Using Computer Simulations for Hypothesis-Testing and Prediction: Epistemological Strategies

Tan Nguyen

Abstract

This paper explores the epistemological challenges in using computer simulations for two distinct goals: explanation via hypothesis-testing and prediction. It argues that each goal requires different strategies for justifying inferences drawn from simulation results due to different practical and conceptual constraints. The paper identifies unique and shared strategies researchers employ to increase confidence in their inferences for each goal. For explanation via hypothesis-testing, researchers need to address the underdetermination, interpretability, and attribution challenges. In prediction, the emphasis is on the model's ability to generalize across multiple domains. Shared strategies researchers employ to increase confidence in inferences are empirical corroboration of theoretical assumptions and adequacy of computational operationalizations, and this paper argues that these are necessary for explanation via hypothesis-testing but not for prediction. This paper emphasizes the need for a nuanced approach to the epistemology of computer simulation, given the diverse applications of computer simulation in scientific research. Understanding these differences is crucial for both researchers and philosophers of science, as it helps develop appropriate methodologies and criteria for assessing the trustworthiness of computer simulation.

Introduction

As computer simulation methods become increasingly common across various fields, concerns about their ability to generate trustworthy knowledge also increase. The crucial question is whether the inferences from a specific computer simulation are warranted, considering their intended application. Simulations can inform us about the expected behavior of real-world systems under certain conditions, serving as tools for prediction. For example, if a simulation is employed to predict climate variables, does it forecast the temperature or precipitation with sufficient accuracy? Alternatively, simulations can help us understand systems and their behaviors. When data about a system's behavior is already available, computer simulations can address questions about what could have occurred to produce the observations. For example, if a simulation of human working memory is utilized to study potential neural mechanisms underlying working memory, how can we trust that the simulated results—that one neural mechanism of working memory is more likely than another neural mechanisms? In short, how can researchers justify inferences drawn from simulations for their intended purpose? (Winsberg, 2010)

Computer simulations can be used for various purposes, and they can be generally categorized into two categories: predicting the behavior of a system (prediction) or generating explanations for the system's behavior by evaluating multiple hypotheses about what could have occurred (explanation via hypothesis testing). This paper argues that the distinct goals of computer simulations, namely explanation via hypothesis testing and prediction (Breiman, 2001; Chirimuuta, 2021; Shmueli, 2010), require separate epistemological approaches to address the specific practical and conceptual constraints unique to each goal. Specifically, this paper seeks to characterize how different goals of computer simulations necessitate distinct strategies that researchers employ to justify their inferences.

Winsberg (2001) argued that knowledge produced by computer simulations is the result of inferences that are downward, motley, and autonomous, and that an account of epistemology of computer simulations should consider these features. Downward inference means that the starting point of computer simulations are well-established scientific theories, which subsequently justify conclusions about real-world systems from simulation results. Similarly, the third strategy to justify inferences from simulations identified by Parker (2008) requires that computer simulations are based on well-confirmed theory. For Winsberg and Parker, computer simulations must be based on well-confirmed theory. This paper argues that well-confirmed theory is in fact a necessary feature of using computer simulations for explanation via hypothesis-testing, but it is not a necessary feature of using computer simulations for prediction. Moreover, when researchers operationalize (translate) a theoretical assumption into computational terms, (Parker, 2010) argued that they need to make sure that these computational terms represent target theoretical constructs and processes. This paper argues that while explanation via hypothesis-testing requires high representational quality of computational terms, prediction does not require that computational terms represent theoretical constructs and processes.

In the sections below, I will first describe the form of inference for two goals and give two scientific examples. Then, I will describe shared strategies that researchers use to increase confidence in their inferences in both explanation via hypothesis-testing and prediction, and I will argue that these are necessary for explanation via hypothesis-testing but not for prediction. Then, I will describe unique strategies researchers use to increase their confidence in each context, explanation via hypothesis testing and prediction.

Explanation via Hypothesis-testing versus Prediction

For explanation via hypothesis-testing, the researchers first propose a set of hypotheses, each posits an alternative way about what could have happened to produce the system behavior. The primary goal of explanation via hypothesis-testing is to determine which hypothesis, out of two (or more) alternative hypotheses, is more likely to be true, thereby giving an explanation about the system behavior. The form of inference goes like this: given two opposing hypotheses D1 and D2, if hypothesis D1 is true, assuming background theoretical assumptions A and B are true, then a computational model embodying A, B, D1 should correspond to experimental observations more than a computational model embodying A, B, D2.

In the context of cognitive neuroscience, researchers might be interested in testing two opposing hypotheses about the neural mechanisms underlying working memory. Hypothesis D1 posits that working memory relies on persistent activity in specific neuronal populations, while Hypothesis D2 suggests that working memory is maintained through synaptic changes in connectivity patterns. To test these hypotheses, researchers could develop two computational models that embody the well-established principles of neural computation, such as:

Theoretical Assumption A: Neurons communicate through action potentials (spikes) and use principles of integration and spiking to process information. This assumption highlights the fundamental role of action potentials and the integration of inputs in neuronal communication and information processing.

Theoretical Assumption B: Neural networks exhibit a balance of excitatory and inhibitory activity, which plays a crucial role in maintaining network stability and modulating information processing. This assumption emphasizes the importance of the interplay between excitatory and inhibitory neurons and their role in shaping the overall dynamics of the neural system.

The two models differ in whether working memory is operationalized as sustained neuronal activity, hypothesis D1, or short-term synaptic plasticity, hypothesis D2. The researchers then operationalize these theoretical assumptions and hypotheses into computational terms. Then, they evaluate how well each model corresponds with human decision-making behaviors or neural activity patterns. If one model fits existing behavioral and neural data better than another model, the researchers might conclude that hypothesis D2 is more likely than hypothesis D1. To explain the mechanism of working memory, the researchers then adopt what hypothesis D2 posits: working memory is maintained through synaptic changes in connectivity patterns.

For prediction, the researchers want to anticipate how real-world systems would behave under certain conditions, particularly when observational data is limited or challenging to obtain. The form of inference goes like this: if a computational model embodying theoretical assumptions A, B, C corresponds to observable data in a range of scenarios X, Y, Z, the model's prediction might be what would actually happen in scenario U (where we cannot collect data). In the context of predictive modeling, there are two primary types of predictions that can be made based on different aspects of generalization:

Within-domain extrapolation: Researchers make predictions for scenarios or conditions that extend beyond the range of data that was used to build or validate the model. This type of prediction relies on the assumption that the underlying patterns or relationships observed in the data continue to hold beyond the observed range. For example, to predict temperature, they could develop a climate model that embodies theoretical assumptions A, B, and C (e.g., mechanisms underlying atmospheric circulation, ocean currents, and greenhouse gas concentrations), based on prior research about these physical processes that drive the climate system. The model can then be tested against observable data from a range of regions X, Y, or Z (Northwest region, Pacific Region, African region) where temperature can be measured. If the model corresponds well to the observable data in these scenarios, researchers might use the model to extrapolate and predict the temperature of the Arctic region, assuming that the relationships between the variables learned from the available data would continue to apply in the unobserved Arctic region (Winsberg, 2010). In this case, X represents the temperature data from the northwest and other regions, and U represents the temperature data from the Arctic region.

Cross-domain extrapolation: Researchers make predictions for scenarios or conditions that involve different, but related, variables or domains. This type of prediction relies on the assumption that the model can capture some underlying climate processes that are applicable to different domains. For example, they could develop a climate model that embodies theoretical assumptions A, B, and C (e.g., mechanisms underlying atmospheric circulation, ocean currents, and greenhouse gas concentrations), based on prior research about these physical processes that drive the climate system. The model can then be tested against observable data from a range of domains X, Y, and Z (e.g., temperature, humidity, and sea level changes in the Pacific Northwest) where these climate variables can be measured. The model could be used to predict precipitation in the Pacific Northwest region. In this case, X represents the temperature data from the Pacific Northwest, and U represents the precipitation data from the Pacific Northwest.

In each case, explanation via hypothesis-testing or prediction, what can the researchers do to warrant their inferences? In the sections below, I argue that for explanation via hypothesis-testing to be trustworthy, simulations should meet these criteria: Empirical Corroboration of Theoretical Assumptions; Adequacy of Computational Operationalizations; Controlling for Confounds and Varying Conditions; Addressing Underdetermination and Model Simplicity. For prediction to be trustworthy, simulations should meet these criteria: Within-domain extrapolation; Cross-domain Extrapolation. The researchers can increase the trustworthiness of predictions by satisfying other two criteria, even though they are not necessary: Empirical Corroboration of Theoretical Assumptions; Adequacy of Computational Operationalizations.

Empirical Corroboration of Theoretical Assumptions

In the context of explanation via hypothesis-testing, researchers should corroborate theoretical assumptions A and B with previous findings (Parker, 2010; E. Winsberg, 2009), demonstrating their validity in the current study. For example, researchers can review the literature to find empirical evidence supporting the idea that neurons use certain computational principles, such as integration and spiking, to process information. By doing so, they can confidently incorporate these assumptions into their computational models to test the competing hypotheses D1 and D2. To corroborate Assumption A (neurons communicate through action potentials or spikes), researchers can consult studies that have measured and recorded the action potentials of individual neurons during information processing. For example, they might look into electrophysiological studies that have captured the spiking activity of neurons in various brain regions and under different experimental conditions. These studies provide direct evidence for the role of action potentials in neuronal communication, lending support to Assumption A.

In the context of climate modeling, researchers can apply this strategy by identifying and examining key assumptions about the climate system, such as atmospheric circulation, ocean currents, and greenhouse gas concentrations. They can then corroborate these assumptions with previous empirical findings and observational data to establish their validity in the context of the model. For instance, one theoretical assumption with respect to atmospheric circulation could be the Coriolis effect. The Coriolis effect is a key theoretical assumption that arises from Earth's rotation and has a significant impact on atmospheric circulation. This effect causes moving objects, such as air masses, to be deflected to the right in the Northern Hemisphere and to the left in the Southern Hemisphere. This deflection influences the development of large-scale wind patterns and the formation of high and low-pressure systems, which are essential components of atmospheric circulation. Researchers can review the extensive body of literature and observational data that supports the role of Earth's rotation in shaping global wind patterns and weather systems. This may include examining historical weather records, satellite data,

and findings from previous studies investigating the impact of the Coriolis effect on the formation of cyclones, anticyclones, and trade winds. By gathering empirical evidence that validates the influence of the Coriolis effect on atmospheric circulation, researchers can confidently incorporate this theoretical assumption into their computational climate models.

Empirical corroboration of theoretical assumptions can increase the trustworthiness of both explanation via hypothesis-testing and prediction. This is related to Winsberg’s notion of downward inference: starting from well-established theory. In prediction, theoretical assumptions usually serve as building blocks in many predictive models, and the verity of these assumptions can increase confidence in model’s predictions. However, though it’s necessary for explanation via hypothesis testing, it’s not necessary for prediction. There is a class of modern data-driven predictive models, such as deep neural networks, which totally ignore foundational theorical assumptions in the domain that the model is intended to predict and perform state of the art level of accuracy. This class of data-driven models extracts statistical regularities in the data used to train these models without relying on theoretical assumptions (Leonelli, 2020). In data-driven models, such as deep neural networks, the primary focus lies in the application of robust statistical and mathematical methods to enable accurate extrapolation, rather than relying on well-established theoretical assumptions. Although these theoretical assumptions may still play a role in the modeling process, they primarily serve as inductive biases, which aid the model in efficiently extracting statistical regularities from the available data. Inductive biases, in the context of data-driven models, refer to the inherent assumptions or predispositions that guide the learning process of a model. While data-driven models may not rely heavily on well-established theoretical assumptions like mechanistic or process-based models, inductive biases offer an alternative way to incorporate domain-specific knowledge and improve the model's ability to learn from data. These biases can stem from the model's architecture, choice of activation functions, optimization methods, or even from the initial parameter settings. By incorporating certain inductive biases, a model can be better equipped to identify relevant patterns in the data and generalize more effectively to unseen scenarios. For example, convolutional neural networks (Krizhevsky et al., 2012) (CNNs), a type of deep learning model commonly used for image recognition tasks, incorporate a specific inductive bias inspired by human perception through their architecture. By using convolutional layers and shared weights, CNNs assume spatial invariance, which means that the model can recognize features regardless of their position in the image. Though this inductive bias helps the CNNs achieve better-than-human performance on various image processing tasks, more recent models such as Vision Transformer (Dosovitskiy et al., 2021) (ViT) performs better than CNNs on various image processing tasks. This highlights the fact that inferences drawn from simulations (prediction in this case) are not necessarily downward.

In explanation via hypothesis-testing, researchers aim to determine whether the likelihood of a particular hypothesis, D1, being higher compared to another hypothesis, D2, within the context of the real world. This determination is made based on simulation results, which indicate that the probability of D1 given theoretical assumptions A, B, C, and available data is greater than the probability of D2 given the same assumptions and data, represented as $P(D1 \mid A, B, C, Data) > P(D2 \mid A, B, C, Data)$. For the inference from the conditional probability to real-world mechanisms be justified, that D1 is more likely than D2 in real-world, it is essential that the theoretical assumptions A, B, and C are well-established and closely aligned with the truth. If these assumptions are not close to the truth, then the researchers could at best conclude that D1 is more likely than D2 in a “hypothetical” situation where A, B, C are assumed to be the forces driving the system behavior.

Adequacy of Computational Operationalizations

In order to warrant an explanation via hypothesis testing, the researchers must ensure that the computational operationalizations of theoretical assumptions A and B are adequate. This can be achieved by looking at past successful computational operationalizations of these assumptions in similar contexts (Parker, 2008, p. 200; Sargent, 2013). For example, to validate the computational operationalization of theoretical Assumption B, researchers can refer to research that has successfully modeled the balance between excitatory and inhibitory activity in neural networks, such as studies investigating the role of this balance in shaping oscillatory activity or stabilizing network dynamics. One such example is the study of gamma oscillations, which are rhythmic neural activity patterns in the frequency range of 30-80 Hz, thought to play a role in various cognitive processes. The researchers can use studies building computational models that have demonstrated that the interplay between excitatory and inhibitory neurons can give rise to these oscillations, with the balance of excitation and inhibition being crucial for maintaining the oscillatory activity. By examining these studies, researchers can validate the computational operationalization of theoretical Assumption B in their models, and gain confidence in the ability of their models to capture the essential dynamics of neural systems when testing their hypotheses about working memory.

Furthermore, they can consider the similarity between a computational operationalization and a theoretical assumption (Oreskes et al., 1994; Parke, 2014; Weisberg, 2013). Let's consider the

computational operationalization of theoretical assumption A that neurons communicate through action potentials (spikes) and use principles of integration and spiking to process information. In a computational model, this theoretical assumption can be operationalized using a spiking neuron model, such as the leaky integrate-and-fire (LIF). Researchers can establish the similarity between a computational operationalization, such as the leaky integrate-and-fire (LIF) model, and a theoretical assumption, such as neurons communicating through action potentials via principles of integration and spiking, by following these steps. First, they identify the key features of the theoretical assumption, which include integration of input action potentials, the generation of action potentials at a threshold, and the propagation of these action potentials along neural pathways. Second, they analyze the computational operationalization, like the LIF model, which simulates a neuron's membrane potential integrating input, leaking voltage over time, and firing a spike upon reaching a threshold. Third, they compare the LIF model's representation of the basic properties of the theoretical assumption (threshold-based firing, integration, and leak), while acknowledging the simplification of certain biophysical mechanisms.

To further validate the computational operationalization, researchers can assess the LIF model's behavior in various scenarios to verify if it accurately represents spiking activity observed in real neurons (Parker, 2010; Sargent, 2013). If the LIF model generates action potentials in response to input and exhibits similar firing patterns to biological neurons, it supports the similarity between the operationalization and the assumption. Additionally, they can evaluate the model's robustness and sensitivity to changes in parameters (e.g., membrane time constant, threshold, and input strength) to ensure that it maintains similarity to the theoretical assumption across various conditions. Finally, they can seek external validation by comparing the LIF model's spiking activity with empirical data from electrophysiological recordings or other well-established spiking neuron models. Following these steps allows researchers to establish the similarity between the computational operationalization (e.g., LIF model) and the theoretical assumption (neurons communicate through action potentials) in the context of neural computing.

To apply the strategy of assessing the adequacy of computational operationalizations in the context of prediction, researchers can use strategies described above: historical successes of the same computational operationalization, similarity between the computational operationalization and the theoretical assumption, and the correspondence between computational simulation with empirical data.

For the climate modeling example, researchers can look at past successful implementations of the Coriolis effect in other climate models to validate the adequacy of their computational operationalization. They can review the literature to find examples of models that have accurately simulated atmospheric circulation, large-scale weather systems, and wind patterns influenced by the Coriolis effect. By analyzing these historical successes, researchers can learn from the established methods and apply similar techniques to their own climate model.

In order to validate the computational operationalization of the Coriolis effect, researchers should ensure that the mathematical representation of the Coriolis force in their model captures the essential aspects of the theoretical assumption. They can first identify the key feature of the theoretical assumption, which is the deflection of air masses due to Earth's rotation. Then, they can ensure that the mathematical representation of the Coriolis force in their climate model captures the essential aspects of the theoretical assumption. If the model's equations capture the fundamental relationship between Earth's rotation and the deflection of air masses, it supports the similarity between the computational operationalization and the theoretical assumption.

To further validate the computational operationalization of the Coriolis effect, researchers can assess the model's behavior in various scenarios to verify if it accurately represents the influence of Earth's rotation on atmospheric circulation observed in real-world data. They can compare the model’s wind patterns output, which is influenced by deflected air masses caused by Earth’s rotation, with observational data, such as historical weather records, satellite data, and other measurements of wind patterns. If the model can reproduce the observed large-scale wind patterns that are influenced by the Coriolis effect, it supports the correspondence between the computational operationalization and the theoretical assumption. Additionally, researchers can evaluate the model's robustness and sensitivity to changes in parameters related to Earth's rotation, such as the rotation rate and the latitude of the simulated region. This helps to ensure that the model's representation of the Coriolis effect is robust and behaves as expected under various conditions.

Adequacy of computational operationalizations can increase the trustworthiness of both explanation via hypothesis-testing and prediction. However, the standards are different. In the prediction scenario, one would operationalize (translate) 3 theoretical assumptions (A, B, C) into computational terms. In the explanation via hypothesis-testing scenario, one would operationalize (translate) 5 theoretical assumptions (A, B, C, D1, D2) into computational terms. To what degree should these computational components in the model represent the theoretical processes? I argue that the computational operationalizations should meet the above criteria in explanation via hypothesis-testing, but the computational operationalizations do not necessarily need to meet the above criteria in prediction.

In prediction, the reason computational terms representing A, B, C does not need to bear similarity with corresponding theoretical assumptions or need to correspond to observational data is that the researchers can build an ensemble of models, each operationalizing these theoretical assumptions differently, and rely on the average of predictions from these models. This ensemble methods can make model’s prediction more robust (Batterman, 2002; Knutti & Sedláček, 2013). Moreover, as discussed above, with data-driven models, the emphasis is on statistical and mathematical methods to ensure robust extrapolation, and researchers do not need to rely on well-established theoretical assumptions.

Though, these theoretical assumptions can serve as inductive biases, helping the model to extract statistical regularities in data more efficiently.

In explanation via hypothesis testing, researchers are assessing the conditional probability of a model representing hypothesis D1 (or D2) given the data and background theoretical assumptions A, B, C: whether P(D1 | A, B, C, Data) and P(D2 | A, B, C, Data). Because the scientific conclusion concerns whether D1 or D2 is more likely in the real world, the researchers must demonstrate that the computational terms standing for D1 and D2 are good representations of hypotheses D1 and D2. Otherwise, the inference from conditional probability to real-world mechanisms (D1 or D2 being more likely) is not warranted. It is also necessary that the computational operationalizations standing for A, B, and C are good representations of theoretical assumptions A, B, C. If these computational operationalizations do not represent theoretical assumptions well, then the researchers cannot conclude that given theoretical assumptions A, B, C, and the data, D1 is more likely than D2 from the conditional probability.

Criteria specific to explanation via hypothesis testing

Controlling for Confounds and Varying Conditions

In order to warrant the inference from conditional probability P(D1 | A1, B, C, Data) > P(D2 | A1, B, C, Data) to the conclusion that D1 is more likely than D2 in the real-world, researchers need to establish that differences between the conditional probabilities are attributable to the differences between D1 and D2 (Parker, 2008; Platt, 1964; Zuidema et al., 2020). Researchers should ensure that both models are built upon the same principles and share the same theoretical assumptions, except for the specific components that represent D1 and D2. Moreover, researchers can test if differences in model performance consistently arise from the differences between D1 and D2. Models often require initializations of parameters, representing initial conditions of a target real-world system. By performing model comparisons under various initial conditions, researchers can ensure that the inference holds across multiple conditions. In other words, hypothesis D1 is more likely than D2 across multiple conditions. For example, researchers might test models of working memory with various types of stimuli or levels of difficulty, to ensure that the differences in model performance are not due to factors unrelated to the hypotheses being tested.

Addressing Underdetermination and Interpretability: Model Simplicity

In some cases, the available empirical evidence may not be sufficient to fully distinguish between competing mechanistic hypotheses, such as D1 and D2, and researchers might need to rely on other criteria such as model simplicity to select the most favorable model (Huemer, 2009; Kieseppä, 1997). This is a well-known problem of underdetermination, tracing back to Duhem and Quine. In some cases, researchers might choose the most parsimonious model, which makes the fewest assumptions or has the least complexity, to favor one hypothesis over another. For example, when comparing two models with similar explanatory power, but one requires fewer parameters or relies on more straightforward mechanisms, researchers might prefer the simpler model as it is more likely to generalize to other scenarios and is less prone to overfitting.

Moreover, it is crucial to ensure that any differences observed between the models can be attributed to the differences between D1 and D2, rather than other factors or assumptions that may have been introduced during the modeling process. Model simplicity plays a critical role in this context, as it allows for more transparent and interpretable results (Myung & Pitt, 1997). Simple models, with fewer parameters and assumptions, are easier to analyze and understand, making it more straightforward to attribute any observed differences in performance to the competing hypotheses themselves. When models are more complex, they may contain numerous interacting components, making it challenging to determine which specific element is responsible for the differences in model performance. In such cases, it becomes increasingly difficult to attribute the performance disparity to D1 and D2, and it may be unclear whether the observed differences result from the core hypotheses or from extraneous factors introduced by the complexity of the models. By maintaining simplicity, researchers can more confidently establish a connection between the model's performance and the underlying hypotheses. This helps to ensure that the conclusions drawn from comparing the models accurately reflect the likelihood of the competing hypotheses.

In prediction scenarios, while simplicity sometimes leads to more accurate or robust models (Geman et al., 1992), more complex models with enough data to estimate their parameters often outperform simpler models (Breiman, 2001).

Criteria specific to prediction

Within-domain extrapolation

In the example of using a climate model to predict the temperature of the Arctic region, researchers can apply the following strategies to ensure that the model's within-domain extrapolation is trustworthy:

Cross-Validate the model in the range of data: To ensure that the model is well-validated within the range of available data, researchers can compare model predictions of temperature in the Northwest, Pacific, and African regions to observed temperature data. If the model accurately reproduces the known data patterns in these regions, it provides evidence that the model is likely to be reliable when extrapolating to the Arctic region.

Identify boundary conditions: Researchers should determine the limits of the model's applicability by identifying the boundary conditions, such as the range of latitudes, temperatures, or other climatic conditions for which the model is expected to provide accurate predictions. For example, a model may be less reliable when simulating extreme climate events or when predicting climate variables at very high or low latitudes. This can help researchers avoid extrapolating beyond the limits where the model's performance is well-established and ensure that the extrapolation to the Arctic region falls within the model's reliable range.

Cross-domain Extrapolation

In the example of using a climate model to predict precipitation in the Pacific Northwest region based on temperature, humidity, and sea level changes, researchers can apply the following strategies to ensure that the model's cross-domain extrapolation is trustworthy:

Analyze commonalities and differences between domains: Researchers should identify the underlying structures, relationships, or processes that are shared across domains (e.g., the role of atmospheric circulation in both temperature and precipitation patterns) as well as the unique features that differentiate them (e.g., the specific mechanisms driving precipitation versus temperature changes). Understanding these similarities and differences can help researchers adapt and refine the model for cross-domain generalizability, making it more suitable for predicting precipitation based on temperature data. One strategy is to ensure that the climate model is built based on theoretical assumptions and principles (e.g., mechanisms underlying atmospheric circulation, ocean currents, and greenhouse gas concentrations) that are applicable to multiple domains, such as temperature and precipitation. This increases confidence in the model's ability to capture underlying climate processes that are applicable to different domains.

Validate the model across multiple domains: To validate the model's performance across different domains, researchers can compare model predictions of temperature, humidity, and sea level changes in the Pacific Northwest to observed data in these domains. If the model accurately reproduces the known patterns across multiple domains, it can provide evidence of the model's ability to generalize across domains, and thus increase confidence in its predictions of precipitation in the Pacific Northwest.

Conclusion: Contrast the two goals of computer simulations

As argued, though empirical corroboration of theoretical assumptions and adequacy of computational operationalizations can increase the trustworthiness of predictions drawn from computer simulations, they are not necessary. However, these two criteria are necessary to explanation via hypothesis-testing.

Moreover, due to its goal, in the explanation via hypothesis-testing scenario the researchers must take measures to make sure that differences between two models embodying D1 or D2 can be attributable to the differences between D1 and D2. This criterion prompts researchers to adopt more parsimonious models, focusing on the core differences between the hypotheses, and avoiding unnecessary complexity that could make interpreting the results more challenging. Furthermore, explanation via hypothesis-testing often suffers from the problem of underdetermination in practice. This practical constraint also prompts researchers to rely on other criteria such as model complexity in addition to its degree of correspondence with empirical data.

Whereas, in the prediction-making scenario, the scientist is essentially making an extrapolation when they make predictions, and for us to believe in extrapolations, the model should predict well in a wide range of scenarios. For example, the degree to which we can believe in a predicted temperature from a climate model depends on how well the model's temperature predictions have been confirmed in the past. If the model is making predictions across multiple domains, researchers need to understand shared processes (e.g., atmospheric circulation's impact on temperature and precipitation) and distinct mechanisms (e.g., those driving precipitation and temperature changes) to build a model that could generalize across domains.

In conclusion, the distinct goals of computer simulations, specifically explanation via hypothesis-testing and prediction, necessitate separate epistemological strategies to address the unique practical and conceptual constraints inherent to each goal. The paper identifies strategies applied in each context and how explanation via hypothesis-testing or prediction requires unique strategies to justify inferences drawn from simulation results. Moreover, even with shared strategies (empirical corroboration of theoretical assumptions and adequacy of computational operationalizations), standards are different.

Ultimately, this paper emphasizes the importance of a nuanced and goal-sensitive approach to the epistemology of computer simulations. The fact that there is not a comprehensive epistemology of computer simulations account in the literature might reflect the multiple purposes of computer simulations, and each account in the literature implicitly addresses a different goal of computer simulations. As computer simulations continue to play an increasingly central role in scientific research across various disciplines, it is crucial for researchers to be aware of the distinct epistemological challenges and considerations associated with different modeling goals. For researchers, recognizing the differences between these two goals is essential for the development and evaluation of computational models, as it enables researchers to adopt appropriate methodologies and criteria for assessing the trustworthiness and appropriateness of their models for the intended purpose. For philosophers of science, recognizing the differences between explanation via hypothesis-testing and prediction is crucial to develop a comprehensive account of epistemology of computer simulations.

Annotated Bibliography

Batterman, R. W. (2002). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press.____Robert W. Batterman emphasizes the importance of stable phenomena and robust models in scientific inquiry. He demonstrates how asymptotic reasoning enables scientists to identify and understand the stability and robustness of these models, leading to a deeper comprehension of complex systems and the emergence of stable phenomena across various scientific disciplines.

Breiman, L. (2001). Statistical Modeling: The Two Cultures. *THE TWO CULTURES*. __Leo Breiman compares and contrasts two approaches to statistical modeling: data models and algorithmic models. He argues that while traditional statistical models focus on making assumptions about data and fitting models, algorithmic models prioritize prediction accuracy while having no assumptions about data generation processes.

Chirimuuta, M. (2021). Prediction versus understanding in computationally enhanced neuroscience. *Synthese*, *199*(1–2), 767–790. https://doi.org/10.1007/s11229-020-02713-0 __Chirimuuta's article explores the epistemic implications of using machine learning, specifically artificial neural networks (ANNs), in neuroscience as opposed to traditional models. She uses the literature on model intelligibility to establish benchmarks for the interpretability of ANNs and examines case studies on motor cortex and visual system modeling. Chirimuuta argues that there is a trade-off between predictive accuracy and the level of understanding these models provide, which is better explained by a non-factivist account of scientific understanding.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, *521*(7553), 452–459. https://doi.org/10.1038/nature14541 __Zoubin Ghahramani presents an overview of probabilistic approaches to machine learning and AI. He emphasizes the importance of incorporating uncertainty into these methods and demonstrates how probabilistic models can lead to more robust and adaptable systems, paving the way for advances in artificial intelligence research.

Knutti, R., & Sedláček, J. (2013). Robustness and uncertainties in the new CMIP5 climate model projections. *Nature Climate Change*, *3*(4), 369–373. https://doi.org/10.1038/nclimate1716 __Knutti and Sedláček assess the robustness and uncertainty of the new generation of climate model projections from the Coupled Model Intercomparison Project Phase 5 (CMIP5). They analyze the consistency of the projections, discuss sources of uncertainties, and emphasize the importance of understanding these factors for making informed decisions about climate change mitigation and adaptation strategies.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam’s razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*(1), 79–95. https://doi.org/10.3758/BF03210778 __Myung and Pitt explore the application of Occam's razor, a principle of simplicity, in cognitive modeling through a Bayesian framework. They argue that this approach allows for more effective model selection and evaluation, leading to a deeper understanding of cognitive processes while maintaining a balance between model complexity and explanatory power.

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641 __Oreskes, Shrader-Frechette, and Belitz discuss the challenges and complexities associated with verifying, validating, and confirming numerical models in earth sciences. They highlight the importance of understanding the limits and uncertainties of these models and emphasize the need for a rigorous and transparent evaluation process to ensure their reliability in scientific research and policymaking.

Parke, E. C. (2014). Experiments, Simulations, and Epistemic Privilege. *Philosophy of Science*, *81*(4), 516–536. https://doi.org/10.1086/677956_____Elizabeth C. Parke examines the relationship between

computer simulations and their target systems. Focusing on the ontological similarity between the two, she explores how simulations can in some cases possess higher epistemic privilege compared to experiments.

Parker, W. S. (2008). Franklin, Holmes, and the Epistemology of Computer Simulation. *International Studies in the Philosophy of Science*, *22*(2), 165–183. https://doi.org/10.1080/02698590802496722_______Wendy S. Parker proposed one approach to epistemology of computer simulation, developing analogies between computer simulations and experiments in order to draw on recent work in the epistemology of experiment. Franklin (1986, 1989) identifies numerous apparatus checks and assessments of experimental results that scientists perform to increase confidence in their results, offering an epistemology of experiment. Parker (2008a) demonstrated that many if not all of Franklin’s confidence-building strategies do have straightforward analogues in the context of computer simulation, offering an epistemology of computer simulation.The strategies are as below:_1. Franklin's strategy of “apparatus gives results that match known results” aims to build confidence in experimental results by showing that the experimental apparatus used in obtaining the results gives accurate results in other relevant instances. In the context of model evaluation, for a climate model, confidence in the predictions of yearly rainfall will be increased if the model can simulate the past evolution of yearly rainfall amounts in the region accurately._2. Franklin's strategy of “apparatus responds as expected following interventions” involves demonstrating that the response of the experimental apparatus following an intervention on the experimental system is as expected. In the context of model evaluation, for instance, one might check whether decreasing the value of the variable denoting friction in the simulation model does affect the values taken by other variables in such a way that the calculated efficiency of the system is increased._3. Franklin's strategy of “apparatus based on well-confirmed theory” states that if the theoretical principles used in designing an experimental apparatus are themselves well-confirmed, then this increases confidence in the apparatus and the observations made with the apparatus during an experiment.. In the context of model evaluation, there is often an attempt to increase confidence by pointing out that key modelling assumptions come from accepted scientific theories. _4.Franklin's strategy of “independent confirmation of results” involves showing that the results of an experiment can be replicated in other experiments using different apparatus. In the context of model evaluation, this strategy involves showing that a simulation result closely matches a result generated in another study addressing the same question about the target system._5. The fifth strategy proposed by Franklin, the “Sherlock Holmes” strategy, aims to increase confidence in experimental results by eliminating plausible sources of error (by ruling out any potential issues with the experimental apparatus) and alternative explanations of the results. In the context of computer simulation studies, this strategy involves ruling out sources of error such as programming mistakes or numerical instability.

Parker, W. S. (2010). Computer Simulation. In *The Routledge Companion to Philosophy of Science*. Routledge. https://doi.org/10.4324/9780203744857.ch13 __Wendy S. Parker provides an overview of the role of computer simulations in modern scientific inquiry. She discusses how models are validated (in the context of verification and validation), by comparing model's output with observations or relying on historical successes of similar methods.

Platt, J. R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*(3642), 347–353.

https://doi.org/10.1126/science.146.3642.347 __John R. Platt advocates for the adoption of strong inference, a methodical and hypothesis-driven approach to scientific investigation. He argues that employing this systematic method can lead to more rapid progress in scientific understanding and discovery, compared to less structured approaches.

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, *7*(1), 12–24. https://doi.org/10.1057/jos.2012.20 ___Robert G. Sargent explores the crucial processes of verifying and validating simulation models to ensure their accuracy and reliability. He discusses various methodologies, techniques, and best practices for evaluating and testing the models.

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3). https://doi.org/10.1214/10-STS330 ___Galit Shmueli discusses the distinction between explanatory and predictive modeling in statistical science. Predictive modeling can be defined as the application of a statistical model or data mining algorithm to data with the objective of forecasting new or future observations. This definition specifically focuses on nonstochastic prediction, in which the aim is to predict the output value (Y) for new observations based on their input values (X). Temporal forecasting is also included, where observations up to time t (the input) are utilized to forecast future values at time t + k, where k > 0 (the output). Predictions can take the form of point or interval predictions, prediction regions, predictive distributions, or rankings of new observations. A predictive model refers to any method that generates predictions, regardless of the underlying approach used: theory-based or data-driven methods. Explanatory modeling refers to the application of statistical models to data for the purpose of testing causal hypotheses about theoretical constructs.

Weisberg, M. (2013). *Simulation and Similarity_ Using Models to Understand the World*. Oxford University Press. __In this book, Weisberg introduced the idea of weighted feature matching. Michael Weisberg's idea of weighted feature matching is an approach to evaluate how scientific models can be used to represent real-world target systems. According to Weisberg, a model represents a target system when there is a significant degree of similarity between the two, particularly in the relevant features for a given context or purpose. Weighted feature matching posits that not all features of a model or target system are equally important for establishing their similarity. Some features are more significant and carry more weight in determining the degree of similarity. These weighted features are context-dependent and depend on the scientific goals or purposes for which the model is being used. In other words, a model's ability to accurately represent a target system relies on the degree of match between the most important, or heavily weighted, features of both the model and the target system. By focusing on these weighted features, scientists can better understand how well a model captures the essential aspects of the target system and assess the model's usefulness for addressing specific research questions or purposes.

Winsberg, E. (2001). Simulations, Models, and Theories: Complex Physical Systems and Their Representations. *Philosophy of Science*, *68*(3), S442–S454.___ _____Eric Winsberg explores the relationship between computer simulations, models, and scientific theories. He investigates the unique aspects of inferences drawn from simulations, focusing on their downward, motley, and autonomous features, and discusses the implications for epistemology of computer simulations.__Downward: The concept of “downward” in the context of computer simulation models refers to the fact that established scientific theories often serve as the foundation for creating these models and play a crucial role in

justifying inferences made from simulation results to real-world target systems._Motley: The term “motley” highlights that simulation results typically depend on various model components and resources, not just theory. These components can include parameterizations, numerical solution methods, mathematical tricks, approximations, idealizations, fictions, ad hoc assumptions, function libraries, compilers, computer hardware, and most importantly, the extensive trial and error involved in the development process._Autonomous: The “autonomous” nature of knowledge produced by computer simulations means that it cannot be entirely validated through comparison with observations. Simulations are often used to study phenomena where data is scarce, and in such cases, they are intended to replace experiments and observations as data sources. This is because the relevant experiments or observations may be inaccessible due to principled, practical, or ethical reasons.

Winsberg, E. (2009). *A Tale of Two Methods*. ____This article explores the essential differences between simulations (both digital and analog) and experiments, despite their shared features. The author examines two proposals: one suggesting that experiments directly investigate nature while simulations only investigate models, and another stating that simulations involve manipulating objects with only formal similarity to their targets. The author rejects both proposals and argues that the fundamental distinction between simulations and experiments lies in the background knowledge used to support the “external validity” of the investigation.

Winsberg, E. B. (2010). *Science in the age of computer simulation*. The University of Chicago Press. In “Science in the Age of Computer Simulation,” ___Eric Winsberg explores the impact of computer simulations on the scientific process and methodology. He delves into the role of simulations as a tool for gaining scientific understanding and discusses the philosophical implications and challenges they bring to the traditional view of experimentation and observation.

Zuidema, W., French, R. M., Alhama, R. G., Ellis, K., O’Donnell, T. J., Sainburg, T., & Gentner, T. Q. (2020). Five Ways in Which Computational Modeling Can Help Advance Cognitive Science: Lessons From Artificial Grammar Learning. *Topics in Cognitive Science*, *12*(3), 925–941. https://doi.org/10.1111/tops.12474___In this article, the authors discuss the potential of computational modeling to advance cognitive science, using artificial grammar learning as a case study. They argue that computational techniques can significantly enhance the design, implementation, and analysis of experiments, ultimately elevating the overall research quality. The paper provides five concrete examples demonstrating how computational modeling can be used to (a) formalize and clarify theories, (b) generate stimuli, (c) visualize data, (d) perform model selection, and (e) explore the hypothesis space in the field of cognitive science.