Describing Models

Before diving into Zollman's models, I want to explore models in general and to define terms that I will continue to use for the rest of this thesis. This section will touch on the vibrant debate about the role of models in science by explicitly presenting the interpretations I'm using to understand models. Overall, I do not endorse this set of definitions as the right or the only way to interpret models, I simply propose that this understanding is the right lens through which to understand the issues salient in computational modeling of the sort I discuss here. To this end, I will propose that I focus on representational mathematical models as a natural lens to view computational modeling. Next, I discuss the differing purposes of modeling and why I opt to focus on explanatory, rather than predictive modeling when discussing computational models. Finally, I will advocate viewing explanation in computational models as *mechanistic* explanation because computational models lend themselves well to elucidating mechanisms because they are highly decomposable.

What Sort of Model Computes?

When I say a *model*, I refer to the mathematical structures used to represent some real phenomena in a target system. If we adopt the ontology for models proposed by Frigg and Hartmann, I refer to *representational* models of phenomena, or models which are positioned relative to a target. When a model M models a target T, we would say "M models T" and mean that M serves as an effective stand in for T in the context relevant to the goals of the model. It is important to distinguish these representational models from models which represent data points. A model of data can certainly inform a model of phenomena, but discussion of modeling data itself is typically discussed in the context of experimentation and empirical evidence, rather than the context of scientific understanding.[1]

In addition to being representational models of phenomena, I focus on mathematical models. By mathematical, I mean they are composed of equations rather than physical objects, set-theoretic structures, textual descriptions or any other kind of model. I focus on this sort of model because this sort of model lends itself best to direct computer simulation. Because computers at their core crunch numbers and move bits around, simulating a model on a computer requires converting that model into equations and mathematical structures that can be described as a computable program. Because the process of programming leads to simulated models being specified as equation-based models, the models-as-equations understanding is frequently used to discuss how computer simulations and their associated models work as a part of scientific inquiry.[2]

Furthermore, these relational models are inextricably tied to the possible worlds they model. A model has a *target*, which I broadly take to be an object or

[1] P. 11 Frigg and Hartmann, "Models in Science."
[2] Winsberg, "Computer Simulations in Science."

phenomena in a real or possible world that a model represents. Any equation-based representational model is going to somehow be about a target in some world, otherwise it would not be representational. To illustrate this, consider a scientist studying a tissue and its constituent cells. Say she has formed a model which represents the behavior of the cells within the tissue which accurately reflects the behavior of actual cells in the actual tissue. Such a model reflects the real world because there is an accurate representational relationship between the equations of the model and a tissue that exists in the real world. Now, say the scientist knows a drug affects cells in a certain way though some empirical means. She could form a model with the modified cell behavior and use it to explore how the tissue would behave under this regime. In this case, the model models the tissue in a hypothetical world in which the cells are affected by the drug.

Finally, I should clarify exactly what I mean by "simulation". I adopt Eric Winsberg's definition as "a program which runs on a computer and uses step-by-step methods to explore the approximate behavior of a mathematical model."[3] This definition stresses both that simulations approximate mathematical models and that they are typically time-dependent. The simulation approximates, rather than mimics perfectly the model because computers deal in discrete data so whenever a continuous model equation is modeled, some fidelity is lost. Consider modeling a flowing river. Where the model equations might define smooth rules for how the water flows at any time scale, a computer might need to simulate the model in second-long chunks of time for the computation to be tractable. If the system is already discrete, the simulation may more exactly approximate the model mathematics. Furthermore, this definition stipulates that simulations concern time-dependent processes, which rules out static computer renderings of geometrical structures, but allows videos of such renderings as could be the case in a virtual reality simulation. In practice, this focuses models on describing and understanding dynamic processes rather than static structure or other unchanging properties.

Example: Modeling a Cannon

To further discuss how equation-based models represent phenomena in the real world, consider how one might formulate and use such an equation. Say an engineer, Alice, builds a cannon and places it on top of a fort. Alice wants to know how far the cannon can shoot to understand how far that cannon can fire a cannonball, so she might decide to use kinematic equations to model the cannonball's flight path. To do this, Anna wants to use as complete and as simple an account of motion as possible to understand what affects the flight path of the ball. Thus, she adapts Newtonian mechanics' kinematic equations to model the cannonball's motion. However, the a theory of mechanics on its own is insufficient to model the cannonball and Alice must chose how she will model the initial height and velocity of the ball, the angle of the cannon as well as the

[3] Winsberg.

acceleration due to gravity. Say Alice knows how tall the fort is (denoted y_0), the initial velocity of the ball for a standard shot (v_0), the acceleration due to gravity on Earth (g) and can set the cannon angle to θ degrees. If she assumes all these values are fixed, she can use the general kinematic equations to relate all these quantities. For example, the ball's distance from the ground could be modeled in relation to time as follows, where $v_0 \sin(\theta)$ is the vertical component of initial velocity:

$$y(t) = y_0 + v_0 \sin(\theta)t - g_0 t^2$$

Using this equation, Alice could solve for the time when the ball hits the ground, t_h. Alice can model this in the kinematic equation as some time after the ball is fired at $t = 0$ (meaning that $t_h > 0$) when the ball ends up on the ground (meaning $y(t_h) = 0$). Once this time is found, Alice can formulate another kinematic equation to model the horizontal motion of the ball, where $v_0 \cos(\theta)$ is the vertical component of initial velocity:

$$x(t) = v_0 \cos(\theta)t$$

Note that in the kinematic equation for horizontal movement does not include an initial position or a term for gravity because Alice models the distance away from the cannon over time and gravity is only modeled in the kinematic equation for vertical position as gravity principally accelerates objects towards the Earth. Now, using this equation-based model of the motion of cannonball, Alice can find that the ball travels a distance of $x(t_h)$.

What is important to note about this model is that Alice uses it as a stand-in for firing an actual cannonball in the actual world. Thus, Alice takes these mathematical equations to represent a real cannon shot despite them being relatively simplistic and ignoring many parts that might have some effect that is not deemed relevant for the behavior Alice wises to understand. For example, she doesn't model the affect of wind or the friction the ball would experience when the ball hits the ground. If we look closely, this means that the modeled ball's distance $x(t)$ increases indefinitely even after the ball hits the ground at t_h. Beyond missing friction from the ground, the ground doesn't even exist in the model at all. We can see this because the equation $y(t)$ will continue to decrease (at increasing rates) after t_h. However, because Alice only is modeling the ball's flight, the target of the model is simply the position of the ball when it is flying between time $t = 0$ and time $t = t_h$.

The target establishes a clear boundary for a model. Alice only cares about modeling the ball in flight, meaning she doesn't evaluate her model at times when that is not the case. After the impact at t_h, the above model stipulates the ball will continue to go into the ground at ever increasing speed. This is obviously not possible in the real world, but this doesn't threaten the model-target relationship because the target simply isn't the behavior of the ball after impact and the

model performs well on the actual target of the ball's flight. Thus, targets can be very constrained, yet still have a justifiable model-target relationship.

If Alice instead wanted to consider a target world where the cannon was placed on the planet Mars, she'd have to change her model such that it represents this new target. Specifically, because we know that objects on the surface of Mars behave in much the same way to objects on Earth, she can use the same equation with g changed to match the gravitational acceleration on the surface of Mars ($g_0 \approx 3.7 m/s$). However, it is imperative that Alice know that the same kinematic equations apply on Mars, otherwise the model won't be a good one for the new target. While the change is small in this case, we could easily imagine a case where Martian objects didn't behave according to classical mechanics. For instance, consider a possible world where gravity on Mars varied significantly from one second to the next due to rapid movements in mass in the planet's core. Alice would need to account for this time dependence in her cannonball motion equations by adding terms to the model equations to ensure they actually replicate this different target.

Now consider what a simulation of this model might look like. In this case, such a simulation is relatively trivial as the equations are closed-form. Alice could simply convert the kinematic equations given above to a computer program, then run them at increasing time-steps $t = 0, 1, 2, \ldots n$. Using this simulation, she could reach a similar answer to the analytical solution reached above. When she notices that $y(t)$ dips below 0, she will know the modeled cannonball has hit the ground and can consult $x(t)$ to see how far the ball traveled. Notice that this answer will be approximate if $y(t)$ is zero at a fractional time, like 1.1s, which is an artifact of discretizing the continuous kinematic equations to discrete time values in the simulation.

Purposes of Models

Now that I've introduced computer simulations and representational models, I turn my attention to the reasons to create a model in the first place. Different types of models can have a wide variety of purposes and a model that is a great model for one purpose might be a very poor model for another. To clarify what sense I mean by "purpose" I'll consider a few examples.

First, consider a globe. It is a physical model of our planet that is intended to educate us about the geography of the planet and employs a tactile and visual representation to do so. However, that same tactile representation makes it a very bad model of the gravitational forces our actual planet imparts on objects around it. Furthermore, it is also a poor substitute for more detailed technical maps which help planners understand the world in which a new bridge or building might inhabit due to a lack of detail. Thus, a globe is a great fit for educating us about the general geographic structure of our planet, yet is a poor fit for understanding our built environment or our planet's gravitational forces.

While educational models are useful, that purpose is somewhat orthogonal to

the purposes of models more directly associated with scientific inquiry. Within science, prediction and explanation are typically more common reasons to craft models. These purposes seem similar on the surface because if we can predict something, wouldn't we understand it? While the two are often closely related and aiming for one often leads to the other, the design of a model can be heavily influenced by the goals of a model.

For example, consider DeepMind's AlphaZero project[4] which resulted in a computer go player which can beat even the best human players in the world at the classical board game. They use deep neural nets trained with reinforcement learning to build this model go player such that the model has superhuman capabilities at selecting the best move in any given game state. However, such models notoriously do not give us much insight into *why* any given move is chosen, meaning the model does not help us better understand what a great move in go is, it simply makes them. This problem of "explainable machine learning" has become a hot topic for research with organizations like DARPA attempting to find ways to build models which do not only predict and classify, but also explain.[5] Thus, models can be fantastic at prediction, while explaining relatively little about how those predictions are made. This type of successful, but not explainable model stands in opposition to classical mechanical models which provide both a deep understanding of physical phenomena and provide exceptional predictive power.

Let this not be construed as a normative claim in support of pursuing only models which further our explanations of the world. Predictions can save lives! For example, Meteorologists build models of the atmosphere principally to predict its behavior in the future because accurate weather prediction gives people enough warning to evacuate. Consider the fact that the National Weather Service mission is not to understand and explain the weather, but to "provide weather, water, and climate data, forecasts and warnings for the protection of life and property and enhancement of the national economy."[6] While understanding the weather might help with this goal, a hypothetical unexplainable weather model which provides exceptional predictions would be a fantastic model for the NWS as it would save lives despite not necessarily satisfying our thirst for explanation.

What is an Explanation?

I now turn my focus to the scientific method and models which do primarily aim to help us craft explanations. However, understanding how models can provide explanations for phenomena, requires a high-level understanding of what we take explanations themselves to be. As is a theme with whole section, explanations are a contested and rich topic of debate, so my main goals are to cover the major themes as they relate to scientific modeling without making any normative claims about what mode of explanation should be used in general. To do this,

[4] Silver et al., "Mastering the Game of Go Without Human Knowledge."
[5] "Explainable Artificial Intelligence (XAI) Program Update."
[6] "About the NWS."

I'll paint the debate loosely along the lines painted in James Woodward's survey of the field.[7]

First, I turn my attention to the general structure of an explanation and why it differs from a a description. Broadly, I take successful explanations to give a satisfactory account of *why* a given phenomena occurred the way it did. This stands in contrast to a description, which is principally tasked with *what* that phenomena itself is. There is some relation between the two in the sense that it seems, for most cases, we cannot explain a phenomena without first being able to describe it. To frame things more formally, I refer to an explanation E of a phenomena P. Furthermore, in each case I'll also handle cases where multiple explanations can explain a given phenomena, so we could say that E_1, E_2 both explain P.

Logical Explanation

The first major account of explanation frames an explanation in terms of logical implication, as proposed by Hempel and Oppenheim, who frame explanation in terms of laws and logical rules. They posit that explanation can be framed in terms of three logical rules:

1. The explanation must imply the phenomena, i.e. $E \to P$.
2. The explanation E must contain "general laws" and at least one of these general laws must be required to deduce the implication.
3. E must be, in principle, empirically verifiable.
4. E must be true.

For a paradigmatic example of this sort of explanation, let us return to our kinematic equations. We might say that those equations are a description of behavior P which can be derived from the foundational laws of motion. These laws, E, state that motion is conserved, that force is proportional to mass and motion ($F = ma$) and that every force has an equal and opposite force. We can use these general laws to derive the above kinematic equations, so $E \to P$; we know the E consists of well-verified general laws, so we can conclude that E explains the kinematics of the cannonball's motion P under this model.

Now, the issue with such an account of explanation within the context of modeling is not with the logical structure but rather that often we take things to be explanations which do not follow from general natural laws. For example, in fields such as epidemiology, convincing explanations are produced without deduction from general laws but rather from inference or induction from observed data. For example, recently there has been much talk about the outbreak of a new strain coronavirus in Wuhan, China which seems to spread person-to-person and cause moderate to sever symptoms.[8] However, why do we speak of "coronavirus" and not an outbreak of severe respiratory dysfunction in China

[7] Woodward, "Scientific Explanation."

[8] Li et al., "Early Transmission Dynamics in Wuhan, China, of Novel CoronavirusInfected Pneumonia."

as well as the discovery of a unique coronavirus species (2019-nCoV) in nearly all those experiencing the symptoms? We seem to treat 2019-nCoV as the explanation for the symptoms but there are no general natural laws from which some epidemiologist has derived this fact as a physicist would derive kinematics from Newton's laws of motion. Instead, epidemiologists rely on observation, induction and careful statistical methods to reach these conclusions.

One additionally note about this example is that under this law-based view, so long as laws don't exist, explanations don't exist. Thus, even if an (unethical) controlled trial were carried out on people to show that 2019-nCoV is the cause of the disease, it wouldn't count as an explanation so long as it doesn't follow directly from general laws. Though, I will cover such methods in more depth in relation to other formulations of explanation.

Because this account relies so heavily on analytical or law-based reasoning, it doesn't naturally fit with a simulation-based approach to scientific inquiry. Simulations, almost by definition, are not deductively solved because if they could be, the scientist would craft solutions analytically. Furthermore, the cases of the simulation discussed here in social science, cognitive science, etc often do not have general laws in the Hempel's sense of the term.

Probabilistic Explanation

While there are ways to directly extend the above logic and law-based to work probabilistically through an inductive formulation of the law-based approach above,[9] I propose instead to focus on a later development of statistical explanation presented by Salmon[10] as *statistical relevance*. I focus on this account because I'll argue later that simulations are well-suited to substantiate such explanations.

Salmon's statistical relevance is built on the concept of conditional probabilities. For events A and B, conditional probability is a way of representing the probability of A occurring *given* B has occurred. To represent this relationship, we write:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Consider the trivial example where A and B are independent such that $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{4}$. The probability both A and B occur is $P(A \cap B) = \frac{1}{2} * \frac{1}{4} = \frac{1}{8}$. Which leads to the conditional probability:

$$P(A|B) = \frac{\frac{1}{8}}{\frac{1}{4}} = \frac{1}{2}$$

[9] Hempel, *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science.*
[10] Salmon, "Statistical Explanation."

Thus, the probability A occurs given we know B occurs is the same as the probability of B, which makes sense because B is independent of A.

Now, this concept of conditional probability leads very naturally to Salmon's statistical relevance form of explanation by essentially formalizing the notion of an explanation E being *relevant* to the phenomena P. In the context of the events A and B, if they are independent and $P(A|B) = P(A)$, then B isn't relevant to A, whereas if $P(A|B) \neq P(A)$. Using this idea and applying it to scientific experimentation we can get the following formulation. First, pick some subject for study A, say all people. Now, we'd like to study whether a fact F about a person is explained by some other property E. To do this we create a study group $A.E$ of people who have the property E and one without the property $A.\bar{E}$ such that $\bar{E}$ and E are mutually exclusive (formally a homogeneous partition). Using this partition, we consider E and explanation for F if and only if $P(F|A.E) \neq P(F|A)$. Essentially, if the conditional probability of the experiment group, $P(F|A.E)$, is different from the control group $P(F|A.E)$, then it seems E, in some sense, explains F in either a positive or negative sense.

This model provides good motivation for how a simulated model can aid understanding. If we take that a model M models its target W well enough to stand in for W for measuring a fact F. If the modeler wants to see if some property \$E of M explains F. The modeler may setup a model $M' = M.E$ where E is modified to be the case in the model code to be simulated. Now, the modeler may measure $P(F|M)$ and $P(F|M')$ and if $P(F|M) \neq P(F|M')$. Simulated models, in principle, make it easy to modify model behavior in code so it is easy to compare how models with different properties behave and compare them to develop explanations in this sense.

However, there is a serious problem with this model of explanation in that it does not require that explanations be *causal*. Consider the case of the ShakeAlert early warning system.[11] Such a system contains many seismic sensors across a wide area near fault lines such that any given earthquake will occur physically close to a sensor so shaking will be detected nearly instantaneously. This system then immediately notifies the rest of the population by push notification or text message. If those notified by the system are further away than the sensor was from the earthquake, they will receive the notification before the shaking begins because the notification travels over communications networks faster then the P-waves propagate shaking.

Where T is all moments, consider the general probability of feeling shaking at any given moment $P(S|T)$ and the probability of feeling shaking after after noticing a push notification from ShakeAlert $P(S|T.A)$. If we assume the system works as designed $P(S|T.A) >> P(S|T)$, which because the alert property is exhaustive (we are always either alerted or not), means A explains S under this scheme. However, it is clear that the alert did not cause the shaking, the earthquake did! There are methods of addressing this issue proposed by Salmon,

[11] Kohler et al., "Earthquake Early Warning ShakeAlert System."

however, they aren't particularly salient for the simulation-based models such that a general, high-level treatment of causality is not sufficient.

Causal Explanation

So then, what does it take for an explanation to be causal? This question touches on an extremely lively debate about what causality fundamentally is. However, for the purposes of dealing with establishing causal explanations it should be sufficient to choose a reasonable, if not unchallenged, account of causality which handles the relevant cases well enough. To this end, I will adopt the counterfactual account of causation most widely attributed to David Lewis.[12] While this theory is certainly not without problems, the high level ideas should prove useful in elucidating what it means to *establish* a causal relationship in the contexts important here.

Lewis was interested in defining counterfactuals formally and approaches the problem by using "possible world semantics". First though, I adopt his notation that the counterfactual of two facts A and C is $A\Box \rightarrow C$, which represents that if A were true then C were also true. More precisely Lewis says $A\Box \rightarrow C$ is true in a world w either vacuously when A is false in w or when an alternate world w' in which A and C are true is "closer" to w than another possible world w'' in which A is true and C is not. For something to be causal, all possible worlds w' must be closer to w than all possible worlds w''.

Now, let's reframe the earthquake alert explanation in terms of this definition of causality to show how causal explanations are more satisfying than simply statistically-relevant ones. Recall that an alert A warns of impending shaking S from an earthquake E and that we found A statistically relevant to S, meaning it could be considered an explanation if we take explanations to be properties which are statistically relevant to other properties. However, $A\Box \rightarrow S$ would not hold. To see this, let w be the world in which I observe a ShakeAlert notification on my phone and thus A is true. This means $A\Box \rightarrow S$ is not vacuously true. Now, consider a possible world w' where the alert is received and the ground shakes a few moments later (A and S both true). By contrast, in the world w'' the alert is received and the ground doesn't shake a few moments later. As we discussed earlier, many such w' worlds exist because the alert is statically relevant to shaking. However, there are also many such w'' possible worlds which are equally "close" to w. Consider the case where the system over-estimates the propagation of the S-Waves and alerts people outside the actual zone of shaking or a simple accidental triggering of the system due to a system test or faulty sensor. Many people would receive an alert, but never feel any shaking, so A holds true while S does not. Thus, the conditions for the alert to be a causal explanation for the shaking don't hold and $A\Box \rightarrow S$ is false.

Now, if we instead focus on the explanation that the earthquake E explains the shaking instead of the alert, we do find a causal relation $E\Box \rightarrow S$. Let world w

[12] Lewis, "Causation."

be one in which there is an earthquake near me (assume within range for S-waves to cause shaking near me). Now, world w' be the clearly possible world in which an Earthquakes waves, E, are below me and I experience shaking. World w'' then becomes the world in which the Earthquake occurs E but I feel no shaking. This seems impossible, save for some exotic worlds in which I'm standing on some dampening structure, etc. Because all the w'' are not obviously possible like the w' worlds are, we can consider all the w' worlds closer to w than the w'' worlds and thus uphold the relation $E\Box \rightarrow S$.

In this example, we see that the causal explanation captures the sense in which the earthquake "explains" the shaking more precisely than an explanation merely built on statistical relevance. While it might seem unsatisfying to say an earthquake explains the Earth shaking, we can always ask deeper questions about what explains the earthquake or what in particular about the earthquake causes the shaking to refine an explanation. However, in any case, we'd like to ensure all our explanations are causal because otherwise we can end up with explanations which don't help us understand anything at all such as the statically relevant but not causal case where the alert explains the shaking.

Establishing Causal Explanation

Now that we've described what a causal explanation is, how might we convince ourselves of a particular explanation? Because this question is quite core to the scientific method, it is heavily studied and debated across different fields. However, there is broad consensus across medicine and the social sciences that randomized controlled trials are the "gold standard" for establishing causal links, standing in opposition to muring observational studies which typically are not taken to establish causal relationships. While randomized controlled trials are a specific methodology originally designed for clinical trials, the general experimental principles are more general and are reflective of a widely applicable way of looking for causal explanations, which I will detail in this section. After explaining how experimentation can generate causal relationships, I'll consider a few alternative cases where causality may be inferred without such experimentation to demonstrate that there is no unique method of determining causality.

Now, consider a generic, idealized experimentation setup where a researcher wishes to determine if some property E explains another property P. The researcher first hypothesizes that E explains P, then begins to carefully design an experiment to tease out this relation. To to this, the experiment must:

1. Hold all properties other than P constant.
2. Allow the observation of P when both when E is true and when it is not.
3. Show that P is true if and only if E is true.

From this sort of setup if the experiment is carried out carefully, the experimenter can conclude that $E\Box \rightarrow P$ or that E *explains* P under the causal understanding of explanation. In the terms of Lewis' possible worlds, this experiment shows

that there are no possible worlds w'' where P is true but E is not which are closer to w than any world w' where both P and E are true. We learn this because this experiment, if carefully designed, isolates the effects of E by changing *only* E between runs and observing the effect on P. If P changed, we know that worlds in which E holds lead to worlds in which P holds because, counterfactually, worlds in which E does not hold do not lead to worlds in which P holds.

Now, this general frameworks works only when experimentation is carefully and flawlessly carried out. This is very difficult and is a core reason science is hard! It is very tricky to know in practice that a given intervention changed *only* one property in the world. For example, consider a trial of a new drug designed to lessen flu symptoms. A naive study trying to show that the drug explains the lack of flu symptoms in users might create two equivalent groups of people, giving one the new drug and the other nothing to serve as a control group. While seemingly the drug is the only difference between the groups, more careful consideration shows that one group takes a pill and knows they are taking a drug while the other does not. As it turns out, the placebo affect shows that this small difference can be enough to explain the perceived or even actual improvement in symptoms. A more careful experiment design which holds this property constant across both groups would give the experiment group pills containing the new drug and the control group identical pills containing no drug at all so both groups have the same modifications to their routine. There are many more such cases where very subtle differences between the experiment group and the control group can end up explaining experimental results instead of the property an experiment was meant to test. Thus, experimentation must be carefully scrutinized to ensure that the variable being tested is in fact the only relevant one which is varied between control groups. In practice, designing careful experiments can involve carful statistical methods which help to extract signal from data made noisy by other present, but unimportant variations between groups.

While experimentation provides a powerful tool for generating causal explanations, causation can be inferred from data in some cases where the structure of causal relations fits certain patterns.[13] This is an exciting area of research because if passive data collection can be sufficient for establishing causation, many new avenues for generating explanations are opened which do not require the often fraught process of designing an experiment which really does isolate the relevant variable. However, in this thesis I discuss simulations, which are experiments, so I will focus instead on the traditional experimentation-based approach.

Mechanism in Science and in Models

Now, if we frame explanation in terms of causal relationships and posit that experimentation can elucidate and support these explanations, how might we connect this account to models specifically? I argue that mechanistic accounts

[13] Hitchcock, "Causal Models."

of science happen to provide a fantastic way of conceiving of models which explain real world phenomena. Mechanisms emerged from the study of biological scientific inquiry because many biologists explicitly frame their work as searching for and understanding mechanisms, though the approach has gained favor in explaining science more broadly. In this section, I draw from Carl Craver and James Tabery's article on this topic to hopefully provide a consensus account of mechanism.[14]

At a high level, I adopt Bechtel and Abrahamsen's definition of mechanism:

> A mechanism is a structure performing a function in virtue of its component parts, component operations and their organization. The orchestrated functioning of the mechanism is responsible for more than one phenomena.[15]

This definition identifies the key parts of a mechanistic model: phenomena, parts, causing and organization.

The phenomena of a model, in the context of a mechanistic explanation, is simply what is being explained by the behavior of a given mechanism. We can take this in a causal sense as we developed earlier saying the phenomena is *caused* by the mechanism. A way to help specify what a phenomena and mechanism is in many cases is in terms of inputs and outputs. For example, an engine (at a high level) takes in gas as an input and outputs rotational energy and exhaust, so we could consider it a mechanism which explains the phenomena of that energy conversion.

Mechanistic models are composed of one or more parts. How this decomposition happens will vary a lot depending on the mechanism and how it can naturally be divided. In the case of the engine example, the engine is composed of pistons, spark plugs, timing belts, exhaust pipes and fuel injectors, among others. It is easy in this case to identify the parts because the engine is engineered by people by composing parts, however, in scientific examples these divisions between parts can become less clear.

Next, these parts causally interact with one another. Because mechanistic models seek to provide causal explanations, we call these interactions *causings*. For example, the spark plugs cause the ignition of the gas mixture. A part may also cause multiple things, for example, a piston both captures the energy from the combustion of the gas mixture and pushes exhaust from that combustion out of the chamber.

Finally, the parts are organized such that phenomena emerges from their combination. For example, any one part of the engine could not convert gas to motion, but the arrangement of the parts creates this emergent phenomena with all parts working in concert. Additionally, the same part may show up multiple times in a mechanism as an engine often has four, six or eight identical cylinders.

[14] Craver and Tabery, "Mechanisms in Science."
[15] Bechtel and Abrahamsen, "Explanation."

If we combine all these parts and can break the phenomena down into parts and understand how those parts relate and combine to *cause* the phenomena, we'd say we understand it pretty well. An engine is a good example because engineers understand them very well because each of these parts and relations were designed with the intent to produce the given affect.

However, mechanism doesn't just apply to things we've designed as mechanism exist all around us. To pick a particularly complicated example: the brain can be viewed as a mechanism. Say it takes sensory data in and outputs movements (ignoring inner cognitive states in this specific example). We might divide the brain into the cerebrum, the cerebellum and the brain stem and somehow categorize interactions between these parts. While we clearly don't have the same level of understanding, we do understand some interactions and some parts parts. Not having fully realist and mechanistic understanding of a phenomena does not imply such an understanding is possible because it can be much harder to generate such an explanation than to understand it. For example, a team of researchers tried to dissect a microprocessor using first principles research techniques, as one would use in neuroscience. Despite the microprocessor being man made and mechanistically understood by some engineering team somewhere, the research team had trouble reaching any sort of mechanistic understanding because of the complexity of the system and limited means they had to test with.[16] Clearly a mechanistic understanding of a microprocessor is possible, it just may be hard to substantiate and discover with no prior understanding.

Models and Mechanisms

Now, I return to the topic of models and explain how simulated models offer mechanistic explanations of phenomena. To understand this relation between models and mechanisms, we'll first discuss models in isolation from their targets, then discuss how the model mechanisms might correspond to real mechanisms in the real world.

To begin to talk about models mechanistically, let a model be specified as the set $M = \{ m_1, \ldots, m_n \}$, where an m_i is a specific part of the mechanistic model. In a broad sense, these pieces of the model are separable portions of the model that causally interact with one another just like a mechanism in real life. Because simulated models must at some point be written down as computer code which is typically quite structured, this structure is often quite explicit. Thus, we can think of each model part m_i may be a variable, an equation, a piece of code or anything else that distinctly interacts with other parts and leads to the emergent behavior of the model. This division of the model into parts can happen at varying levels of granularity. From transistors at the lowest levels to subroutines for addition, multiplication and memory management somewhere in the middle to subroutines which describe model behaviors, a computational model can be mechanistically described due to running on a well-understood

[16] Jonas and Kording, "Could a Neuroscientist Understand a Microprocessor?"

machine. However, we typically focus on the higher-level mechanisms specific to the simulation program rather than those generic to all computation.

For a toy example, consider a physics simulation of many small bouncy balls bouncing in potentially chaotic ways. I'll call this model of balls M. Such a model will likely have some representation of a ball in the model $m_b \in M$ which contains properties like position, velocity, mass and whatever else is salient for the kinematic equations used. Furthermore, the model will likely contains some parts pertaining to the ball's environment, for example there might be a part pertaining to a floor, $m_f \in M$, which has a position and a certain coefficient of restitution to determine how bouncy the surface is. In some cases this floor may even move in space being modeled is an elevator. Furthermore, computational models often have significant bits of code which manage the model in general ensuring the right parts of the model run at the right times and that results from simulation is collected effectively. We can call all this glue the orchestration part of the model $m_o \in M$.

Once the parts of M are chosen, the modeler will have to spell out how the parts of the model interact with one another. By specifying these rules as code, the model gains much of its substance and much of the model's behavior is defined. In a model, the mechanism's causings still exist, they are just simply spelled out in great detail by the programmer of the modeler. Finally once all behavior is defined, all the parts must be organized in a certain way to setup how the parts will interact with each other. For the model M, this might mean choosing the initial positions and number of balls to subject to the simulation. A key property of the computational model's mechanism vs. mechanisms found in the real world is the model is *defined* in a certain way by the modeler. As I discussed earlier, it is much easier to understand a model which is designed by engineers piece-by-piece, like a microprocessor, than one which is given to us fully formed, like a brain. However, this ease of understanding computational model mechanisms seems to create tension: how can a model composed of trivially understandable mechanisms represent something which we don't yet understand?

I posit that the answer to this question lies in the relations between the model and the modeler's understanding of the target. As the modeler learns more about the target system through traditional experimentation or review of other research literature, the modeler can modify the relationships in the model to codify that understanding and make that understanding more rigorous. If those relations are strong, the simulations can lead to new findings about the target system by placing it into situations which are interesting, but difficult or even impossible to observe in the real world. In the case of the model M of balls, we can test how the model behaves in cases where we might have millions of balls, when that is very difficult to replicate in the real world. If the model's mechanism replicates the phenomenas of the real phenomena in all the relevant respects, we learn something. However, the model lives and dies by these relations because if the model mechanism produces different phenomena than the real world mechanism, we don't learn anything at all.

If we can find a way to make models that we are confident replicate model behavior, the model could be an incredible tool for experimentation because it is nearly infinitely malleable. To create a controlled experiment, all the experimenter need do is modify some code and compare the modification to the original. It is then quite easy to know that no other changes occurred and that any change in phenomena between the two was *caused* by the change in code. As we saw earlier, real mechanisms can be really difficult to manipulate in a controlled way due to confounding factors like the placebo affect which require complicated experimental procedures to account for. Thus, a simulated model gives the researcher a scalpel to ask questions about a mechanism in the real world, potentially identifying areas of interest and trying out interventions virtually before committing to the development of a potentially complicated and expensive real-world intervention.

Learning about Mechanism with Mechanism

So now we are left with mechanistic models as malleable, but questionably accurate surrogates for a real-world phenomena. This section proposes vocabulary for talking about the relationships between a model's mechanism and and the target's mechanism. In later sections, I will substantiate that relations between mechanisms are a key way of establishing that a model's phenomena matches that of the targets while maintaining the a model that is manipulable to allow for simulated experimentation.

Using similar notation as for the model mechanism, we can denote the target world as the set $W = \{w_1, \ldots, w_n\}$. As discussed in the context of model targets in general, W as a target system, may represent the actual world or a hypothetical mechanism of the modeler's choosing. As a model target, W is picked by the modeler and the model M designed in relation to it.

Say again that the modeler wants to model the mechanisms at play for a potential experiment dumping one thousand bouncy balls into an elevator to determine how the bouncing balls affect the center of mass of the elevator. The target mechanism could be decomposed into the parts $W = \{w_b, w_f\}$ which correspond to the balls and the floor respectively. The two are distinct, yet do interact with one another to set where the center of mass of the elevator ends up. Now, to model this we can use the mechanistic model M from above which consists of the parts m_b, m_f and m_o which correspond to simulated balls, simulated floor and walls and orchestration code which ties everything together. Note that m_b and m_f seem to model virtual versions of w_b and w_f in the target mechanism, where m_o seems to represent part of the model mechanism that supports the overall model but does not have a direct counterpart in the target.

Now, take a step back and consider what it would take for M to faithfully recreate the relevant aspects of W. This requires the phenomena that M creates be similar in all ways in which the modeler measured both M and W. Ideally, an observer could not tell the difference between data from W and data from

M, though in practice we might just hope that they are highly correlated.

To show that such a correspondence exists, there is no substitute for measuring the real world in similar scenarios to what the model is modeling. We'd need to really throw all those balls in the elevator and measure the center of mass to know for sure. But, in many cases, that puts the cart before the horse in the sense that the model is supposed to be a way of avoiding real experimentation which might be impossible or difficult. So now consider what it might look like to leverage knowledge about the parts of a model to try to ensure the mechanism and its internal relations model the target as much as possible.

For a model's mechanism to faithfully represent a target mechanism, that model must relate to the target by some relation R which is defined over $\mathcal{M} \times \mathcal{W}$, where $\mathcal{M}$ is all possible models and $\mathcal{W}$ is all possible targets. Now, for some model M and target world W, we find $R(W, M)$ if:

1. All parts of the target $w_i \in W$ have counterparts $m_i \in M$ in the model.
2. The salient causal relations between the w_is must be replicated between the m_i
3. The organization of the mechanism of M must be similar to realistic organizations of W.

If $R(W, M)$ holds, I'll argue we have a pretty good idea that M is a good stand in for W and that while this relation is not a sure thing, it is useful when true experimentation is difficult, expensive or unethical.

Once such a relation is established, a researcher can run experiments modifying individual properties of individual parts of the mechanism M to easily form controlled experiments. If we replace m_b with a bouncier virtual ball m'_b, we can be pretty sure that the results would be reflective of when the target system is modified correspondingly to include bouncier balls w'_b. Furthermore, this is a causal link because of the same possible worlds argument given above in support of controlled experimentation. If all else is constant, which in a computational model we know for sure, then the change must be responsible for any change in output. If that causal link holds in the model and we have good reason to believe its mechanisms mirror those of the target, we get a decent signal that the target does in fact behave like that.

Conclusion

This section frames the problem of computational representational modeling in mechanistic terms and presents a potential way mechanism in models can help generate understanding in a causal sense in the real-world targets of models. In the coming sections, I'll move back to Zollman's models to show how Zollman takes a mechanistic approach to modeling, then use this mechanistic model framework to highlight the strengths and weaknesses of his models in generating understanding about the target.

"About the NWS." https://www.weather.gov/about/, n.d.

Bechtel, William, and Adele Abrahamsen. "Explanation: A Mechanist Alternative." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 36, no. 2 (June 2005): 421–41. https://doi.org/10.1016/j.shpsc.2005.03.010.

Craver, Carl, and James Tabery. "Mechanisms in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University, 2019.

"Explainable Artificial Intelligence (XAI) Program Update." DARPA, November 2017.

Frigg, Roman, and Stephan Hartmann. "Models in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2018. Metaphysics Research Lab, Stanford University, 2018.

Hempel, Carl G. *Aspects of Scientific Explanation, and Other Essays in the Philosophy of Science*. New York, Free Press, 1965.

Hitchcock, Christopher. "Causal Models." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2019. Metaphysics Research Lab, Stanford University, 2019.

Jonas, Eric, and Konrad Paul Kording. "Could a Neuroscientist Understand a Microprocessor?" *PLOS Computational Biology* 13, no. 1 (January 2017): e1005268. https://doi.org/10.1371/journal.pcbi.1005268.

Kohler, Monica D., Elizabeth S. Cochran, Doug Given, Steve Guiwits, Doug Neuhauser, Ivan Henson, Renate Hartog, et al. "Earthquake Early Warning ShakeAlert System: West Coast Wide Production Prototype." *Seismological Research Letters* 89, no. 1 (January 2018): 99–107. https://doi.org/10.1785/0220170140.

Lewis, David. "Causation." *The Journal of Philosophy* 70, no. 17 (October 1973): 556. https://doi.org/10.2307/2025310.

Li, Qun, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, et al. "Early Transmission Dynamics in Wuhan, China, of Novel CoronavirusInfected Pneumonia." *New England Journal of Medicine* 0, no. 0 (January 2020): null. https://doi.org/10.1056/NEJMoa2001316.

Salmon, Wesley C. "Statistical Explanation." In *Statistical Explanation and Statistical Relevance*. Pitt Paperback, Vol. 69. [Pittsburgh]: University of Pittsburgh Press, 1971.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, et al. "Mastering the Game of Go Without Human Knowledge." *Nature* 550, no. 7676 (October 2017): 354–59. https://doi.org/10.1038/nature24270.

Winsberg, Eric. "Computer Simulations in Science." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2019. Metaphysics

Research Lab, Stanford University, 2019.

Woodward, James. "Scientific Explanation." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2019. Metaphysics Research Lab, Stanford University, 2019.