a genuine agent that could potentially offer genuine naive explanations of their actions - we simply don't treat the relations between stages of an action as causally linked. Humans will often fail to be the kinds of systems where we can assume that the Rationalization relation holds. Just as there are systems where Causal Faithfulness fails, this situation simply means that the additional analytical tools based on that assumption cannot be deployed. In modeling human behavior that fails to be adequately rational, our predictions made using the Rationalization condition will be less accurate than in systems where it holds. But this does not indicate that the Rationalization condition is fundamentally unusable. If one is modelling predator-prey relationships, then the Lotka-Volterra model might be apt. If one is not modelling such relationships, there is no reason to even consider the L-V model; yet surely the fact that there are modelling situations where it doesn't apply does not mean that the L-V model is *never* correct. Just as surely, that there are many cases where humans fail to be agents in the requisite way does not mean they are *never* agents in the requisite way.

By only relying on the Rationalization condition when we are justified in holding that genuine rational action is taking place within the system to be modeled, we are using a selection method for systems. By enforcing the appropriate selection criterion for systems to which we apply the rationalization condition, we know before we ever find such a system that certain features will obtain and can draw on this in making inferences.

VI. Using the Rationalization condition: turning left

There are many, many instances of actions that turn out to be quite prosaic but which clearly demonstrate that not only *could* we use naive action explanation in prediction much like using causation, but that we already *do* this, so effectively that it provides a cornerstone of modern living: driving.

Consider first the general structure of driving somewhere in the context of naive action explanation. Imagine we are driving down a particular street, and someone asks, Why are you driving down this street? Our responses, in cases like this, are not of the form that a causal explanation would require, even a very general or abstractly described one. It would be weird and tedious to answer such a question by saying that we had been driving on the street back there, and turned right at the corner onto this street, and that was why we were driving on this one. Knowing the path that we took to be on this very street just does not answer the question of *why* we are on it; instead it

[2] This proposal thus differs from other extensions of causal modeling such as Schaffer (2016): instances of grounding will violate the independence condition, and fail D separation, in way that is avoided by the Rationalization condition.

answers something more like *how* we got to it. More fitting answers involve situating our driving down this street in the larger encompassing trajectory of our drive. We are driving *to* that place over there, and this is the only connecting street between where we were and where we are going. We are driving to some further destination and thought this was a more scenic road than the other alternatives. Our map directed us here as part of the fastest journey from starting point to destination.

All of these are kinds of naive action explanations. We take our driving down the street to be akin to reaching for the flour on the shelf, and explain this part, driving on this very street, by encompassing it into a trajectory that presupposes our end in driving is to arrive at a set destination. We *are here* because we are going *to there*. Any particular part of the drive is explained as a stage in a longer drive defined by end points. We consider it to be the exceptional case when we really aren't going to anywhere, just driving around for no reason. It is only in such cases that there is no unifying action under which to subsume our current driving. Indeed, we still usually give such explanations a naive flavour, explaining this drive with respect to the absence of such an arrival-at-destination plan into which it is a stage, situating it into something like an entertainment-or-diversion end instead.

Consider next how turn signals, used properly, display a driver's intentions so that we consistently rely on them to predict how to safely navigate roads shared with other drivers. If I am at a stop sign, and a car on the other side is also stopped, and we both have our left turn signals on, I confidently pull into the intersection when there is space, on the knowledge that the other car will not be driving straight (and thus, into my own car) but instead turning left. I don't have to see inside the window to the driver, much less peer into their secret intentions, in order to predict what they will do next. At this particular stage of their drive, they intend to turn left, and we confidently, and with high success, predict that they will be turning left when the opportunity such as a break in oncoming traffic arises. Recall the nonfinality of actions that may serve as rationalizing ends, from section II. We need not know where they are *really* going to know that they are turning left here.

When facing, at a stop sign or red light, a car with a left turn signal blinking, nothing about having a turn signal on *causes* the driver to turn left. Yet having the turn signal on, and subsequently actually turning left, are both highly correlated. Interestingly though unsurprisingly, having no turn signal on is less highly correlated with going straight than having a left or right signal on is correlated with actually turning left or right. Driving behavior demonstrates very clearly that we already have a myriad of ways of thinking about genuinely intentional behavior, replete with actions and full-fledged intentions, about which we have no qualms reasoning.

If one wants to model a system for car movements in a given intersection, one could include variables like Turn Signal [Left, Right, None], Car Movement [At rest, Continues straight, Turns Left, Turns Right], Stoplight [green, yellow, red], etc. There will be R superscripts on the first two variables, but not on the Stoplight variable. There will be a causal connection between Car Movement and Stoplight, which need not be treated with a superscript R. The stop light is not an R variable: a requirement for using the Rationalization relation are that it hold only between two R labeled variables. Between Turn Signal and Car Movement, there will be a new arrow added to the graph, with an R superscript on it. Adding the superscript R allows for a few additional variables and arrows to be introduced to a graph that would not otherwise be possible, and which allow for more predictions about behavior in the system to be made.

Recall from the section on naive action explanation that such rationalization relations can only transpire between two actions. This limits the extent to which such additional arrows will be

added to the DAG. They can only connect R variables, of which there will be a limited number as well. If each and every variable in a system is an R variable, then in a real sense one is not doing causal modeling, and should switch to using straightforward naive action explanation instead. It is only when mixing clearly causal variables with action related variables that the Rationalization condition will come in handy.

In general, we must treat other drivers as if their driving related actions are causally connected, in order to rely on each other to follow traffic rules and thus stay safe on the road. We already know, when pressed, that these are not strictly causal relations in the way that a dropped glass of water and a wet floor are causal. But we have to interact with other drivers at a mass scale that makes it easier to treat these *as if* they were causal. A breakdown in road rules, where we cannot predict what other drivers will do, leads to a worse situation for everyone; that is what happens when R fails.

What happens in modeling such cases when the Rationalization condition fails? Two comparisons are illuminating. First, compare this to failures of the Causal Markov and Faithfulness conditions. If we have inadequate justification to treat the system as containing actions subject to naive action explanation, then the Rationalization condition either just isn't used, or holds trivially by not applying to any of the variables or arrows. If there is only one R variable, then there will be no R relations in the graph, and no need to invoke Rationalization.

Second, compare the failure to the two ways of applying a model, from section III. If we suspect that we have a case where there are genuinely no actions susceptible to naive action explanation (perhaps we are looking at badly programmed driverless cars), then we have two available moves. We could take the first approach to using a model and reject Rationalization, sticking with the system at hand and developing some other set of variables to better reflect its causal structure. Or, we could take the second approach, and reject the system: if we want to model genuine driving behavior, we would reject such a system and continue on to find a better system that illustrates the Rationalization condition.

VII. Conclusion

The internal connection between means and end exhibited in naive action explanation has a modal strength that is more like that of distinctively mathematical explanations than that of causal explanations. Yet, because it can be treated in DAGs, and meets criteria like D-separation, it can be used to strengthen inferences that can be drawn from causal models. This chapter aimed to motivate incorporation of the Rationalization condition into causal modeling practices, where it is apt for the system(s) being modeled, and to provide the basics for incorporating R variables into systems of variables and R arrows into DAGs.

The proposal developed here fits in a longer trajectory of discussion of mental action and causation that goes back to Davidson (1963) and Anscombe (1981). Since Kim's (1998) Causal Exclusion problem, the issue of causation and action, or or causation and the mental in any guise, has been construed in terms of causation relating higher and lower levels, rather than competing descriptions of the very same relata, as Davidson originally discussed. It has also led to a widespread sense, mostly among those working in the philosophy of science, that interventionist causal modeling has supplanted any genuinely causal role for something as ephemeral and internal as reasons or action. The ways in which many philosophers have attempted to eliminate or reduce

action explanation using interventionist or related causal modeling approaches involves a deep misunderstanding of the character of action.

Once we note that rationalization and causation behave differently, we could decide to reduce action and insist it be replaced with causal explanation. The fact that kneading dough does not cause one to let it rise could mean that there is nothing more to connect them than the tenuous possibility of some weak physical causal chain. On the hand, we could modus tollens instead of modus ponens, and conclude that the failure of causation to accommodate the connection between stages of actions like breadmaking means that causation itself is insufficient to handle such a connection, and look to *supplement* causal analysis with action analysis where it is apt. Reliance on the Rationalization condition where it is appropriate can be justified by its own usefulness. It also paves a better path forward in bringing together these distinctive forms of explanation to enhance rather than replace one another.

Acknowledgements: Much thanks for helpful discussion and comments from Shimin Zhao, Varsha Pai, Matthew Maxwell, Cem Erkli, Zili Dong, and Weixin Cai. I am grateful for the opportunity to live and work on the unceded territory of the Musqueam, Squamish, Tsleil-Waututh, and Kwikwetlem nations.

References

Andersen, H. (2016). Complements, not competitors: causal and mathematical explanations. *The British Journal for the Philosophy of Science*, *69*(2), 485-508.

Andersen, H. (2013). When to expect violations of causal faithfulness and why it matters. *Philosophy of Science*, *80*(5), 672-683.

Anscombe, G. E. M. (1981). Metaphysics and the philosophy of mind.

Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues*, *20*(1), 64-79.

Cartwright, N. (1999). Causal diversity and the Markov condition. *Synthese*, *121*(1), 3-27.

Cartwright, N. (2002). Against modularity, the causal Markov condition, and any link between the two: Comments on Hausman and Woodward. *The British journal for the philosophy of science*, *53*(3), 411-453.

Davidson, D. (1963). Actions, reasons, and causes. *The journal of philosophy*, *60*(23), 685-700.

Forster, M., Raskutti, G., Stern, R., & Weinberger, N. (2017). The frugal inference of causal relations. *The British Journal for the Philosophy of Science*, *69*(3), 821-848.

Hausman, D. M., & Woodward, J. (1999). Independence, invariance and the causal Markov condition. *The British journal for the philosophy of science*, *50*(4), 521-583.

Hausman, D. M., & Woodward, J. (2004). Modularity and the causal Markov condition: a restatement. *The British journal for the philosophy of science*, *55*(1), 147-161.

Hitchcock, C. (Fall 2018 edition). Probabilistic Causation. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2018/entries/causation-probabilistic/>.

Lange, M. (2012). What makes a scientific explanation distinctively mathematical?. *The British Journal for the Philosophy of Science*, *64*(3), 485-511.

Kim, J. (2000). *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT press.

Ryle, G. (2009/1949). *The concept of mind*. Routledge.

Pearl, J. (2009). *Causality*. Cambridge university press.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical studies*, *173*(1), 49-100.

Spirtes, P., Glymour, C. N., Scheines, R., Heckerman, D., Meek, C., Cooper, G., & Richardson, T. (2000). *Causation, prediction, and search*. MIT press.

Thompson, M. (2008). *Life and action*. Harvard University Press.

Weisberg, M., & Reisman, K. (2008). The robust Volterra principle. *Philosophy of science*, *75*(1), 106-131.

Von Wright, G. H. (2004/1971). *Explanation and understanding*. Cornell University Press.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

Zhang, J., & Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, *18*(2), 239-271.

Zhang, J., & Spirtes, P. (2016). The three faces of faithfulness. *Synthese*, *193*(4), 1011-1027.