**Modeling Action: Recasting the Causal Theory**

Megan Fritts and Frank Cabrera

**\*\*DRAFT—Please Cite Published Version\*\***

**Abstract:** Contemporary action theory is generally concerned with giving theories of action ontology. In this paper, we make the novel proposal that the standard view in action theory--the Causal Theory of Action—should be recast as a “model”, akin to the models constructed and investigated by scientists. Such models often consist in fictional, hypothetical, or idealized structures, which are used to represent a target system indirectly via some resemblance relation. We argue that recasting the Causal Theory as a model can not only accomplish the goals of causal theorists, but also give the theory greater flexibility in responding to common objections.

**1. Introduction**

 Action theory as a philosophical subdiscipline has, since Donald Davidson, been mostly concerned with identifying the correct theory of action ontology. And indeed, Davidson’s Causal Theory of Action (hereafter: CTA)[[1]](#footnote-1) has remained (as dubbed by Velleman 1992) “the standard theory” of the ontology of human action, albeit with myriad versions of the basic theory springing up.[[2]](#footnote-2) It has remained thus, however, largely due to a relative dearth of popular positive alternatives, and despite numerous confounding objections that the theory fails to capture an accurate picture of human action (Fritts 2021). In order to advance the debate in action theory, we make the novel proposal that the CTA ought to be re-imagined as a “model”, akin to the models constructed and investigated by scientists, such as the ideal gas theory in physics.[[3]](#footnote-3)

 The rest of the paper is structured as follows. In section 2, we draw on the literature in the philosophy of science concerning models. In particular we examine, by way of example, what models are and how they function in scientific practice. Our primary focus is *theoretical* (rather than *material*) models whose function is to represent phenomena, but unlike other forms of theorizing, theoretical models represent phenomena *indirectly*. Roughly, this form of indirect representation proceeds by establishing that some hypothetical, fictional, or idealized model system *resembles* some aspect of a real-world system. In section 3, we review the central features of the CTA, including its dual role in providing an ontology of and an explanation for human action. In section 4, we sketch out what it would look like for the causal theory of action to be regarded as a model rather than a direct representation of human action. In addition, we discuss two compelling arguments for the claim that the causal theory ought to be re-imagined in this way: 1) construing the CTA as a model better serves the project of naturalizing philosophy of action, and 2) recasting the CTA as a model provides some novel responses to several extant objections to the CTA that have been put forward by non-causal theorists. In section 5, we consider and respond to objections to our model-based account of the CTA, the most compelling of which is that our re-conceptualization would not allow the CTA to satisfy some of the key desiderata for any viable theory of action. Finally, in section 6 we offer some concluding remarks, including some suggestions for further research.

**2. Model-Based Scientific Inquiry**

 Increasingly, philosophers of science have been concerned with the nature and function of models in scientific investigation.[[4]](#footnote-4) As much recent discussion has emphasized, there are many prima facie distinct entities to which scientists give the name “model”, and moreover, according to an “emerging consensus” the function of models in the scientific process appears to be many and varied (Downes 2011, p. 757). Perhaps the most familiar and accessible kind of model is the *material* model, examples of which include the 3-dimensional double-helix model of DNA built by Watson and Crick in the 1950s, or scaled-down versions of physical systems, e.g. a replica of an airplane in a wind-tunnel, frequently used by engineers (Sterrett 2004). Other examples of material models include so-called, “model organisms”, living things such as the fruit fly *Drosophila* *melanogaster*, the bacterium *Escherichia* *coli*, or the common house *Mus* *musculus*. Such organisms play an essential role in much contemporary biological and biomedical research, being used, for example to study biomolecular processes such as the development of cancer, genetic inheritance, or protein synthesis (Leonelli and Akeny 2013). In contrast to material models, either animate or inanimate, many models are *theoretical* or *abstract*. For example, evolutionary biologists often represent the genealogical relationships between groups of organisms, such as species, on branching tree-like diagrams called “phylogenetic trees”. Such diagrams, and the very idea of the “tree of life” have been considered by biologists to be general models of evolutionary history; however, alternative models to that of a tree, such as a web or network also receive discussion (Cabrera 2017: 79-80). Moreover, many theoretical models that are used in the sciences are mathematical in nature, e.g. the Lotka-Volterra model of predator-prey interaction, which attempts to represent the dynamic growth rate of a single predator and a single prey population by means of a pair of non-linear differential equations. The use of mathematized theoretical models in physics is well-known, including the Ideal Gas model, which equates various phenomenological properties of gases by means of the simple equation pV=nRT, where p is Pressure, V is Volume, n is the amount of the substance, R is a constant, and T is the temperature.

 From this brief survey, we can see that there are many different entities that are called “models”; and just as there are many different things that are called models, so too models also perform various functions. The function of some models is pedagogic, aiding the student or the non-specialist in understanding some phenomenon. Some of the material models canvassed above, e.g. the 3-dimensional model of the double-helix serve this role. In other cases, a model might be employed as a kind of proof of possibility (Grüne-Yanoff 2013). For example, one oft-discussed model, the “checkerboard model” of racial segregation, constructed by Thomas Schelling (1978) is intended to show that micro-level individual preferences, such as the desire to live in a neighborhood where at least 30% of one’s neighbors are of the same race, can in principle lead to macro-level patterns of immense racial segregation. The Schelling model is highly abstract and contains several simplifications and omissions and is thus not intended to be an actual representation of any real-world neighborhood. Rather, the point of the model is to show that overtly racist preferences are not a necessary condition for the formation of patterns of racial segregation; the mere preference of not being in the minority can, in principle, aggregate in unexpected ways leading to undesirable social patterns.

*2.1 Modeling as Indirect Representation*

 Despite these and other uses of models, perhaps the paradigmatic function of models is to *represent* the world. Scientists often construct and investigate the consequences of models largely because they are interested in depicting some aspect of a real-world system—the “target system”—to some degree of accuracy. While the idea that scientists attempt to accurately represent the world is a commonplace one, the philosophically interesting feature of model-based science is that representation of some target phenomenon is achieved *indirectly*. Following Giere (1999), Godfrey-Smith (2006), and Weisberg (2007), it is helpful to analyze the modeling process by way of three steps.

 First, the modeler begins by defining or specifying some model system with the use of some representational device, e.g. mathematical equations, a pictorial diagram, etc. The model system often contains simplifications, idealizations or outright falsehoods, and on some accounts constitutes an entirely fictional or hypothetical world, akin to a literary novel (e.g. Godfrey-Smith 2009; Frigg 2010; Toon 2012; Cartwright 1983, etc.). For instance, the ideal gas model makes several assumptions, which are known to be false if taken to be literally descriptive of actual gases, e.g. the motion of molecules is random, collisions are perfectly elastic, the volume of molecules is zero, etc.

 Second, after specifying a model system, the properties of the model are investigated, frequently with the aid of mathematical or computational techniques. The results of such investigations can often include surprising or unanticipated consequences. For example, although Schelling’s checkerboard model was not intended to represent any real-world system, the model itself is, of course, subject to theoretical inquiry. We can run simple computer simulations on Schelling’s checkboard model, adjusting some of parameters contained in the model, e.g. the population size ratios, to see how this variation affects the rapidity of segregation.[[5]](#footnote-5) Importantly, these are investigations of the properties and consequences of the model system *itself*. The ideal gas law should be regarded, in the first instance, as a law that describes the nature of *ideal* gases, i.e. those that satisfy the various assumptions of the model, as opposed to directly describing *real* gases.

 Finally, after the properties of the model are sufficiently mapped out, the model system needs to be brought into contact with the real world. To the extent that the model system resembles in some way or another—e.g. by being *similar* to—the more complex target system, then the model can be thought of as representing, even if only partially so, the target system.[[6]](#footnote-6) If the model represents the target system sufficiently well, then at the very least the model can be used to predict aspects of the target system.[[7]](#footnote-7) For example, even though real gases do not satisfy the modeling assumptions of the ideal gas model, real gases behave like ideal gases at low pressures and high temperatures, which is why the ideal gas law can be used as a partial representation of a real world gases. It is crucial to keep in mind, however, that whatever epistemic goods are achieved by model-based inquiry, the modeler always takes, as Godfrey-Smith aptly puts it, “a deliberate detour through fiction” (2009, p. 103).

*2.2 Modeling vs. Abstract Direct Representation*

 Although some philosophers of science concerned with the nature and role of models have attempted to argue that *all* scientific theorizing relies upon models (e.g. Giere 1999; Suppes 1960), this universal generalization does not square well with actual scientific practice. As an example of scientific theorizing that does not rely upon models, consider the case, discussed by Weisberg (2007), of Dmitri Mendeleev’s pioneering work in theoretical chemistry, which ultimately culminated in the modern Periodic Table of Elements. By organizing the elements by their atomic weight, Mendeleev realized that certain sets of salient chemical attributes periodically recur. In what has now become a familiar representational tool, Mendeleev arranged these elements with common properties in various groups:



Far from being merely an organizational device or classificatory scheme, Mendeleev’s theory of elemental ordering successfully predicted the existence of then-unknown elements, including gallium (Ga), scandium (Sc), and germanium (Ge). These elements filled the gaps that were present in Mendeleev’s original table. Although subsequent discoveries would later modify the Periodic Law that undergirded Mendeleev’s original system—e.g. organizing elements based on atomic number, rather than atomic weight—his work stands as important and impressive theoretical achievement.

 What is most significant for our purposes is that although Mendeleev’s work was abstract, theoretical, and intimately connected to a diagrammatic device, i.e. the above table, his is not an instance of model-based scientific reasoning. Mendeleev was attempting to *directly* describe the patterns of properties that characterize the various elements in a systematic and unified way. The patterns or laws that Mendeleev uncovered, which are encapsulated in his periodic table, are intended to represent the world, but the table itself is not a model in the sense that we employ the term. The sort of inquiry that does not rely on taking a detour through a model is what Weisberg calls “abstract direct representation” (2007, p. 210). Such inquiry typically takes the form of attempting to directly describe true, lawful generalizations about phenomena, sometimes involving the postulation of unobservable entities. In conjunction with our litany of examples of models in science, what the case of Mendeleev illustrates is that there are at least two ways in which to engage in theorizing while doing science. One style of theorizing relies heavily on models and represents the phenomena indirectly, whereas another style represents the phenomena directly without the mediation of a model. Whether one sees it fit to employ model-based reasoning will often depend on highly specific contextual factors and perhaps the goals of inquiry.[[8]](#footnote-8) As we will argue below, recasting the causal theory of action (CTA) as a kind of *theoretical* model better serves the goals of inquiry of action theory.

**3. The Causal Theory of Action**

In order to recast the CTA as a model, it is important to first understand the details and motivations of the causal theory. The CTA is really two theories intertwined: a theory of action ontology (CTA-O) and a theory of action explanation (CTA-E). These two parts of the CTA correspond to two different desiderata that causal theorists believe a theory of action ought to satisfy. The two desiderata are as follows: 1) a theory of action should be able to distinguish action events from non-action events, and 2) a theory of action should be able to generate explanations of why someone acted the way that they did, and explain these actions in terms of *reasons to act*.

*3.1 Action Ontology*

 The CTA-O states that actions are those bodily movements that are “caused by the right mental states and events in the right way” (Schlosser 2019), and typically holds beliefs and desires as the two causal factors needed for action to take place. The basic idea of the CTA-O is this: humans have many desires, and our actions are determined by 1) what we desire, and 2) our beliefs about how we can satisfy those desires. Once we have a desire, and a belief about how to achieve desire-satisfaction, the combination of our desires and our beliefs about how to satisfy our desires determine what we will do. While some nuanced versions of the CTA employ other folk-psychological concepts such as judgments, intentions, temptations, or goals (Holton 2009; Mele 2009), the causes of actions are always a *representing* state combined with a *motivating* attitude. Importantly, when the requisite attitudes cannot be found in the causal antecedents of some event, this indicates that the event was not an action. The CTA-O thus provides a way of fulfilling the first theory desideratum: it gives us a principled way of metaphysically distinguishing action events from non-action events (Foley 1977, p. 58; Frankfurt 1978, p. 157). In particular, to fulfill this desideratum, the CTA attempts to give a set of necessary and sufficient conditions for something’s being an intentional action.

*3.2 Action Explanation*

 Once we have the appropriate set of necessary and sufficient conditions, the story goes, we will not only be able to identify actions, but we will understand the nature of the explanatory relationship that holds between reasons and actions. [[9]](#footnote-9) Accordingly, the CTA-O is usually accompanied by a causal theory of action explanation (CTA-E)—actions are explained by citing their causes (beliefs and desires), and these causes also rationalize the action (Roth 1999, p. 841). The CTA-E holds that these causes are the appropriate constituents of reasons-explanations of actions, where reasons-explanations of an action render the action intelligible by revealing the intention of the agent. The second thesis held by causal theorists, then, is that the mental states cited in the reasons-explanations of actions are the efficient causes of the action in question, and that actions are explained by citing the agent’s reasons that are also a particular kind of cause. The CTA-E thus purports to satisfy the second desideratum for theories of action: a good theory of action should tell us why someone acted the way that they did, in terms of their reasons to act. We demand these sorts of explanations about token actions—both simple and complex—all the time: “Why did Sarah move out of state?; “Why did Frank pour two glasses of wine?”; “Why did Gwen turn on the light?”, etc. According to the CTA-E, all of these why-questions will be correctly answered by citing the relevant belief-desire pairs.

**4. Recasting the Causal Theory of Action as a Model**

 What we take to be clear from our survey of model-based inquiry and the standard formulation of the CTA is that the latter is not an instance of the former. Proponents of the CTA do not take themselves to be describing and working out the implications of some fictional, idealized, or hypothetical system, which is then used to represent the real world of human action. Rather, proponents of the CTA take themselves to be directly describing what actions are, and how actions are to be explained. With the distinction between different ways of theorizing discussed in section 2 in mind, we can see that the CTA as it is typically presented by action theorists should be classified, like Mendeleev’s theoretical work, as an instance of abstract direct representation. Just as Mendeleev attempted to uncover true, universal generalizations connecting atomic weight and the periodic recurrence of salient chemical properties, so too the proponent of the CTA attempts to formulate true, universal generalizations connecting human action with mental states such as beliefs and desires. Thus, the standard formulations of the CTA lack the key feature of model-based inquiry upon which we have focused, namely, to represent a target system indirectly. Although the standard formulation of the CTA is an instance of abstract, direct representation, we are free to consider whether the CTA ought to be re-imagined as a model, and what that would imply for central debates in the philosophy of action.[[10]](#footnote-10) As it happens, it is a common feature of scientific practice for previous attempts at a direct representation to later be reworked as a model, e.g. Newtonian physics (Godfrey-Smith 2006, p. 734).

*4.1 A Simple and a Complex Causal Model of Action*

 To regard the CTA as a model, we need to first insist that it is not a direct description of human agents, but rather, in the first instance, a description of some *ideal* agent—just as the ideal gas theory is description of ideal gases, not real gases.[[11]](#footnote-11) We can imagine a “causal model of action” (CMA) including several assumptions along different dimensions. An example of a simple CMA is as follows:

**Beliefs:** Agents possess categorical beliefs; that is, for any proposition P, agents either believe P or believe ~P.

**Desires:** Agents have no conflicting desires.

**Number of agents:** Agents are singular, not a collective, and not composites of other agents.

**Rationality:** Agents’ belief-desire pairs always cause means-ends rational actions. [[12]](#footnote-12)

Typically, simpler models are more theoretically tractable (Batterman 2001), and this is true of our simple CMA above. For instance, one logical consequence of this model is that there are no “weak-willed” agents. Since agents do not have conflicting desires, it is not possible for an agent in this CMA to act on a less strongly held desire. Additionally, this simple CMA does not allow for the possibility of group agency, and so the complex relationship between individual action and group action is avoided. An example of a more complex CMA is as follows:

**Beliefs:** Agents have beliefs that come in degrees; that is, for any proposition P, agents have a credence toward P ranging from 0 to 1.

**Desires:** Agents may sometimes have conflicting desires.

**Number of agents:** Agents may be singular or composites of multiple agents.

**Rationality:** Agents’ belief-desire pairs may sometimes cause irrational actions.

This is a more complex CMA, and some of the logical consequences of the simpler CMA do not follow. For instance, this CMA leaves open the possibility of weakness of will and group agents. This demonstrates a significant difference between the simple and the complex model, which is that the more complex model introduces greater uncertainty. In the simple model that disallowed for weakness of will, the answer to questions about whether weak-willed behavior “counts” as action *par excellence* is answered easily: no. The more complex model shows us a more realistic model system, but one with more gray-areas. In applying the more complex model to the real-world agents, we are deprived of an immediate answer to whether weak-willed behavior is an instance of action.

 Both the simple and the complex CMA amount to distinct and incompatible model systems, wherein all agents conform to the specified assumptions. It is not the case, for instance, that the simple CMA simply omits, leaves out, or abstracts away from group agency; rather, group agents are ruled out by the modeling assumptions. While the simple CMA and the complex CMA differ in substantive ways, we may regard both as members of the same “family of models”, both having maintained the core claims of the traditional causal theory of action. We have offered only two toy examples of a CMA. But in principle more CMAs retaining the core features of the causal theory can be constructed. We can regard decision theory, for instance, as of causal model of human action. Decision theory represents practical reasoning mathematically by attributing to agents a utility function, which captures the agent’s preferences, and a probability function, which captures the agent’s subjective degrees of belief. Importantly, much of Davidson’s earliest work was in decision theory (Harnay and Rème 2017). Although Davidson was at times critical of decision theory as a description of human action, he nonetheless concedes that, insofar as it combines some cognitive element (subjective probabilities) with some conative element (subjective utilities), decision theory “answers to our intuitions about how actual decisions are made” (Davidson 1999: 10).

We may find that different CMAs are useful for dealing with different problems. While a model of action should, of course, be used to shed light on the phenomena of action and agency, it remains an open question, once we consider the resemblance relations of the model system to the target system, whether the model represents all aspects of human action with maximum accuracy.

There are several reasons to prefer our modeling account of the causal theory to the standard CTA—a theory of the ontology of actions and reasons explanations. These reasons are: 1) the modeling approach helps us to further naturalize action theory; 2) conceiving of the CTA as a model allows us to give better responses to some of the most prominent objections to the causal theory.

*4.2 Naturalizing Action Theory*

 It is a widely accepted desideratum of action theory that it be, to the furthest extent possible, naturalized. One method of naturalization that many philosophers of action find appealing is, to the furthest extent possible, the incorporation of the methods of science and empirical data in the development of a theory of action.[[13]](#footnote-13) Myles Brand (1984) once called for action theory to enter its “third stage”—a stage marked by continuity with the empirical sciences—which he believed would fully naturalize the discipline. Bence Nanay (2014) follows suit in calling for the completion of the work begun by Davidson toward naturalizing action theory. To this end, Nanay proposes a deeper dive into the empirical science of neurophysiology, arguing that incorporating more realistic ideas of mental representation can allow us to fully naturalize our investigation into action.

 Adopting the modeling approach in action theory would allow action theory to become more continuous with the empirical sciences. This is because the CTA, by positing phenomena, such as beliefs and desires, as both the causes and the rationalizing explanation of actions, presupposes “folk psychology”; but as Godfrey-Smith (2005) convincingly argues, we should regard folk psychology, as a *model* in the sense that we employ the term, rather than as an abstract, direct representation. In particular, the familiar array of concepts such as beliefs, desires, hopes, plans, the will, etc. and their various connections, which are often taken for granted by working psychologists, are better understood, not as an attempt to describe the machinery of human brains, but rather as a model system in its own right, which might then be applied (or “construed”) in different ways for the purposes of performing different tasks. According to Godfrey-Smith (2005, pp. 7-8), this model-based conception of folk psychology has the advantage of greater flexibility and makes better sense of the wide range of philosophical perspectives surrounding folk psychology. Those who are inclined to take an instrumentalist stance toward folk psychology, e.g. Dennett (1978), construe the folk-psychological model (merely) as a predictively useful input-output device, whereas those who are inclined to take a realist stance toward folk psychology, e.g. Fodor (1987) construe the model as a map of the inner causal workings of the mind. Crucially, both applications of the folk-psychological model “are *available* but none is *mandatory*” (Godfrey-Smith 2005, p. 10). Since the folk psychology oft-discussed by cognitive scientists and psychologists is best construed as a model, and since the CTA clearly makes essential use of folk psychological concepts, the goal of naturalizing action theory would be better served by regarding the CTA, in turn, as a model, rather than an abstract, direct representation.[[14]](#footnote-14)

*4.3 The Strategy of Model Variation*

 Another reason in favor of our modeling approach in action theory is that models are more flexible in face of recalcitrant data. In discussing folk psychology as model, Godfrey-Smith remarks: “[T]here is a single core […] model, plus a range of variants and elaborations. The core can be used alone, or it can be used as the foundation for more complex structures” (2005, p. 10). A consequence of the flexibility of working with a core model that admits of variations is the ability to provide a wider range of responses to potential objections. There are many objections to the standard CTA in the action theory literature, and in this paper we will only survey a few of the most prominent ones.

 The first is the objection from deviant causal chains, considered the most “serious and persistent problem” for the CTA (Schlosser 2019). The most well-known example of deviant causal chains comes from Donald’s “Freedom to Act”:

A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope he could rid himself of the weight and danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do it intentionally (Davidson 1973, p. 79).

In this scenario, though the climber’s mental states did cause the rope loosening, and though those mental states also rationalized the rope loosening, the causal connection between the mental states and the loosened grip is “deviant”—something about the connection fails to individuate an intentional action. The problem of deviant causal chains is normally understood as an objection to a certain picture of the ontology of intentional actions—an objection to the claim that mental states that rationalize and cause an action will explain the resulting act *qua* action. However, as Wilson et al. (2006) note, “There have been many attempts by proponents of a causal analysis of intention in action […] to spell out what ‘the right kind(s)’ of causation might be, but with little agreement about their success.”

 This objection becomes easier for causal theorists to respond to if the CTA is conceived of as a model. This is because, if the CTA is a model, then variations of the model, with various ways of filling in “right” causation, can be proposed and used in the investigation of action. Such an objection need not become an objection to the entire project—as it would if the CTA were a theory of action ontology—but rather an invitation to consider a different, perhaps more complex model of action.

 Another common objection to the CTA is referred to as the “disappearing agent objection” (Lowe 2008). The idea behind this objection is that the CTA reduces actions to things that *happen to us* (beliefs and desires) instead of thinking about actions as *things that we do*, rendering human agency incidental or thin (Schlosser 2019). Causal theorists have responded to some versions of this objection by arguing that the objection mistakenly blurs the distinction between the “causal order” and the “natural order.” Simply put, just because the agent disappears on an examination of the efficient causal order of the action-event, we should not thereby conclude that the agent is no meaningful part of the natural order (Mele 2003). Other versions of this objection, however, are raised by philosophers arguing for alternate conceptions of agent-causation, in favor of a conception that is naturalistic and compatible with (though not identical to) event-causation (e.g. Velleman 1992)

 If conceived of as a theory of action ontology, the CTA must settle on an account of causation—in particular, it must decide either that the causal order does not include “agents” as things that cause events in themselves, or that the causal order does include agents as *loci* of causation. If the former, the disappearing agent objection feels weighty because it is difficult to imagine actions without agents. If the latter, the whole premise of the CTA appears to be in jeopardy, as mental states are now no longer the only contenders for causal explanations of actions.

 If the CTA is conceived of as a model of action, however, variations on this model may be used to test and investigate different hypotheses on the location of the agent in the causal chain of events that leads to action. Recasting the CTA as a model can allow for multiple variations of the core ideas favored by causal theorists, with some variations modeling an agent (as either an independent causal node or as a collection of event-causal nodes) and other variations excluding the agent.

 Finally, a recent objection to the CTA turns on its purported exclusion of certain instances of group action (Fritts 2020). If some group actions cannot be explained by citing group beliefs and desires that cause the group action, then the CTA cannot account for those actions and thereby fails to metaphysically delineate actions from non-actions. This argument, if it works, puts the CTA (construed as an abstract, direct representation) in an awkward position—the causal theorist must either deny the reality of group actions, or admit that the CTA’s ontological component is mistaken.

 However, if the CTA is construed as a *model* of action, causal theorists again have greater flexibility in how they can respond to these challenges. Counterexamples are a much less significant threat to models than they are to directly-representing theories, since models are not expected to perfectly capture the phenomena. As Williamson notes “what defeats a model is not a counterexample but a better model” (2017, p. 168), and causal theorists may propose new and improved CMAs whenever the need arises. For example, we may conceive of the group action argument as an argument against, say, the simple CMA. However, such an objection does not apply to the complex CMA. Likewise, one may opt to fill in the causal antecedents of actions in different CMAs differently, to account for new data, replacing old conceptions of “mental states” with functional accounts of agential attitudes. Having different models of action for different sorts of agents would not be out of the question; scientists often exploit multiple, incompatible models. Such incompatible models might contain, for example, incompatible idealizing assumptions. One reason for relying on multiple models is that there appear to be several desirable properties, such as simplicity, generality, realistic assumptions, precision, etc., that no single model can possess to the utmost degree (Levins 1966).

 We can appeal to an oft-invoked analogy to map-making to help clarify the function of multiple, conflicting models (Giere 1999; Kitcher 2001; Longino 2002). As Kitcher (2001) helpfully puts it, scientists are like map-makers and “[m]ap-makers are invariably selective” (p. 56). Consider the case of a subway map. Such a map is a representation of the underground train system, depicting the number and order of stops between stations, the connections between different train lines, etc. However, such a representation is, of course, partial and, in most cases, distorted. The distances between the stops are typically not to scale and the respective sizes of the stations are not represented accurately, all of which are shown as dots or circles of equal size. Despite these omissions and distortions, the correctly constructed subway map is still a useful tool for a traveler navigating the transit system. If one would like a representation of some of the features that the subway map leaves out or distorts, then one will have to opt for a different map, e.g. a blueprint of the interior of whatever subway station is of interest. Importantly, in order for this second map to be useful for its task, it might contain its own distortions or omissions. But this is no problem, as both maps can accurately represent some aspect of the subway system, even if strictly-speaking, the maps are incompatible. Armed with multiple maps, each of which may be concerned with different aspects of the world, we will be better able navigate reality. As the analogy goes, much the same is true with models. Just as no one map can represent everything that we are interested in mapping, so too no one model can represent all aspects of a target system. Different models can partially represent different aspects of a target system, and moreover, can do so even if the models are, strictly-speaking, incompatible. This should be of value for action theorists interested in theorizing about both individual and group agents: two very different kinds of agents may instantiate radically different structures of agency.

**5. Objections**

*5.1 Does the modeling approach give up the ontology game?*

The first objection to our thesis that we anticipate is that in conceiving of the CTA as a model, are we not throwing in the towel in our investigation into the nature of action? To recall from section 3, contemporary action theory is largely concerned with formulating a theory that fulfills two desiderata: it should provide the basis for a metaphysical distinction between action and non-action events, and it should generate explanations for why an agent performed some action. If the CTA is a model of action, then it is not a directly-representing theory of the ontology of action. But then, one might worry, the modeling approach that we defend does not include a set of necessary and sufficient conditions for an event’s being an action, thereby running afoul of the first desideratum.

There are several things to say in response to this. First, it is not universally accepted in action theory that there exists a set of necessary and sufficient conditions for something’s being an action. Wittgenstein’s widely-cited discussion of action is one in which he explicitly rejects the possibility of such a set of conditions. In *Philosophical Investigations* sections 621 he writes:

“Let us not forget this: when ‘I raise my arm’, my arm goes up. And the problem arises: what is left over if I subtract the fact that my arm goes up from the fact that I raise my arm?

 ((Are the kinesthetic sensations my willing?))”

This fragment is often used as a tool or “computation device” (Velleman 2000) to pinpoint the feature that distinguishes the action from the surrounding non-action events—a way of identifying the set of necessary and sufficient conditions. Contrary to this reading, P.M.S. Hacker writes that “Here W. brushes this suggestion aside. When one raises one’s arm, one does not usually try to do so” (Hacker 2000, p. 605). Wittgenstein’s reason for asking this question was to show that this way of delineating actions from non-actions is doomed to failure, because the proposed subtraction is impossible—a conceptual confusion. On Wittgenstein’s conception of action (and related ones, e.g. Anscombe 1958) there is no set of necessary and sufficient conditions for something’s being an action.[[15]](#footnote-15)

As history has shown us, giving up the pursuit of a reductive conceptual analysis does not entail abandoning philosophical inquiry into that concept altogether. To see this, consider the historical development of epistemology concerning the conceptual analysis of knowledge. In a seminal paper, Gettier (1963) challenged the prevailing analysis of propositional knowledge as justified true belief, and thereby spurred decades of epistemological inquiry into the nature of knowledge. Repetitions of a familiar cycle ensued. Many plausible analyses were put forward, and not long after some clever counterexample was brought to light. In response, a revision of the initial analysis was proposed to deal with that counterexample. But then more elaborate counterexamples were put forth in turn. Owing to decades of failure to solve the Gettier problem, Williamson (2000) argued that the project that Gettier’s paper launched in 1963 is misconceived. Knowledge does not admit of a reductive analysis in terms of informative individually necessary and jointly sufficient conditions. Knowledge is a primitive relation. The quest for a conceptual analysis of action may be thought of as occupying a similar position as the quest to solve the Gettier problem did 20 years when Williamson proposed the idea of “knowledge-first” epistemology. It is not beyond the pale to wonder if perhaps a similar sort of reorientation in action theory is in order. Considering multiple models of action is one tool that can help us determine if this elusive set of conditions even exists.

Additionally, it is important to note that conceiving of the CTA as a model frustrates, at most, a grand and all-encompassing theory of the ontology of actions. What is not hindered by the modeling approach is investigation into the ontology of *aspects* of action: for example, the nature of causation (event and agent), the nature of mental states and attitudes, free will and determinism, etc. Instead of thinking of action theory as a single theory of the ontology of action, it may be more profitable to see it as engaged in a project of theory-unification, similar to projects in the empirical sciences. Cabrera (2021) writes, “it is important to first note that unification has always been an important goal of natural science, ever since Thales of Miletus (624-546 B.C.E.) theorized that everything is water.” (p. 3673). Dawid (2013) concurs, “[t]he evolution of fundamental physics can be construed as a series of unifications”. Projects of unification range in scope, from domain-specific unifications to what is sometimes romantically referred to as a “theory of everything” (the potential unification of models of the four fundamental forces in physics). Conceiving of action theorizing as action modeling could be seen as a sort of project of unification, where different models of action, with different assumptions and parameters, can be tinkered with like puzzle pieces and potentially fitted together.

*5.2 Can models generate action explanations?*

 Another potential objection stems from the second desideratum for theories of action: can the CTA as a model generate reasons-explanations of actions? As we discussed in section 3.2, the CTA is the standard theory largely because it is widely considered uniquely capable of providing truth-apt explanations of actions in virtue of citing the reasons’ causal connections to the resulting actions. Donald Davidson’s(1980) argument for this can be formalized as follows:

 P1: An agent may have rationalizing reasons for her action even if she does not act for these reasons.

P2: Given (1), an action-explanation requires more than reasons-rationalization to connect the agent’s reasons to her actions.

P3: A causal connection is the best candidate for the “more” required of an action explanation, to connect the agent’s reasons to her actions.

 C: So, causally efficacious reasons are the best candidates for action explanations.

Davidson’s influential challenge to non-causal views cemented the CTA as the standard view of action ontology and explanation. Because reasons must explain actions, our reasons for acting must have some explanatory connection to what we do, and allegedly a causal connection is the only sort of connection that could distinguish our candidate reasons for acting from our actual reasons for acting.

 A concern, then, is that the shift away from a direct-representing theory of action ontology impedes the ability to give truth-apt explanations of actions in virtue of failing to cite an actual causal connection between an agent’s reasons and her actions. As we have presupposed throughout, the primary function of theoretical models is to represent some aspect of a target system. Broadly, this involves providing information about the phenomena being modelled. It is often the case that scientific models are used to provide predictive information about the future development of the target system. Can models be used to provide explanatory information as well? On its surface, this seems impossible. A compelling argument for this claim derives from the widely-endorsed thesis that the explanatory relation is *factive* (Bird 2006, p. 44). A paradigmatic example of a factive relation is the knowledge relation. For any proposition *p,* a subject S knows that *p* only if *p* is true. An instance of a non-factive relation is that of *hope*. It is not the case that S hopes that *p* only if *p* is true. According to the standard view, explanation is like knowledge and unlike hope: A explains B only if A is true. But model systems, as we have indicated, are descriptions of hypothetical, idealized, or fictional structures, in which case either it makes no sense to speak of the model itself as being either true or false, or all models come out false. If models are not factive, then the causal relations cited in CTA reasons-explanations would not be factive. This general feature of theoretical models then seems to undercut their ability to explain the phenomena.

There are several responses we can make to this objection. First, although models are in the first instance descriptions of idealized or fictional structures, there is no reason that statements about reality constructed on the basis of models need to be false, provided such statements are carefully circumscribed. Despite the idealized nature of models, one can use a model to formulate statements about the world, what Giere calls “theoretical hypotheses” (1999, p. 51). A theoretical hypothesis consists in some claim that the target system is similar to the model system in certain specified respects, and such a statement can very well be true. To claim, without qualification that real gases obey the ideal gas law would amount to saying something false. But, crucially, to say that real gases behave in a way that is similar to ideal gases within the model’s standard domain of application, would amount to saying something true. In general, claims of the form: “Model M is similar to Target T in Respects R1, R2, R3,.., Rn” can be true. Those true statements that are derived from the application of the model to the target can then be used to offer explanations of phenomena that respect the widely-endorsed constraint that explanations are factive.

Similarly, while on our modeling account it may be strictly false to say something like “belief-desire pairs cause actions”, it may be true that the series of causes that produce actions looks quite similar, or relevantly similar, to some specified CMA. In fact, many note a distinction between “Davidson’s dogma”, which “construes beliefs, themselves, as producing causes of behavior” (Curry 2018, p. 4) and Davidson’s actual thesis. While Davidson did not endorse the popular dogma strictly-speaking, he held to something like an indirect thesis: “belief explanations must [...] highlight an inner cause that produced the action under scrutiny.” (Curry, p. 5) This qualifeid thesis stems from Davidson’s “anomalous monism”, which denies the existence of causal laws that hold between mental events (i.e. beliefs, desires, intentions) and physical events (i.e. actions, brain states) in an effort to retain what is unique about the mental while preserving the important aspects of physicalism (Malpas 2019). Any actual cause of a physical event like an action, on this view, must be another physical event. The actual cause is, therefore, only highlighted (and not directly denoted) by the belief-desire pair.

Second, much recent work on the role of models in the scientific process has emphasized that idealized models *can* provide explanations (e.g. Bokulich 2011). Although there are many different accounts of scientific explanation, one crucial feature of successful explanations is that they allow us to grasp the myriad ways in which the *explanandum* would have been different, had the *explans* (e.g. some set of causal factors) also been different (Strevens 2008; Woodward 2003). As Woodward puts it, successful explanations provide understanding by correctly answering such “what-if-things-had-been-different questions” (2003, p.11). Crucially, in exploring the consequences of a model system, e.g. by varying the adjustable parameters of a mathematical model, the modeler can acquire such counterfactual information about the model itself. Now, provided that the modal structure of the model system sufficiently resembles the modal structure of the target system, we can use our knowledge of the counterfactual dependencies of the model system to shed light on the counterfactual dependencies exhibited by the target system. Since idealized models often do provide this modal information, then we should still regard such models as explanatory.

Connecting action explanations to counterfactual information is not unheard of in action theory, though such an approach is more common in non-causal accounts of action (Sehon 2005 p. 158). Relying on counterfactuals to ground explanations, however, is not a problem for our proposal to adopt a modeling approach to theorizing about action. As we pointed out above, it is worthwhile to consider different CMAs, which may need to be applied to different problems. One possibility is to consider a model variant that respects the Davidsonian insistence on reasons as *producing* causes. Alternatively, we can construct a CMA that includes a Lewisian understanding of causes as “counterfactual difference-makers” (1986), which may also do the work of meeting Davidson’s challenge of causally connecting reasons to actions.

*5.3 Can the modeling approach allow for moral evaluation?*

Another objection that one might have to our central proposal to treat theories of action as akin to idealized scientific models is that such a proposal is not suitable for moral evaluation. Our practices of assigning praise and blame normally require that we first settle the question of *who* acted. An agent X cannot be truly blameworthy for φ-ing unless it is true that X actually did φ. Ordinarily, we regard claims attributing actions to agents to be straightforwardly true or false.

 In addressing this objection, we must reiterate that, although models are properly understood as descriptions of idealized or fictional structures, we can still make true statements about reality based on those models. For instance, consider a Kantian model of action whose key idealizing components include “maxims” and “willings.” Within the model, an agent acts because they will a particular maxim. Since we are dealing with an idealized model, it may not strictly be true that an agent in the real world wills a maxim. Perhaps there are no such things as willings. Even so, we can say the causes that produce the agent’s actions are sufficiently similar to the willing of a maxim as specified in the Kantian model. Such a statement can be straightforwardly true or false, as required by our moral discourse and practice.

 It might seem that this response only trades one problem for another, however. Suppose two distinct models with incompatible idealizing assumptions M1 and M2 can both capture a wide range of actions, but which give conflicting verdicts in a few particular cases. Perhaps M1 does not classify event E as an action, whereas M2 does. If E is some event that one wishes to morally evaluate, then one might very well be troubled by this possibility.

 There are several responses that we can marshal on behalf of our proposal here. First, if we look at traditional action theory there is good reason to believe that the scenario posed above, where different models issue different verdicts about whether an agent acted at all, will be quite rare. Some hard cases notwithstanding, the debate in philosophy of action usually centers on *what explains* action, not the extension of the term “action.” Second, contra Anscombe’s (1958) insistence that ethical inquiry first requires a theory of action, as it actually happens contemporary ethical theorizing proceeds largely independent from action theory. It is quite rare for ethicists to directly invoke a theory of action before undertaking their moral evaluation. This may be because ethicists do not normally disagree about whether an agent *acted* at all. The locus of debate is typically over whether the person acted rightly or wrongly. Most parties in a moral debate tend to agree about who acted. Third, even if two ethicists disagree about whether an agent acted, this disagreement should not matter, at the end of the day, for the purposes of moral evaluation. For instance, suppose ethicist *A* believes the agent performed an action, whereas ethicist *B* believes the agent’s behavior does not constitute an action, but rather an *omission*. Whether this dispute matters for the purposes of moral evaluation depends upon first-order ethical claims about the moral relevance of the action vs. omission distinction. Certainly, we are sometimes morally responsible for our omissions, and convincing arguments have been raised against the moral relevance of the action vs. omission distinction (Rachels 1975). So, even if the two ethicists armed with different models disagree about whether some event is an action, plausibly this will not really alter the moral assessment of the event. Ethicist *A* can simply assume, *arguendo*, that the agent’s behavior constituted an omission and then proceed to show why that omission was morally wrong. Finally, although we suggest the idea that action theorists appeal to multiple models for understanding action, we are open to the possibility that there is one single best model. If we land on the one true model, that will plausibly be the result of an extended and complicated process of reflective equilibrium, where mutual adjustments between our model and pre-theoretic intuitions take place.

**6. Concluding Remarks**

In 1963, Davidson’s “Actions, Reasons and Causes” emerged as a proposal for a radical new direction in action theory—the now-common approach that marks what D’Oro labels the “ontological turn” (2008). In the face of several pervasive problems it has remained the standard theory of action, due largely to a lack of competition and an extremely prolific research program. In this paper, we have proposed another shift in the study of action which, following D’Oro, may be called a “modeling turn”. The key feature of the modeling turn is that action is studied indirectly, via idealized models, rather than via abstract direct representations of the ontology of action.

As we have argued, there are some compelling considerations in favor of this proposal. First, the modeling approach helps us to further naturalize action theory, making the study more continuous with the empirical sciences. This is because the vast web of folk psychological concepts and their connection, which the CTA employs as the primary *explanatia* of actions, are themselves be better understood as a model of cognition. Second, understanding the CTA as a model allows for a wider array of responses to common objections. As a model, the CTA is free from the constraint of having to account for every instance of every type of action, and some data that cannot be accounted for with one model may be accounted for with another. We believe that all these reasons, taken together, provide a strong argument in favor of recasting the CTA as a model of action. Of course, there are many questions of interest to action theorists that we do not here have the space to discuss—questions about distinguishing reasons *for* an action from the reasons an agent in fact *acted on*; questions about free will and voluntariness; questions about non-human agents; the list goes on. We consider these questions exciting challenges, and believe that future work on action modeling can help advance this literature.

References

Anscombe, G. E. M. (1957). *Intention.* Harvard University Press.

Anscombe, G. E. M. (1958). “Modern Moral Philosophy.” *Philosophy*, 33(124), 1–19.

Ankeny, R. & Leonelli, S. (2011). “What’s So Special about Model Organisms?” *Studies in History and Philosophy of Science Part A*, 42 (2):313-323.

Aristotle, & Irwin, T. (1999). *Nicomachean Ethics*. Hackett Publishing.

Bailer-Jones, D. (2009). *Scientific Models in Philosophy of Science*. Pittsburgh, PA: University of Pittsburgh Press.

Batterman, Robert W. (2001). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press.

Bird, A. (2006). *Philosophy of Science*. London: Routledge.

Bokulich, A. (2011). “How Scientific Models Can Explain,” *Synthese*, 180(1): 33-45.

Brand, M. (1972). “A Theory of Human Action.” *Journal of Philosophy,* 69 (9):249-257.

Bratman, M. (2000). "Valuing and the Will." *Philosophical Perspectives* 14: 249-65

Buss, L. (1987). *The Evolution of Individuality*. Princeton: Princeton University Press.

Cabrera, F. (2017). “Cladistic Parsimony, Historical Linguistics and Cultural Phylogenetics,” *Mind and Language*, 32(1): 65-100.

Cabrera, F. (2021). “String Theory, Non-Empirical Theory Assessment, and the Context of Pursuit,” *Synthese* 198: 3671–3699.

Cartwright, N. (1983). *How the Laws of Physics Lie.* Oxford: Clarendon Press.

Curry, D. S. (2018). “Beliefs as inner causes: the (lack of) evidence.” *Philosophical Psychology* 31 (6):850-877.

Davidson, D. (1963). “Actions, Reasons, and Causes.” *Journal of Philosophy* 60 (23):685.

Davidson, D. (1973). “Causal relations.” *Journal of Philosophy*, 64, 691-703.

Davidson, D. (1980). *Essays on Actions and Events.* Oxford University Press.

Davidson, D. (1999). “The Emergence of Thought,” *Erkenntnis*, 51(1), 7-17.

Dawid, R. (2013). “String Theory and the Scientific Method.” Cambridge University Press.

Dennett, D. C. (1978). *Brainstorms: Philosophical Essays on Mind and Psychology*. MA: MIT Press

D’Oro, G. (2008). “The Ontological Backlash: Why did Mainstream Analytic Philosophy Lose Interest in the Philosophy of History?” *Philosophia* 36 (4):403-415.

Downes, S. (2011). “Scientific Models,” *Philosophy Compass*, 6(11): 757–764.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Enç, B. (2004). “Causal Theories of Intentional Behavior and Wayward Causal Chains.” *Behavior and Philosophy* 32 (1):149 - 166.

Fodor, J.A. (1987). *Psychosemantics*. Cambridge, MA: MIT Press

Foley, R. (1977). “Deliberate action.” Philosophical Review 86 (1):58-69.

Frankfurt, H. (1997). “The problem of action.” In Alfred R. Mele (ed.), *American Philosophical Quarterly.* Oxford University Press. pp. 157-62.

Frigg, R. & Nguyen, J. (2016). “Scientific Representation,” *The Stanford Encyclopedia of Philosophy (Spring 2020 Edition)*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/ spr2020/entries/scientific-representation/>.

Frigg, R. (2010). “Models and Fiction,” *Synthese*, 172(2): 251–268.

Frigg, R. (2022). *Models and Theories: A Philosophical Inquiry*. London: Routledge.

Fritts, M. (2020). *Actions and Causes: A Methodological Inquiry into the Explanation of Action*.(PhD Dissertation).

Fritts, M. (2021). “Reasons explanations (of actions) as structural explanations,” *Synthese* 199(5-6): 12683-12704.

Gelfert, A. (2016). *How to Do Science with Models: A Philosophical Primer*. Cham, Switzerland: Springer.

Gettier, E. (1963). “Is Justified True Belief Knowledge?” Analysis, 23(6): 121–123.

Giere, R. (1999). “Using Models to Represent Reality,” in *Model Based Reasoning in Scientific Discovery.* (eds.)m . L. Magnaiv, N. Nersessian, and P. Thagard. New York: Plenum Publishers, 41–57.

Giere, R. (2006). *Scientific Perspectivism*. Chicago, IL: University of Chicago Press.

Godfrey-Smith, P (2009). “Models and Fictions in Science,” *Philosophical* Studies, 143: 101–16.

Godfrey-Smith, P. (2005). “Folk psychology as a model.” *Philosophers' Imprint* 5:1-16.

Godfrey-Smith, P. (2006). “The Strategy of Model Based Science,” *Biology & Philosophy*, 21(5): 283-299.

Goodman, N. (1976). *Languages of Art*. Indianapolis and Cambridge: Hackett.

Grüne-Yanoff (2013), T. “Appraising Models Nonrepresentationally,” *Philosophy of Science*, 80(5): 850-861.

Hacker, P. M. S. (2000). *Wittgenstein: Mind and Will, Volume 4 of an Analytical Commentary on the Philosophical Investigations.* Wiley-Blackwell.

Harnay, P.-V., & Rème, P. (2017). “Davidson: Decision and Interpretation,” *Argumenta*, 3(1), 39-52.

Holton, R. (2009). *Willing, Wanting, Waiting.* Oxford University Press UK.

Kitcher, P. (2001). *Science, Truth, and Democracy*. Oxford: Oxford University Press.

Levins, R. (1966). “The Strategy of Model Building in Population Biology,” *American Scientist*, 54(4): 421-31.

Lewis, D. (1986a): “Causation.” reprinted in *Philosophical Papers*, Vol. II, New York: Oxford U.P, pp. 159-213.

Longino, H. (2002). *The Fate of Knowledge.* Princeton: Princeton University Press.

Lowe, E. J. (2008). *Personal Agency: The Metaphysics of Mind and Action*. Oxford University Press.

Malpas, Jeff, "Donald Davidson", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/davidson/>.

Maynard Smith, J. & Szathmáry, E. (1995). *The Major Transitions in Evolution*. Oxford: Oxford University Press.

Mele, A. R. (2002). *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.

Mele, A. R. (2003). *Motivation and Agency*. Oxford University Press.

Mele, A. R. (2009) *Effective Intentions: The Power of Conscious Will*. Oxford University Press.

Morgan, M & Morrison, M. (eds.), (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge: Cambridge University Press.

Nanay, B. (2014). “Naturalizing action theory.” In Mark Sprevak & Jesper Kallestrup (eds.), *New Waves in Philosophy of Mind*. Palgrave.

Paul, L. A. (2012). “Metaphysics as modeling: the handmaiden’s tale.” *Philosophical Studies* 160 (1):1-29.

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago, IL: University of Chicago Press.

Rachels, J. (1975). “Active and Passive Euthanasia,” *New England Journal of Medicine*, 292:78-80.

Roth, A. S. (1999). “Reasons explanations of actions: Causal, singular, and situational.” *Philosophy and Phenomenological Research* 59 (4):839-874.

Schelling, T. C. (1978). *Micromotives and Macrobehavior*. New York: Norton.

Schlosser, M. "Agency", *The Stanford Encyclopedia of Philosophy* (Winter 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2019/entries/agency/>.

Sehon, Scott R. (2005). Teleological Realism: Mind, Agency, and Explanation. Cambridge MA: Bradford Book/MIT Press.

Sterrett, S. (2004). “Models of machines and models of phenomena,” *Studies in the Philosophy of Science,* 20(1): 69-80.

Stoutland, F. (2008). “Indeterminism, Intentional Action, and Bodily Movements.” in: C. Sandis (ed.), *New Essays on the Explanation of Action*, Basingstoke.

Strevens, M. (2008). *Depth: An Account of Scientific Explanation*. Cambridge, MA: Harvard University Press.

Suárez, M. (2004). 2004, “An Inferential Conception of Scientific Representation,” *Philosophy of Science*, 71(5): 767–779.

Suppes, P. (1960). “A Comparison of the Meaning and use of Models in Mathematics and the Empirical Sciences,” *Synthese*, 12: 287–300.

Titelbaum, M. (2012). *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief.* Oxford University Press.

van Fraassen, B.C. (1980). *The Scientific Image*. Oxford: Oxford University Press.

Velleman, D. (1992). “What Happens When Someone Acts?” *Mind*, 101(403): 461–481.

Velleman, D. (2000). *The Possibility of Practical Reason*. Oxford University Press.

Weisberg, M. (2007). “Who is a Modeler?” *British Journal for Philosophy of Science,* 58: 207–33.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford: Oxford University Press.

Williamson, T. (2000). *Knowledge and Its Limits.* Oxford: Oxford University Press

Williamson, T. (2017). “Model-building in philosophy.” In R. Blackford & D. Broderick (Eds.), *Philosophy’s future: The problem of philosophical progress* (pp. 159–172). Hoboken: Wiley.

Wilson, G., Shpall, S. and Piñeros Glasscock, J. S., "Action", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/action/>.

Wittgenstein, L. (1963) *Tractatus Logico-Philosophicus*. Trans. D.F. Pears & B.F. Mcguinness. Routledge and Kegan Paul.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

1. Although many causal theorists trace the root of the theory to Aristotle, Davidson is the source of the contemporary causal theory with which we are most familiar today. [↑](#footnote-ref-1)
2. Dretske (1988), for example, seeks to naturalize Davidson’s account of beliefs and desires in a way that still renders them explanatory. Mele (2002) aims to add types of explanatory mental events to Davidson’s framework, such as “willings” and “intentions”, which he argues are irreducible to beliefs and desires. [↑](#footnote-ref-2)
3. To be sure, it is uncommon to think of philosophical investigation as sometimes involving model-building but see Paul (2012) and Williamson (2017) for some notable exceptions. As Williamson points out, much formal epistemology and philosophy of language already relies on “models” in the sense that scientists understand the term (2017, p. 164-6). Thus, although model-building is much more common in the sciences, there is nothing essentially “scientific” or “non-philosophical” about employing models. As we argue, central debates in action theory lend themselves particularly well to the modeling approach. While we think it would be fruitful for philosophers to consider model-building in domains outside of philosophy of action, we wish to remain neutral with respect to ambitious methodological claims, e.g. that philosophy ought to *primarily* involve model-building. Rather, we presuppose only the modest thesis that model-building is *a* legitimate way of doing philosophy. [↑](#footnote-ref-3)
4. See, for example, Giere (2006), Bailer-Jones (2009), Gelfert (2016), Weisberg (2013), Morgan and Morrison (1999), Potochnik (2017), etc. For a recent general overview, see Frigg (2022). [↑](#footnote-ref-4)
5. See, for example, http://nifty.stanford.edu/2014/mccown-schelling-model-segregation/ [↑](#footnote-ref-5)
6. There are many competing accounts about how this last stage, the representation stage, works (Frigg and Nguyen 2016). Many of those who agree on the basic outline of the modeling approach disagree on the details of how models represent. For Giere (1999), an informal relation of *similarity* is good enough to capture how it is that models represent, notwithstanding the stock objections that similarity is too vague, or everything is similar to everything in some respect (e.g. Goodman 1976). Drawing on the structuralist tradition, others appeal to precisely defined logico-mathematical relations, such as *isomorphism* (van Fraassen 1980). Still, others adopt a deflationary account of the representation, ultimately cashing out the representational capacity of models in terms of their ability to allow agents to draw *inferences* concerning some target (e.g. Suarez 2004). We will remain neutral on this internal debate about representation in this paper, however. [↑](#footnote-ref-6)
7. Later in section 5, we will consider to what extent models can be used to *explain*. [↑](#footnote-ref-7)
8. In fact, it is possible for different scientists to study the same phenomena, one with and one without the aid of models See the helpful comparison made by Godfrey-Smith (2006, p. 730-2), of important works in evolutionary theory: Buss’ *The Evolution of Individuality* (1987) and Maynard Smith and Szathmáry’s *The Major Transitions in Evolution* (1995), where the latter of which relies heavily on mathematical models, while the former does not. [↑](#footnote-ref-8)
9. For a representative instance of the CTA, which fits with our general characterization, see Enç (2004, p. 157) [↑](#footnote-ref-9)
10. Bratman (2000) discusses action theory in a way that he describes as “modeling”. This sort of modeling, however, has a therapeutic end. That is, the different “models” Bratman discusses are fictional agents representing different levels of potential agency, and the purpose of presenting them in succession is to bring the reader to specific insights about agency. The goal of Bratman’s project is still to produce an abstract directly- representing theory of agency. This is different from the sort of indirect modeling that we are interested in here. Godfrey-Smith (2005, p. 14) gestures at something more closely related to our project here: “What consequences would it have if some of the rather baroque psychological structures discussed in action theory […] were explicitly understood as ‘model-bound’ entities, in a sense analogous to the scientific case?” [↑](#footnote-ref-10)
11. As a helpful analogy, consider that the Bayesian research program involves positing ideal agents in order to theorize without the constraints of e.g. limited memory capacity (Titelbaum 2012, p. 73). [↑](#footnote-ref-11)
12. Although we draw on the early Davidsonian picture of reasons for actions as belief-desire pairs, we recognize that other causal theories (including Davidson’s later work) incorporate other attitudes as causal reasons for actions such as intentions, judgments, plans, etc. Such attitudes could be modeled as easily as those used in our examples. [↑](#footnote-ref-12)
13. In this paper we lack the space to argue specifically for this concept of naturalization, and we recognize others may have different approaches in mind for “naturalizing” action theory, but we believe this idea of naturalization enjoys broad appeal in the discipline. [↑](#footnote-ref-13)
14. Thanks to Marcos Picchio for crucial discussions about the nature of folk psychology. [↑](#footnote-ref-14)
15. Wittgenstein and Anscombe were, of course, not causal theorists. Still, causal theorists may consider such a proposal regarding the quest to secure a set of necessary and sufficient conditions for action. [↑](#footnote-ref-15)