Epistemology of synthetic data in Machine Learning: An alternative perspective based on the theory of model/data symbiosis

Boet Bouten

October 3, 2025

Graduation Thesis
Media Technology MSc program, Leiden University
Thesis Advisors: Max van Duijn (First supervisor) and Rodrigo Ochigame (Second supervisor)

Abstract

Advances in the field of machine learning in recent years have dramatically increased the potential of generative models to produce synthetic data. Synthetic data sources provide powerful advantages for training machine learning models, as they seem to circumvent many of the fundamental challenges of conventional data collection - and processing for this task. Much of the current excitement and concern about the application of synthetic data for model training, especially in the scientific context, revolves around recent breakthroughs in generative modelling using deep learning architectures. However, there is at present still little theoretical ground available that links this ML-oriented conception of synthetic data to prior conceptions of synthetic data, especially in the context of simulation modelling. I argue that such a wider scope of analysis is crucial to understanding how synthetic data can be used to build more accurate and informative machine learning models, and where the practical and theoretical limitations lie from an epistemic perspective. This thesis aims to contribute to ongoing research on this matter by applying the theory of model/data symbiosis to synthetic data applications in machine learning. Model/data symbiosis states that models and data shape and transform each other in a mutually beneficial manner if performed correctly. If model/data symbiosis is executed wrongly, artefacts are introduced into both models and data and the results produced through these modelling methods could potentially be misleading. Developed in the context of simulation modelling, the extension of this theory into Machine Learning requires a conceptual separation of synthetic data into 'simulated' and 'generated' data. The core of this thesis is dedicated to substantiating the epistemic concerns involved when applying simulated and generated sources of data to the machine learning-pipeline. The end of the thesis looks ahead and considers what theoretical shifts in model epistemology might be required to account for the increasingly ’symbiotic’ relationships between generated data and AI-models that are currently under development.

1 Introduction

The production of synthetic data by generative models for the purpose of model training is currently a topic of much academic interest and concern. In broad terms, synthetic data are computationally produced data sources that mimic the statistical properties of a particular set of real data. Current day advances in deep learning methods have led to a significant qualitative and quantitative upscaling of the capacity of generative models to produce synthetic data. Synthetic data presents itself to researchers and engineers as an especially powerful data source for model training. The appeal of synthetic training data is twofold: firstly it provides a material advantage in terms of time and resources invested in data collection and processing, and secondly it allows for the creation of dynamic and context-specific data sets which could further previously unfeasible modelling goals. Running counter to this anticipation of the possible advantages of synthetic data are concerns over the negative effects these developments might have on the safeguarding of both fairness and integrity in relation to ethical and scientific standards. The ethical concerns in particular are targeted towards the risk that over reliance on synthetic data sources might perpetuate biases already present in the model. However, while the ethical concerns have garnered widespread debate in academic circles, assessments of synthetic data for model training embedded in philosophy of science perspectives are noticeably lacking.

From the position of the latter, analysis of the application of synthetic data for model training revolves around epistemology: to what extent do such data sources contribute beneficially to producing more accurate and reliable models, and thus lead to more accurate and reliable information about the real-world systems and phenomena that we are interested in? In this thesis I propose a particular approach to this question centred around the theory of model/data symbiosis, first developed by Edwards (Edwards 1999). The core premise of this theory is that there is an interdependent and mutually beneficial relationship between data and models; large amounts of data are involved in the construction of more accurate and reliable models, while on the other hand data sets are often heavily processed using models in order to make these data sets more consistent and informative. In this way, model-processed data sets aid in producing more accurate models, models which in turn could be utilised to greater effect when processing data sets. If we then view model/data symbiosis as a gradient, we encounter synthetic data towards the model-heavy end as data sources that are not just processed by, but fully produced using models. As such, model/data symbiosis presents itself as a useful lens through which we might assess how synthetic data sources play beneficial or detrimental roles in training more informative machine learning models

However, available literature on model/data symbiosis comes from a specific disciplinary niche, that of geosciences and climate modelling, and in-depth theoretical and practical accounts of it are scarce (Bokulich 2020; W. Parker 2020; Edwards 1999). Furthermore, this literature is firmly anchored in discussions on the epistemology of simulation models, which is not the modelling context we are interested in in this thesis. It is therefore crucial that we first establish a comprehensive theoretical overview of model/data symbiosis so that we have a good grasp of its implications. This will serve as the first chapter of this thesis. The main point of reference here will be the account of synthetic data as proposed by Bokulich, as part of her taxonomy of model-based data processing techniques (Bokulich 2020). In his chapter we furthermore establish the most important criteria present in the literature for assessing when model/data symbiosis is executed correctly, and under what circumstances we might identify detrimental relationships between data and models. These criteria depend on a conceptual separation of synthetic data from 'real' data. To define what kinds of data constitute as 'real data' the literature on model/data symbiosis places the emphasis on there being some form, however mediated, of physical interaction of measurement with a system or phenomena of interest. There is much room for nuance in this perspective however. This will be addressed at various points throughout the thesis.

With this groundwork in place we can effectively theorise how we could transfer the principles of model/data symbiosis across the 'model gap' from the simulation to the machine learning context, which will be the topic of the second chapter. In this chapter I work to extend Bokulich's original taxonomy by proposing a delineation of synthetic data into 'simulated' data and 'generated' data. I argue this distinction is warranted because simulation models and generative models respectively produce instances of synthetic data that exhibit different kinds of properties. Consequentially, both data types thus require different epistemic considerations when used for model training. The third chapter will be dedicated to discussing generative models and generated data, as generated data does not appear in the literature on model/data symbiosis and thus requires additional analysis. Practical examples of how real data, simulated data and generated data can be applied to the machine learning pipeline are the subject of the following chapter. The examples in this chapter show concrete opportunities and obstacles for applying synthetic data to machine learning pipelines, with the intention of substantiating how model/data symbiosis provides a useful framework for assessing the epistemic concerns present in these cases.

As we discuss the above-mentioned cases, it will become apparent that there is a theoretical limit to how well model/data symbiosis can deal with the most complex cases of synthetic data integration in machine learning that are currently available for study. This can be attributed partially to the limitations of the original literature in terms of scope and time of publication. However, I wish to argue that there is a more fundamental issue at play here, which revolves around the standards of epistemic reliability on which the theory of model/data symbiosis itself is built. Approaching this problem requires a more comprehensive examination of how we expect models in themselves to be informative. In doing so I aim to establish some grounds for thinking about a different epistemic context for complex machine learning models by drawing upon both prior and contemporary work on the philosophy of modelling. The last chapter of this thesis will be dedicated to this discussion. In this chapter I propose that complex machine learning models are capable of constructing representations of systems or phenomena that we cannot theoretically constrain or predict using principles that proved adequate in the context of simulation modelling. As a result, we might benefit from looking critically at our expectations of epistemic reliability when responding to the knowledge produced by such models.

2 Theoretical foundations of synthetic data: model/data symbiosis

To arrive at a concrete understanding of why synthetic data can epistemically be best understood in the context of a model/data symbiosis, it is important to have a clear idea of its theoretical origins. In essence, the concept of the model/data relationship itself as proposed by Edwards draws on foundational discussions in philosophy of science on the relationship between theory and observation, primarily on the directionality in which one validates the other. Although an extensive overview of this foundational work is beyond the scope of this thesis and can thus only be briefly addressed here, it will be useful to subsequently focus on one crucial period of transformation in the history of scientific practice, namely the emergence of statistical simulation modelling and its implications for our understanding of the epistemic relationship between models and empirical data.

2.1 Simulation modelling: From Pure-Theory Modelling To Model/Data Symbiosis

The practice of testing theoretical models of physical systems against available observational data is foundational to the empirical sciences and predates the extensive use of computational models. 'Pure-theory modelling', following Edwards, supposes a model/data relationship that is predominantly one-directional; empirical data tends to hold epistemic privilege over the predictions of a theoretical model (Edwards 1999). The emergence of computer simulation in particular catalysed important shifts in the theorisation of the model/data relationship as it introduced a markedly different modelling objective; instead of theoretical models being validated against empirical data as a means to test the underlying theory, simulation models apply rather than test theoretical structures with the goal of reproducing a physical system's expected behaviour (Winsberg 1999 in Edwards 1999). As noted by Winsberg, the integration of simulation models into scientific practice was made possible by virtue of the development of more powerful computers, as the study of such systems required complex mathematical models to act as mediators between the theoretical structures of the target system and arriving at any concrete knowledge about the system (Winsberg 1999). Edwards calls disciplines that make heavy use of such modelling techniques 'computational sciences' (Edwards 1999). Here it will be helpful to discuss a practical example of the kinds of changes these technological developments brought to the practice of scientific modelling, taken from the field of climate sciences from which most of this prior work originates.

2.1.1 A closer look at an example from climate modelling

Much crucial work on the epistemology of scientific modelling has been done by philosophers of science in the context of climate modelling. It is safe to say that climate as a physical system cannot be studied within the confines of a laboratory experiment. Instead, global climate models require mathematical simulations in order to study long-term changes in atmospheric conditions both backwards and forwards in time (Edwards 1999). What this requires in practice is the integration of immense amounts of empirical data taken from a wide variety of instruments, taken from sources ranging from satellites to weather stations, into global data sets. Once compiled, these data sets can then be leveraged for the construction of global simulations of climate systems. Naturally, the empirical data to be integrated into such global sets is collected under an enormous variety of different conditions, often containing notable gaps in terms of their accuracy and completeness both spatially as well as temporally. For these disparate data sources to be informative and useful for the purpose of constructing climate models, climate scientists require the use of 'intermediate models' to process the initial data sets, models based on theoretical laws governing the various meteorological subsystems involved in a global simulation of climate (Edwards 1999). In short, while large amounts of empirical data are included in the construction of climate models, and these models themselves contain semi-empirical parameters, this data itself is also heavily subjected to methods that optimise the data using other kinds of theoretical models. Since the target system is not accessible for experimentation directly, complex mathematical models act as mediators between empirical data and concrete information about the target system. In this process, models, data and theory are all connected in a network of interdependencies.

2.1.2 Model/data symbiosis

Compared to pure-theory modelling, this above situation presents a much blurrier and dynamic scenario; data and models continuously inform each other through multiple inferential layers and by means of diverse statistical optimisation methods. The central epistemic question becomes how well researchers can account for the presence of artefactual elements in both the models as well as the data (Edwards 1999; Winsberg 1999). In this context we can understand artefactual elements as any potential source of error or disturbance that might sanction the reliability or accuracy of either the model or the data (Winsberg

1999; Bokulich 2020). If this is done appropriately, then models and data in that particular modelling context are in a mutually beneficial relationship with each other, and there can be confidence in the reliability of both the model output and the empirical data given a specific modelling problem. In this view, models and data can be seen as symbiotic rather than as oppositional forces in their classical conception. This approach to understanding the epistemology of the model/data relationship is what defines model/data symbiosis (Edwards 1999). Edwards further argues that the premise of model/data symbiosis suggests a degree of *circularity* between models and data, which as noted above, can be considered beneficial if that circularity does not introduce artefacts into the data or the model. In the example of climate modelling, it is the degree to which various sources of data and multiple independently developed models are all constantly monitored and compared to each other that safeguards this, thus still allowing much ground for artefactual elements to be controlled for (Edwards 1999).

2.1.3 General and local variants of model/data symbiosis

An important delineation that needs to be established here is one addressed by Parker, where she argues for separately considering *general* and *local* varieties of model/data symbiosis (W. Parker 2020). Parker observes that Edwards' original conceptualisation of model/data symbiosis involves a widely distributed network of interdependencies between various models and data sets. This extensive picture of model/data symbiosis takes as reference research happening at the scale of entire research programs, for example in the context of constructing global climate models. Parker argues that model/data symbiosis also takes place at a local scale, involving a two-way dependency between '*particular* simulation models and data sets' (W. Parker 2020, p.808). In such cases, the relationship between the model and the data set is closer and more concrete, thus also potentially exposing this relationship to a higher degree of problematic circularity. Parker argues that even in these local cases of model/data symbiosis, problematic circularity is the most likely when simulation results *directly shape* the contents of a specific data set which is then used again to evaluate a simulation model (W. Parker 2020). The issue of circularity as discussed here by Parker will be crucial when we discuss the application of synthetic data in machine learning models later on in this thesis. More generally, local model/data symbiosis also aids in formulating more concrete examples of model/data relationships and their particular constraints, which will be important when discussing Bokulich's taxonomy of model-ladenness in the next few sections.

Overall, as I will argue further in this thesis, it is especially this local model/data symbiosis variant that will provide the most fruitful epistemological ground for evaluating the various applications and use cases of synthetic data in the context of machine learning. However, before we can begin to outline these cases, we ought to first address how the concept of synthetic data first emerged from these debates on the epistemology of computer simulation. This will be important so that we can establish proper ground on which both realisations of synthetic data, as products of different kinds of model/data relationships, can be compared to each other.

2.2 Epistemology of measurement in light of computer simulation

As computer simulation became more widespread in scientific research for the purpose of modelling complex and/or inaccessible systems of interest, philosophers of science were confronted with the following question: if simulation models are in essence computer systems solving complex series of dynamical equations, what kind of epistemic status can be attributed to their output? Simulation modelling thus unravelled a host of new epistemic concerns regarding how simulation models contribute to providing novel information about the physical systems they simulate (Morrison and Morrison 2015 ; W. S. Parker 2017). This could be seen as a consequence of model/data symbiosis as applied to computer simulation pushing against the very notion of what can be considered empirical. Given the logic of model/data symbiosis, which states that empirical data is often more useful for scientists when filtered through appropriate data-modelling techniques, to what extent can this process rely on mathematical abstraction before the output of the model loses its empirical provenance?

In short, the changes that computer simulation brought to the practice of scientific modelling, from which model/data symbiosis emerged as an epistemic framework for understanding these particular model/data relationships, also brought the epistemology of measurement itself under renewed scrutiny. An important aspect of these debates concerned how the output of simulation models ought to be epistemically compared against the output of measurement practices which 'directly' measure the system of interest. What we see emerging in this situation is a need to categorise data-processing techniques by virtue of the extent to which they make use of theoretical or computational models. From a perspective of the epistemology of measurement, this puts the logic of model/data symbiosis on a gradient, with simulation output on the one extreme and 'raw' data or instrument readings on the other extreme. These categorisations of data are all in themselves products of specific model/data relationships, and to reiterate, might all prove useful and informative for a given modelling problem given that artefacts in both the data and the model are controlled for.

2.3 Synthetic data as 'virtual measurements': Bokulich's taxonomy of model-ladenness of data

The integration of complex computer simulation models into the empirical sciences required there to be solid epistemic grounds on which the contributions of such models in producing novel information about real systems could be evaluated. As the general purpose of simulation models is to provide predictions about the behaviour of a target system through mathematical abstraction, based on strong theoretical laws and principles, it was not at all clear how epistemically reliable such abstract data sources truly were. It is in the context of these debates that we see the concept of synthetic data emerge as a way to classify simulated data as the product of a specific kind of model/data relationship, where the data is fully shaped by a computational model, albeit it based on theoretical foundations but without any physical interaction with the target system (Bokulich 2020; W. S. Parker 2017).

We find both the ideas of model/data symbiosis and the earlier mentioned need for more concrete categorisations of the data products of specific model/data relationships encapsulated in the taxonomy of model-ladenness of data, as outlined by Bokulich (Bokulich 2020). In this taxonomy, Bokulich addresses the need for more concrete knowledge about how data can be model-laden, as described in the model/data symbiosis framework, given specific data-processing techniques. Going into each aspect of the taxonomy would be too exhaustive for its intended purpose in this thesis. Instead, it would be most productive to give a general overview of the taxonomy and highlight some specific aspects of it that would further our discussion.

2.3.1 'Static' assessments: From data conversion to data fusion

The first five entries in the taxonomy concern data-modelling techniques that produce 'static' assessments of the data. What this means in practice is that such techniques are applied to data sets as a concrete, bounded operation with the goal of improving the accuracy and reliability of the data set. These methods tend to involve fairly conventional statistical optimisations to correct potential sources of error in the data, extrapolate measured data units to fill in gaps in the data set, or to combine multiple heterogenous data sets for the purpose of creating a more globally useful data sets, just to name three of the discussed techniques (Bokulich 2020). The products of these operations all result in model-laden data sets that, as long as a good account is made of the uncertainties present in the collected data and the modelling process, are more informative and useful for scientists. Although these entries in the taxonomy in particular are not that applicable to our discussions on synthetic data, what is most important to take away from this part of the taxonomy is that these data-modelling techniques are still relatively concrete and contain a fairly limited amount of layers of inference to arrive from the original data sources to the modelled data. As noted by Parker, any kind of measurement process necessarily involves layers of inference, informed by theory and models that apply this theory, to produce epistemically useful data (W. S. Parker 2017). As such, all the data-modelling techniques in this taxonomy involve such layers of inference; the more complex the operation in terms of scaling or extrapolating from the available data sources or combining various data sources together, the more layers of inference are involved. Parker argues that inferring accurate measurement outcomes from such operations requires a clear assessments of the possible sources of error in that specific modelling procedure. This will become more apparent when we discuss the last two entries in the taxonomy; data assimilation and synthetic data.

2.3.2 'Dynamic' integration: Data assimilation and synthetic data

The last two entries in the taxonomy are significant for this discussion as they involve much more prominently the application of simulation models. These techniques are essential to model highly complex and non-linear systems, with the end goal of producing not just model-based optimisations of data, but to produce dynamical models that can provide estimations on the behaviour of a particular system. *Data assimilation* is a data-modelling technique which integrates dynamical models with combinations of heterogenous data sources (often modelled using a variety of the techniques that precede it in the taxonomy), in order to construct sophisticated models capable of producing estimates about the behaviour of the target system (Bokulich 2020). The climate modelling examples as discussed previously make heavy use of data assimilation. Although Parker mentions that data assimilation is simulation-dependent and often comes in the form of simulation output (W. S. Parker 2017), in the taxonomy Bokulich delineates it from synthetic data outright as data assimilation still relies heavily on observational data. Data assimilation is thus a data-modelling technique in which simulation output is assimilated with measurements from various other data sets in order to provide accurate and reliable predictions for a target system.

Lastly, following data assimilation, Bokulich defines *synthetic data* as 'virtual measurements', constructed fully from the output of a simulation model. While simulation models that produce such data might thus be based on theories and principles that are derived from empirical observations, they produce output only by virtue of a mathematical abstraction of such theoretical principles. As these simulation models lack a direct measurement process that involves the target system, the output they produce are in themselves not to be treated as measurement outcomes (Bokulich 2020; W. S. Parker 2017). However, synthetic data sets are still highly useful for testing and calibrating many of the other data-modelling techniques mentioned in the taxonomy (Bokulich 2020). Furthermore, Parker argues that the output of simulation models in fact *can* be seen as measurement outcomes but only in cases of these models being embedded in non-simulation measurement practices, which we see in data assimilation (W. S. Parker 2017). This proper embedding of synthetic data that accounts for its status as 'virtual measurements' further contextualises our earlier discussion on circularity in local model/data symbiosis; if data sets consisting of synthetic data do not directly shape the data which is used to validate a simulation model, then the risk of circularity between synthetic data and simulation models might be mitigated.

In this taxonomy we thus see synthetic data emerge as the product of a model/data relationship where the data is model-laden to an extreme degree. Following the logic of model/data symbiosis, Bokulich and Parker state that synthetic data is primarily useful either as a means to help produce dynamical estimation models based on observational data (as is the case in data assimilation) or to test and calibrate other kinds of data-modelling methods. The three main concerns regarding the epistemic reliability of synthetic data can be summarised as follows:

1. **Materiality**: as simulation models are mathematical abstractions of theoretical principles associate with a target system, their output cannot be considered the result of an empirical measurement process as there is no physical interaction with the target system.

2. **Quantifying uncertainty**: simulation models and other modelling techniques that utilise these models (such as data assimilation), often involve highly complex and non-linear systems that possess complicated error structures (Bokulich 2020). As discussed earlier, a foundational idea behind model/data symbiosis is that scientists can adequately account for artefactual elements in both the data and the model for the two to be in a mutually beneficial relationship. As Parker notes, modelling techniques involving simulation require multiple inferential layers to be corrected for in order to arrive from the empirical data to the modelled data (W. S. Parker 2017). If the associated uncertainty involved in the model is not well-understood, and correction for error can thus not be propagated properly through the model, its epistemic reliability is compromised.

3. **Problematic circularity**: depending on the manner in which synthetic data sets are applied in local model/data symbiosis, there is risk of the model/data relationship ceasing to be mutually beneficial if synthetic data ends up being treated as real observations.

In the conclusion of her paper on the taxonomy of model-ladenness of data, Bokulich notes the following:

> My claim is not that model-laden data are always more epistemically reliable; the model-based processing of data, if not done appropriately, can introduce artefacts into the data and systematically mislead researchers (Bokulich 2020, p.10).

As such, the possibility of poorly understood or executed model/data symbiosis can thus prove a significant danger to epistemic reliability. Given the status here assigned to synthetic data as 'virtual' measurements, the extent to which this data provides accurate and reliable information about the real world is uncertain. This necessitated synthetic data to be in constant juxtaposition to sources of 'real' data which satisfy the two criteria mentioned above: data gathered through a measurement process physically involving the target system, and of which the potential sources of error are well-understood, quantified and appropriately corrected for.

2.3.3 What makes real data 'real'?

The literature on epistemology of measurement as discussed above places much emphasis on the importance of not equating the output of simulation models with real observations (Bokulich 2020; W. S. Parker 2017). It will be useful here to expand on what constitutes 'real' data. It is important to stress that real data does not imply an epistemological status of objective realism, reminiscent of the conception of data in pure-theory modelling. Instead, in our current discussion real data gives an indication more so of the methodology of its production, as a measurement process in physical interaction with the real world. This real data might still be heavily mediated by theoretical and simulation models depending on the particular model/data symbiosis context, as is for example the case in data assimilation as discussed previously (Bokulich 2020). However, this terminological delineation indicates the importance of clear knowledge on the epistemological constraints of synthetic data; in any context in which it is applied, its epistemological reliability is very much consolidated through a comparison to a trusted scope of real data.

3 Bridging the model gap: from simulation models to machine learning pipelines

In the above sections, we have seen synthetic data emerge as the product of a model/data relationship constrained by the epistemic limitations of computer simulation. In the taxonomy of model-ladenness as established by Bokulich, we find synthetic data as the final entry in the taxonomy, positioned all the way to the 'model-side' of the model/data symbiosis spectrum. Given its status as 'virtual measurements', scholarly work on model/data symbiosis tends to evaluate the epistemic reliability of synthetic data as strongly conditional in terms of how data sets produced through simulation are utilised. Following this, much of this work requires synthetic data to be embedded in a model/data symbiosis context in such a way that two important caveats can be avoided: 1) That potential sources of error inherent to the simulation model remain poorly understood and can thus not be accounted for and 2) Problematic circularity, in which synthetic data sets and simulation models feed into each other in a way that propagates error. Generally speaking, in this simulation context we can characterise the identity and usefulness of synthetic data as follows:

- Synthetic data are *virtual measurements*: they pertain data points relating to a specific target system, produced through solving dynamical equation without explicit interaction with this system.
- Synthetic data sets are most useful as when employed to *initialise simulation models but not to evaluate them* (avoiding problematic circularity), or to *test other data-processing methods* (W. Parker 2020).

Up until this point, our discussion on model/data symbiosis and synthetic data has been confined to the modelling context of computer simulation. The epistemic concerns regarding the integration of computer simulations into the empirical sciences gave the impetus for the framework of model/data symbiosis, and its associated theorisation of synthetic data, to be developed in the first place. The application of its principles in turn has also generally been confined to the scientific discipline in which it was first established, namely that of the geosciences, although there are no grounds for considering its ideas as uniquely applicable to this discipline (Bokulich 2020). It is thus not at all clear how the principles of model/data symbiosis are to be most informatively applied to synthetic data in the context of AI. As such, we ought to establish a useful comparison between simulation models and machine learning pipelines as distinguishable model/data relationships, so that in the following sections we can extend the taxonomy of model-ladenness to include a realisation of synthetic data particular to the machine learning pipeline.

3.1 Examining simulation models and machine learning pipelines as deductive and inductive approaches to modelling

To a certain extent, machine learning models follow a similar modelling purpose as simulation models when compared to pure-theory modelling: rather than act as a *test* of theory, machine learning models similarly aim to *reproduce* and/or *predict* the characteristics or behaviour of a particular system but crucially go about this in a different way (Edwards 1999; Von Rueden et al. 2020). The crux of this difference lies in the reasoning process behind *how* in each respective modelling architecture the model is arrived at; in simulation modelling this process is primarily *deductive*, while in machine learning pipelines this process is primarily *inductive*. This warrants further explanation.

As mentioned before, Edwards and Winsberg describe simulation models as *applications* of theoretical structures in order to reproduce the behaviour of a particular system (Edwards 1999; Winsberg 1999). The theoretical structures underlying the simulation are often the result of extensive research and experimentation, eventually arriving at a deductive model which describes the causal relationships within the system in the form of differential equations (Von Rueden et al. 2020). Approximate solutions to this series of equations given a set of starting conditions are then arrived at by running the simulation. As such, the simulation models that form the basis for our prior discussions on model/data symbiosis are in this sense 'knowledge-based' models (ibid.). In short, a simulation model is primarily derived from *theory*. Important to stress here is that a simulation often concerns a target system that is explicitly *known* to the modellers.

In contrast to this deductive and theory-driven approach, machine learning models are *inductive* in nature and primarily *data-driven* (Von Rueden et al. 2020). In a typical machine learning context, a model of the system is arrived at inductively through iterations of training on large amounts of data. The model is inductive in the sense that it aims to describe previously unknown patterns in data, for which the explicit causal relationships are not known. Machine learning models could thus be seen not so much as applications of theory like in simulation models, but more so advanced applications of data-processing techniques that produce representations of models of the data in the absence of a strong theoretical model of the system. As such, a machine learning model is primarily constituted through *data*.

3.2 Comparing simulation models and machine learning pipelines as model/data relationships

In what manner do these more fundamental differences between simulation models and machine learning models show in their respective architectures? As we have already discussed in earlier sections, simulation models are generally built on a mathematical abstraction of a solid theoretical model of the target system, usually in the form of a sequence of differential equations (Von Rueden et al. 2020). After this derived model is in place, the parameters for the simulation are defined, which concern the model's initial starting conditions and the boundary conditions that constrain the simulation. Running the simulation involves the actual computation of this series of differential equations, constrained by the parameters defined beforehand. This simulation procedure can then be repeated iteratively while tuning the associated parameters to achieve the most optimal result.

The construction of a machine learning model is typically done through a series of data-processing steps called a *pipeline* (Munde 2024). A machine learning pipeline (ML-pipeline from here on out) generally speaking defines the sequence of steps necessary for an inductive model to be learned from data. This process typically starts with the pre-processing of the training data, after which a hypothesis set is selected for the purpose of mapping the input features to the target variables of the system of interest. The hypothesis set refers to the collection of all possible mapping functions that the algorithm can consider. This hypothesis set is based on the architecture of the model, which can range from fairly simple linear regression models to complex multi-layer neural networks commonly utilised in deep learning. The parameters of this hypothesis set are then iteratively fine-tuned and optimised by a learning algorithm. Eventually a final model, or a final hypothesis of the mapping function the algorithm has decided fits the data best, is learned which can then be applied to new data sources (Von Rueden et al. 2020; Rueden et al. 2021).

It is important to stress that these above distinctions should not be taken too exclusively. As we have already noted before, the model/data symbiosis framework shows that complex simulation models often contain semi-empirical parameters that take in observational data. In the same vein, ML-pipelines often incorporate domain-specific knowledge meant to contextualise the data and subsequently also contextualise the patterns identified by the algorithm in order to make its implementation more suitable to its target domain (Miller et al. 2024). This is all in accordance with our model/data symbiosis framework. Furthermore, the above descriptions of simulation and machine learning modelling approaches are by no means meant as exhaustive and fully generalisable accounts of their implementation. We have already established that simulation models are often not employed in isolation but are in interdependent relationships with other models and data sets. On the other end, the above description of the ML-pipeline is most applicable to supervised machine learning, and as such is not fully applicable to all other machine learning and deep learning strategies.

However, this comparison helps to illustrate some of the fundamental differences in modelling philosophy between simulation and machine learning. As we have seen in practice with the case of simulations, despite their primarily deductive nature from theory such models are often very much 'data-laden' as they can include semi-empirical parameters. With ML-pipelines, we see a modelling approach much more anchored in induction from data. As a result, these models are much less constrained by theoretical assumptions, but possibly show deeper intertwined 'local' interdependencies between the hypothesis set and specific data sets that serve as training data. From the perspective of model/data symbiosis, there might be very different kinds of interdependencies between models and data in both modelling contexts. This in turn makes the data they produce, and how this data is then applied for the purpose of constructing more informative models, subject to distinctive epistemological constraints.

3.3 Synthetic data in the context of the ML-pipeline

Given this latter conclusion, it would be beneficial to extend the taxonomy of model-ladenness as established by Bokulich to further include a rendition of synthetic data produced through the ML-pipeline as distinctive from simulated data. Furthermore, as we established that the ML-pipeline concerns a substantially different model/data relationship by virtue of its construction primarily from data as opposed to theory, it will also be crucial to identify what kinds of interdependencies exist between particular kinds of data and their application in the ML-pipeline. As we have established that the ML-pipeline concerns a model/data relationship that is primarily inductive in nature, it will be crucial that aside from evaluating synthetic data produced by such pipelines as the product of a model/data relationship *in itself*, we also evaluate the constraints of such data as *training data in the ML-pipeline.* Given these nuances, the extended taxonomy will need to account for both of these situations.

Given the wide scope of possible model configurations and architectures achievable in both simulation modelling and machine learning, it would not be feasible to give an overarching account of these interdependencies that generalises to all possible methods for producing and applying synthetic data. Nonetheless, by handling a handful of particular cases studies related to the application of synthetic data in ML-pipelines, we can at the very least identify some of these interdependencies and discuss their properties in terms of epistemic reliability. The three criteria for epistemic reliability of synthetic data mentioned in the previous sections, namely *materiality*, *quantification of uncertainty* and *circularity*, can serve as useful principles to apply these cases.

Lastly, it is important to recognise that within a ML-pipeline, different types of data can play various roles in the construction of a model. Synthetic data produced though simulations can be employed in the training process of a ML-model, which in turn also produces synthetic data but through a different kinds of model/data relationship, as we established in the previous sections. To avoid terminological confusion and to concretely recognise the substantial differences between these types of data, in the following sections I will delineate the data types that will be discussed across three domains:

1. **Real data**: data produced through some form of mediated interaction with the real system of interest. As noted before, real data can still be heavily shaped by theoretical or simulation models. In complex situations, real data is realised from 'raw' data through various layers of inference and various layers of modelling. It is therefore essential not to equate real and raw data with each other, nor should the moniker of 'real' data be seen as this data representing 'reality' in some unquestionable form. As such, it will be most useful to look at real data as data whose error structures and possible artefacts are well understood, and as such can function as data to which other kinds of data can be validated against.
2. **Simulated data**: synthetic data sets produced using simulation models. It will be important to define this kind of synthetic data as a separate entity, as simulation data can also be applied to ML-pipelines, where we will need to be able to distinguish it from generated data. Simulated data, though lacking a substantial 'physical' measurement process, generally differs from generated data as the parameters for simulation models can be tightly controlled for, and are thus data produced in a more 'closed' environment.
3. **Generated data**: data produced through ML-pipelines, distinct from simulated data as this data is produced through a different model/data relationship. Given the inductive nature of ML-models, the data that can be generated by such models is a lot less theoretically explainable than their simulation counterparts. However, the capacity of ML-models to learn from a wide variety of data sets also provides much more detailed extractions of features of data. This allows for ML-pipelines to model data for which the associated 'real' system is largely unknown from a theoretical perspective.

With all aspects of the taxonomy now covered, the extended version is represented in figure 1.

In next section, we will examine generated data as a distinct form of synthetic data. In this section we will primarily take a closer look at the characteristics of this kind of synthetic data and what sets it apart from previous conceptions of synthetic data. In the section after, we will look at some particular cases involving generated data in the ML-pipeline under various kinds of circumstances, through which we can get a clearer picture of what kinds of interdependencies exist between generated data and the ML-pipeline itself, as well as the other kinds of data that might be involved in the construction of the model .

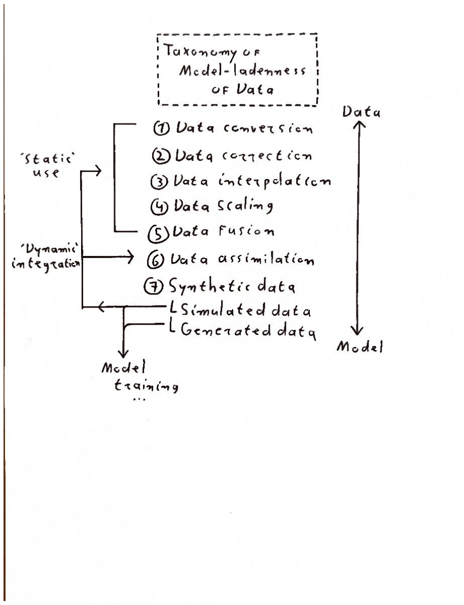

Figure 1: Diagrammatic representation of the taxonomy of model-ladenness.

4 'Generated data': producing synthetic data using generative model pipelines

In the previous chapter we have laid out a broad theoretical framework which we can apply to the study of synthetic data, based on the principles of model/data symbiosis. Following the taxonomy of model-ladenness as established by Bokulich, we have encountered synthetic data as a product of a particular model/data relationship inherent to simulation modelling. To bridge the gap between this conception of synthetic data and one more applicable to the context of contemporary AI models, we have established some crucial differences between simulation models and ML-pipelines as distinct model/data relationships, in itself resulting in realisations of synthetic data likely to be subject to different kinds of epistemic constraints following the principles of model/data symbiosis. As a conclusion to this chapter I have argued for the extension of the category of synthetic data in the aforementioned taxonomy in such a way as to account for 'generated data' as a substantially different synthetic data source compared to simulated data. In this chapter we will further lay out some essential characteristics of generated data. We will do so by giving an overall account of 'generative models' and how generated data is produced through such models following some of the inductive principles we have already laid out regarding ML-pipelines in the previous sections.

4.1 Generative models

As we have established previously, the goal of inferring a model from data using a machine learning approach is to eventually employ this model either for prediction on or for generation of new data sources. At a general level, this difference in modelling objective comes down to a difference between *discriminative* and *generative* model objectives. Discriminative models follow what Harshvardan and colleagues call a *discriminative classification process*: in essence, the model learns to distinguish between and identify correctly classes of instances by learning from the training data (Gm et al. 2020). During this process, the model does not learn the underlying distribution of the training data in its entirety, as the modelling objective is solely to map the input-output function of a specific task (Jebara 2004). This approach lends itself well to a variety of classification and detection tasks. Generative models in turn do not exclusively model the 'decision boundary' in the data space based on the modelling task, but treat the entire data space itself as a probability distribution from which the data is created (Gm et al. 2020). In other words, generative models learn the 'overall distribution of the data', and subsequently produce a modelled distribution resembling the original distribution (Gm et al. 2020). This modelled distribution of the data can subsequently be leveraged to create samples containing the extracted features of the original distribution based on specific input variables. In this manner, generative models are capable of generating new instances of data from this modelled distribution that resembles the original distribution of the training data. Generally speaking, it is this generative approach to machine learning through which synthetic data sources are created. Here again we see the difference in modelling philosophy in comparison to simulation modelling. Generative models start from data and generate a model through learning the probability distribution of a set of data. Simulation models in turn start from theoretical assumptions about the nature of a system, after which running the simulation produces a data space (Von Rueden et al. 2020). The full scope of possible generative model architectures is exhaustive and for the purpose of scoping out a working definition of generated data, it will not be necessary to discuss them in detail. Generally speaking, generative model architectures range from relatively simple clustering algorithms such as Gaussian Mixture Models to advanced multi-layered deep neural networks (Gm et al. 2020). This latter category of generative models currently stands as the cutting-edge of synthetic data generation using ML-pipelines, and its practical use cases also exhibit the most complex and thought-provoking applications of generative data in the ML-pipeline itself. As such, in most of our examples in the following chapter we will be dealing with deep generative models.

4.2 Generated data as a synthetic data source

In our discussion on synthetic data in the context of simulation models, we have established through the literature a general account *what* constitutes synthetic data and *why* synthetic data is useful for the purpose of scientific modelling:

synthetic data constitute *virtual measurements* of a system of interest, and their application can be beneficial mostly for testing, calibrating and improving data-processing models primarily making use of real data (Bokulich 2020). To what extent do these above two conditions apply to generated data?

4.2.1 Generated data and measurement

It is not immediately clear if generated data could reasonably hold the same status of 'virtual measurements' as we have established for simulated data. Generated data could similarly be assessed as a data source construed without any physical interaction with the system of interest, fully shaped by a computational model. What complicates the issue is determining exactly what a particular data point *represents* in relation to the system that it attempts to model. This is an issue that is not completely divorced from the simulation case either, as in cases involving data assimilation the question still remains open as to what exactly one can expect the simulation model to give 'virtual measurements' of (W. S. Parker 2017). However, in the case of simulation modelling, it is the nature of the system itself of which modellers have a clear idea at the very least. As we have previously established, simulation models are primarily constructed in a deductive manner from theoretical principles, despite their interdependencies with empirical data sets. This fundamental theoretical underpinning of the system of interest allows for some manner of confidence that even in the face of uncertainties in the interpretability of simulation results, one at least has a clear picture of what constitutes the system of interest and what the theoretical boundaries of this system are.

Given the inductive nature of the modelling process in ML-pipelines, this matter becomes more complicated. As ML-models are primarily data-based and are not in themselves derived from a strong theoretical model of the system of interest, the question regarding what system is actually being modelled is therefore a lot less trivial then it might sound. As noted before, generative models that are capable of producing synthetic data sets in essence model the overall probability distribution of the data space, and return a distribution similar to the original one. Exactly of what system or phenomena this generative model returns measurements is therefore constrained by the data space from which it modelled this distribution. As a result, what constitutes the system of interest is thus to some extent inferred from the data space similarly to the model itself. This presents a markedly different scenario of measurement than is typical at least in the empirical sciences, and exactly what implications this has for the epistemology of measurement goes beyond the scope of this thesis. For the purpose of this discussion, we can take away from this observation that the absence of strong theoretical constraints regarding the nature of the system or phenomenon being modelled by a generative model will have important consequences for evaluating the data it produces.

Now we have gotten a better understanding of some the characteristics that set generated data apart from simulated data. We ought to now shift the focus to the ML-pipeline as a model/data relationship in which real, simulated and generated data can find various kinds of applications in the process of model-training. For each type of data, I will discuss examples of their application and address possible epistemic concerns in light of the principles of general and local model/data symbiosis.

5 Real, simulated and generated data in the ML-pipeline

In the previous sections we have established that particular configurations of the ML-pipeline, namely that of generative models, are model/data relationships that produce synthetic data with different properties than the synthetic data produced using simulation models. Given the distinctive modelling context and philosophy underlying generated data, we can also expect there to be differences in function compared to prior accounts of the usefulness of synthetic data. We have already given extensive attention to the latter in chapter 2: from prior work on model/data symbiosis we have taken that simulated data is thought to be most useful as 'virtual measurements' which can be deliberately employed to either initialise the simulation model (or related models) or to test and evaluate other kinds of data-processing methods. We have further identified that improper applications of synthetic data in this manner might lead to problematic circularity, which in the context of model/data symbiosis implies a mutually harmful relationship between data and models which introduces artefactuality and jeopardizes epistemic reliability. With this is mind, we can investigate the ways in which generated data can be applied in ML-pipelines and to what extent we can find similar or novel epistemic concerns in light of the principles of model/data symbiosis. In addition, it will be useful to also delineate how simulated data can be made useful in this modelling context, and how such applications might highlight important differences in terms of how simulated and generated data are conceptualised when applied in ML-pipelines.

5.1 Real data within the ML-pipeline

The prior literature on model/data symbiosis differentiates real data from synthetic data primarily on appeals to materiality: real data, while often mediated by a variety of data-processing methods, is a product of a measurement process that involves the target system in some physical manner (W. S. Parker 2017; Bokulich 2020). This emphasis on materiality underscores the importance of instrument calibration and estimating uncertainty in the results. As such, real data serves as the best possible informative account of a target system in most cases, involving inferences from raw data to useful models of data that are well-understood and calibrated. When considered in the context of simulation modelling, real data therefore serve as the epistemic 'backdrop' against which the usefulness of synthetic data can be carefully considered and its implementation in tandem with this real data carefully monitored.

To what extent do these considerations transfer to the ML-pipeline context? Generally speaking, the provenance of real data for ML-models need not be considered much differently: data for model training can be collected and compiled into data sets in many different ways, once again potentially mediated by various data-processing techniques along the way depending on the specific modelling problem. The most crucial difference lies in the implications of the data-driven approach to modelling inherent to machine learning. Considering the general lack of theoretical modelling prior to the construction of the model, it is the accuracy and most importantly, the *representativeness* of the real data that gives researchers confidence in the distribution that the model learns from this data.

5.1.1 From real data to the 'real domain': the domain shift problem

The inductive nature of ML-modelling invests the merit of a model primarily in the following question: how well does this model that has learned a distribution of a particular set of data of a target system generalise to data that is has not seen before (Stacke et al. 2021)? Ideally, we would assume that this unseen data follows the distribution of the training data, but due to variance in available training and test data this often needs to be accounted for (Chakrabarty, Sreenivas, and Biswas 2023). If the distributions between training and test data differ significantly, we encounter a *domain shift*: the performance of the model is compromised because essentially, the distribution that it has modelled does not match the 'target distribution' on which it is validated (Stacke et al. 2021; Chakrabarty, Sreenivas, and Biswas 2023). The challenge thus becomes transforming models and data across domains in ways that mitigate this problem.

Stacke and colleagues define domains essentially as *types* of data with possibly different kinds of distributions, with the successful adaptation of a model to these distributions serving as a evaluative benchmark for the performance of a model (Stacke et al. 2021). Such differences in distributions might be minute, statistical differences (for example in colour or contrast when the data consists of images) between sets of real data. More importantly for our discussion, domain shifts are also crucial beyond this preoccupation with real data sets, namely when the domains in question involve both real and synthetic data. As such, models and data also transfer between synthetic and real 'domains'. Domain shifts speak to the larger issue of understanding the relationships between model and data in ML-pipelines. In the pipeline, data is required both to learn the model and to test if the model has learned a distribution that is useful and informative for making predictions on new data or to produce similar instances of said data. However, different types of data can have both known and unknown characteristics that effect the pipeline in both beneficial and detrimental ways.

5.2 Simulated data within the ML-pipeline

For relatively small ML-problems, for which high-quality annotated data sets can be leveraged from real domains without much issue, there is yet little reason to depend on synthetic data for training. With the development of more complex machine learning structures, namely deep neural network architectures, the strain on available data from real domains to train such large models becomes more severe. Utilising simulated data for model training is one potential avenue for combatting these challenges (Melo et al. 2022; Tremblay et al. 2018; Deist et al. 2019). In addition to possibly addressing issues of data scarcity, simulated data might also in themselves prove preferable over real data in specific modelling scenarios. This is primarily due to the properties of simulated data as virtual measurements; the degree of oversight and control one can exercise over data produced by simulations opens up possibilities for model training that are difficult to achieve solely in the real domain. Such data sources might prove useful for training and improving a model, yet if done carelessly without appropriate assessment of how this might adversely impact the model's performance in the real domain, the results might be misleading. Van Breugel and colleagues call this the 'naive synthetic data approach'; treating synthetic data as if it were real data (Breugel, Qian, and Schaar 2023). Here it will be essential to consider if and in what way simulated and generated data are substantially different enough that they also warrant different considerations in this regard.

5.2.1 Calibrating simulated data for ML-pipelines: synthetic-to-real domain shift

As discussed in the previous sections, much of the epistemic uncertainty surrounding the relationship between ML-models and training/testing data is characterised by domain shifts: as these models learn distributions of particular data sets, their performance is 'intrinsically tied to the quality, volume and relevance of their training data' (Wang et al. 2024). If ML-models are trained primarily on data outside of the real domain in which they ought to provide useful and informative predictions, then some kind of transformation is required to facilitate this transition between domains: this is commonly known as the 'synthetic-to-real' domain shift (Reddy et al. 2024; Nikolenko 2019). Techniques to account for this shift involve closing the 'gap' between the available data from both domains. Two such approaches involve either diversifying the simulation data that the model is exposed to in domain randomisation (Tobin et al. 2017), or statistically aligning the simulated data with the real data to minimise the potential error in transitioning the model from one domain to the other in domain adaptation (Melo et al. 2022). In general, such methods benefit from the fact that certain characteristics of the simulated data can be exactly known through the configuration of the simulator. A good example of this is the utilisation of simulators for training machine vision models: in this case, one could quite literally simulate a highly variable synthetic 'domain' where the data from the simulated scenes becomes training data for the model (Melo et al. 2022). Central to any such approach is caution for the propagation of error inherent to the shift in domains: the less carefully this shift in domains is accounted for, the higher the chance that the simulated data introduces artefactuality into the distribution learned by the model, leading to misrepresentation in the real domain.

5.2.2 Integrating simulators into the learning pipeline: a comparison with data-assimilation

Up until this point, the models that produce synthetic data are separate entities from the pipeline itself. What this means in practice is that the simulated data sets are static in nature, with the ML-model only able to adjust its learned distribution over successive training episodes as it is exposed to different data sets. In other words, the model cannot learn from such data sets in a dynamic and continuous manner. To facilitate this, the simulator itself needs to be integrated into the ML-pipeline; in this manner, a model could learn continuously from a dynamic simulated domain over much longer periods of training time (Melo et al. 2022). Such a pipeline configuration would enable model training that is practically beyond anything that can be achieved with data sets from real domains alone.

Here it will be illustrative to draw a comparison to our earlier discussion on data assimilation. Data assimilation, which involves integrating numerous heterogenous data sets with simulation models, made it possible to produce dynamical model estimates beyond the time-and-space ordered boundaries of conventional measurement. We see a similar pattern of reasoning take place in the context of integrating simulators in ML-pipelines; just as dynamical models of complex physical systems are not feasible to construct without elaborate simulation models, it similarly appears unfeasible to train models that form dynamic, robust generalisations of a system without means to expose the model to task-and-context specific, continuously updated streams of data. Given the limits of data collection and processing in real domains, simulated data presents itself as a viable alternate path. However, the scenario presented by integrating simulators directly into the model pipeline encroaches upon the concerns for problematic circularity between data and models as expounded by local model/data symbiosis. Similarly, it is not entirely clear how such closely related integration of simulated data and ML-pipelines effects the effectiveness of methods meant to account for domain shifts.

Despite the above concerns, if the models in question are mostly limited to performing classification or recognition tasks then we have not quite yet arrived at the point where the data produced by a model and the data involved in model-training and model-testing becomes entirely circular. As stated before, discriminative models do not have the capacity to produce new instances of data in the same way that generative models can. In the following section we will discuss synthetic data produced by the latter, where we will dive further into this and other potential complexities of integrating generated data into ML-pipelines.

5.2.3 An in-between case: AlphaGeometry's synthetic data language

One curious edge case of synthetic data production can be found in DeepMind's AlphaGeometry project, a groundbreaking AI system designed for automated geometric theorem proving (*AlphaGeometry* 2025). AlphaGeometry's training data consists of nearly one billion synthetic, randomly sampled theorem premises, but the methods behind the construction of this data set provide valuable context for understanding the transition from simulated to generated data. The training data requirements for AlphaGeometry far exceed the availability of human-designed geometric problem data sets. To produce enough synthetic data to train the model, the modellers developed a diagram constructor language capable of producing theorems based on a general set of actions the constructor could execute. Note that the authors state that at no point any existing problems sets were utilised (*AlphaGeometry* 2025). No training data was involved in the production of this synthetic data set, and thus this synthetic data cannot be considered 'generated data' for the purpose of our discussion. In this sense, AlphaGeometry's diagram builder functions more akin to a simulator as its behaviour is bounded by defined theoretical principles; the validity of its output can be controlled for by deductive means. At the same time, it is not clear if this constructor language can rightfully be called a 'geometry simulator' along the same standards as that we have discussed simulation models so far. There is no need to think in terms of mathematical 'abstraction' given that the system of interest in itself lies in the mathematical domain to begin with. As such, AlphaGeometry's constructor could be seen as lying somewhere in between a geometry simulator and a geometry generator.

The primary reason why this approach to synthetic data production worked for AlphaGeometry is that the underlying rules for what constitutes useful synthetic data are exactly known, properties more similar to that of simulated data. However, many empirical domains lack such strict standards of verification. If knowledge about the properties of data and the system of which this data is thought to convey useful information are insufficient, then the alternative is to infer these properties inductively if they cannot be theoretically deduced, recollecting our previous discussion on the differences in modelling philosophy between simulation and machine learning.

5.3 Generated data within the ML-pipeline

The usefulness of synthetic data produced by simulators is constrained by the extent to which one can define and adjust the properties of the simulator so that it produces data adequate for the training task. This adequacy is mostly safeguarded using methods to account for domain shifts, with which the disparities between simulated and real data domains for a given modelling problem can be identified and minimised. In the previous section we have examined cases in which this is feasible. This largely depends on the available knowledge of the properties of the data and its underlying system which, as we have discussed earlier, in the case of modelling problems suitable for machine learning is often lacking. As such, the most effective way to produce synthetic data for such cases is for a generative model to extract the most essential features and characteristics out of data available from the real domain with the goal of generating new instances of data similar to the target data (Hao et al. 2024). Current deep learning architectures have achieved incredible rates of success at such tasks for a variety of different formats. From the perspective of model/data symbiosis however, the application of such data sources for model training presents radically different concerns in terms of how potentially unexpected and detrimental relationships between data and models can come about, and to what extent they can be accounted for.

5.3.1 'Modes of failure': identifying uncertainty in generated data

A crucial principle of model/data symbiosis involves the careful assessment of uncertainty in both models and data sets, as to ensure that the relationship between data and models is beneficial and down the line leads to more epistemically reliable information. Parker noted that in the context of measuring practices involving computer simulation, a major hurdle in determining the epistemic reliability of such model results lay in the difficulties associated with their calibration (W. S. Parker 2017). These difficulties arose in part due to, at the time, the relative novelty of such modelling practices, with more conventional instruments having undergone much more rigorous calibration over longer periods of time. Parker also notes that the complex and non-linear fashion in which such models operate and integrate different sources of data makes the already challenging process of calibration even more daunting (W. S. Parker 2017). Given the scarcity of further publications on model/data symbiosis, it is unclear to what extent these challenges might have been successfully addressed in the case of data assimilation. In the case of generated data however, similar concerns are highly relevant for essentially the same reasons; generated data is only a recent development in data synthesis for which trusted methods of calibration are still few and far between, a process which is complicated by the complex, non-linear nature of its production.

A core principle of model/data symbiosis concerning synthetic data is that such data sources can serve various functions depending on the specific modelling context; if and how synthetic data might aid in producing more epistemically reliable models or data sets depends on the 'details of the case' (Bokulich 2020). This applies to generated data just as much as simulated data. In a relatively early publication on the evaluation of generative models, Theis and colleagues state the following:

> Probabilistic generative models can be used for compression, denoising, inpainting, texture synthesis, semi-supervised learning, unsupervised feature learning, and other tasks. Given this wide range of applications, it is not surprising that a lot of heterogeneity exists in the way these models are formulated, trained, and evaluated (Theis, Oord, and Bethge 2016).

Such a wide range of applications does not just present many opportunities for generated data to be useful, but also many opportunities for this data to be inappropriate or unreliable for the task at hand. Alaa and colleagues call the latter a model's potential 'modes of failure' (Alaa et al. 2022). The authors suggest that evaluation metrics that are too general in nature, often evaluating the quality of a generative models output on the entire distribution by means of a single metric, gloss over how even individual generative samples can contain errors which might negatively impact the modelling task (Alaa et al. 2022). In the context of model/data symbiosis, this would indicate potential propagation of artefacts in the model output. It is important to state that as generative models improve and even more applications for generated data are found, more in-depth evaluation metrics across domains continue to be developed (Betzalel, Penso, and Fetaya 2024; Goyal and Mahmoud 2024). However, I argue here that the calibration of generative models is made more difficult due to there being little to no theoretical grounds on which one can formulate expectations about these degrees of uncertainty. Given that an ML-model is in essence an inductive representation of the patterns an algorithm can identify in large amounts of data, identifying the potential sources of bias and error in both the models and the data becomes one more so of empirical rigour than of theoretical derivation. These epistemic concerns are magnified to an even greater degree if the model/data relationship starts exhibiting circularity.

5.3.2 Problematic circularity: local and general forms of model autophagy

Up until this point, the epistemic concern of circularity has not yet featured prominently in our discussions on synthetic data in the ML-pipeline. We have established previously that problematic circularity is an especially noticeable concern when it involves a close two-way interdependency between a particular simulation model and data set, or *local* model/data symbiosis. It becomes immediately clear that these same risks are present for generative models: if the output of a generative model enters into a circular relationship with its training data, this could be considered a clear breach of this principle. However, problematic circularity of generated data can also take place at the scale of *general* model/data symbiosis. In this scenario, generated data from multiple models are captured in training data sets used to train the same or similar models, potentially unknowingly and unintentionally. The process by which ML-models, particularly generative models, are trained on generated data has received a novel term in recent literature: *model autophagy*. Given the scope of this thesis I will constrain our discussion of model autophagy to what is relevant in the context of model/data symbiosis.

Although model autophagy does not by definition imply a negative relationship between model and generated data, the term came to prominence because of such connotations. As sources of generated data became more accessible for model training, so did accounts in the literature of the harmful effects of overexposing models to generated data; if repeated over successive generations of model training, the model risks diverging from its original distribution to such an extent that it can no longer reproduce it (Alemohammad et al. 2023). This state is known as *model collapse* (Statsenko, Andriyanov, and Shishkin 2024). In the context of model/data symbiosis, model collapse is a result of problematic circularity between generated data sets and ML-models. Here we must recall how generative models produce instances of data; as the model is inferred inductively, its output approximates a distribution of real data that is in itself inherently incomplete (Statsenko, Andriyanov, and Shishkin 2024). Excessive amounts of generated data flatten out the learned distribution's diversity curve, reducing the capacity of the model to make diverse predictions and converging its output to a distribution that is not representative of the target system. The worst case-scenario of model autophagy identified in the literature is a fully synthetic loop, where a model is predominantly trained on its own generated data (Xing et al. 2025). These examples involve a close relationship between a particular model and its output serving as its own training data; in correspondence with model/data symbiosis, I will call this scenario *local model autophagy*. Although much research already has been dedicated to combat its effects with success in controlled interventions, current reviews of this literature indicates that although it can be mitigated, it seems unlikely that it can be fully eliminated (Xing et al. 2025).

It is crucial to stress that not all applications of generated data in the model pipeline lead to model collapse. However, techniques that do make successful use of generated data for model training, such as in knowledge distillation (Statsenko, Andriyanov, and Shishkin 2024), require that the balance between real data and generated data in the pipeline can be tightly monitored and that effective interventions can be made. This brings us to what I call here *general model autophagy*, where generated data from various models indiscriminately proliferate in training data sets with little to no oversight or effective means to filter this data out. As with general model/data symbiosis, this kind of model autophagy happens at much larger scales. As the most powerful deep generative models require massive data sets often compiled from databases which scrape huge swathes of the Internet, the indiscriminate and often unregulated manner in which this happens provides ample opportunity for generated data to proliferate in such data sets (Xing et al. 2025). In addition, as more general-purpose generative models are often used as baselines to train smaller models, this uncertainty propagates even further. This essentially implies a collapse of the real and generated domains. In earlier accounts of model/data symbiosis, problematic circularity at the general scale formed less of a concern as independently developed models and data sets could be cross-compared to each other following discipline-specific standards (Edwards 1999). However, as model training at large scales necessitates more so the appropriation of primary sources of data of haphazard origins than the curation of data specifically for one directed (scientific) purpose, it is still very much unclear to what extent the harmful effects of general model autophagy can be accounted for or even confidently assessed (Knuuttila, Carrillo, and Koskinen 2024).

5.3.3 Integrating generated data into the pipeline: DeepSeek R1's reinforcement learning approach

As alarming as the findings on the detrimental effects of model autophagy are from an epistemic perspective, this has not by any means deterred research into the further integration of generated data into ML-pipelines. In a manner similar to the integration of simulators into the pipeline as we have discussed earlier, the most powerful strategy from a model training perspective would be to integrate the generation of training data directly into the pipeline. Aside from optimising the training process for more continuous learning, there are additional motivations here that are not as salient in the simulation example that need to be addressed, Most importantly, integrating data generation within the pipeline makes possible the scenario where a model can learn reflexively from its own output.

We can see this idea in practice in the model pipeline of DeepSeek-R1, a large language model developed by DeepSeek. DeepSeek-R1 is a powerful state-of-the-art generative model for producing textual output. DeepSeek-R1 is based off an earlier DeepSeek model, DeepSeek-V3-Base. These models require exceedingly large amounts of training data, and thus there is little concrete, useful information on the nature of the data that the base model is trained on. However, what's crucial here is not the training process of the base model, but how DeepSeek-R1 interacts with data produced by the base model. To improve DeepSeek-R1's reasoning capabilities, an extra step was added to the pipeline: DeepSeek-R1 performs reinforcement learning on output from the base model (DeepSeek-AI et al. 2025). What this means in practice is that the model trains itself in an iterative manner by producing its own problems that it then attempts to solve and learn from. It is important to state that this is a grossly oversimplified explanation, but the significance of this strategy is clear; this integration of generated data into the pipeline not only circumvents the material strain of compiling large amounts of real data for model training and fine-tuning, it also allows a generative model to model more complex distributions of data by learning from its output. At this point, it becomes increasingly problematic to apply the principles of model/data symbiosis as they are understood in its original modelling context to models such as DeepSeek-R1. The distribution learned by the model is now subject to so many layers of inference that a comprehensive account of the associated uncertainty that might propagate through these layers is likely intractable. Furthermore, allowing for a model to produce its own data sets from which to learn seems to be a grave example of problematic circularity. Yet in the context of this modelling problem, this circularity within the model actually is beneficial to the evaluation of its output. How to reconcile all these seemingly contradictory observations is far from trivial.

In this chapter we have seen the converging of various kinds of synthetic data with the model pipeline increase in complexity. Yet at the same time, we also see the epistemic concerns associated with model/data symbiosis, namely estimating uncertainty in models and data and problematic circularity between models and data, magnified in theory yet disappearing out of sight in terms of having concrete methods and frameworks to assess their influence. These developments invite a broader perspective on the symbiotic converging of generated data with ML-models that goes beyond our established expectations of epistemic reliability. Instead, we ought to consider how these developments could inform our understanding of model epistemology itself. In the final chapter of this thesis I will approach a first perspective on this question.

6 The limits of epistemic reliability: model/data symbiosis and the representational capacity of ML-models

Towards the end of the previous chapter, we have encountered contemporary ML-strategies that require an especially close relationship between models and data, to the point that model-input and model-output converge within a single pipeline. As generative pipelines grow more powerful, more elaborate and efficient methods for producing generated data are likely to emerge, in turn paving the way for ML-pipelines that leverage this generated data to greater extents. On the one hand, this scenario could be interpreted as perhaps model/data symbiosis in the most *literal* sense. On the other hand, the requirements for successful model/data symbiosis as outlined in the literature which, following Bokulich, 'require that the uncertainties associated with both data and models be quantified and appropriately propagated', appear more and more untenable (Bokulich 2020). This warrants a critical look at *how* we could expect generated data and output from models trained on generated data to be informative, and to what extent this breaks with the notions of epistemic reliability that we have associated with model/data symbiosis until now.

6.1 The representational function of models

In our discussions on model/data symbiosis, we have generally stuck to assessing specific modelling contexts and philosophies, and the interdependencies between data and models within those contexts. In this final leg of the discussion, it will be fruitful to take an additional step back and take a more holistic look at what the purpose of models are from an epistemic point of view. Throughout this thesis we have referred to models as representations of a specific target system. This target system can be a concrete physical system with well-understood theoretical foundations, such as in our earlier examples from climate modelling. We can also attempt to model more ambiguous systems for which we do not have a deductively derived theoretical model but instead have large amounts of data we believe are related to this system, such as in the case of large language models. Despite the differences in modelling philosophy, it could be argued that their underlying purpose remains essentially the same, that is, to act as *mediators* between the systems and phenomena of the world and our understanding of them (Morrison and Morrison 2015). Knowledge in this context is mediated because it comes from the *representation* of this system; the representational function of a model allows it to convey information (Morrison and Morrison 2015).

Given the importance of this representational function, it might appear critical that we have a reliable theory of what constitutes a representation. However, the extreme diversity of methods and devices for constructing models makes such a general theory largely infeasible (Morrison and Morrison 2015). Nonetheless, what makes a model informative in a given context is not arbitrary. At least in the case of scientific modelling, the constraining factor is *theory* (Morrison and Morrison 2015; Suárez 2003). To quote Morrisson (ibid.):

> Theory determines what is the right or best way to model/represent the system in question, not the users of theory. The latter determine what they want from the representation, but which representation will deliver on that request is not a decision that rests solely with them.

It is this constraining factor of theory that motivates the emphasis on epistemic reliability that we find in model/data symbiosis; this reliability is greatly invested in the means we have at our disposal to formulate concrete expectations about how a model might communicate information. These constraints then also make it possible to estimate possible sources of uncertainty and error and deal with them appropriately. As such, this assessment of epistemic reliability appears to depend strongly on such constraints being in place. I would like to argue here that the case of model/data symbiosis between generated data and ML-models invites to approach this notion of representation in models in an alternative way, one that takes into account the peculiar mediating role that such models play in communicating information about the world.

6.2 'The third knowledge dimension': AI and model epistemology

In the context of model/data symbiosis, we see the mediating role of models play out in their application as data-processing vehicles; in order to produce more informative data about the world of greater complexity and scale, models are our means to account for the layers of inference we need to traverse when validating, combining and in the case of synthetic data, eventually producing data that aids in this goal. Whether this process of mediation actually makes data more informative depends on to what extent we can assess its epistemic reliability. This degree of success is invested in our knowledge of and means to account for the ways in which model-based data processing might introduce artefacts into the data. And as mentioned in the previous section, what makes a reliable model is largely constrained by theory.

Given our account of the intractabilities of applying these same standards of epistemic reliability to complex ML-models, and especially to the application of generated data in their training pipelines, how do we reconcile this with the unmistakable potential such models have shown in making previously unfeasible research objectives possible? One possible approach I wish to expand on here is to consider the epistemic status of such models and its products in an entirely new, dedicated context. Here I wish to draw upon the work of Ruth Edith Hagengruber, who has suggested this possibility in her work on the 'third knowledge dimension' (Hagengruber 2023). In short, adjacent to the first knowledge dimension of the empirical world and the second knowledge dimension of subjects that can produce knowledge about the world (in other words, human beings), the third knowledge dimension is dedicated to 'machine knowledge'. This machine knowledge, the product of systems capable of constructing their own representations of systems and phenomena in the world, physical or of our own creation Hagengruber argues, will be the end-result of advances in artificial intelligence. As such, inductive models capable of learning distributions of data in ways that we cannot theoretically constrain or estimate with confidence, end up becoming additional 'determinants' in producing knowledge about the world. Taken to its extremes, this perspective essentially treats knowledge produced by AI-systems as epistemically independent and its constituent factors largely inaccessible to us.

The concept of the third knowledge dimension thus serves as an intuitive future scenario in which our established expectations for epistemology, or what we consider accurate and reliable knowledge, might warrant serious reconsideration. However, it is also difficult to imagine how this would play out in the short term, and what kind of tangible consequences this would have for scientific practice altogether. In the last section of this discussion I wish to critically examine this idea in light of current developments in research on the internal representational mechanisms of AI-systems.

6.3 Treating the model *as* the system: model/data symbiosis and mechanistic interpretability

In this third knowledge dimension we thus find systems capable of communicating information about the world that are not simply yet another extension of our means to represent the world around us, but an independent source of representation that communicates *with* us and that we respond to (Hagengruber 2023). How then do we position ourselves against such systems? Here we can refer back to a necessary aspect of representation that, although understated in the literature on model epistemology, in this context acquires new significance: representation, above anything else, requires *cognitive potential* (Morrison and Morrison 2015). Morrisson follows Van Fraassen in suggesting that, regardless of the method behind or shape of a model that has representational capacity, there needs to be an agent capable of cognitive action in order to execute the method or use the objects involved in making the representation happen (Morrison and Morrison 2015; Fraassen 2010). However, as we noted earlier, if the goal is to produce informative representations of systems or phenomena, this process is not arbitrary. Instead, Suarez argues that informative representations should provide 'cognitive value'; they ought to give specific information about a target system/phenomena that could not readily be given by any other arbitrary representation (Surez 2004).

If we were to take the stance that AI-systems have some form of cognitive potential, and if in addition we abandon the belief that we can confidently constrain and estimate what representations such systems develop, then it is not a stretch to assign to such systems an agentic quality. Following Hagengruber's vision, we would thus have to contend with such systems possessing a sense of agency in deciding what is the best possible representation of a system or phenomena based on available knowledge. Here also, again, this is not an arbitrary process, but in this case constrained by the architecture of the pipeline and the nature of the training data. Nevertheless, a crucial point that I argue Hagengruber does not emphasise enough is that the systems themselves are *accessible*. With the appropriate computational means, and even the application of other kinds of models, it is possible to study the internal representational mechanisms of AI-systems, not (yet) in a deductive manner, but in an experimental, empirical manner. Exactly this is the objective of mechanistic interpretability, a field of research in AI that aims to develop a more global functional understanding of how complex inductive models learn, solve problems and generalise to form representations (Sharkey et al. 2025). I argue that in this vein of research we see a more practical application of a third knowledge dimension; in this scenario, complex models themselves are systems of interest, open to empirical study with the goal of gaining insight into the nature of their representational mechanisms. Epistemic reliability in this context thus becomes a much messier affair, more exploratory and experimental, quite unlike the precise quantifications and theoretical deductions of the simulation context.

How does this relate to model/data symbiosis? Going back to where we left off in the previous chapter, it could be said that with the development of ML-pipelines that make more and more circular use of generated data, we are seeing model/data symbiosis happen at level of interdependency that goes beyond what the early literature could possibly account for. In this sense, the principles of model/data symbiosis, that of close relationships between data-laden models and model-laden data as a driving force especially for the epistemic aims of science, are perhaps more important than ever. However, assessing whether these relationships are beneficial or detrimental, whether they lead to reliable and accurate information or whether they lead to artefactual, misleading information, the available theory currently cannot answer. Nevertheless, developments in field such as mechanistic interpretability could go some way in gathering experimental evidence as to how data with different kinds of properties, whether real, simulated or generated, play a role in shaping the representational mechanisms of AI-systems.

7 Conclusion

In this thesis I have worked to apply the theory of model/data symbiosis to applications of synthetic data in ML-model training. Current-day conceptions of synthetic data production driven by generative models stand as the latest modelling context in the longer history of synthetic data applications, tracing back all the way to the theoretical foundations of simulation modelling. I have argued that the theory of model/data symbiosis is a useful framework through which we can analyse the epistemic challenges of employing synthetic data to construct more accurate and reliable models. Model/data symbiosis allows for the careful assessment of how models and data shape each other in the modelling process, yet no prior interpretations of this theory in the context of machine learning exist. For this reason I have laid out the theoretical groundwork necessary to bridge the modelling gap from the simulation context to the machine learning context, in the process establishing how the principles of model/data symbiosis could best be applied to cases in machine learning. This discussion defined the early parts of this thesis.

Extending the theory of model/data symbiosis to include ML-generated sources of synthetic data required a further partition of synthetic data into simulated - and generated data, each associated with their respective modelling architectures. Simulation modelling and machine learning harness two distinctive modelling philosophies, generally corresponding to deductive and inductive approaches to constructing models. Given these differences, I have argued that it is important to recognise that synthetic data produced using either approach yields data with different properties. As both types of synthetic data can be employed for model training, it follows that different kinds of epistemic challenges might emerge from this application. Much of the latter parts of the thesis revolved around discussing specific cases of synthetic data applications in training machine learning models, through the framework of model/data symbiosis. The purpose of discussing these cases is to highlight how model/data symbiosis helps us to make sense of the different epistemic challenges that emerge when using synthetic data to construct models. These challenges primarily concern the two most critical hurdles to successful model/data symbiosis, which are the comprehensive estimation of uncertainty in models and data and the risk of problematic circularity between models and data.

Although the principles of model/data symbiosis do shed light on a number of curious problems associated with synthetic data, such as domain shift and model autophagy, I concluded that it does not seem feasible to assess the most high-level cases of synthetic data in model training with our current knowledge of model/data symbiosis. We thus find ourselves in somewhat of a paradoxical situation: in these aforementioned cases we find arguably the most deeply symbiotic relationships between models and data, yet at the same time model/data symbiosis seems to fall short of dealing with the complexity of these interdependencies. In the last leg of this thesis I suggested that an epistemology of models in the absence of solid theoretical constraints (which characterises to some extent the epistemic transition from simulation to machine learning modelling) might benefit from looking critically at how we engage with the informative potential of current and future AI-models. Placing such models in a 'third knowledge dimension', following Hagengruber's concept, significantly alters the mediating role such models play in our wider system of knowledge production. At this point in the research I cannot yet give any concrete leads on how model/data symbiosis might be integrated in this viewpoint.

To further develop this research in the future, it will be essential to set out more rigorous criteria for how model/data symbiosis might be assessed in the more high-level cases especially. Literature on model/data symbiosis is still sparse and has to this point never been developed outside of its original disciplinary niche. Although Bokulich notes that it would be a mistaken view to limit model/data symbiosis to this discipline of geosciences (Bokulich 2020), it is not clear yet how to substantiate the principles of model/data symbiosis in a way that it can be applied with more confidence to modelling cases from other disciplines. In this thesis I have worked to take some first steps in this process, but more comprehensive studies of a wider variety of cases are necessary. Furthermore, to aid in this I believe that it would be beneficial to gather more qualitative reports regarding how scientists from different disciplines interpret the epistemic risks and payoffs of using synthetic data to construct models. We currently have only a narrow scope of insight into how such issues are discussed and moderated in the actual research process. Applications of synthetic data in model training are currently being explored at immense speeds and in the absence of strong theoretical fundaments through which we can make sense of how these developments affect our understanding of scientific practice and of knowledge production itself. Therefore, detailed ethnographic material might be immensely insightful in painting a picture of this subversive, volatile period from the perspective of philosophy of science.

References

Alaa, Ahmed M. et al. (July 2022). *How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models*. en. arXiv:2102.08921 [cs]. DOI: 10.48550/arXiv.2102.08921. URL: http://arxiv.org/abs/2102.08921 (visited on 05/02/2025).

Alemohammad, Sina et al. (July 2023). *Self-Consuming Generative Models Go MAD*. en. arXiv:2307.01850 [cs]. DOI: 10.48550/arXiv.2307.01850. URL: http://arxiv.org/abs/2307.01850 (visited on 12/05/2024).

AlphaGeometry (Mar. 2025). *AlphaGeometry: An Olympiad-level AI system for geometry*. en. URL: https://deepmind.google/discover/blog/alphageometry-an-olympiad-level-ai-system-for-geometry/ (visited on 03/24/2025).

Betzalel, Eyal, Coby Penso, and Ethan Fetaya (Sept. 2024). "Evaluation Metrics for Generative Models: An Empirical Study". en. In: *Machine Learning and Knowledge Extraction* 6.3. Publisher: Multidisciplinary Digital Publishing Institute, pp. 1531–1544. ISSN: 2504-4990. DOI: 10.3390/make6030073. URL: https://www.mdpi.com/2504-4990/6/3/73 (visited on 08/11/2025).

Bokulich, Alisa (Dec. 2020). "Towards a Taxonomy of the Model-Ladenness of Data". en. In: *Philosophy of Science* 87.5, pp. 793–806. ISSN: 0031-8248, 1539-767X. DOI: 10.1086/710516. URL: https://www.cambridge.org/core/journals/philosophy-of-science/article/abs/towards-a-taxonomy-of-the-modelladenness-of-data/E00CE80C1EBF60CC80047D04073B2DF2 (visited on 04/04/2025).

Breugel, Boris Van, Zhaozhi Qian, and Mihaela Van Der Schaar (July 2023). "Synthetic Data, Real Errors: How (Not) to Publish and Use Synthetic

Data". en. In: *Proceedings of the 40th International Conference on Machine Learning*. ISSN: 2640-3498. PMLR, pp. 34793–34808. URL: `https://proceedings.mlr.press/v202/van-breugel23a.html` (visited on 10/03/2025).

Chakrabarty, Goirik, Manogna Sreenivas, and Soma Biswas (Oct. 2023). "A Simple Signal for Domain Shift". en. In: *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. Paris, France: IEEE, pp. 3569–3576. ISBN: 979-8-3503-0744-3. DOI: `10.1109/ICCVW60793.2023.00384`. URL: `https://ieeexplore.ieee.org/document/10350521/` (visited on 08/05/2025).

DeepSeek-AI et al. (Jan. 2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. en. arXiv:2501.12948 [cs]. DOI: `10.48550/arXiv.2501.12948`. URL: `http://arxiv.org/abs/2501.12948` (visited on 05/23/2025).

Deist, Timo M et al. (Oct. 2019). "Simulation-assisted machine learning". In: *Bioinformatics* 35.20, pp. 4072–4080. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btz199`. URL: `https://doi.org/10.1093/bioinformatics/btz199` (visited on 08/06/2025).

Edwards, Paul N. (Dec. 1999). "Global climate science, uncertainty and politics: Data-laden models, model-filtered data". en. In: *Science as Culture* 8.4, pp. 437–472. ISSN: 0950-5431, 1470-1189. DOI: `10.1080/09505439909526558`. URL: `http://www.tandfonline.com/doi/abs/10.1080/09505439909526558` (visited on 06/02/2025).

Fraassen, Bas C. van (July 2010). "Scientific Representation: Paradoxes of Perspective". In: *Analysis* 70.3, pp. 511–514. ISSN: 0003-2638. DOI: `10.1093/analys/anq042`. URL: `https://doi.org/10.1093/analys/anq042` (visited on 08/18/2025).

Gm, Harshvardhan et al. (Nov. 2020). "A comprehensive survey and analysis of generative models in machine learning". In: *Computer Science Review* 38, p. 100285. ISSN: 1574-0137. DOI: `10.1016/j.cosrev.2020.100285`. URL: `https://www.sciencedirect.com/science/article/pii/S1574013720303853` (visited on 07/15/2025).

Goyal, Mandeep and Qusay H. Mahmoud (Jan. 2024). "A Systematic Review of Synthetic Data Generation Techniques Using Generative AI". en. In: *Electronics* 13.17. Number: 17 Publisher: Multidisciplinary Digital Publishing Institute, p. 3509. ISSN: 2079-9292. DOI: `10.3390/electronics13173509`. URL: `https://www.mdpi.com/2079-9292/13/17/3509` (visited on 07/03/2025).

Hagengruber, Ruth Edith (Aug. 2023). "The Third Knowledge Dimension: From a Binary System to a Three-limbed Epistemology". en. In: *Women Philosophers on Economics, Technology, Environment, and Gender History: Shaping the Future, Rethinking the Past*. Ed. by Ruth Edith Hagengruber. De Gruyter, pp. 119–128. ISBN: 978-3-11-105180-2. DOI: `10.1515/9783111051802-013`. URL: `https://www.degruyter.com/document/doi/10.1515/9783111051802-013/html?srsltid=AfmBOor0DwIOrE0ggP6ZL20wRTKQ2Usrs92z006msDVwhYlmhacpfRGR` (visited on 03/14/2025).

Hao, Shuang et al. (Jan. 2024). *Synthetic Data in AI: Challenges, Applications, and Ethical Implications.* arXiv:2401.01629 [cs]. DOI: 10.48550/arXiv.2401.01629. URL: http://arxiv.org/abs/2401.01629 (visited on 03/23/2025).

Jebara, Tony (2004). "Generative Versus Discriminative Learning". en. In: *Machine Learning: Discriminative and Generative.* Ed. by Tony Jebara. Boston, MA: Springer US, pp. 17–60. ISBN: 978-1-4419-9011-2. DOI: 10.1007/978-1-4419-9011-2_2. URL: https://doi.org/10.1007/978-1-4419-9011-2_2 (visited on 07/15/2025).

Knuuttila, Tarja, Natalia Carrillo, and Rami Koskinen (Aug. 2024). *The Routledge Handbook of Philosophy of Scientific Modeling.* en. 1st ed. London: Routledge. ISBN: 978-1-003-20564-7. DOI: 10.4324/9781003205647. URL: https://www.taylorfrancis.com/books/9781003205647 (visited on 08/11/2025).

Melo, Celso M. de et al. (Feb. 2022). "Next-generation deep learning based on simulators and synthetic data". English. In: *Trends in Cognitive Sciences* 26.2. Publisher: Elsevier, pp. 174–187. ISSN: 1364-6613, 1879-307X. DOI: 10.1016/j.tics.2021.11.008. URL: https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(21)00293-X (visited on 05/02/2025).

Miller, Tymoteusz et al. (Jan. 2024). "AI in Context: Harnessing Domain Knowledge for Smarter Machine Learning". en. In: *Applied Sciences* 14.24. Number: 24 Publisher: Multidisciplinary Digital Publishing Institute, p. 11612. ISSN: 2076-3417. DOI: 10.3390/app142411612. URL: https://www.mdpi.com/2076-3417/14/24/11612 (visited on 07/11/2025).

Morrison, Margaret and Margaret Morrison (Jan. 2015). *Reconstructing Reality: Models, Mathematics, and Simulations.* Oxford Studies in Philosophy of Science. Oxford, New York: Oxford University Press. ISBN: 978-0-19-938027-5.

Munde, Anjali (2024). "The Machine Learning Pipeline: Algorithms, Applications, and Managerial Implications". In: *Deep Learning Concepts in Operations Research.* Num Pages: 18. Auerbach Publications.

Nikolenko, Sergey I. (Sept. 2019). *Synthetic Data for Deep Learning.* en. arXiv:1909.11512 [cs]. DOI: 10.48550/arXiv.1909.11512. URL: http://arxiv.org/abs/1909.11512 (visited on 05/02/2025).

Parker, Wendy (Dec. 2020). "Local Model-Data Symbiosis in Meteorology and Climate Science". en. In: *Philosophy of Science* 87.5, pp. 807–818. ISSN: 0031-8248, 1539-767X. DOI: 10.1086/710621. URL: https://www.cambridge.org/core/journals/philosophy-of-science/article/local-modeldata-symbiosis-in-meteorology-and-climate-science/7A0ECB7DE2FDFCE81EBD2EB05CE7B3B5 (visited on 04/10/2025).

Parker, Wendy S. (Mar. 2017). "Computer Simulation, Measurement, and Data Assimilation". In: *The British Journal for the Philosophy of Science* 68.1. Publisher: The University of Chicago Press, pp. 273–304. ISSN: 0007-0882. DOI: 10.1093/bjps/axv037. URL: https://www.journals.uchicago.edu/doi/10.1093/bjps/axv037 (visited on 04/04/2025).

Reddy, Arun V. et al. (Aug. 2024). *Synthetic-to-Real Domain Adaptation for Action Recognition: A Dataset and Baseline Performances*. en. arXiv:2303.10280 [cs]. DOI: 10.48550/arXiv.2303.10280. URL: http://arxiv.org/abs/2303.10280 (visited on 08/07/2025).

Rueden, Laura von et al. (2021). "Informed Machine Learning – A Taxonomy and Survey of Integrating Knowledge into Learning Systems". In: *IEEE Transactions on Knowledge and Data Engineering*. arXiv:1903.12394 [stat], pp. 1–1. ISSN: 1041-4347, 1558-2191, 2326-3865. DOI: 10.1109/TKDE.2021.3079836. URL: http://arxiv.org/abs/1903.12394 (visited on 07/08/2025).

Sharkey, Lee et al. (Jan. 2025). *Open Problems in Mechanistic Interpretability*. en. arXiv:2501.16496 [cs]. DOI: 10.48550/arXiv.2501.16496. URL: http://arxiv.org/abs/2501.16496 (visited on 08/22/2025).

Stacke, Karin et al. (Feb. 2021). "Measuring Domain Shift for Deep Learning in Histopathology". In: *IEEE Journal of Biomedical and Health Informatics* 25.2, pp. 325–336. ISSN: 2168-2208. DOI: 10.1109/JBHI.2020.3032060. URL: https://ieeexplore.ieee.org/abstract/document/9234592 (visited on 08/05/2025).

Statsenko, Ilya, Nikita Andriyanov, and Oleg Shishkin (Oct. 2024). "Leveraging Knowledge Distillation to Mitigate Model Collapse". en. In: URL: https://openreview.net/forum?id=8TbqoP3Rjg (visited on 03/22/2025).

Suárez, Mauricio (Jan. 2003). "Scientific representation: against similarity and isomorphism". In: *International Studies in the Philosophy of Science* 17.3. Publisher: Routledge _eprint: https://doi.org/10.1080/0269859032000169442, pp. 225–244. ISSN: 0269-8595. DOI: 10.1080/0269859032000169442. URL: https://doi.org/10.1080/0269859032000169442 (visited on 08/20/2025).

Surez, Mauricio (2004). "An Inferential Conception of Scientific Representation". In: *Philosophy of Science* 71.5. Publisher: [The University of Chicago Press, Philosophy of Science Association], pp. 767–779. ISSN: 0031-8248. DOI: 10.1086/421415. URL: https://www.jstor.org/stable/10.1086/421415 (visited on 08/22/2025).

Theis, Lucas, Aäron van den Oord, and Matthias Bethge (Apr. 2016). *A note on the evaluation of generative models*. en. arXiv:1511.01844 [stat]. DOI: 10.48550/arXiv.1511.01844. URL: http://arxiv.org/abs/1511.01844 (visited on 08/11/2025).

Tobin, Josh et al. (Sept. 2017). "Domain randomization for transferring deep neural networks from simulation to the real world". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. ISSN: 2153-0866, pp. 23–30. DOI: 10.1109/IROS.2017.8202133. URL: https://ieeexplore.ieee.org/abstract/document/8202133 (visited on 08/07/2025).

Tremblay, Jonathan et al. (Apr. 2018). *Training Deep Networks with Synthetic Data: Bridging the Reality Gap by Domain Randomization*. en. arXiv:1804.06516 [cs]. DOI: 10.48550/arXiv.1804.06516. URL: http://arxiv.org/abs/1804.06516 (visited on 08/06/2025).

Von Rueden, Laura et al. (2020). "Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions". en. In: *Advances in Intelligent Data Analysis XVIII*. Ed. by Michael R. Berthold, Ad Feelders, and Georg Krempl. Vol. 12080. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 548–560. ISBN: 978-3-030-44583-6 978-3-030-44584-3. DOI: 10.1007/978-3-030-44584-3_43. URL: http://link.springer.com/10.1007/978-3-030-44584-3_43 (visited on 05/14/2025).

Wang, Yinong Oliver et al. (June 2024). "Domain Gap Embeddings for Generative Dataset Augmentation". en. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Seattle, WA, USA: IEEE, pp. 28684–28694. ISBN: 979-8-3503-5300-6. DOI: 10.1109/CVPR52733.2024.02710. URL: https://ieeexplore.ieee.org/document/10657264/ (visited on 08/06/2025).

Winsberg, Eric (July 1999). "Sanctioning Models: The Epistemology of Simulation". en. In: *Science in Context* 12.2, pp. 275–292. ISSN: 1474-0664, 0269-8897. DOI: 10.1017/S0269889700003422. URL: https://www.cambridge.org/core/journals/science-in-context/article/sanctioning-models-the-epistemology-of-simulation/DD7E62F804B04867B8D16043310761AE (visited on 06/02/2025).

Xing, Xiaodan et al. (Feb. 2025). "On the caveats of AI autophagy". en. In: *Nature Machine Intelligence* 7.2. Publisher: Nature Publishing Group, pp. 172–180. ISSN: 2522-5839. DOI: 10.1038/s42256-025-00984-1. URL: https://www.nature.com/articles/s42256-025-00984-1 (visited on 08/06/2025).