HOW CAN FICTIONAL MODELS BE EXPLANATORY?

Lindsay Brainard

A thesis submitted of the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Philosophy.

Chapel Hill
2013

Approved by:
Marc Lange
Matt Kotzen
John Roberts
Abstract

LINDSAY BRAINARD: How can Fictional Models be Explanatory?
(Under the direction of Marc Lange, Matt Kotzen, and John Roberts)

Fictional scientific models include falsehoods that are not captured under the categories of simplification, approximation, or idealization. Recent commentators have found it puzzling that these literally false models are still capable, in some cases of furnishing scientific explanations. In this paper, I spell out a paradox that arises from accepting the explanatory capacity of fictional scientific models and provide a way of understanding these models that resolves the paradox. My proposed solution is that we should understand fictional scientific models as a type of paraphrasable metaphor. I apply my solution to three case studies and conclude on the basis of its successful application that my proposal provides a better way of understanding these models than the competing accounts I consider.
Table of Contents

Introduction .......................................................................................................................... 1

Chapter I: Identifying Explanatory Fictional Models .......................................................... 3

1. What are fictional models? .............................................................................................. 3

   A. Fictions contain falsehoods that are not limited to simplifications, approximations, and idealizations ........................................................................... 3

   B. Fictions do not represent our best guess at what the target system is really like .................................................................................................................. 6

   C. A Note of Qualification ............................................................................................ 7

2. Some Non-fictional models ............................................................................................ 8

3. Three cases of Fictional Models .................................................................................. 11

   A. Electron orbitals in chemistry ............................................................................... 12

   B. Memory as a pushdown stack in neuroscience ..................................................... 17

   C. Maxwell's mechanistic model of the electromagnetic field .................................. 20

Chapter II: On the Explanatory Capacity of Fictional Models ....................................... 23

1. Introduction .................................................................................................................. 23

2. The Paradox of Explanatory Fictional Models ............................................................ 23

3. Resolving the Paradox by Rejecting the Truth Requirement: Bokulich's Account .............................................................................................................. 25

4. Fictional Models as Metaphors .................................................................................. 30

   A. Primary versus Secondary Subject ....................................................................... 31

   B. Paraphrasable Metaphors versus Unparaphrasable Metaphors .......................... 32
C. Understanding Fictional Scientific Models as Metaphors ........................................ 36

5. Fictional Models as Metaphors: Implications for Explanatory Capacity ............ 42
   A. Lesson One: The paraphrase, not the literally false model, really explains ....... 43
   B. Lesson Two: Only the true parts of fictional models explain ......................... 45
   C. Conclusions from (L1) and (L2) .................................................................. 47

6. Final Analysis of the Cases .................................................................................. 49
   A. Electron Orbitals Model: Explaining valence .............................................. 49
   B. Maxwell's Mechanistic Model of the Electromagnetic Field:
      Explaining a Possibility .................................................................................. 52

7. Conclusion ............................................................................................................ 57

Bibliography .............................................................................................................. 59
Introduction

Recently, philosophers of science have devoted considerable attention to the presence of fictions in scientific modeling.\(^1\) Some philosophers have suggested that there are instances in which the explanatory capacity of models with fictional elements is not diminished by the presence of these fictional elements.\(^2\) Often, these discussions focus on questions of whether one can both accept fictional models as explanatory and be a scientific realist.\(^3\) For present purposes, I will assume the truth of scientific realism.\(^4\) Moreover, though there is reasonable debate to be had on this point, I will take for granted in this paper that models can provide scientific explanations. I propose to answer the question: How can it be that a fictional scientific model furnishes a genuine scientific explanation?

In Part I, I will first demarcate the class of cases that interest me (I.1). For comparison's sake, I will consider some examples of models that admit of some

---


\(^3\) For a helpful overview of the purported tension between fictions and realism, see Levy (2012).

\(^4\) I take it that the success of my account does not depend on the truth of realism. However, the challenge of accounting for how fictional models can be explanatory is greater within a realist framework, so I will seek to answer this greater challenge. It would, presumably, be easier within an anti-realist framework to account for the phenomenon of explanatory fictions, both because anti-realists tend not to consider explanation to be of great importance in scientific practice and because anti-realists do not take science to regard its theories about unobservables as true.
familiar types of falsehoods without exhibiting the more radical falsehoods that qualify models as fictional in the sense that interests me (I.2). I will then present three cases of fictional models and draw out some common features (I.3). In Part II, I will present a paradox that arises from considering these fictional models to be explanatory (II.2). I will then consider one recent proposal – from Alisa Bokulich – that would resolve the paradox, but at what I take to be too great a cost (II.3). I will then provide my own account of how fictional models function (II.4) and show how it resolves the paradox (II.5). In brief, I will say that we should interpret models with fictional elements as a class of metaphors, and that doing so will show that no genuine paradox arises concerning the explanatory capacity of these fictional models. Finally, I will show that my analysis applies seamlessly to the three cases I considered in I.3 (II.6).
Chapter I: Identifying Explanatory Fictional Models

1. What are fictional models?

Despite the considerable attention that has been paid to the role of fictions in scientific modeling, there is no widespread consensus about the precise definition of a fictional model. Thus, it is first necessary for me to stipulate the class of cases to which I will apply the term. While I think my account lines up with most of the cases that philosophers have labeled fictional, it may exclude some models others categorize as fictional and may include some models other categorize as non-fictional. To delineate the models I consider fictional, I will now present two criteria. If a model meets either criterion, it is not a fictional model on my account. Models that fail to meet both criteria, but appear nonetheless to provide or feature in good scientific explanations, are the fictional models I will examine.

A. Fictions contain falsehoods that are not limited to simplifications, approximations, and idealizations.

In a trivial sense, all models are false in some ways. For instance, they necessarily leave out some details of their target systems, the scientific phenomena the models seek to represent. All models feature simplifications, and many feature idealizations, and approximations as well. By simplifications, I
mean omissions of some details of target systems that are irrelevant to the purposes for which individual models are devised. For example, a model of the human circulatory system devised to explain blood oxygenation will leave out some capillaries. I take approximations to be mathematical imprecisions, usually in the form of estimations or probabilities. For example, a model of the solar system might construe the earth as a sphere, even though its actual shape is not perfectly spherical. The term “idealization” is a bit more amorphous. It applies broadly to distortions of the target system that serve various pragmatic purposes (e.g. to make aspects of the system more easily comprehensible, or to include only those features necessary for a close approximation of mathematical features).

I will distinguish idealizations that I consider fictional from idealizations that I consider non-fictional by relying on a taxonomical tool introduced by Michael Weisberg (2007). Weisberg helpfully divides models sometimes categorized as idealizations into three categories: Galilean idealizations, minimalist idealizations, and multiple-models idealizations. He differentiates these categories primarily on the basis of their representational aims. When I speak of idealizations in this paper, I will refer to Galilean idealizations and minimalist idealizations only. All the models I consider will be single models taken to be internally consistent, whereas multiple-models idealizations are combinations of incompatible models of the same phenomenon. Thus none of the models I discuss will be eligible for categorization as multiple-models idealizations.

Galilean idealizations are those that distort the target system in order to make it more “computationally tractable” (2007, p. 3). They admit falsehoods only insofar as doing so is necessary for the model to be useful in fulfilling its representational aims.
According to Weisberg, the defining representational aim of a Galilean idealization is *completeness*, the accurate and comprehensive depiction of the target system. Whatever distortions are made in Galilean idealizations, they are made to serve this aim (though, of course, truly complete representation is presumably unreachable for a model). As he puts it, “models generated by Galilean idealization are [...] approximate, but carry with them the intention of further revision, ultimately reaching for a more precise, accurate, and complete model” (p. 19). Minimalist idealizations, on the other hand, include distortions that purport to isolate the features of the target system that are causally responsible for the phenomenon in question. In Weisberg's words, “a minimalist model contains only those factors that make a difference to the occurrence and essential character of the phenomenon in question” (p. 5). Their representational aim, then, is to provide causal information. Specifically, Weisberg claims, they are designed to isolate *one* causal factor (p. 19).

This digression into Weisberg's taxonomy of idealization is meant to illustrate the sorts of representational aims that are conventionally taken to be characteristic of idealizations. Although the broad definition of idealization may appear to subsume the cases of fictions I will present, I hope to show, with reference to these typical representational aims of idealizations, that the cases I have in mind are importantly different from these conventional idealizations. Their aims are not the aims we typically understand as characteristic of idealizations. In my discussion of fictional models, I will try to illustrate respects in which the aims of such models diverge from the aims of completeness or causal isolation.

It is important to note that these categories – simplification, approximation,
idealization – are not mutually exclusive. A single model may feature simplifications, approximations, and idealizations, or any combination thereof. The relevant feature that unifies these three categories is that, in spite of some disagreement about which models fall into which categories, the fact that all models will fall into one or more of these categories is widely accepted as an inevitability of scientific modeling. Although it is not universally agreed upon, a common intuition is that these are unproblematic features of a scientific model. If there are fictional models of the sort that interest me in this paper, they will be inaccurate in some further way. Specifically, they will feature representations of entities that are not taken to exist – even in approximate, simplified, or idealized forms – by the scientists who employ the models. Thus, I will have to show that the falsehoods in any example of a scientific model with fictional elements that I present are not falsehoods merely in virtue of being simplifications, approximations, or idealizations.

Therefore, my first criterion that renders a model non-fictional is that the model's only falsehoods are simplifications, approximations, or idealizations (understood as distortions made with the representational aims noted above). I will call this the simplification, approximation, or idealization (hereafter SAI) criterion.

B. Fictions do not represent our best guess at what the target system is really like.

My second criterion is, at least prima facie, more straightforward than my first: a fictional scientific model must contain components that are acknowledged by those who employ the model and subscribe to the accompanying theory not to correspond to our best guess about the reality of the target system. In many cases, fictional models
straightforwardly contradict accepted scientific theories (e.g. some renderings of electron orbitals). In other cases, fictional models contain elements that fail to constitute our best guess about the target system because we have no real evidence to support elements of the model (e.g. Maxwell's mechanistic model of the electromagnetic field). So, fictional models can fail to represent scientists' best guess in at least two ways: they can be false by the governing theory's lights, or they can, as Eric Winsberg puts it, be *unconcerned* with the truth (2009). Here again, divergence from the theory's best guess must not merely consist of simplifications, approximations, or idealizations as explained above, in order for the model to be classified as fictional on my account.

If every component of a model is consistent with the governing theory's best guess, the model will satisfy what I call the *best guess criterion*. If a model satisfies the best guess criterion, it is non-fictional.

C. A Note of Qualification

It is important to note that I will be focusing on particular instances in which scientists put forth models with fictional elements and properly regard them as explanatory. Whether some models qualify as fictional or non-fictional on my account will depend upon facts about the particular contexts of their introduction. For instance, it may be that scientists sometimes speak fictionally about electron orbitals in chemistry, and sometimes speak non-fictionally about them. Whether a model is rightly deemed fictional will depend on whether the model in question is introduced in such a way that it satisfies my two criteria. As I will show in section IV, in some cases, as with electron
orbitals models, scientists vacillate, sometimes employing a fictional model of the phenomenon in question, sometimes employing a non-fictional model.

2. Some Non-fictional models

As a foil to the cases of interest, it is helpful to think of a few models that are straightforwardly non-fictional. Consider, for instance, modeling practices in ideal gas theory. Ideal gases are modeled as collections of volume-less point particles that move randomly and do not exert forces on one another. Ideal gas models do not meet the best guess criterion. Our best guess about the reality of the target system is that gas molecules have volume, do exert forces on one another, and so on. However, models depicting how ideal gases behave are non-fictional because they are only false insofar as they are idealizations. Specifically, I take ideal gas models to be Galilean idealizations. The distortions made in ideal gas models are made with the aim of completeness. Ideal gas models are, so to speak, steps on the way to our best theory. Scientists set some values to zero that will turn out only to be very near to zero; gas particles are modeled as though they have no volume, when in fact they do; and so on. These omissions serve the aim of completeness because the models “carry with them the intention of further revision” (Weisberg 2007, p. 19). Thus, ideal gas models meet my SAI criterion. They are non-fictions.

There are also non-fictional models in which the SAI criterion is not met, but the best guess criterion is. However, of necessity, we can only identify these in hindsight, after the falsehoods have been identified and the prior best guess replaced by a new best
guess. If such a model had been acknowledged to be false at the time of its employment, it would not have satisfied the best guess criterion. Though models featuring Ptolemaic epicycles may have satisfied the best guess criterion at one time, the same model presented today would not meet this criterion and thus not qualify as non-fictional. Because the falsity of Ptolemaic astronomy was not known to the scientists who subscribed to such models, models employing Ptolemaic epicycles to explain celestial movement were non-fictions at the time of their acceptance. These models had false elements – namely epicycles – that were false not merely in virtue of being simplifications, approximations, or idealizations. That epicycles were responsible for observations of retrograde planetary motion was considered astronomy's best guess at the time. So these models met the best guess criterion. This case highlights an important aspect of how I mean to apply the terms “fiction” and “non-fiction.” Namely, the correct application of these terms will be relative to the scientific contexts in which the models are introduced.\(^5\)

Many cases of non-fictional models will satisfy both of my criteria. For instance, Thomas Schelling produced a series of models to explain the phenomenon of inevitable racial segregation in neighborhoods whose residents did not want to be in a racial minority, even when they were motivated by a preference to live in racially integrated communities (1969). Here he describes the simplest explanation presented by the

---

\(^5\) It is worth noting that some models could qualify as fictional without being of interest to the present investigation because they are not put forth as genuine explanations. For instance, if models featuring Ptolemaic epicycles were still employed today as a means of making predictions – though acknowledged not to feature in answer the relevant “why?” questions typically associated with scientific explanation – they would fail to satisfy both my SAI criterion and my best guess criterion, but there would be no mystery about how they remain explanatory in spite of being fictions. They would not be explanatory because they would not be put forth to serve the purpose scientific explanations are put forth to serve.
simplest of his models:

Within a given set of boundaries, not both groups (colors, sexes) can enjoy numerical superiority. Within the population as a whole, the numerical ratio is determined at any given time; locally, in a city or neighborhood, a church or a school, either blacks or whites can be a majority. But if each insists on being a local majority, there is only one mixture that will do it: complete segregation. (p. 489)

Schelling showed that, in spite of their holding preferences for integration, as long as the individual residents were guided in their housing choices by an overruling preference not to be in a racial minority with respect to their immediate neighbors, segregation would develop over time.

Schelling's models involve mapping out hypothetical neighborhoods on grids and lines, a practice that constitutes an instance of idealization. In this case, the model features both Galilean and minimalist idealizations:

For illustration, define everybody's “neighborhood” as extending four neighbors on either side, and suppose that everyone is content if half his “neighbors” are the same color as he. If fewer than half are his color, he moves in either direction to the nearest point (measured in the number he passes on the way) at which half his eight nearest neighbors are the same color as he. (1969, p. 490)

We know that most real neighborhoods are not perfectly grid-like, nor are they usually arranged on a straight line. But because actual neighborhoods come in all manner of configurations, some stipulations and assumptions (distortions) must be made to make the phenomenon computationally tractable. In this respect, Schelling's model presents a Galilean idealization. The representational aim of modeling neighborhoods as grids or straight lines is still completeness. It is not that there is a more accurate way to represent the configuration of neighborhoods that is being ignored for some other purpose. Rather, we do not think there is some uniform arrangement of neighborhoods that we could
identify and add to the model to make it more complete. With the aim of representing the phenomenon of interest in this case as completely as possible, a grid or a line is perhaps the best we can do.

Furthermore, people do not make decisions to move in such simple and predictable ways. Human decision making is multi-faceted and complex. Even if we hold the preference ranking identified by Schelling, this ranking cannot hope to exhaust all the relevant preferences humans have about where they live. We might not move to the closest available house in which we are not the minority if we do not like the size of its backyard, for example. However the model is designed in part to isolate and develop one causal factor that explains racial segregation. There are assuredly other causal factors that this model does not seek to represent. The model thus contains a minimalist idealization. The salient falsehoods in Schelling's models are not false in some further way (beyond simplification, approximation, and idealization), and thus the models satisfy my SAI criterion. Nothing about these models contradicts the background theoretical framework or represents something other than our best guess about the reality of the target system, so they also satisfy the best guess criterion.

3. Three cases of Fictional Models

I will now present three cases of models that satisfy neither my SAI criterion nor my best guess criterion and thus qualify as fictional on my account. In this section, I will describe the models and provide some evidence that scientists take (or took, in the case of my historical example) the models to be both fictional and explanatory.
A. Electron orbitals in chemistry

Scientists often employ fictions in modeling atomic and molecular electron orbitals. Such models are used to explain, among other things, chemical valence. Scientists regularly speak as though orbitals have features that are not consistent with the accepted theoretical framework to which they are committed, namely, one in accord with most standard interpretations of quantum mechanics. Critically, it is often the case that scientists depict orbitals as having properties inconsistent with quantum mechanics while simultaneously acknowledging this inconsistency. Moreover, they take the very model they acknowledge to be fictional to furnish genuine scientific explanations.

The depiction of electron orbitals as configurations of electrons in atoms and molecules is a fundamental starting point for contemporary students of chemistry. In the simple case of atomic models, electron orbitals are generally modeled as “shells” or “energy levels” surrounding nuclei. Each orbital is meant to be a clearly demarcated region in which either one or two electrons is likely to be located (Keeler and Wothers 2003, p. 36). Orbitals correspond to wavefunctions that describe the probabilistic arrangement of electrons in the system (p. 39). Electrons are said to “fill” orbitals in a designated order, and the electron configuration of a particular atom or molecule will specify which orbitals are occupied (p. 35). Once we have the electron configuration of an atom, we can determine its valence, and thus make predictions about chemical reactions.

---

For some time now, it has been acknowledged that the literal posit of orbitals featuring in some of the models that contemporary chemists put forth as explanatory is inconsistent with aspects of quantum mechanics. Eric Scerri explains:

It is now known that the view of electrons in individual well-defined quantum states represents an approximation. The new quantum mechanics formulated in 1926 shows unambiguously that this model is strictly incorrect. The field of chemistry continues to adhere to the model, however. (1991, p. 317)

It is inaccurate, according to most interpretations of quantum mechanics, to describe individual electrons as being in individual, specifiable spatial locations. Chemists acknowledge that we must view the wavefunctions associated with orbitals as approximations rather than as fixed delineations of an electron's spatial position, but these wavefunctions are still treated as approximations of a specifiable electron configuration, as though they provide us with our best guess about the spatial location of the electrons (Keeler and Wothers 2003, p. 39). Yet, there is no such specifiable spatial arrangement of electrons, according to the governing theory. There is no fact of the matter about where the electrons in a system are located. Rather, the wavefunctions associated with orbitals designate the probability of detecting electrons in regions of the system. Contemporary chemists are, of course, keenly aware of this. As one common introductory chemistry text – *Why Chemical Reactions Happen* by James Keeler and Peter Wothers – clarifies: “the electron is not localized to a particular region nor is it going round a particular orbit; rather it is smeared out over space” (2003, p. 41). Notice, however, that even this more careful description of an electron's being “smeared out over space” could easily be

---

7 There are interpretations of quantum mechanics that do attribute positions to electrons. On David Bohm's interpretation, for example, all particles have positions (1965).
interpreted as a description of a determinate spatial location (i.e. it is natural to read the latter part of the sentence as suggesting that the electron is smeared out over a particular region of space).

The critical point to note about the orbitals case is that, though scientists recognize that orbitals are merely approximations of a sort, they do not always model them as such. The cases in which scientists model orbitals in some further false way (beyond modeling them as simplifications, approximations, and idealizations) are the cases I consider fictional. Orbitals are not always modeled by scientists as fictions, but in many cases they are. Models featuring orbitals often reflect properties (such as spatial location) that are no longer part of our best guess about the properties of electrons, though they may have been our best guess at one time (for instance, in the classical Bohr model of the atom, in which orbitals were presented as the literal, definable trajectories of electrons around the nucleus of an atom).  

Orbitals tend to be modeled fictionally when put to certain explanatory uses (e.g. in modeling electron configurations to explain valence). See, for example, the following explanation in a popular contemporary chemistry textbook:

> When we consider the chemical reactions of atoms, our attention is usually focused on the distribution of electrons in the outer shell of the atom [...] This is because the outer electrons (those in the outer shell) are the ones that are exposed to the other atoms when the atoms react. The inner electrons of an atom, called the core electrons, are buried deep within the atom and normally do not play a role when chemical bonds are formed. (Brady and Senese 2004, p. 331)  

---

8 Alisa Bokulich provides a helpful historical overview of how scientists (including Neils Bohr) came to view the classical orbitals model as fictitious (2009).

9 This fictional modeling of chemical orbitals appears not to be limited to textbooks; Scerri argues that orbitals are treated as real entities by contemporary chemists both in atomic spectroscopy and in the calculation of an atom's energy (1991, p. 319-9). In at least the latter case, a model that treats orbitals as
In this passage, electrons are overtly described as being spatially located. Chemists could speak non-fictionally here instead and explain the phenomenon of valence without providing a fictional model, but it seems there are incentives to retaining the fictional language. For instance, while Scerri cautions against chemists putting too much faith in the continued success of this fictional model, he acknowledges that it does a better job than any other available (and perhaps didactically accessible) model of conveying reliable information about the chemical properties of atoms and molecules (1991, p. 321). One clear practical benefit of modeling chemical valence in this fictional way is that doing so provides a more intuitively comprehensible picture. It is, it seems, more difficult to convey complicated interactions between probability density functions (the current best guess about what orbitals actually are) than it is to inure someone to science's best guess about the nature of electrons by initially portraying them as spatially located objects moving between clearly defined energy levels that resemble planetary orbital trajectories. There is thus a didactic benefit of speaking fictionally in this context, and doing so may be unproblematic because the fictional model has a great deal of predictive power – in fact just as much as our best non-fictional model of valence would.

Could these cases of modeling electron orbitals actually just be approximations, idealizations, or simplifications? As I noted above, the wavefunctions associated with orbitals are acknowledged to be approximations of a sort. However, as they are presented in these models, orbitals are not approximations. They are not merely mathematically imprecise. Rather, they are a false description of the target system. They are not our best real entities appears to be the only practicable approach in current use (p. 321).
guess at the way the target system actually is because we know, given standard interpretations of quantum mechanics, that there are no specifiable spatial-defined arrangements of electrons. Such arrangements are what orbitals in these models purport to depict, and so we must conclude that they are fictional.

Put simply, orbitals cannot merely be approximations of electrons' definite positions because we know that electrons do not have definite positions. One might ask, then, why an orbital cannot be an approximation, simplification, or idealization of something else, like the likelihood of detecting an electron at a specified spatial location if we took a measurement. This is certainly how orbitals are sometimes described. When chemists describe what they think orbitals really are, they say they are really probabilistic wavefunctions whose shapes are formed by the distribution of regions in which we would likely find an electron over many measurements. Speaking of orbitals this way is non-fictional because both the best guess and SAI criteria are met. But this is consistent with chemists modeling orbitals fictionally in other cases. It seems that chemists often use orbitals as fictions and tell fictional narratives about them (while at the same time acknowledging the non-fictional story that is more accurate). When we are specifying the electron configuration of an atom to determine its valence, we rely on orbitals (in the classical sense) and a story about their shapes, how they fit together, and how they get filled in a certain order. We want to know about the configuration of electrons in the outer, highest energy level “shell” so we can explain bonding. Speaking in this way, about inner and outer shells and the definition locations of electrons within those shells, is speaking of orbitals as fictional entities rather than as probability density functions. In contexts for which it is made clear that orbitals are being spoken about in ways that are
consistent with chemistry's best guess about electrons, not as entities with spatial properties, orbitals are spoken about non-fictionally. The details of a particular case settle the matter of whether the model in question is fictional.

B. Memory as a pushdown stack in neuroscience

The model of human memory as a pushdown stack, a “serial self-terminating process” was originally proposed by John Theios, et al (1973). The model is designed to explain the hierarchical, last-in-first-out, patterns observed in experimentation on human memory. Consider the following description of the model from an introductory text on probabilistic modeling of psychological processes:

The model derives its name by assuming that memory is arranged as a pushdown stack, that is, that there is a hierarchical ordering to memory with the more frequently presented items tending toward the top of the stack. The popular analogy here is to the almost magical stack of lunch trays so often encountered in cafeterias. We pick a tray up and the rest of the stack rises until its place is taken by the newly exposed tray. Similarly, when a tray is replaced on top of the stack, the whole stack settles down somewhat so that the topmost tray is almost always at the same level. In a pushdown stack, trays are accessed on a last-in, first-out basis. (Townsend and Ashby 1983, p. 134)

This model has had a great deal of predictive success in experimental contexts, especially in experiments about short-term memory. In particular, it explains well the probability that a particular stimulus will be recalled, taking into account the serial order in which it was presented to a subject. Thus, the model is taken by neuroscientists to be explanatory, to accurately depict something about the structure of memory retention and recall that accounts for why we observe the patterns we do.
However, though scientists take the model to proffer a genuine scientific explanation, they acknowledge that it is a false model in a number of ways. For instance, Juliana Goschler points out that no one thinks we've ever encountered anyone whose memory is *full*, though the model indicates that memory has a finite capacity (2007 p. 13). So though a pushdown stack can be full, scientists do not, by and large, think the same is true of human memory. Also, items in memory are importantly different from trays in a stack; you can re-represent to a subject an item that is already in her memory, for instance, whereas trays in a stack each have one instance only. And when you do re-represent something to someone, that item does not necessarily move to the “top of the stack,” as it would have to in the tray case (Townsend and Ashby 1983, p. 135).

Moreover, this model is an instance of a theme in neuroscientific modeling in which brains are depicted as containers and memory depicted as storage space that is filled. Goschler argues that this language is pervasive in neuroscientific modeling, even though scientists, when pressed, would not be able to earnestly ground the attribution of spatial properties to memory and other brain functions (pp. 11-14).

Does the pushdown stack model for memory satisfy either of my criteria for a non-fictional model? Are all the falsehoods I have identified in this model merely simplifications, approximations, or idealizations? Perhaps the model's depiction of memory as finite in capacity represents an approximation of a sort, but to think this, we would have to think there was something like “the capacity of human memory” that the model could be approximating. As this seems to rely on the suspect depiction of a brain as a container, it seems that this manner of calling this falsehood an approximation misses the mark. Moreover, the model includes further falsehoods about memory that cannot, by
any stretch, be construed as mathematical estimations, such as the depiction of memory as a process for which each item can be represented only once in the hierarchical ordering.

Could this model be an idealization – a distortion with the representational aim of completeness or isolating a cause? It is difficult to see how either of those representational aims would map onto this model. It is hard to see how models of memory as a pushdown stack could be distortions that “carry with them the intention of further revision” (Weisberg 2007, p. 19). There does not appear to be a way to revise the model such that the fictions are removed without the features the model is introduced to communicate – such as its qualities of being serial and self-terminating – being also jettisoned. Likewise, it is not clear a single cause is being identified by the model. Rather, it seems that the representational aim of the model is to draw out several features of memory's structure.

Does the model satisfy the best guess criterion? The answer to this seems to be a straightforward no. To name just one respect in which it violates this criterion, it is part of scientists' best guess about the target system that memory items can be re-represented. This is contradicted by the model. Moreover, it is not part of scientists' best guess about memory that it is finite, so the model represents something other than the best available guess about the reality of the target system. Thus, I take scientists who apply the pushdown stack model to memory to be presenting a fictional model.
C. Maxwell's mechanistic model of the electromagnetic field

James Clerk Maxwell, in his 1861 paper "On Physical Lines of Force," modeled the electromagnetic field as an arrangement of simple machines including gears, cogs, and idle wheels (1997). Here is a passage from the paper, in which he discusses the introduction of an idle wheel to the model:

It is difficult to understand how the motion of one part of the medium can coexist with, and even produce, an opposite motion of a part in contact with it.

The only conception which has at all aided me in conceiving of this kind of motion is that of the vortices being separated by a layer of particles, revolving each on its own axis in the opposite direction to that of the vortices, so that the contiguous surfaces of the particles and of the vortices have the same motion.

In mechanism, when two wheels are intended to revolve in the same direction, a wheel is placed between them so as to be in gear with both, and this wheel is called an "idle wheel." The hypothesis about the vortices which I have to suggest is that a layer of particles acting as idle wheels, is interspersed between each vortex and the next, so that each vortex has a tendency to make the neighbouring vortices revolve in the same direction with itself. (1997, p. 158)

The aim of the model is to explain the apparently action-at-a-distance phenomena we observe in electricity and magnetism. Maxwell is describing a physical system made up of simple machines that allows us to explain the character of the phenomena we observe in these contexts (e.g. why a magnet (or an iron filing) moves in a particular direction when placed in a particular relation to another magnet).

On what appears to be the most plausible interpretation of his work in this paper, Maxwell did not think there were actual material mechanisms in magnetic fields. Thomas Simpson explains:
Maxwell indulges in the invention of wheels and gears to fill the vacuum and do his bidding in a way that Rube Goldberg might envy. The reader must be the judge, but I submit that Maxwell has tongue in cheek here and does not seriously intend that the devices he conjures up would in fact be found to exist in the form he is describing [...]. Maxwell is claiming latitude to think things through to the end, to pursue the question whether any imaginable physical system could accomplish what electromagnets and pith balls are regularly observed to do. What is at stake may be the very concept of an intelligible relation of cause and effect in nature. (Simpson 1997, p. 40)

Taking Simpson's interpretation for granted, the point of Maxwell's model was not to accurately describe what is going on in the electromagnetic field. It was to show that it is possible to have a system that obeys all the fundamental laws of physics and functions like the electromagnetic field does. The point was neither to make predictions nor to provide, say, causal information about a particular empirical phenomenon. It was to give an argument that something analogous to the electromagnetic field is possible. Maxwell shows that there is at least one intelligible explanation of the electric and magnetic phenomena we observe, even if the actual model does not advance our understanding of the field beyond “something we know not what” that meets the basic demands of physical possibility.

I take it to be quite clear that Maxwell's model does not satisfy the best guess criterion. It does not appear to be the case that Maxwell means to foster any confidence that the electromagnetic field actually contains idle wheels. It seems equally clear that this model is not a simplification (a mere omission of detail) or an approximation (a mathematical imprecision). Could it then be an idealization with a representational aim either of completeness or of identifying a single cause? These seem inapt descriptions of completeness or identifying a single cause? These seem inapt descriptions of completeness or identifying a single cause? These seem inapt descriptions of completeness or identifying a single cause?

---

10 Eric Winsberg provides an overview of some “something we know not what” fictional posits (e.g. silogen atoms) employed in models of nanoscale solids for purposes of simulating and designing nanoelectromechanical systems (2009).
the model's representational aims. The aim is to represent a possibility for the sake of establishing that such a possibility exists, not to advance our understanding of the particular features of the electromagnetic field. Moreover, given that Maxwell does not take gears and cogs to be real causes of the electric or magnetic phenomena we observe, it does not seem that this model could aim to isolate a cause of these phenomena. Thus, I conclude that Maxwell's model meets neither the best guess criterion nor the SAI criterion, and is therefore rightly categorized as a fictional model.
Chapter II: On the Explanatory Capacity of Fictional Models

1. Introduction

In what follows, I will sketch a way of understanding fictional models that dispels the air of mystery some recent commentators have sensed in the fact that some such models appear to be explanatory in spite of being fictional. I will suggest that the best way to understand the class of fictional models I have identified is to see them as a type of metaphor. A brief investigation into how metaphors of this type function will shed some light on the respects in which these models are fictional and the respects in which they are explanatory. I aim to show that considering fictional models from this more perspicuous vantage point ultimately dissolves the oft-perceived tension between their categorization as fictional and their categorization as explanatory. When properly understood, these models turn out not to constitute a special class of explanatory entities that requires a special notion of explanation, as some theorists have suggested.

2. The Paradox of Explanatory Fictional Models

In Part I, I discussed some widely accepted ways in which models can be false without losing their explanatory capacity. These were cases in which the models were false only in virtue of involving simplifications, approximations, or
idealizations (SAI). While there remain unanswered puzzles about how falsehoods of these sorts feature in scientific explanations, it is generally taken for granted that models can be explanatory despite containing these types of falsehoods. In my brief examinations of three putative fictional models, I concluded that it looks like there really are scientific models – both ones currently in use and ones that have served important roles in the history of science – that are (or were) acknowledged by those who employ (or employed) them to be false in some deeper way. The questions of central interest to the remainder of my investigation are whether and how something with a deeper sort of falsehood can nevertheless count as explanatory. After all, as I demonstrated in the previous section, there is an intuitive sense in which some fictional models really do appear to be explaining phenomena in their target systems. Given the persistence in scientific practice of models acknowledged to be fictional, it also appears as though the scientists who employ them take them to be capable of furnishing scientific explanations.

One way to understand the puzzle I hope to solve is by spelling it out as a paradox. There are three mutually inconsistent claims, each of which appears true on its own, that constitute the paradox of explanatory fictional models. These are (1) that the fictional models are false, (2) that the fictional models explain, and (3) that only true depictions of target systems can explain.\footnote{I owe this formulation of the paradox to Matt Kotzen.} \footnote{In (1), I take “false” to mean false in some further way than being one of SAI. Likewise, in (3), I take “true” to mean true apart from including inevitable instances of SAI.}

The most straightforward resolution of this paradox would be to give up one of these premises altogether. An approach that is gaining some momentum in the literature is to abandon (3). Alisa Bokulich takes this path, for instance (2008, 2009, 2011, 2012). As I
will discuss in section 3, she argues that we should see fictional explanation as a special category of genuine scientific explanations in which genuinely false (in a deeper way than merely being one of SAI) depictions of the target system are nonetheless explanatory. I will seek to provide an account of how fictional models can be explanatory that avoids the abandonment of (3). I will assume that, given the *prima facie* plausibility of (3), we should prefer an account of explanatory fictions according to which (3) is preserved unless, on pain of contradiction, we are unable to reconcile (3) with our understanding of the important roles these fictional models play in science.

In sections 4 through 6, I will articulate my own account of how we should understand cases of fictional models like the ones I considered in Part I. Rather than abandoning any of these premises altogether, I will argue that (1) and (2), as stated here, are imprecise and therefore obscure what is actually going on in cases of fictional models. I will advocate the adoption of a clearer articulation of the sense in which the models are fictional as well as the sense in which the models explain. Once I present alternate accounts of (1) and (2), it will be clear that the modified versions of (1) and (2) will not generate a paradox with (3). Thus, I conclude that there is no reason to abandon (3) in order to accommodate the role played by fictional models in scientific explanation. I take it as a virtue of my account that it does not call for the rejection of (3).

3. Resolving the Paradox by Rejecting the Truth Requirement : Bokulich's Account

One way to resolve the paradox – a resolution I will ultimately reject – is to remain committed to (1) and (2), but to abandon (3). This is to say that genuinely false
models can be genuinely explanatory, and this is a reason to give up the strict requirement in (3) that only true depictions of target systems can explain. As I noted above, this is the general strategy Alisa Bokulich takes in her account of how fictional models explain (2008, 2009, 2011, 2012).

In her defense of the explanatory capacity of fictions, Bokulich devotes much of her attention to the explanatory capacity of models depicting orbitals classically, as spatially-defined electron trajectories (2008, 2011, 2012). In addition to referencing one of the same cases that I have examined, she appears to be using roughly the same definition of a fictional model that I spelled out above, taking care to differentiate genuine fictions from idealizations, simplifications, and approximations (2008). Thus, I take her to be understanding fictional models in at least the same general sense that I have been understanding them.

Bokulich abandons the true depiction of a target system as a necessary condition for a scientific model's being explanatory (2012, p. 18). Borrowing a term from Philip Kitcher, she reasons that a scientific fiction is explanatory as long as it is legitimately included in the scientific community's “explanatory store” at the time of its employment (2012, p. 17). Because she does not take truth to be a necessary condition for inclusion in the explanatory store, she abandons truth as a necessary condition for a fictional model's being explanatory as well. Without being able to appeal to truth to determine which fictional models are explanatory, Bokulich must provide some other desiderata.

According to Bokulich's account, a fictional model explains only if it meets three

---

13 Kitcher introduces this term to refer to all the scientific argument patterns that scientists legitimately have at their disposal for explaining phenomena (1981). These argument patterns are legitimately employed when the set of things in the explanatory store is maximally unifying.
criteria (taken together, she refers to these three criteria as the “justificatory step”) (2008, 2011, 2012). I take Bokulich to be specifying these three criteria as necessary, but not sufficient, for a fictional model's explanatory success (2012, p. 19). She provides these three criteria as part of more general account in which she identifies further criteria models must meet in order to be explanatory (2011). I will omit these details from the present investigation because they do not address the explanatory capacity of fictional models as such.

The first criterion Bokulich advances is that fictional models, rather than being true, must bear some sort of “contextual relevance relation set by the current state of scientific knowledge, which specifies what sort of entities, states, and processes are potentially relevant to the explanation of the explanandum phenomenon” (2012, pp. 20-21). This is the explanatory store inclusion criterion, and it makes what counts as an explanation on her account epistemically relative. The second criterion is that the scientific community must have specified the domain in which the model is applicable, “specifying where – and to what extent – the model can be trusted as an adequate representation of the relevant features of the world for the purpose(s) in question” (2012, p. 21). The third criterion is that scientists must have some sort of “key” that will allow them to “translate statements about the fictional [...] elements in the model into correct conclusions about the target system” (2012, p. 21). She does not provide an explicit account of what such a key must involve, but she does specify that “the translation key is from statements about the fictions to statements about the underlying structure or causes of the explanandum-phenomenon,” indicating that the translation must not be merely a statement of the explanandum, for instance (2012, p. 20).
By applying these three criteria, Bokulich thinks we will come to understand that classical periodic orbitals are an explanatory fiction whereas fictions that we are antecedently committed to denying as explainers, like Ptolemaic epicycles, are not explanatory fictions (2008, 2009, 2011, 2012). The distinction, in most cases she considers, is determined by the first condition – orbitals remain in the explanatory store in contemporary science whereas Ptolemaic epicycles do not.

So, according to this picture, fictions themselves are explanatory even though they are falsehoods. Bokulich's presentation of the third criterion does not make what she means by “correct conclusions about the target system” explicit, but a plausible reading is that she means for the translational key to translate false statements into true ones somehow. If this is right, fictions can be explanatory only if they are somehow translatable into truths, so they must bear a special relation to the truth of the target system. However, Bokulich insists that it is the fictions themselves, not merely the truths into which they are translated by the key, that are doing the explaining. In this insistence, she gives up premise (3) of the paradox outlined above.

Something worth noting about Bokulich's account of how fictional models explain is that she appears to have a more expansive definition of “explain” in mind than I do. She often defends the explanatory capacity of fictional models by insisting that they are capable of generating genuine knowledge. For example, she says: “Although it is certainly not the case that all fictions can explain, I believe that some fictions can give us genuine insight into the way the world is, and hence be genuinely explanatory and yield real understanding.” (2009, p. 94). She says this specifically in defense of the explanatory capacity of the orbitals case: “although these classical trajectories are also useful
calculational tools, they are not *mere fictions*. Insofar as these closed orbits are giving us genuine insight into the structure of the quantum dynamics, they are *explanatory fictions*” (2009, p. 106). Indeed, throughout her discussions of the explanatory capacities of fictional models and scientific fictions broadly construed, Bokulich differentiates the explanatory fictions from the non-explanatory fictions (in her words, the “mere fictions”) by identifying as explanatory any fiction that gives us “genuine insight” into the truth of the target system. She reiterates repeatedly that a fictional representation is explanatory only if it is able to generate genuine knowledge. I take this to be evidence that she is requiring for explanatory success something like rhetorical success – a model's success at *conveying* the genuine insight, in point of fact.

But for the purpose of this discussion, I want to stipulate that when I ask whether models are explanatory, no part of what I am asking includes whether they are rhetorically successful. There are all sorts of functions that are easily confused with scientific explanation in the strict sense. There is no doubt that fictions can perform a host of important functions in science apart from being explanatory in the strict sense I mean to employ. For instance, they can be didactically useful; that is, they can make a process comprehensible to someone for whom it was incomprehensible before. This seems to be much of what is going on when the orbitals model is presented fictionally to explain chemical valence in chemistry textbooks. However, such functions are not the ones that I would need the fictions to perform in order to deem them scientifically explanatory. In general, I take a model to be explanatory if and only if it provides an adequate answer to the right sort of “why?” question about the target system (where the right sort of question will usually be in the form “Why did x occur?” or “Why does x have some property?”). If
my analysis of Bokulich's usage is correct, she is not unique in using the term “explain” more loosely than I intend to use it in the present investigation. It is thus possible that her rejection of premise (3) in the paradox is a rejection of a much weaker claim than the one I take (3) to be conveying. In any case, I take any account that must reject (3) to be making a significant conceptual sacrifice, either in over-extending the usage of the term “explain” or in rejecting a highly intuitively plausible governing principle of scientific explanation.

4. Fictional Models as Metaphors

I will now present my own account of the explanatory capacity of fictional models, in which I argue that premise (3) from the paradox of explanatory fictional models can be preserved without undermining the explanatory value of these models.

The idea that something strictly false can nonetheless explain is not an alien notion or one unique to scientific contexts. In fact, there is a widely-acknowledged common occurrence of literal falsehoods conveying truths both in literature and everyday language. Metaphors are statements that are literally false but ultimately unhindered in their communicative success by their falsity. It is illuminating, I will show, to understand fictional models of the sort I have described above as a class of metaphors. Thinking of fictional models as metaphors reveals how they can be literally false depictions of a system and at the same time communicate true claims about that system. Given that it is widely accepted that metaphorical language can convey something true without being literally true, it should be the case that, if we can show fictional models to be metaphors, this lends plausibility to the claim that fictional models can be both false and explanatory.
In this section, I will defend the proposal that fictional models can provide scientific explanations because, though they are literally false, they may be metaphorically true.

What would it mean for a fictional model to provide an explanation by serving as something akin to a metaphor? To answer this question, it will first be useful to introduce some terminology introduced to describe how metaphors function to convey meaning.

A. Primary versus Secondary Subject

Philosophers and linguists who study metaphor distinguish the primary subject in a metaphor from the secondary subject. The primary subject is the topic actually under consideration in the context in which the metaphor is introduced. The secondary subject is the subject introduced to suggest an analogy with the primary subject. The secondary subject becomes a sort of tool for saying something about the primary subject. Consider for example the extended metaphor, drawn out in Jacques' famous monologue in Shakespeare’s *As You Like It*, that begins: “All the world's a stage, and the men and women merely players” (1997). Here, the primary subject is unscripted human life and the secondary subject is a theatrical production. Obviously, Jacques is not saying that all human activity is literally scripted or literally taking place in a theater. He is drawing out relevant similarities between life as depicted theatrically and life as it is experienced by people outside of a theatrical context. The similarities amount to a series of claims about the types of experiences humans have over the course of a life (e.g. uncertainty in childhood, moodiness in romantic love, and wisdom in later life) and the many roles an actor is expected to play. Jacques' metaphor conveys true claims about the primary

---

14 This terminology was introduced by Monroe Beardsley (1962).
subject, even though the monologue, taken literally, falsely attributes properties of a staged production (the secondary subject) to unscripted life (the primary subject).

The entire content of a metaphor can be broken down into claims about the primary subject and the secondary subject. In the three cases of putative fictional scientific models discussed above, the primary subjects are the mechanisms underlying chemical valence, short-term memory, and the electromagnetic field. So we should understand the fictional models themselves as being presented in terms of secondary subjects, introduced to reference and suggest analogies with the primary subjects. Both the references to the secondary subjects and the references to the primary subjects are included in the models. According to the analysis I will provide, the falsehoods that render explanatory models fictional will turn out to be attributions of features of secondary subjects to primary subjects.\footnote{It could be that a metaphorical model falsely attributes a property to the secondary subject. I will remain agnostic with respect to whether this would necessarily constitute a failure of the metaphor to convey something true. In at least some cases, a metaphor containing a false attribution of properties to the secondary subject would be the reason the metaphor failed to the have content its author intended. For example, imagine that a metaphor's author knew that dynamite often consists of sawdust soaked in nitroglycerin, but was under the mistaken impression that the sawdust rather than the nitroglycerin was the explosive component. Further imagine that the author, in attempting to convey information about the irascible character of a friend, described him as “sawdust.” In this case, the metaphor does not have the content its author intended. By illustrating such an example, I mean to suggest that it is not the case that all property attributions about the secondary subject are trivially true in virtue of being stipulated. Falsely identifying properties of the secondary subject can be one reason a metaphor fails to have the content it was intended to have. There might be some cases of benign falsehoods in a metaphor's attribution of properties to the secondary subject (ones that do not undermine the metaphor's having the content its author intended), but I am unable to think of any such examples.}

B. Paraphrasable Metaphors versus Unparaphrasable Metaphors

Another important feature of metaphors is that they must be either\textit{paraphrasable} or\textit{unparaphrasable}. In other words, either the content about the primary subject can be
fully expressed in non-metaphorical language, or the content about the primary subject cannot be fully expressed in non-metaphorical language. In many cases, if not in all cases, metaphors seem to provide one way of saying something that could be stated in another way, specifically in non-metaphorical language. Often, it turns out that paraphrasable metaphors employ strictly false language (e.g. property attributions that may be true of the secondary subject but are false of the primary subject) as a kind of shorthand for some other content (about the primary subject) that can be stated in literal terms, but that would be more complicated or tedious to state literally. The benefits of choosing the metaphorical language might be aesthetic, pragmatic, or determined by myriad other interests. To think that that all metaphors function in this way would be to think that all metaphors are paraphrasable.\textsuperscript{16}

Some theorists deny that all metaphors are paraphrasable. These theorists typically identify a richness of expression or evocative character that is not translatable into literal language as the basis for their contention that some metaphors are unparaphrasable.\textsuperscript{17} Some part of the meaning of unparaphrasable metaphors would inevitably be lost in any attempt to translate such a metaphor into literal language, they claim. For instance, you might think that “Juliet is the sun” would lose some of its meaning if translated in any non-metaphorical language because the richness of the comparison would not be exhausted by any literal translation. A metaphor might be unparaphrasable in virtue of being “open textured,” or as Friedrich Waismann put it when

\textsuperscript{16} Stanley Cavell, for instance spells out a version of this claim, according to which we must understand metaphors as paraphrasable in order to understand them as meaningful, though he does not ultimately advocate that all metaphors are \emph{entirely} paraphrasable (1969, pp. 78-79).

\textsuperscript{17} For a helpful discussion of three ways of holding that not all metaphors are paraphrasable, see Stewart (1971).
he introduced the term, the relevant concept is impossible to define “with absolute precision” (1951, pp. 119-124). The meaning of an open textured concept is somehow always open to revision; no specification of its meaning can be taken to be exhaustive. To hold that some metaphors are unparaphrasable is not to deny that they can communicate true claims, but rather to deny that the true claims they communicate can be stated in language that is itself literally true.

For the purposes of the present investigation, I am interested only in determining whether a fictional scientific model, understood as a metaphor, is translatable into some literal description (paraphrasable) in principle or not translatable into some literal description (unparaphrasable) in principle. For other purposes, someone might be interested in determining whether a particular agent offering the model in a particular instance is actually able to produce a literal paraphrase. However, I am not interested in exploring the sense of paraphrasability that has to do with the current epistemic status of the agent providing the metaphor. I am interested in whether there is a literal paraphrase that fully captures the meaning of the metaphor in question. Someone offering a model may not know whether it is paraphrasable in principle. But I assume there is a fact of the matter. It would be a mistake to think that whether a model explains is determined by whether a particular agent is in a position to relay the entire content of the model. So an account of how fictional models might be explanatory should not hinge on features of an agent's epistemic status such as whether she is able to produce a paraphrase of the model. It might be reasonable to think that whether a particular agent has been successful in performing the speech act of explaining some phenomenon depends on some features of the agent's epistemic status, but the present investigation does not seek to evaluate the
success of such speech acts. I am interested in whether the fictional models *themselves* can be explanatory.

Thus, I will proceed with the assumption that a fictional model must be either *paraphrasable in principle* or *unparaphrasable in principle*. I will use the terms “paraphrasable” and “unparaphrasable” only to refer to paraphrasability in principle and unparaphrasability in principle, respectively.

I propose that fictional models, understood as metaphors, must be paraphrasable in order to be scientifically explanatory. As I will illustrate in section 6, it is, at the very least, clear that the cases of fictional models I considered in Part I are paraphrasable. Conceivably, one might hold the view that fictional models are genuine explanatory fictions, but are unparaphrasable metaphors. This would require a commitment both to the view that some metaphors can be truly unparaphrasable and to the view that unparaphrasable metaphors can be explanatory. To use one of the models under consideration as an example, you might hold that “short-term memory is a pushdown stack” is relevantly like “Juliet is the sun” in that there is some richness of the metaphor that cannot be conveyed non-metaphorically. Such an account would attribute some mysterious character to fictional models (of the sort that might motivate a rejection of premise (3) in the paradox above), but such an account strikes me as mistaken, simply on the grounds that I see no difficulty, at least for the examples I have considered, in drawing out the content of the models in non-metaphorical language. Without any evidence to the contrary, I will take for granted that unparaphrasable metaphors, if there are any, are not the sort that could constitute scientific explanations.
C. Understanding Fictional Scientific Models as Metaphors

If the right way to understand fictional scientific models is to understand them as paraphrasable metaphors, as I have proposed, this requires that everything the model conveys in terms of the secondary subject can be stated in terms of the primary subject. In other words, we can, at least in principle, give a literal, non-metaphorical paraphrase of the model's content about the primary subject. A scientific model's being paraphrasable means that it is possible to communicate all the claims the model makes about the primary subject without including any non-SAI falsehoods. This is a critical assumption on which my account relies. Though the model itself does include falsehoods, the model's real content and the paraphrase of the metaphor, contain no falsehoods (in a successful explanatory case). Paraphrasability is the feature by which metaphors function to convey truths while being literally false. Applying the concept to fictional scientific models reveals how they can be explanatory.

It is worth noting that one important detail about my view will not be completely filled in at present. As I noted above, the paraphrase of an explanatory fictional model will include only true claims about features of the primary subject. I take it that the model's content is enhanced by background information that indicates which features of the secondary subject are put forth as analogous to features of the primary subject and which features are falsely attributed to the primary subject by the model. It would be an implausible result if all of the features of the secondary subject that happened to be analogous to features of the primary subject, even if the author of the model had no idea that such an analogy obtained, ended up as part of the paraphrase, as part of the
explanatory content of the model. I take it as most plausible that the matter of which features are highlighted as analogous in an explanatory fictional model will be determined by contextually relevant background information (which is included in the content of the model). As I mentioned above, I do not have in mind a full account of how background knowledge plays this role, though I acknowledge that a complete defense of my view would ultimately require such an account. For present purposes, it suffices to say that the relevant background knowledge makes up part of the model's content, though not part of the model itself.

An assumption underlying my account is that the model, the content of the model, and the paraphrase of the model are all propositional. That is, they can be fully articulated in terms of true and false claims. The model is comprised of claims about the secondary subject (e.g. “a pushdown stack orders items serially”) and claims that serve to attribute the properties of the secondary subject to the primary subject (e.g. “short-term memory is a pushdown stack” and the property attributions this statement entails when combined with the model's description of a pushdown stack). The model's content is more expansive than the model itself. It includes a description of all the claims in the model, the contextually-relevant background information (noted above) that indicates which of the model's claims are false, and the claim that all of the remaining claims made by the model (i.e. those that are not deemed false by the contextually-relevant background information) are true.\textsuperscript{18} The paraphrase contains only the parts of the model's content that

\textsuperscript{18} Within the model's content, the description of the model's claims includes both its claims attributing properties to the primary subject and its claims attributing properties to the secondary subject. As I noted in footnote 13, it is unlikely in a successful explanatory case that any of the model's claims about the secondary subject will be false. If they turn out to be, however, this would not be a problem for the account. Presumably, the contextually-relevant background information in the model's content would indicate that they are false.
are true claims about the primary subject. The paraphrase is the model's explanatory content.

Applying my account of how fictional models function as paraphrasable metaphors to the pushdown stack model, the reason that the false depiction of the relations as spatial in the secondary subject does not impugn the model's explanatory capacity is that the false claim (when augmented with relevant background knowledge) entail true claims, which end up in the paraphrase. The model makes false claims the relations are spatial in order to convey true claims about the actual relations that obtain among memories. Of course, the claim that memories bear spatial relations to one another is not among the model's true claims. Rather, we think recall relations consist of a temporally based order, in which a subject's memories of the most recently presented stimuli are more likely to be forgotten than a subject's memories of stimuli that were presented less recently. In a physical pushdown stack (the secondary subject of the metaphor), the relations we are interested in highlighting by employing the metaphor (last-in-first-out structure, serial quality, etc.) obtain because of the spatial arrangement of items in a stack. The serial quality of the recall relations we think obtain among short-term memories does, necessarily, obtain among the items in the physical pushdown stack employed to represent them. In other words, claims about the serial recall relations scientists actually attribute to memories when employing this model are entailed by claims about the spatial relations of the items in the physical stack, along with background knowledge, included in the content of the model, indicating that certain features of the secondary subject (including, in this case, the serial structure of the physical stack) are not true of the primary subject. Claims about spatial relations do not
make it into the paraphrase, but claims about the structure of the serial recall relations that they entail does. Thus, the model entails literally true claims without being comprised of literally true claims itself.

To put the point another way, the falsehoods in these fictional models are not preserved, in any sense, in the literal paraphrase; only the true claims about the primary subject that are entailed by the fictional model have a place in the paraphrase. The pushdown stack metaphor depicts the relations among items presented to someone's memory (serial memory recall relations) as spatial relations. The model depicts items in the physical stack as being located beneath or on top of other items in the stack. It would be a mistaken interpretation to think that the scientists who put forth this model of short-term memory consider the relations between actual memories to be spatial. They clearly see memories as being serially accessible, but not in virtue of the spatial relations one memory bears to another. The attribution of serial accessibility to short-term memory makes it into the paraphrase (the model's explanatory content about the primary subject), but the attribution of spatial structure to short-term memory does not. The process of paraphrasing filters out all the falsehoods. Critically, the falsehoods, as such, are neither in the model's content nor in the paraphrase. And, as I will argue in section 5, for this reason, we can see they are not doing any of the explaining. They are not part of the literal paraphrase of the model, and thus not part of the explanatory content of the model (the truths about the primary subject in the paraphrase that are explanatory).

D. How Fictional Scientific Models Explain as Metaphors

What makes a fictional model explanatorily successful is that the model has
specifically identifiable, paraphrasable entailments that are literally true and that answer
the relevant “why?” questions about the target system. If the model constitutes a
successful scientific explanation, the paraphrase must be suited to explain the phenomena
in question. The reason that these false models can explain is not that their falsehoods are
somehow transformed into truths or that their claims are somehow both true and false.
Rather, as I noted above, it is that they actually entail the truths that do the explaining.

One way to understand what I am suggesting about how the fictional model
entails its paraphrase is to understand the paraphrase as something like a collection of
Ramsey sentences. Ramsey sentences were introduced originally by Frank Ramsey and
later employed by Rudolf Carnap as a tool to identify a sense in which a theory that
posits unobservables can be accepted as true and explanatory without committing its
subscribers to accepting the existence of the unobservables posited.¹⁹ Rather than saying,
as theories positing unobservables do before being stated in Ramsey sentences, that there
exist unobservable entities with various qualities, a Ramsey sentence says that there exist
some things (we know not what) that bear relations corresponding to the relations of the
unobservable entities posited by the theory. The upshot of this, for an anti-realist, is that
all of the scientifically useful content of the theory is contained by the corresponding
Ramsey sentence. This allows for a respect in which one can remain noncommittal
regarding the unobservable posits of a theory without this undermining one's acceptance
of the theory as an explanation of observed phenomena. The Ramsey sentence thus
conveys structural information about the relations obtaining among the entities without

¹⁹ For a clear discussion of Ramsey sentences and their role in the realism versus anti-realism debate,
see Psillos (1999, pp. 40-69). For their original description, see Ramsey (1929).
identifying the entities. I want to suggest that, in a similar respect, the paraphrase of a fictional model contains only some of the content from the original metaphor, but captures all of the explanatory capacity of the model.

The pushdown stack model literally says that each memory has a place in the stack, either above or below some other memory. Thus, it literally attributes a false property to memories. The paraphrase of the model (understood as something relevantly like a Ramsey sentence) might say that there is some relation such that each memory bears that relation to other memories and identify some structural features of the relation without identifying exactly what the relation is. After all, it would be strange to think that the pushdown stack model, properly understood, entails an account of just what the relations are that obtain among memories, given that there is no such widely accepted account among scientists. Similar references by the model to “something we know not what with some identifiable features” are present in each of the cases I have examined. Both Ramsey sentences and the account of paraphrases I am offering preserve this character.

Importantly, I do not mean to suggest that paraphrases of fictional models are comprised of Ramsey sentences. There are constraints on Ramsey sentences – for instance, that they omit all posits of unobservables – that I see no reason to think must be placed on paraphrases of fictional models. Rather, I mean to suggest that considering paraphrases as analogous to Ramsey sentences helps highlight features of the relationship I take to obtain between a fictional model and its paraphrase. Just as a Ramsey sentence is an entailment of a theory that is weaker than the theory, a paraphrase follows from the metaphorical model's content and is weaker than the metaphorical model's content.
Moreover, just as Ramsey sentences entail all the empirical consequences of their corresponding theories, paraphrases of metaphorical models entail all of the explanatory content of the metaphorical model. My intention has been to draw out these features of my account by way of an analogy to Ramsey sentences – a more familiar philosophical tool. I will now show how applying my account resolves the paradox introduced above without making any significant conceptual sacrifices.

5. Fictional Models as Metaphors: Implications for Explanatory Capacity

Once we have identified these fictional scientific models as paraphrasable metaphors, there are two straightforward lessons to be drawn regarding their explanatory capacity. I will develop these two ways of describing how fictional models can furnish scientific explanations, (L1) and (L2), in this section. These lessons are different ways of capturing the account I developed above (according to which explanatory fictional models are paraphrasable metaphors), so there is no real difference between them in terms of the accuracy with which they depict the phenomenon of fictional models furnishing explanations. They differ only in their prescriptions for how we ought to apply the relevant terminology to describe fictional models in light of the analysis I have offered and in the prescriptions they offer for how to resolve the paradox of explanatory fictional models.

Given that the two lessons lend themselves to different regimentations of the relevant vocabulary, terminological preferences and pragmatic aims will determine whether one or the other provide a better framework for understanding the explanatory capacity of fictional models. At present, I do not see a decisive reason to prefer one over
the other, though there may be such reasons, either in general or in specific contexts. The crucial point is that to accept either (L1) or (L2) as correct is to accept that there is no mysterious respect in which falsehoods are explanatory. As long as at least one of these two ways of stating the verdict of my account suffices to accurately characterize the role fictional models play in scientific explanation, my account resolves the paradox.

A. Lesson One: The paraphrase, not the literally false model, really explains.

(L1) Because we have identified fictional models as paraphrasable metaphors, we can see that it is really the paraphrase that is doing the explanatory work. So these are not cases of genuine scientific explanation by false models.

Those who defend the claim that scientific fictions can be explanatory do not, of course, think that the fictional models themselves are true, but they do seem to think that the fictions get at something true. As Bokulich frequently puts it, we take only those fictional models to be explanatory that generate “genuine insight” or “genuine knowledge” (2008, 2009, 2011, 2012). Thus, even those who insist that the fictional models qua fictions are explanatory seem to be committed to thinking that the thing the models “get at” has to be a truth for the explanation to be successful. If we understand fictional models as metaphors, what has to be true is the description they provide of the primary subject. And if we understand them as paraphrasable metaphors, we see that this description of the primary subject is not inexorably bound up in falsehoods. There is a literal paraphrase that is doing the real explaining. And that description in literal language, the paraphrase, is one for which we can evaluate explanatory success in all of the standard ways (e.g. by asking whether it provides causal information, is unifying,
answers the right contrastive “why?” questions, and so forth).

According (L1), it is not really the fiction itself that is doing the explaining. The fiction is adverting to the real explainer in some way (as I have suggested, by being part of the model that entails it). The false descriptions might be good enough (or even better) for various pragmatic purposes, but really the falsehoods explain only by conveying literal truths about the target system that do the real explaining. The real explainer, the paraphrase, is not fictional, so these cases do not appear to be genuine cases of scientific explanation by fictions.

Applying (L1), we could reject either premise (1) or (2) of the paradox identified in section 2, depending on how we choose to define the key terms. We could reject premise (1) (which says that the fictional models are false) if we interpreted it as saying that the model's content is false. After all, all of the claims in the model's content turn out to be true when fictional models are rightly understood as paraphrasable metaphors. However, if we choose to read premise (1) as saying that the fictional models themselves are literally false, we ought not abandon it. We could instead reject premise (2) (which says that the fictional models explain) if we understood it as saying that the literally false model, not the model's content or the paraphrase, is doing the explaining. If we preferred to read premise (2) as saying that the fictional model conveyed explanatory content, then we ought not reject it. If you read both (1) and (2) in the ways that are incompatible with (L1), you would reject both (1) and (2) in applying (L1). Whichever of these options you choose, applying (L1) appears to dissolve the paradox.
B. Lesson Two: Only the true parts of fictional models explain.

Now I will consider a second way of answering the question of whether fictional models can explain, given my claim that they are best understood as paraphrasable metaphors:

(L2) Fictional models can genuinely explain, in virtue of the fact that part of the content they convey can explain (namely, the true claims about the primary subject that comprise the paraphrase). However, their explanatory success is derivative from the explanatory success of the paraphrase.

(L2) is meant to capture the intuition that there is some sense in which it is correct to say that fictional models are explaining because they communicate the paraphrase. It therefore offers another option for regimenting the application of term “explanatory” with respect to these models. Whereas (L1) says these models do not explain because their only role in the explanatory enterprise is to communicate content that entails the paraphrase (which does the explaining); (L2), by contrast, says that communicating the explanatory paraphrase counts as explaining. True, (L2) must allow, these models derive their explanatory success from the explanatory success of the paraphrase, but that does not mean it would be false to say these fictional models are explanatory.

(L2) tells us that, though it may be correct to apply the term “explanatory” to the fictional model rather than exclusively to the paraphrase, we should not see this as a case of explanation by falsehood. According to my account, the false claims in the model do not themselves end up being part of the explanation the model furnishes. For instance, if we thought the false claim that short-term memory has a finite capacity explained something about short-term memory, it seems we would have to think that there was
some relevant “why” question to which those who subscribe to the model take the finite capacity of the stack to provide an answer. But this is not one of the explanatory uses to which the model is put. Rather, the model is put forward to answer questions such as “why did the subject forget the piece of information most recently presented to her rather than the piece of information presented to her a longer time ago?” The true claim that short-term memory has a last-in-first-out structure, which is part of the paraphrase, does provide an answer to the relevant “why?” questions, the questions for which the model is put forth to provide an answer. Only the true claims in the model's content are explaining.

If it were really the case that the false claims in fictional models are explanatory as falsehoods (as, for instance, Bokulich appears to think), we would expect to see this happening in at least one of the cases of fictional models I considered in part I. But as I will show in section 6, applying my account to the three cases reveals this not to be the result. In all three of the fictional models under consideration, the true claims that end up in the paraphrases do all of the explanatory work. The false claims in the models do not themselves explain, even according to the scheme of terminological regimentation endorsed by (L2). (L2) permits the application of the term “explanatory” to something only part of whose content is explanatory, but it does not follow from (L2) that false claims merit this epithet.

To put the conclusion of (L2) simply, when we understand what each of the claims of the metaphorical model contributes explanatorily, we see that the explanation and the fiction come apart. When genuine explanation is given by a fictional model, it is not being given by the false claims in the model, the claims that make the model fictional. The claims attributing false features to the target system are genuinely fictional (they are
false without being one of SAI), but they are not genuinely explanatory. Each of the model's claims is either explanatory, but not fictional or fictional, but not explanatory.

How does (L2) apply to the paradox introduced in section 2? You might reject premise (1) (which said that the fictional models were false) if you thought that it entailed that they were false through and though, so to speak. In other words, you might abandon premise (1) in light of (L2) if you thought that the inclusion of some falsehoods in the model did not justify deeming the entire thing “false.” Perhaps a revised version of premise (1) informed by (L2) would say simply “the models have false parts.” Likewise, applying (L2) to premise (2) (which stated that the fictional models are explanatory), you might reject this premise if you thought it entailed that every part of the model was explanatory. In other words, you might abandon premise (2) in light of (L2) if you thought that all parts of the model had to be explanatory in order for the model to be correctly considered explanatory. A revised version of premise (2) informed by (L2) might say “fictional models have explanatory parts.” Accepting such a revised versions of premises (1) or (2) eliminates the contradiction with (3) that gives rise to the the paradox. Here again, you might also resolve the paradox by rejecting both (1) and (2) in favor of these revisions.

C. Conclusions from (L1) and (L2)

For present purposes, as I mentioned above, I want to remain agnostic with respect to whether (L1) is the best way to understand what is going on in fictional models or whether (L2) is. While the two lessons will be equally viable for many cases of fictional models, in some cases, it may be more natural to apply one than the other. This
sort of consideration may weigh in favor of one regimentation of the relevant vocabulary over another. I will not explore this in any depth in the present investigation, but I leave open both the possibility that one lesson may turn out to be terminologically preferable and the possibility that the ease with which the lessons accommodate cases of interest may turn out to rule decisively in favor of one of the two.

These lessons appear to exhaust the plausible options for making sense of the explanatory capacity of fictional models if we understand metaphors as paraphrasable metaphors, as I have argued we should. By adopting either, the paradox dissolves. Thus, if the fictional models I have considered are accurately described as paraphrasable metaphors, and if other fictional models are accurately described as paraphrasable metaphors, these models do not seem to present us with a genuine explanatory paradox. In both cases, there is a respect in which the model explains, and that respect is not the same respect in which the model is fictional.

Finally, I want to guard against a possible misreading of my account that these lessons might call to mind. I do not mean to suggest that, if you simply removed all the falsehoods from the model, you would be left with an explanation of the phenomenon in question. When I say that only the true parts of the model are explaining, I do not mean to say that the true parts of the model that are explanatory, all by themselves, constitute a complete explanation, but rather that they convey explanatory information.\(^{20}\) That is to say, the paraphrase of a successful explanatory fictional model fills in some of the information that would be included in a complete explanation of the phenomenon and, in

\(^{20}\) The distinction between explanatory information and a complete assemblage of all the information that would feature in an ideal explanation (an ideal explanatory text) is defended by Peter Railton (1981).
doing so, answer some of the relevant “why?” questions about the phenomenon's occurrence. The matter of determining how much explanatory information it takes to provide an explanation is notoriously fraught.\textsuperscript{21} I will not attempt to answer this question; however, I take it that the true claims about the primary subject that end up in the paraphrases of explanatory fictional models provide at least explanatory information (they may or may not provide enough of this explanatory information to count as explanations on their own). This is what makes it accurate to apply the term “explanatory,” either to the paraphrases of the models only (in accordance with (L1)) or to the models themselves as well as the paraphrases (in accordance with (L2)).

6. Final Analysis of the Cases

I noted in section 4 that all the fictional models I examined in Part I are best understood as paraphrasable metaphors. In this section, I will describe how my account applies to these cases in more detail. In illustrating my account, I have already explained how it handles the pushdown stack model for short-term memory, so it will not be necessary to discuss it further. I will now turn to the other two models I considered in Part I and describe how my account handles the electron orbitals model of valence and Maxwell's mechanistic model of the electromagnetic field.

A. Electron Orbitals Model: Explaining valence

\textsuperscript{21} Peter Railton, for instance, argues that a complete explanation, one that identifies all the information relevant to the occurrence of a phenomenon (which he calls an ideal explanatory text) is unattainable (1981, p. 239-41).
For various pragmatic purposes, when scientists offer the fictional orbitals model, they communicate truths about the target system through falsehoods, describing it literally as a system in which particles have valences in virtue of having, say, one extra electron or one fewer electron in a determinate spatial region (in an orbital). This falsehood makes the model fictional, but it falls away in the paraphrase. Moreover, the model says that an orbital can accommodate a certain number of electrons, and (among other things) that chemical reactions happen on the basis of particles having complementary numbers of electrons in these determinate spatial locations. However, applying tools from my account, we can say that this literally false model, when augmented with the relevant background knowledge included in the model's content, entails the true claim that there is some physical state or other in virtue of which particles with a certain number of electrons are attracted to particles with complementary numbers of electrons, and so on. This state may be specifiable in terms of wavefunctions, or it may not be, according to the paraphrase of the model. The paraphrase does not include the claim that orbitals are actually wave functions. Rather than specifying precisely what the property is, it indicates that there is such property that plays the identified roles. In this way, it bears many similarities to a Ramsey sentence.

In contexts for which the model is presented as a genuine fiction, we acknowledge that it is not in fact true that electrons have determinate spatial locations. This is one of the falsehoods that falls away, that does not make its way into the literal paraphrase of the model. Only truths remain in the paraphrase, and these truths can furnish an explanation of valence. The paraphrase conveys explanatory information because it identifies some constraints on what the property may be, the claims about those constraints are both true
and answers to relevant “why?” questions.

I take it as a virtue of my account that it portrays the fictional orbitals model as being fairly non-committal about the actual underlying structure of the quantum phenomena. Indeed, the right way to understand this model appears to be that it is making the fairly modest claim that there is some relation (we do not know exactly how it is realized) that has the basic structure conveyed by the fictional orbitals model. The model functions to convey true claims about the general structure underlying valence, and it is worth noting that, even if it turns out that the quantum mechanical picture of electrons is later found out to be false, this would not challenge the truth of the paraphrase of the fictional orbitals model. It is consistent with the fictional orbitals model of valence, properly understood as a metaphor, that the full explanation involves probabilistic wave functions. But the fictional model may also be consistent with other accounts of the underlying property. Given that this model was employed before orbitals were understood as wavefunctions, it would be a strange result if my analysis implied that the model was committed to identifying orbitals as wavefunctions.

Of course, the model's paraphrase is a literally true depiction of the target system that turns out to be compatible with the claims of quantum mechanics. So we do not need to worry that the actual meaning of the model is really inconsistent with quantum mechanics. When we see what the model actually says about the primary subject – when we attend to its paraphrase, which contains only true entailments of the model's content about the primary subject – we see that the orbitals model (when it is presented fictionally) is perfectly compatible with the claims of most standard interpretations of quantum mechanics regarding the properties of electrons. Moreover, the paraphrase of the
model is compatible both with standard interpretations of quantum mechanics that deny that electrons have determinate spatial locations and with non-standard interpretations that attribute to electrons determinate spatial locations.

Applying (L1) to this model, we would conclude that the false model itself is not explanatory because the paraphrase that it communicates is doing the explaining. Applying (L2) to this model, we would conclude that it fine to say that the model is explanatory, given that it functions to communicate the explanatory paraphrase.

B. Maxwell's Mechanistic Model of the Electromagnetic Field: Explaining a Possibility

One way to understand the explanatory aims of Maxwell's mechanistic model of the electromagnetic field is to understand it as a model that seeks to establish that some phenomenon is possible. On this way of understanding the model, it is importantly different from the other two fictional models I have considered. The target *explanandum* is not an actual empirical phenomenon, but rather a modal claim. The *explanandum* is not the actual existence of the electromagnetic field; it is the possibility that an electromagnetic field, understood with certain constraints, exists. However different it may be from the pushdown stack model and the orbitals model, it *does* seem to qualify as a fictional model according to the criteria I provided in Part I. So ideally, my account will also be able to make sense of how this model can be explanatory in spite of being fictional. As I will show, my account can indeed account for the explanatory capacity of Maxwell's model.

Before I apply my account to this case, however, it is worth considering some of
the concerns one might have about classifying this case as I intend to. One reason you might doubt the Maxwell case's eligibility to be classified as a paraphrasable metaphor, or even as a fictional model in the first place, is that you might doubt that it actually contains falsehoods. A tempting reading is to say that each of the claims in the model is somehow modally presented from the outset. You might think that the model's content about the mechanisms it describes (what I have identified as content about the secondary subject) is not “here is a gear, here is an idle wheel, etc.” but rather, “it is possible that there is a gear here, it is possible that there is an idle wheel here, etc.” If this is the right way to think of the model's content, it really does not seem like a metaphor. It is a collection of true modal claims.

To this worry, I respond that this is not how the model is actually presented; however, if it had been presented in these terms, it would not be a fictional model of the sort my account is meant to encompass. When Maxwell gives the model, he does not make it clear that its constitutive claims are somehow inherently modal. Rather, he offers a description of how parts in a mechanistic system fit together. This description is literally false. But the false description can still furnish an explanation of something, something that happens to be modal, namely that a field with certain properties is possible. This model, not presented in modal terms, explains the modal truth that an electromagnetic field is possible. There is some room for interpretation in this case, due to the fact that Maxwell was not explicit about what he actually had in mind when presenting this model. As Thomas Simpson noted in his commentary on Maxwell's model, “the reader must be the judge” of what Maxwell took the ontological status of the model's components to be (1997, p. 40 ). However, if one accepts the interpretive
consensus, it appears this is a case of a genuinely fictional model. As I noted above, if the aforementioned reading (in which each of the model's claims is a modal claim) turns out to be the right one, it would rule out the model's classification as fictional, and this would be no problem for my analysis.

Another reason one might think that the Maxwell case resists analysis as a paraphrasable metaphor becomes clear once I apply my analysis. Part of what ends up in the paraphrase of this model's content is “at least one configuration of a field with such-and-such properties that does not violate the laws of physics is possible.” If you think that the paraphrase contains only this claim, this presents a major difference between this case and the others I have considered. Namely, in the Maxwell model, the paraphrase does not look like an *explanans* as much as an *explanandum*. You might still think there is a sense (however trivial) in which “at least one instance of x is possible” explains “x is possible,” and so my account still succeeds in depicting what is going on here accurately. The model is literally false. The model entails the paraphrase and the falsehoods fall away in the paraphrase. The paraphrase is true, and the paraphrase explains what we set out to explain, the modal claim. So even though it looks like the Maxwell case is fundamentally different form the other two, on this reading, my account handles it just as well. I take this to be a virtue of my account.

If the only claim that seems to make its way into the paraphrase of the model is that an electromagnetic field that is consistent with the laws of physics is possible, one worry might be that (L2) applies awkwardly or does not apply to this case, given that there are no true claims to identify at the level of the unparaphrased metaphor. My reply to this general worry is that, as I noted in section 5, there are some cases for which one of
the two lessons I provided for resolving the paradox of explanatory fictions fits more naturally than the other. I take this to be true of the Maxwell case. (L1) seems to fit the case quite well, whereas (L2) seems to fit less naturally. Applying (L1), the Maxwell model would not come out as a case of genuine scientific explanation by a false model. The paraphrase of the model (whose content is exhausted by “such and such a field is possible”) is true and it is the paraphrase that is doing the explaining. The paraphrase is the explanatory information, and the model literally entails it. Thus, it really is only the paraphrase that is doing the explaining. The mechanical elements of the model play no part in the actual explaining, apart from having entailed the explainer.

However, in addition to the claim identified above as the paraphrase of Maxwell's model (“at least one configuration of a field with such-and-such properties that does not violate the laws of physics is possible.”), one might think that further information must end up in the paraphrase for my account to accurately capture the meaning of the model. To think this would be to resist a paraphrase in which none of the claims about the structural features depicted in the model – claims about the gears and idle wheels, etc. – entail claims in the paraphrase that attribute properties to the target system. One might think that the paraphrase of the model will include claims that there are properties in any possible system of the sort identified by the model that play roles analogous to the roles played by the mechanistic elements (e.g. idle wheels and gears) depicted in the model. Adopting this much richer understanding of what is included in the paraphrase might be better suited for an alternative way of understanding the explanatory aims of Maxwell's model that I have not mentioned up to this point. It may be more accurate to see the explanatory aim of Maxwell's model as being to establish that a certain sort of
explanation for some phenomenon is possible. If this is the correct way to understand Maxwell's model, perhaps the paraphrase of the model ought to include a series of claims that constitute such an explanation rather than merely the claim that the phenomenon is possible. Thus, additional claims, such as claims about properties that are analogous to the mechanistic elements, might end up in the paraphrase.

On this reading, both (L1) and (L2) appear to apply naturally. Applying (L1), of course, would be to say that the model does not explain that a particular sort of explanation is possible, but its paraphrase does. Applying (L2) would yield the verdict that the model as a whole does explain that a particular sort of explanation is possible, though this explanatory success is derived from the explanatory success of the paraphrase. For the paraphrase to convey successful explanatory information in this case would be to communicate some constraints (e.g. claims about the existence of properties that play certain roles in the possible system) on an explanation and also to communicate that such an explanation is possible. Applying either lessons seems to make sense of how the model functions to communicate the possibility of a certain sort of explanation, and so I conclude that my account also fits the Maxwell model on this understanding of its explanatory aims.

It is worth noting that various dimensions of ambiguity in the Maxwell model make it difficult to see whether any of my attempts at applying my account to this case accurately capture its explanatory aims. However, I take it as a virtue of my account that it can answer, at least in a general way, all of the possibilities for how to best understand the model that I have considered thus far.

Given that my account applies to all three of the cases I considered in Part I, I
conclude that a significant virtue of my account is that I can give an analysis of fictional models' explanatory capacities without giving up on (3) from the paradox above, the claim that only truths can be explanatory.

7. Conclusion

In Part I of this paper, I sought to show that there really are cases of scientific models that are genuinely fictional and also genuinely explanatory. To this end, I considered the sorts of falsehoods accepted as inevitabilities in scientific modeling (simplifications, approximations, and idealizations) and contrasted them with the sorts of falsehoods that scientific practice appears to permit in this special class of models. I took the result that falsehoods other than simplifications, approximations, and idealizations appear to be accepted as parts of some scientific models that are put forth to provide explanations to be a truly puzzling phenomenon in scientific practice that called for some account. The paradox of explanatory fictional models that I introduced at the start of Part II illustrated just what is so puzzling about these cases and why some analysis is needed to resolve the apparent contradiction in their acceptance by the scientific community.

In Part II, I offered a positive account of how we should understand the function of fictional models. Specifically, I argued, we should see them as a class of metaphors, a widely accepted semantic phenomenon in which literal falsehoods convey truths. I do not, of course, take my account to have shed any light on the extant puzzles concerning the semantic function of metaphors. The general puzzles that exist for the idea that a metaphor can convey a truth in spite of being literally false are puzzles for my account as well. However, I take it that the ubiquitous employment of metaphors in our language to
this end of communicating truth through falsehood is evidence that they do in fact accomplish this task, however the specific details ought to be filled in. Insofar as this is accepted, the account I have offered, according to which fictional models are metaphors, should be accepted as well.

In the final three sections of part II, I articulated my positive account, provided two options for regimenting the relevant vocabulary to describe how fictional models furnish explanations, showed that adopting my account resolves the paradox of explanatory fictional models without making any important conceptual sacrifices, and demonstrated that the account successfully accommodates the three cases of fictional models I considered in Part I. On the basis of this success, I conclude that understanding fictional scientific models as paraphrasable metaphors in the way that I have described is not only a useful way to make sense of how they function as explainers, but also a plausible account of why their presence in science has not given scientists cause for alarm. These fictional models are not any more mysterious than ordinary cases of metaphor, and are equally viable means of communicating true information.
Bibliography


