The Pitfalls of Predictive Modeling: Investigating How Inaccurate Models Can be Useful Through the Lens of Data Relativity

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Colin Crowe

Fall 2022

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Kathryn A. Neeley, Associate Professor of STS, Department of Engineering and Society

Introduction

Computer modeling as a field involves creating a better understanding of complex systems through the use of computer-generated simulations. Such simulations have always been relevant across domains, but the global outbreak of COVID-19 gave computer models additional importance as a means of gauging the impact of the pandemic and informing possible responses. Even at the University of Virginia, the Biocomplexity Institute uses predictive models to create weekly briefings for both state and federal governments so policymakers are informed on the possible ways the situation could progress (Biocomplexity Institute, Homepage). Clearly, computer models are not just useful tools when it comes to solving theoretical problems, but also for creating real-world policy as well.

There are limitations to the power of predictive modeling, however. Obviously, an inaccurate model wouldn't be very useful for creating good policy, but determining whether or not a model is accurate can be challenging. Published about half a year into the COVID-19 pandemic, one illuminating case study came out of a team at the University of Michigan analyzing the pandemic response in India. In short, the Indian government used predictive models to determine how much of a response was needed. As the pandemic progressed, however, they found that they had vastly underestimated how bad the situation would actually be, and did not have adequate supplies to meet what was demanded. This team of researchers concluded that the issue here had nothing to do with the models technically, but rather the data they were being built on top of. They pointed to the massive underreporting of cases causing the early COVID data to skew below its true values, concluding that "[insufficient data] limits modelers' ability to predict the course of the pandemic, gauge its impact, and estimate health care resource needs—including oxygen supplies and hospital beds" (Zimmermann, 2021, p. 560).

Computer models are obviously a powerful tool, as the discussion of the biocomplexity institute shows. However, as the discussion about India’s pandemic response shows, their results cannot be taken at face value because there are myriad reasons why their predictions might be skewed. So, if models cannot be taken at face value, then there must be some way we can interpret their results to avoid the same situation Zimmermann et al. described. The goals of my research are therefore to determine how exactly the knowledge gained from computer models should be best put to use, why some models succeed in providing useful insight for good real-world action, and why others fail to accomplish those same goals.

Background

Computer modelers of all fields have already grappled with these kinds of questions, and so it is only reasonable to draw upon that wealth of knowledge here. This section will be devoted to exploring pre-existing arguments and establishing what the current literature has to say.

i. Subjectivity in the Model-Making Process Leads to Inaccurate Models

It might seem simple to state that a useful model is one that accurately reflects the system it models. One might even go as far as to say that, if only those researchers in India had considered double-checking their data and accounted for it, their model would have been more accurate, the right amount of supplies might have been ordered, and people’s lives would have been saved. Trying to maximize accuracy as a means of maximizing usefulness may be a reasonable solution, but it is not as achievable as it sounds. Namely, creating a perfectly accurate model is difficult due to the process of “parameterization”, which involves taking some real-world phenomenon and transforming it into a series of numbers and equations that a computer can work with.

This process of parameterization has the result of injecting an element of subjectivity into model creation. Parameterization involves simplifying the thing being modeled as much as possible; after all, if all the complexities could be included, it would no longer be a model, it would simply be the actual thing. As Paul Edwards states in his book about climate modeling, “a parameter is kind of a proxy – a stand-in for something that cannot be modeled directly but can still be estimated, or at least guessed” (Edwards, 2010, p. 338). The “at least guessed” part of that is where issues arise, because proving that those guesses are reasonable is not an easy task. Evidence can be given to justify making such an assumption, of course, but it still ultimately falls upon the subjective discretion of the modeler whether or not they wish to make those assumptions.

The fact that model creation involves some level of guesswork or subjectivity can cause problems, as the assumptions modelers make can have huge effects on the accuracy of their models. Systems of human behavior serve as illuminating examples. One paper published during the 2019 Winter Simulation Conference sought to investigate the degree to which assumptions about human behavior impacted the results of computer models. To do this, they constructed an experiment in which they built several simulations of online dating under several different assumptions, and tested them each to see if a smaller pool of available partners led to more couples forming. The results they found were that some models supported the claim that a smaller pool led to more couples, while other models supported the exact opposite claim; a smaller pool was observed to lead to fewer couples (Turner, 2019).

In sum, creating computer models requires making assumptions and guesses about how the underlying system works, and those assumptions and guesses can hold huge sway over the end results of the model. This greatly complicates the notion that careful construction of computer models can avoid inaccuracies. Inaccuracies are baked into the creation process directly; we must therefore look to some other method if we want to guarantee accuracy.

ii. Several Techniques Exist to Evaluate Model Accuracy

At this point, it has been established that guaranteeing accuracy when making models is not possible due to the subjectivity involved in parameterization and the huge effects those assumptions can have on the results of the model. The goalpost now moves to trying to argue that, if bad models are the result of subjective assumptions, and subjectivity can't be removed from the model creation process, then modelers should instead make as many assumptions as they like and justify that the resulting model is accurate after the fact. At face value, this seems like a solid plan.

Elisabeth Lloyd, in a paper published to the journal Philosophy of Science, identified four major strategies for evaluating model correctness. First, models can be measured for "robustness", or comparing many independently created models to see if they agree. Second, models should display "variety of evidence", meaning they display accurate behavior with regard to many independent variables. Third, consider "independent support", or how well the model matches with data that wasn't considered when constructing it. Fourth and finally, there's "model fit". This one only works for predictive models, and it involves waiting some time in real life before observing the real world to see if the model's predictions came true (Lloyd, 2010).

The methods Lloyd identifies are widely used across domains. Model fit especially seems to be very popular among meteorological models. For example, one study took simulations made in 1979 modeling polar regions and compared them to actual data collected throughout the 1980s. They found that their models had "relatively poor estimation of precipitation", and concluded that a lack of understanding in cloud formation was a significant factor in why their models were so off base, spurring more research into that area (Chen, 1995, p. 89). Other models choose to rely on other techniques, some identified by Lloyd, and some not. Another study sought to find a way to incorporate media influence into an epidemiological model, owing to the aforementioned difficulty in translating some systems into computer models. They evaluated this model using a combination of parameter sensitivity analysis (determining the extent to which small changes in the initial parameters yield disproportionately different results) and independent support, applying their model to two different "case studies" to determine if their methodology made sense in both contexts (Kim, 2019).

It seems at this point that the discussion can end. We have identified robustness, model fit, variety of evidence, and independent support as a suite of seemingly solid evaluation techniques that see frequent use throughout computer modeling. However, to stop here after having found these evaluation techniques would be incomplete. Just as it cannot be taken for granted that our computer models are accurate, it also cannot be taken for granted that our evaluation methods for those models are accurate, either.

iii. There is Disagreement on the Effectiveness of Model Evaluation Techniques

In actuality, it seems that model evaluation techniques are a mixed bag. Similarly to Elizabeth Lloyd, Wendy Parker also published a paper in Philosophy of Science, this time assessing the extent to which robustness is an effective indicator of model accuracy in climatology. She, unfortunately, ultimately reached the disappointing conclusion that "while there are conditions under which robust predictive modeling results have special epistemic significance, scientists are not in a position to argue that those conditions hold in the context of present-day climate modeling" (Parker, 2011, p. 597). Robustness, therefore, is controversial at best for evaluating model accuracy.

The other evaluation methods are no less contentious. In their 2003 book, Luc Bovens and Stephan Hartmann cast doubt onto the concept of variety of evidence as a method of increasing confidence in a hypothesis, arguing that "evidence being less varied as defined on the correlation approach is no guarantee that the hypothesis will receive less confirmation" (Bovens, 2005, p. 106). This rebuttal against variety of evidence has itself been rebuked in yet another paper published in Philosophy of Science, arguing that their approach has enough limitations that they should not be able to conclude that variety of evidence is ineffective as a means of improving confidence. Instead, they propose modifications to their approach that would reveal that a variety of evidence would improve confidence, albeit with the caveat that "There are special epistemic situations wherein more independence does not give more confirmation." (Claveau, 2013, p. 112). It seems that variety of evidence works, but only as long as it isn't under "special epistemic situations".

Ultimately, this suggests that model evaluation techniques cannot be relied upon to guarantee accuracy. Several techniques have been developed and have support within the scientific community, but there is also considerable disagreement on exactly how effective those methods really are. An argument for model usefulness must therefore look to some other concept.

iv. Models can be Useful Even if they Aren't Accurate

Thus far, I have demonstrated that maximizing "accuracy" as a means of ensuring computer models can be effectively applied is difficult at best. This should seem confusing since, as established by the introduction, computer models have and continue to be used effectively. There must therefore be some other quality to models that allows them to be effectively employed - perhaps there is a difference between a model's accuracy and its usefulness.

There is, in fact, precedent for believing that accuracy and usefulness are different concepts. George Ellison, in a paper published about half a year into the COVID-19 pandemic, reflected on the role of computer models in controlling disease spread. He evaluated a set of predictive models used for policy and found many of their predictions never came true, yet, strangely, these models were used effectively by policymakers to combat the disease. From this, he concludes that "While all of these models are bound to be 'wrong,' some will be 'useful'; and, together, the best of them offer complementary insights into the nature of the disease" (Ellison, 2020, p. 510). If predictive models can be wrong and still be considered successful, then that is very strong evidence for separating model usefulness from model accuracy.

Despite this, it seems that the discourse around computer modeling focuses more on establishing accuracy over usefulness. That's not to say that computer modelers never try to establish model usefulness separately from model accuracy; one study on probabilistic choice models, for instance, created equations to measure all three of a model's accuracy, usefulness, and significance. (Hauser, 1977). Rather, it is simply that analyses like Hauser (1977) are a minority. While it seems to be widely acknowledged that inaccurate models are useful, it is uncommon to formally argue for model usefulness. Perhaps a socio-technical framework can help to close this gap in understanding.

The Relativistic View

Defined by Sabina Leonelli in her 2015 paper, a relativistic view of data is one that treats data as inherently meaningless. Instead, data "…consist of a specific way of expressing and presenting information, which is produced and/or incorporated in research practices […] and whose scientific significance depends on the situation in which it is used." (Leonelli, 2015, p.

811). In other words, data is a form of information that depends entirely on how it was collected and how it is used. Breaking this down further, there are two components at play.

First, data is an inherently scientific accomplishment. Collecting it involves processing it in some capacity, whether it be through scientific instruments, methodologies, processes, and so forth. Even the simple act of taking a picture of a natural phenomenon is subject to this scientific processing. Despite seeming like the closest one can get to making an observation without imposing any framework or biases upon it, merely taking the picture requires a researcher viewing the phenomena, deeming it important enough to document, and for future researchers to see that photograph and deem it important enough to include in their research efforts. As Leonelli herself puts it in a 2019 paper "...there is no such thing as 'raw data,' since data are forged and processed through instruments, formats, algorithms, and settings that embody specific theoretical perspectives on the world." (Leonelli, 2019, p. 2). This leads to the first major component, which is that the methods in which data are collected matter just as much as the actual data itself. The data on its own is meaningless; it only gains meaning through an understanding of the methods that made it.

The second major component deals not with how the data is collected, but rather how it is used. Once again returning to Leonelli, "data are defined in terms of their function within specific processes of inquiry, rather than in terms of intrinsic properties." (Leonelli, 2015, p.818). In other words, data is collected with the intent that other researchers will use that data down the line. The way that researchers use data as support for their claims and theories matters when discussing what exactly that data means and represents.

An example of how to apply this framework is given in the table below. On the left is a seemingly innocuous statement that treats data as an objective fact about reality (a "data objectivist" perspective, one could call it). Then, in the middle columns, relativistic revisions are introduced to the statement to better account for the processes that surround the data. Notice that relativism does not say that data cannot be relied upon; indeed, the final column still relies heavily on the data. Rather, a relativistic view states that data cannot be taken at face value.

Table 1: Revising a Flawed Statement with Data Relativity *(Created by Author)*

Hypothetical Statement	**Important Questions**	**Relativistic Revision**	**Revised Statement**
"Over X% of Americans aged 16-18 have a driver's license, so we should make the driving exam harder."	When was this data collected?	"According to a 2009 study…	"According to a 2009 study, a poll conducted in Y town found that over X% of teenagers aged 16-18 had a driver's license. If this sample is representative of the whole town, this indicates a large population of inexperienced drivers, so Y town should give them more experience through a harder exam. That is, of course, assuming that younger drivers are less safe than older drivers."
	How was it collected?	"...a poll conducted in Y town…"	
	Are there any problems with how this data was collected?	"...if this sample is representative of the whole town…"	
	How are you using or addressing the data?	"...this indicates a large population of inexperienced drivers…"	
	Why is this data relevant to how you want to use/address it?	"...so, Y town should give them more experience through a tougher exam before issuing a license…"	
	Are there any problems using the data this way?	"...assuming that younger drivers are less safe than older drivers."	

Leonelli defines this view in the abstract, choosing to be agnostic about which fields it may be applied to. Since this paper intends to use this view to discuss computer modeling, some work must be done to adapt this view to our specific topic. Thankfully, this is relatively simple.

The first half of a relativistic view is concerned entirely with the input data being used. In computer modeling, this asks how the computer modelers use input data to create their models. This can be done through considering the source of the data being used, the methods by which it was collected, and the applicability of that data to the model being considered. If computer scientists are going to comply with data relativity, then they must account for these questions.

The second half, in contrast, asks about how the model's output data will be used. It asks if the policyholder, not the researcher. is thinking about the relativistic nature of the data. This can be done through an awareness of the nature by which this data was collected, how applicable it is to the current situation, and how inaccuracies in the model will impact the usefulness of policies derived from it. Together, these two halves give us a lens with which to analyze the previous situation in more depth and attempt to gain a better understanding.

Insights Gained From the Relativistic View

Earlier, we identified quite the confusing scene when it comes to computer modeling. We discussed whether evaluation techniques were effective, and found reasonable evidence for both sides of that discussion. We also found that it seemingly doesn't matter if models are accurate at all, since they can still be useful regardless. Through an application of data relativity, we gain insight into how these seemingly contradictory statements can somehow coexist.

i. Evaluation Methods are Controversial Because Data is Relative

Unlike many other evaluation methods, the discussion around model fit did not yield any significant problems with the technique from an epistemological standpoint. Perhaps a relativistic view can help explain why this is. Model fit relies on collecting data twice - first to create the model, and then second to evaluate it. Ideally, the only thing that separates these two sessions of data collection would be time; under a relativistic view, as long as the methodology for collecting both sets of data remains the same, then any biases imposed by those collection methods would be neutralized through the fact that the same set of biases are imposed both times. That is exactly how Chen et al. (1995) made use of this technique in the aforementioned study of polar climates; there is no indication that the data collection methodologies significantly changed between when the model was created and when it was evaluated. So, when they discovered a disconnect, that strongly suggests that the difference must be accounted for by inaccuracies in the model. By sourcing the data the same way both times, they were adhering to data relativity, and their evaluation was successful.

In contrast, returning to Zimmermann (2020)'s investigation on pandemic modeling in India, we can see that the data used to create the model and the data used to evaluate its effectiveness were disjoint. As the researchers themselves pointed out, India suffered from massive underreporting of cases while the model was being created. Assuming that this issue was solved (or at least accounted for) by the time the model was evaluated, this means that this model was not being evaluated in a method consistent with data relativity, since it would be evaluated on data that was obtained by completely different methods compared to the data it was created on. Of course, even if they were collected using the same methodology, it is possible or even likely that the model would have still been revealed to have massive issues, but the different techniques used essentially guaranteed that the model wouldn't have a fighting chance.

On the topic of different data sets, return to variety of evidence as an evaluation technique. Recall that this technique was particularly controversial, with the conclusion being that it works, as long as it isn't used under "special epistemic situations", as Claveau (2013) put it. Once again, data relativity can help explain why this method is so controversial. First, the notion that consistent behavior with respect to many different data sets is an indicator of accuracy is, for the most part, congruent with relativistic thought. The caveat is, however, that those data sets should be collected using the same techniques. If a model is evaluated on two data sets that were themselves collected in two completely different ways, then that model yielding vastly different results on those sets would not only be possible, but expected under a relativistic view.

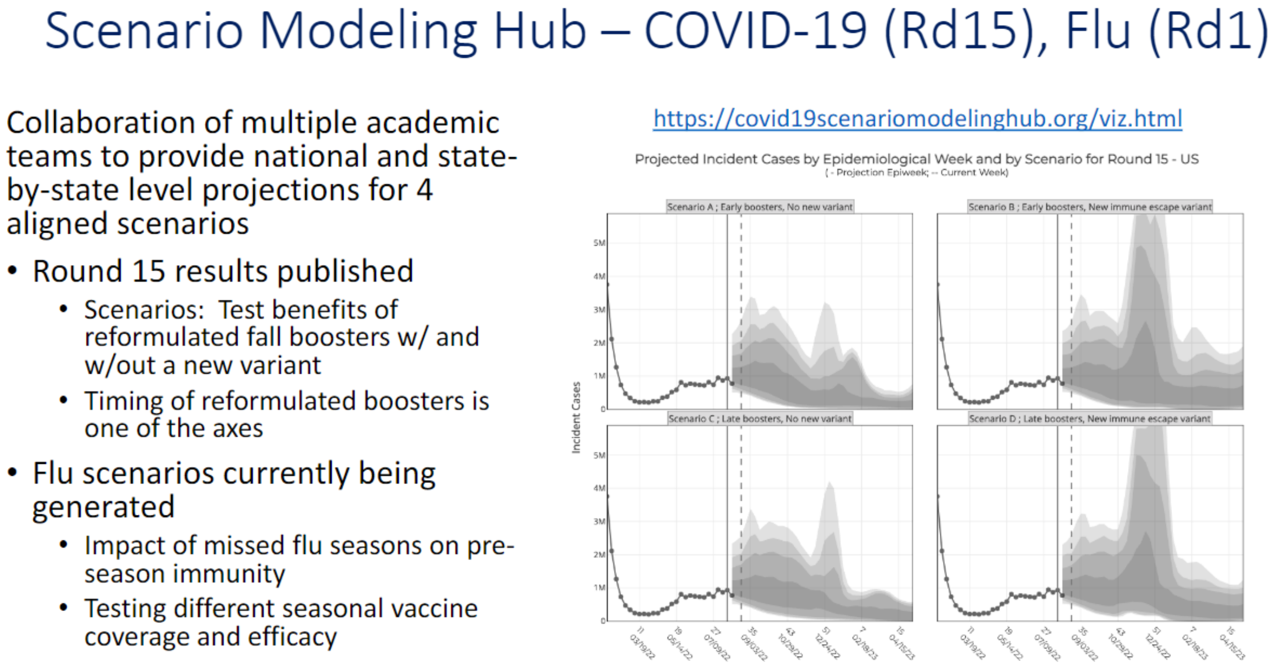
The study incorporating media influence into an epidemiological model by Kim (2019) serves as an illuminating example here. They had two case studies using large cities, one using a dataset from Mexico City, and one from Washington, DC. Datasets for both were obtained via their respective governmental institutions, both were scaled to reflect the population, and both used HealthMap to inform the media bias portion of their model (p. 5-6). The fact that these datasets were sourced from different governments is a cause for concern, of course, but it is still clear that ensuring consistency in how these different data sets were collected was a top priority. This way, any inaccuracies found would be solely the fault of the model, and not a result of biases injected during the data collection process.

Thus far, the entire discussion has focused on the first bullet point for applying the relativistic framework to computer modeling. We will now shift focus and discuss the second one, this time concerning ourselves with how models should be used by policymakers.

ii. Adaptability is More Important than Accuracy when Determining Usefulness

As discussed in the introduction, the Biocomplexity Institute at the University of Virginia creates weekly briefings about the state of the COVID-19 pandemic so government officials can be informed when they create policy responding to it, and those briefings are quite illuminating for answering this question. As seen below in Figure 2:

Figure 2: Excerpt from the 8-31-2022 weekly report *(Biocomplexity, 2022, Slide 24)*



Scenario Modeling Hub – COVID-19 (Rd15), Flu (Rd1)

Collaboration of multiple academic teams to provide national and state-by-state level projections for 4 aligned scenarios

- Round 15 results published
  - Scenarios: Test benefits of reformulated fall boosters w/ and w/out a new variant
  - Timing of reformulated boosters is one of the axes
- Flu scenarios currently being generated
  - Impact of missed flu seasons on pre-season immunity
  - Testing different seasonal vaccine coverage and efficacy

https://covid19scenariomodelinghub.org/viz.html

Projected Incident Cases by Epidemiological Week and by Scenario for Round 15 - US
(- Projection Epiweek; -- Current Week)

Scenario A ; Early boosters, No new variant
Scenario B ; Early boosters, New immune escape variant
Scenario C ; Late boosters, No new variant
Scenario D ; Late boosters, New immune escape variant
Incident Cases
0
1M
2M
3M
4M
5M
11 03/19/22
19 05/14/22
27 07/09/22
35 09/03/22
43 10/29/22
51 12/24/22
7 02/18/23
15 04/15/23

The Biocomplexity Institute gives multiple predictions rather than one singular result.

We can see that they present not one singular result for their model, but multiple different results that correspond to the different assumptions (or “scenarios”, as they call them) that could have been made.

Presenting results in this way addresses the issues discussed during the literature review, where the assumptions made during model creation can drastically affect the results of that model. Multiple results that correspond to different assumptions alleviates the issues that arise when parameterizing, and in exactly the way that data relativity says it should. This style of presenting results ensures that it is impossible to interpret the data coming out of this model as objective and independent of any context, since, once the statement is made that this graph corresponds to the scenario in which, say, a booster shot is rolled out and no new disease variant is found, those results are now forever tied to a context in which they can be understood. In other words, the results of this model can only be evaluated relative to the assumptions - the methods - by which they were generated.

In addition, even if the assumptions that they made led to an inaccurate model, these results can still be used. Even if the modelers assumed that the pandemic wouldn't be as bad as it turned out to be, policymakers can simply switch over to a different scenario with a worse outlook to inform their policy. Perhaps this is what model usefulness really means - not accuracy, but adaptability.

Overall, while there are many seemingly ambiguous or controversial aspects when it comes to evaluating and using computer models, clarity can be brought through an application of a relativistic lens. These results are summarized in the table below, with two broad sections covering the two halves of the relativistic view.

Table 2: Major Insights Gained through Data Relativity *(Created by Author)*

Difficult Topic	**Reasons for Difficulty**	**Relativistic Insights**
Evaluating Models	It is unclear if known evaluation methods are effective.	Evaluation methods are only effective if they account for the data they use.
	Variety of evidence isn't effective under certain circumstances.	Effective use of variety of evidence requires that the different datasets be sourced the same way.
	Models evaluated through model fit can still be inaccurate.	Model fit is best used when the data collection process does not significantly change between creation and evaluation.
Using Models	Model results are often taken at face value.	Present model results relative to the parameters used to create them.
	Model accuracy seems to have little to do with model usefulness.	Model usefulness has more to do with adaptability than accuracy.

Conclusion

Through this investigation, it has been shown that many uncertainties in model creation and model application can be offset, at least in part, through an understanding of data relativity. While the nature of parameterization ensures that uncertainties will always be present, understanding the data being used and being transparent about the assumptions made can ensure that models can still be useful even if they are not accurate.

Like any analysis, however, this conclusion has its limitations. Looking again at the table of results, it can be seen that more insights were found on the model evaluation end rather than on the model use end. This is to be expected; I am, after all, a computer scientist, and so I adopt that perspective when discussing computer modeling. This does not mean that investigating how computer models should be interpreted by non-scientists is unimportant; in fact, quite the opposite is true. It is the non-scientist, the policymaker, who ultimately uses models in the end. A future investigation should seek to gain more ground in that area as well.

Regardless, the insights gained here are quite relevant for computer scientists looking to ensure the usefulness of their models. As long as modelers are conscientious of data relativism when constructing and presenting their models, the chances that yet another faulty model could lead to disaster are that much smaller.

Works Cited

Biocomplexity Institute. (2002, August 31). *Estimation of COVID-19 Impactin Virginia* [Presentation Slides]. University of Virginia.

Bovens, Luc, and Stephan Hartmann. (2005, January 20). '4 Confirmation', Bayesian Epistemology. Oxford Academic, 89 - 111

Chen, B., Bromwich, D., Hines, K., & Pan, X. (1995). Simulations of the 1979–88 polar climates by global climate models. *Annals of Glaciology*, 21, 83-90.

Claveau, F. (2013). The Independence Condition in the Variety-of-Evidence Thesis. *Philosophy of Science*, 80(1), 94-118. doi:10.1086/668877

Edwards, P. N. (2010). A vast machine: Computer models, climate data, and the politics of global warming. MIT Press.

Ellison, G. T. H. (2020, September 1). COVID-19 and the epistemology of epidemiological models at the dawn of AI. *Annals of Human Biology*, 47(6), 506 - 513.

Hauser, J. R. (1978, May 1). Testing the Accuracy, Usefulness, and Significance of Probabilistic Choice Models: An Information-Theoretic Approach. *Operations Research*, 26(3), 406 - 421.

Kim, L., Fast, S. M., & Markuzon, N. (2019, February 4). Incorporating media data into a model of infectious disease transmission. *PLoS ONE*, 14(2), 1 - 13.

Leonelli, S. (2015, December 1). What Counts as Scientific Data? A Relational Framework. *Philosophy of Science*, 82(5), 810 – 820.

Leonelli, S. (2019, October 4). DataGovernance is Key to Interpretation: ReconceptualizingData in Data Science. *HarvardData Science Review*, 1.1

Lloyd, E. (2010). Confirmation and Robustness of Climate Models. *Philosophy of Science*, 77(5), 971-984. doi:10.1086/657427

Parker, W. (2011). When Climate Models Agree: The Significance of Robust Model Predictions. *Philosophy of Science*, 78(4), 579-600. doi:10.1086/661566

Turner, Andrew and Kuczynski, Jennifer. et al. (2019, December). Impacts of behavioral modeling assumptions for complex adaptive systems: an evaluation of an online dating model. In Proceedings of the Winter Simulation Conference (WSC '19). *IEEE Press*, 668–679.

Zimmermann, Lauren V. et al. (2021, July 27). Estimating COVID-19– Related Mortality in India: An Epidemiological Challenge With Insufficient Data. *American Journal of Public Health*, 111(S2), S59 – S62.