2020, 6 (1)

ARGUMENTA

The Journal of the Italian Society for Analytic Philosophy

First published 2020 by the University of Sassari

Produced and designed for digital publication by the *Argumenta* Staff

Argumenta is the official journal of the Italian Society for Analytic Philosophy (SIFA). It was founded in 2014 in response to a common demand for the creation of an Italian journal explicitly devoted to the publication of high quality research in analytic philosophy. From the beginning *Argumenta* was conceived as an international journal, and has benefitted from the cooperation of some of the most distinguished Italian and non-Italian scholars in all areas of analytic philosophy.

Contents

Editorial

Imagination and fiction play so pivotal a role both in our lives and in philosophy that it hardly needs stressing. For one thing, if we had a limited imagination, we would lead impoverished ethical lives. For another, if we were unable to respond to fiction, our lives would soon prove unbearable.

No wonder, therefore, that fiction and imagination have held the interest of philosophy down the centuries, receiving particular attention in recent decades. The Special Issue that opens the present number of *Argumenta*, entitled *Fiction and Imagination: Counterfactual Reasoning, Scientific Models, Thought Experiments* and edited by Carola Barbero, Matteo Plebani and Alberto Voltolini, represents an up-to-date discussion of the most pressing aspects of both themes, and finds in fiction and imagination the thread that binds different phenomena such as counterfactuals, thought experiments, and scientific models.

The present number is then topped off by the section of Book Reviews. In this section, readers will find careful assessments of three very interesting recent books—*Divine Omniscience and Human Free Will: A Logical and Metaphysical Analysis* by Ciro De Florio and Aldo Frigerio, *Musical Ontology: A Guide for the Perplexed* by Lisa Giombini, and *Just Words: On Speech and Hidden Harm* by Mary Kate McGowan.

Finally, I would like to thank all the colleagues who have acted as external referees, the Assistant Editors, the Editor of the Book Reviews, and the members of the Editorial Board. All of them have been very generous with their advice and suggestions.

Argumenta 6, 1 (2020): 3-4
ISSN 2465-2334

As usual, the articles appearing in *Argumenta* are freely accessible and freely downloadable, therefore it only remains to wish you:

Buona lettura!

Massimo Dell'Utri
Editor-in-Chief

Argumenta 6, 1 (2020)
Special Issue

Fiction and Imagination:
Counterfactual Reasoning, Scientific Models, Thought Experiments

Edited by

Carola Barbero, Matteo Plebani, Alberto Voltolini

The Journal of the Italian Society for Analytic Philosophy

Contents

Fiction and Imagination: Introduction

Carola Barbero, Matteo Plebani, Alberto Voltolini

University of Torino

Abstractly speaking, counterfactuals, thought experiments, and scientific models seem to be utterly different phenomena. First, counterfactuals are those conditionals whose antecedents are false, for they describe situations that are merely possible, or even impossible. Second, thought experiments are mental experiments performed both in philosophy and in natural sciences that, instead of relying on concrete actual procedures ultimately grounded upon observations, merely rely on hypothetical considerations. And third, scientific models are patterns, sometimes made of actual things—consider Rutherford's model of the atom, or the three-dimensional eliocentric model of the Solar System that has inspired that model—which, *qua* props that are proxies of the intended reality to be studied, simulate or idealize the behavior of the concrete items constituting that reality. Following Walton (1990), actual truths about the props can be exploited in order to get truths in the model (for example, actual truths about the spatial distribution of certain balls can be exploited in order to get truths in the eliocentric model about the spatial relations of planets in the Solar System). Note that in accordance with Walton (1993)'s idea of *prop-oriented* games of make-believe, things can also go in the other direction. That is, truths in the model can be exploited in order to get actual truths about the props themselves. Cf. on this Caldarola and Plebani 2016.

Yet appearances notwithstanding, there is a family resemblance among such phenomena. First, as Recanati (2000) remarked by following an original suggestion of Mackie's (1973), a conditional, hence a counterfactual as well, is the contracted form of a kind of reasoning moving to a conclusion under the scope of a supposition ("Suppose that *p*. Then *q* ensues"). Second, that kind of reasoning may also be present in telling a thought experiment, especially a philosophical one yielding an argument in favor of a thesis, yet couched in a narrative form. Third, the sort of tale occurring in that experiment may be similar to the one constituting that kind of scientific model that, instead of using actual props, resorts to descriptions.

Perhaps that family resemblance is not just a mere coincidence of overlapping traits, but it has a reason. Indeed, all such phenomena may be seen as forms of imagination, notably the kind of imagination that is exploited when doing fiction: *make-believe*. The phenomenon of make-believe may be conceived in the *normative* terms appealed to by Walton (1990), who resorts to games of make-believe based on (prop-exploiting) principles of fiction generation: it is fic-

Argumenta 6, 1 (2020): 9-13
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.int

tional that *p* iff (in the relevant game using certain props) it is prescribed to imagine that *p*. But it may also be conceived in *cognitive* terms, by appealing either to *multiple representational models*, the reality model and the imaginary model, in order to distinguish the representation of a real situation from the representation of a fictional situation, which is represented in an off-line form detached from behavioral consequences (Perner 1991, Nichols and Stich 2003), or to a *metarepresentational* structure that involves a metacognitive factor aimed at blocking the confusion between fiction and reality: the situation represented in the imaginary model is fictional precisely because it is so represented (Leslie 1987, Meini and Voltolini 2009, Voltolini 2016). Either way, first, the content of a fiction may be described in counterfactual terms. Indeed, the ability of understanding fiction and the mastery of counterfactuals developmentally go hand in hand (Weisberg and Gopnik 2013). As a matter of fact, a sort of counterfactual knowledge is part of what we learn when we learn something from fiction: a knowledge of possibility, as Putnam (1987) remarked. Actually this is what could be seen as *conceptual knowledge*, where what "I learn is to see the world as it looks to someone who is sure that hypothesis is correct. I see what plausibility that hypothesis has; what it would be like if it *were* true; how someone could possibly think that it *is* true" (Putnam 1976: 488), hence a kind of knowledge not to be seen as the possession of information, but rather as Lewis (1983) underlined, as the ability to imagine, to recognize, to predict one's behavior by means of imaginative experiments. Indeed, when knowing that in a certain story something is the case, we know how things would unfold if we were in such a situation, the situation affecting the fiction's protagonists (Currie 1998, Barbero 2017).

Second, the telling of a thought experiment is a kind of short fictional tale that has a real import, to be grasped by science: it is fictionally the case that *p* in order for something to be really the case. As some put it, one may read a thought experiment both as having a fictional content and as having a corresponding real content (Voltolini 2016). Third, scientific models may be compared with games of make-believe, even literary ones, insofar as the latter respectively mobilize physical objects and descriptions as props for imaginary characters, just as scientific models themselves may do in describing an idealized, or even nonexistent, form of reality—frictionless planes, pure distributions of gases, the ether out there (Frigg 2010). And models themselves can be compared to fictional stories, which can further be seen as a sort of abstract objects that amount to cultural artifacts (Thomasson 1999, Salis 2019).

In recent times, all such phenomena have individually been the target of several books (to quote just the most important ones, cf. Lewis 1973, Gendler 2000, Suarez 2009). Yet there is a growing interest in also exploring their connections. As a follow-up of the SIFA Midterm Conference / Graduate Conference of the FINO Ph.D. Programme held in Turin on June 17-18 2019, this issue intends to scrutinize such connections more thoroughly and widely.

The seven essays collected in this issue address central questions for the contemporary debate on counterfactuals, thought experiments and scientific models from new and thought-provoking perspectives.

Conrad Aquilina's "Simulation Modelling in Fiction" draws a comparison between scientific models, or models more in general, and narrative fictions that can be understood in a similar way. This comparison relies on the idea of simulation. As Frigg himself (2010) originally underlined, scientific models do not work as such unless they are *used* as models. According to Aquilina, this use in-

volves a simulation process in which a source world is simulated by another world. This also happens in narrative, insofar as one can literally take the idea of a fictional world generated by the narrative insofar as this world opportunely simulates, in phenomenologically involving terms, the real world from which it departs. Of course, this comparison does not mean coincidence, since scientific models are finally intended to describe portions of the real world and refer to its objects, while narrative fictions typically concern just imaginary scenarios and imaginary individuals (representation of reality is typically not among their purposes).

In a series of paper, James Nguyen and Roman Frigg have developed an account of how scientific models represent certain aspects of the world, the so-called DEKI account. In their contribution to this issue, "Unlocking Limits", Nguyen and Frigg elaborate upon one aspect of the DEKI account: the use of keys, rules that connect the features of the model with the features that should be attributed to the target system. The paper analyzes a kind of keys that play an important role in physics, i.e. limits keys, where the features of the model are the result of taking to the limit certain features of the target. It is argued that limit keys can be used only under certain circumstances, and that analyzing how limits keys work deepens our understanding of how models are used in the actual scientific practice.

Frederick Kroon's paper "Fiction, Models and the Problem of the Gap" starts from a problem that appealing to models as bits of fiction, as in Frigg's (2010) fiction view of models, raises: since the protagonists of a fiction are unreal, they do not really have the properties by means of which they are characterized (they only have such properties in the fiction); so, how can they represent real things by ascribing to them real features? Kroon's answer starts from the fact that we can have *de re* imagining about real objects in which we ascribe them in fiction properties they do not really possess. To this *de re* imagining, a *de dicto* imagining corresponds in which we merely pretend-refer to someone, who is not the real individual, but just a surrogate of it. Ditto for models. We can export, as concerning the target, what in the model only concerns the nonexistent objects that surrogate the real objects in the target. Just as in the aforementioned prop-oriented games of make-believe, this practice makes the model as externally oriented, not as content oriented.

Fiora Salis' essay, "Learning through the Scientific Imagination", analyzes the fundamental role of (constrained uses of) imagination in the development of plausible hypothesis concerning reality. Make-believe is seen as the notion of imagination at work when theoretical models are used as ways of knowing reality and an overarching taxonomy of types of constraints on scientific imagination enabling that kind of knowledge is sketched. Two main kinds of knowledge are hence identified: first, the knowledge of the imaginary scenario specified by models, and second, the knowledge of reality itself.

"Spoiler Alert! Unveiling the Plot in Thought Experiments and Other Fictional Works" by Daniele Molinari explores the connection between thought experiments and literary works. In Molinari's view, the use of spoilers is a necessary condition for a piece of text to be a thought experiment. For a thought experiment is supposed to widen our knowledge of reality. Thus, it is right that a literary work can play the role of a thought experiment, as people following Elgin (2007) hold. Yet in order for this to be the case, one must locate the work in

the proper foretaste context, in which it is settled how to properly appreciate a text.

Starting from the connection between fictional disagreements and thought experiments, Louis Rouillé's paper "From Fictional Disagreements to Thought Experiments" analyses the "great beetle debate" (what did Gregor Samsa metamorphosed into? A beetle or a big cockroach?) as a paradigmatic case. Actually, fictional disagreement is interesting in order to understand what has to be considered as the informational content of a fiction. There is a distinction that needs to be recognized between what is meant by the author (the fictional foreground) and what is inferred by readers (the fictional background). Actually, the fictional background seems to be filled by the reader's representations of reality and other shared (and often conventional) beliefs. The idea is that what happens when we learn from fiction is analogous to what happens when we perform a thought experiment, because in both cases the same informational structure is exploited: instead of filling the fictional background, one informs one's non-fictional representations using the same informational channels in reverse direction.

A much-debated topic in the literature on counterfactuals is whether counterfactuals with impossible antecedents (so-called *counterpossibles*) are vacuously true or not. In "Game Counterpossibles" Felipe Morales Carbonell analyzes various examples of *chess-counterpossibles*: counterfactuals whose antecedents describe a position on the chessboard that is not permitted by the rules of chess. Morales Carbonell defends the view that these examples count as genuine, non-vacuously true, counterpossibles and argues that this kind of counterpossibles are used to think about the consequences of certain changes in the rules of a game.

Finally, Malvina Ongaro's paper, "Fiction, Imagination, and Normative Rationality", addresses the question of how a fictional character, the Rational Agent described in Microeconomics models, can act like a role model for real economic agents and prescribe how they should behave. The paper focuses on the question of how the Dutch Book argument, an argument supporting the conclusion that the degrees of belief of the economic agents should respect the laws of probability, can have normative force. The narrative structure of the Dutch Book argument is analyzed and it is argued that the argument involves the use of the imagination to compare the outcomes of different courses of action. The analysis of the Dutch Book argument presented in the paper leads to the conclusion that imagination plays an important role in decision-making.

References

Barbero, C. 2017, "Aprender de la Imaginación", *Hybris*, 8, 129-41.

Caldarola, E. and Plebani, M. 2016, "Caricatures and Prop-Oriented Make-Believe", *Ergo*, 3, 403-19.

Currie, G. 1998, "Realism of Character and the Value of Fiction", in Levinson, J. (ed.), *Aesthetics and Ethics: Essay at the Intersection*, Cambridge: Cambridge University Press, 161-81.

Elgin, C. 2007, "The Laboratory of the Mind", in Huemer, W., Gibson, J. and Pocci, L. (eds.), *A Sense of the World. Essays on Fiction, Narrative, and Knowledge*, New York: Routledge, 43-54.

Frigg, R. 2010, "Models and Fiction", *Synthese*, 172, 251-68.

Gendler, T.S. 2000, *Thought Experiment: On the Powers and Limits of Imaginary Cases*, New York: Garland (now Routledge).

Leslie, A.M. 1987, "Pretense and Representation: The Origins of 'Theory of Mind", *Psychological Review*, 94, 412-26.

Lewis, D. 1973, *Counterfactuals*, Oxford: Basil Blackwell.

Lewis, D. 1983, "Postscript to Mad Pain and Martian Pain", in his *Philosophical Papers Volume I*, New York: Oxford University Press, 122-33.

Mackie, J. 1973, *Truth, Probability, and Paradox*, Oxford: Clarendon Press.

Meini, C. and Voltolini, A. 2010, "How Pretence Can Really Be Metarepresentational", *Mind and Society*, 9, 31-58.

Nichols, S. and Stich, S. 2003, *Mindreading: an Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*, Oxford: Oxford University Press.

Perner, J. 1991, *Understanding the Representational Mind*, Cambridge, MA: MIT Press.

Putnam, H. 1976, "Literature, Science, and Reflection", *New Literary History*, 3, 483-91.

Putnam, H. 1987, *The Many Faces of Realism*, La Salle, IL: Open Court.

Recanati, F. 2000, *Oratio Obliqua, Oratio Recta: An Essay on Metarepresentation*, Cambridge, MA: The MIT Press.

Salis, F. 2019, "The New Fiction View of Models", *The British Journal for the Philosophy of Science* (online first), https://doi.org/10.1093/bjps/axz015.

Suarez, M. (ed.) 2009, *Fictions in Science: Philosophical Essays on Modeling and Idealization*, London: Routledge.

Thomasson, A. 1999, *Fiction and Metaphysics*, Cambridge: Cambridge University Press.

Voltolini, A. 2016, "The Nature of Fiction/al Utterances", *Kairos*, 17, 28-55.

Walton, K.L. 1990, *Mimesis as Make-Believe: On the Foundations of the Representational Arts*, Cambridge, MA: Harvard University Press.

Walton, K.L. 1993, "Metaphor and Prop Oriented Make-Believe", *European Journal of Philosophy*, 1, 39-57.

Weisberg, D. and Gopnik, A. 2013, "Pretense, Counterfactuals, and Bayesian Causal Models: Why What Is Not Real Really Matters", *Cognitive Science*, 37, 1368-81.

Simulation Modelling in Fiction

Conrad Aquilina

University of Malta

Abstract

This essay assesses the claim that model structures have features in common with narratology and fiction-making. It proposes that simulation—a form of modelling—is amenable to literary narratives which are hypermimetic, in the sense that their cognitive or material reception by the reader demands a phenomenology attained through the heightening of a mimetic secondary reality. Simulation models construct frames of reference for target systems through self-validating mechanisms, and the same is true of narratology. I specifically argue that the modelling of a world out of text, one which is written and read into being, needs to be discussed in simulationist terms. To an extent, narratives or entire fictional worlds, are modelled by an author and a reader since properties, laws and behaviours are imputed on the basis of tacit agreement and shared knowledge. Readers self-identify (or not) with the author's fictional world, and its constructs. A process of verification and validation, analogous to the modelling and testing of simulations, follows. I conclude this essay by proposing a model in which elements from simulation modelling are carried over to narratology to demonstrate permeation between both representational systems.

Keywords: Simulation, Modelling, Narratology, World-Construction, Reader-Reception.

1. Introduction

Simulation is a process which involves modelling, a form of scientific representation that is highly mimetic, function-driven and outcome-oriented. Various simulation theorists, such as Jeff Rothenberg and Pau Fonseca i Casas, have specified most, if not all, of these aspects in their definitions. For Rothenberg, "simulation is a process in which a model of any kind is used to imitate (some aspect of) the behavior of its referent" (1989: 80), while Fonseca i Casas explains that "the act of simulating something first requires that a model be developed [to represent] the system itself, [with] the simulation represent[ing] the operation of the system over time" (2014: 265). Both definitions crucially distinguish between the system model, as mimetic or representational system, and the finished model, the simulation run itself. Furthermore, the dynamic rather than the descriptive aspect of modelling has been noted by John Casti, who specifies that dynamic simulations can be manipulated

Argumenta 6, 1 (2020): 15-30
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.aqu

"so as to modify the reality the model tries to represent" (1997: 19). We can therefore assert that, as far as simulation modelling goes, a finished model simulates X by functionally representing it. It is also the case that while all simulations are necessarily models, not all models are capable of simulation (such as small-scale models or three-dimensional rotatable illustrations of the solar system).

Thus, any representation which functionally models behaviour must also maintain a number of correspondences between the physical (or source) system and the system model (Kheir 1996: 5). These correspondences are chosen based on the referential aspect being modelled, since a model is never fully identical to its source but is always an abstraction based on design selection (Rothenberg 1989: 78). Therefore, choosing which model best simulates the referent means making a conscious choice on how the referent will be re-presented while maintaining a number of fidelity conditions (Van Fraassen 2010).

A model therefore refers to, and substitutes for, a source through a series of referential moves—according to Van Fraassen, "Z uses X to depict Y as F" (2010: 21), even if the source is fictional. This is particularly interesting for fictional narratives which are constructed and enacted along similar principles of representation and referentiality. However, when the model does not ostensibly represent or refer to an actually existing object, such as pseudoreferents in fiction, the model structure must be such that it lends itself to permissible simulation (Rothenberg 1989: 78). In this case, the real-world and its laws sanction the non-actual object which is in turn synthesised through simulation modelling within the narrative in much the same way that simulation permits replicatively and predictively valid behaviours in non-fictional models (Kheir 1996: 5-6).

Roman Frigg's argument that models construct frames of reference for target systems through make-believe mechanisms which validate their truth as fictions (2010a) is also true of narratology. In Frigg's fiction view of modelling (2010a), a system only becomes a model when it is deliberately used as such, and as in literature, combines actual and non-actual elements within the model for which the reader extrapolates content and rules. Fictional worlds are more than mimetic narrative constructs; they are foremost approaches to narrative phenomenology and simulation. This means that the textual model adopted must construct—in some cases, even simulate—its narratives in such a way that its reader feels or experiences the text-world as possible (Ryan 2001; Gerrig 1998). This passage from a primary (source) to a secondary world (target) requires a near-instantaneous decoding of words into semantically and phenomenologically relevant content (Birkerts 1994; Ryan 2001).

An objection to the idea of narrative-as-simulation might be the claim that fictional counter-factuality, in which descriptions or propositions which may not be true to fact are nonetheless used, is not the end objective of any simulation and neither does a simulation run on counter-factual rules. I counter this objection by explaining that textual distancing (where the reader 'travels' from the world of origin) does not warrant ontological distancing and that in the simulation of narrative worlds, "suspension of disbelief", to use Samuel Coleridge's term (1817: 168-174), does not imply a suspension of primary reality but merely the heightening of a secondary one.

In making my argument for simulation modelling in literary fiction, I adopt the following positions, supported by the relevant literature:

i. Common principles underpin narrative fiction and simulation modelling. These are explained in section 2 ('A Fiction View of Modelling').

ii. The construction of a textual world model, with its properties, behaviours and laws, involves various make-believe mechanisms that need to be tacitly agreed upon by a minimum of two parties. In this respect, the successful modelling of a fictional world is an act of joint authorship involving an author and a reader. These mechanisms are discussed in section 3 ('Make-Believe Mechanisms').

iii. The modelling of a world out of text, one which is written and one which is read into being, needs to be discussed in simulationist terms in cases of narratives which make additional mimetic demands from the reader. Certain fundamentals of world-modelling are discussed in section 4 ('World-Building as Simulation Modelling') while their reception is discussed in section 5 ('Reader-Centric Modelling'). Since readers self-identify (or do not) with the author's fictional world, approaches similar to verification and validation processes present in simulation theory are also quite evident.

iv. I conclude this essay in section 6 ('Simulation-Type Modelling in Literary Fictional Worlds') by proposing a model for the construction of a fictional world in which elements from simulation modelling are carried over to narratology to demonstrate permeation and overlap between both representational systems.

2. A Fiction View of Modelling

Roman Frigg's "fiction view of model-systems" (2010a: 99) can be used to explain what common principles underpin narrative fiction and simulation modelling beyond figurative analogies. Frigg's concept of modelling as fiction serves a dual purpose: it relates scientific representation to fictional/semi-fictional constructions such as those found in literary texts and it does so precisely by establishing prescriptive rules typical of narratives. Frigg's fiction view of models, and one which has Kendall Walton's prop theory (1990) as its basis, thus indirectly provides further evidence for a mode of simulation that is quite amenable to narratology. Central to Frigg's argument is that model systems are often composed of fictional and non-fictional elements, which come together through an imaginative exercise in pretense. Hans Vaihinger, Nancy Cartwright, Peter Godfrey-Smith and others have also construed scientific modelling in terms of "intellectual construction", as-if philosophy, and "epistemic practices" (2010b: 255) shared by artistic and imaginative fiction.

Frigg departs from the assumption that scientists adopt models which are abstractions of more complete physical systems. They are "hypothetical systems", distinct from the "target system", the actual source reality which is being represented or simulated (2010b: 253). Hypothetical systems or hypothetical entities "would be physical things if they were real" (Frigg 2010b: 253), yet they are not, and neither do these models—proffered in lieu of a target system—represent the world per se; they represent only their own structures.

A model therefore can only start representing its referent (its target, in Frigg's discourse) once its underlying structure has been "endowed with representative power [enclosed in] a physical design" (2002: 3). But this is not apparently what structures can do on their own—a structure must be made to become a model. Frigg's concept of a model requires "(at least) a structure, a physical design and a process that hooks up the two" (2002: 3). In this manner, Frigg discounts struc-

turalist model theories where a structure and its attributes have direct correspondence (isomorphism) with the object they model, mainly because structures "are not representations of anything in the world" but "pieces of pure mathematics, devoid of empirical content" (2002: 5). Since representation is based on a substitution-for principle (representing X as Y), it requires "semantic content" (2002: 5) in order to stand for something else. Only then will a model acquire representational status since "structures per se do not stand for anything at all [and] do not indicate any real-world system as their object" (2002: 5).[1]

We can posit the same rules for literary fiction. Like Frigg's model-systems, which are an "ensemble" of "things that do and […] do not exist" (2010c: 257), literary plots "are mixtures of existent and non-existent elements" (2010c: 257) whose design prescribes to the reader how they ought to engage with them, despite not characteristically portraying an actual state of affairs. A model system is introduced in the same way literature is introduced, "by giving a description [through] sentences specifying its features" (Frigg 2010c: 257), although a good number of model systems are 'described' non-textually through the use of diagrams and so on. This description is not intended to denote real persons or objects and may or may not have "counterparts in the real world" (Frigg 2010c: 257), yet the reader is aware of this when they engage with the storyworld (a fictionally narrated world/reality) or with a model system for that matter. Moreover, the description of a model system, of which a fictional storyworld is an example, "specifies only a handful of essential properties, but it is understood that the system has properties other than the ones mentioned in the description" (Frigg 2010b: 258).

Essentially, what Frigg is stating here is that model systems—and by extension, fictional worlds—operate on principles of implicit or "extra content" (2010b: 258) which are generated when the reader extrapolates from the model system/narrative itself. (Narrative or genre-models therefore contain self-inscribed or pre-written 'rules' or conditions for their own readability or interpretability, the same as simulations). This extrapolation is also carried out, inevitably, with the target system, and although Frigg has made a case for model systems not being structurally isomorphic to real world counterparts, he concedes that "on every account of representation one has to compare features of the model system with features of the target at some point, even if only to assess how good an approximation the former is of the latter" (2010b: 258).

3. Make-Believe Mechanisms

Both model systems and fictional narratives are nevertheless presented (read: 'structured') as descriptions which function as props in games of make-believe (Frigg 2010b: 260), in which a conscious form of non-deceptive pretension (New 1999: 69-73) is adopted. This analogy is important to keep in mind as conditions of truth or factuality are waived, according to Christopher New, when one considers the nature of fictional texts as "invented narrative[s], consisting of sentences which the author invites the audience to make-believe are true, or to make-believe

[1] Frigg treats scientific modelling as a conceptual rather than material process, in which case the assertion that structures on their own have zero semantic or representational value until they become invested as models is true. Models are contained in the head rather than the hands. However, Frigg does not discount the presence and use of material models, which decidedly requires less structuring.

are authentic utterances of a real or imaginary utterer" (1999: 48). To give one over-cited example, we know that there is no actual historical person called Anna Karenina, yet this person exists in the world of Leo Tolstoy's titular novel. This Anna Karenina is therefore "fictionally true" (New 1999: 108) while claiming that Anna Karenina is not Alexei Vronksy's lover is fictionally false. As readers, we accept the conditions imposed by the game of make-believe, which leads us also to infer fictional truths through logical implicature rather than explicit description when information is deliberately withheld. Thus, Tolstoy writing that "at the very moment when the midway point between the wheels drew level, she threw away her red bag, and […] threw herself forward on her hands under the truck" logically implies Anna Karenina's suicide, albeit a fictional one (New 1999: 109). Therefore, according to New, "fiction involves nondeceptive pretending to oneself, or make-believe", inviting a form of "voluntary imagining" (1999: 69-73) in which we remain somewhat in control of the fictional scenario (unlike a dream or a delusion) and willingly accept the events portrayed (by another), while in the knowledge that they are fictional.

Frigg advances a similar theoretical starting point for his fiction view of modelling, basing it on Kendall Walton's pretense theory in which fictional truths are generated by props, prompting readers (or designers of models) to indulge in a consensual 'game' of intentional pretense where they imagine objects as possessing certain attributes for the duration of this game (2010b). For Walton, fiction and fictional propositions are contingent on props as they act as "generators of fictional truths" (1990: 37). Thus, for example, to claim that a snow construction represents a fort is to say that the snow fort acts as a fictional prop of a real fort, complete with turrets and a moat.

One other condition of a prop is that it is capable of generating fictional truths regardless of people's ability to imagine or not imagine these fictions as long as this prop is prescribed a function and there is social agreement on what this function is. Children may pretend to 'use' the snow fort as the real thing while to a disengaged passerby the snow fort remains a pile of drift (Walton 1990: 38). This highlights the functional aspect of modelling. Props (even within their theatrical context) serve specific functions and are denotative, treated as literal. In Walton's pretense theory, the "principle of generation" (1990: 38) describes what is going to serve as a prop, how it is going to be used, and by whom. If in a game of make-believe, a tree stump is taken to represent a bear, the tree stump acts as a prop only for this particular game and not for others. If a tree stump can be a 'bear' in one (private) game, a 'dragon' in another, and a 'portal to a fantasy world' in yet another game, then the principle of generation becomes what Walton calls "ad hoc" (1990: 51). Frigg adds: "games based on public rules are 'authorized'; games involving ad hoc rules are 'unauthorized'" (2010b: 259). Both involve pretense and imagination, the generation of fictional propositions, yet only in the case of authorised games does a prop acquire stable representational status. (Frigg eventually extrapolates this to mean modelling, whether scientific or, in the case of fictional narratives, the writing of a literary text whose reception depends on sanctioned principles of generation as a prop).

This aspect of fictional pretension is a matter of belief, rather than imagination, since although in ordinary circumstances "we are free to imagine as we please", "we are not free to believe as we please" (Walton 1990: 39). Fiction therefore necessarily places strictures and mandates on the imagination. In Tolstoy's Anna Karenina, the literary conventions of the novel prescribes the kind of props

it utilises—in this case, a train is a train is a train—and we are meant to believe and imagine that Anna Karenina intends to commit suicide and in fact (or in fiction) succeeds. It could not be otherwise.[2]

The way truth statements operate in fiction is seen by Frigg to have correlations with model systems. If fictional truths can exist "independently of people's actual imaginings" (2010b: 262), as long as there are props to sustain them with generational rules then model systems can be similarly constructed. This occurs by: i. replacing fictional propositions (such as 'Macbeth is the only person to see a floating dagger') with claims about the model; ii. replacing descriptions of the type of fictional work (text, play, performance, film etc.) with descriptions of the model system (what Frigg calls the hypothetical model), and iii. replacing the principles of generation innate to that particular work with principles assumed to be operational within that model system (2010b: 262).

While decidedly interesting, Frigg's fiction view of modelling presents various problems for simulation modelling in general, especially since it cannot (just) be considered a conceptual form of modelling, which is what Frigg bases most of his arguments on. On the other hand, the fiction view of modelling proves to be perfectly amenable to discussions of narrative simulation, which this essay seeks to advance. Before proceeding further, however, it might be appropriate to explain which of Frigg's claims are problematic, and why.

That models or literary fictions "are not defined in contrast to truth" (2010b: 260) is only partially correct. A model is not constructed as distinct to what it is held to be true (fidelity principle), so much so that a two-tiered process of verification and validation of the model (especially in functionally accurate simulations) is typically carried out before the model can be called 'good' (see section 5). Likewise, it is true that in fiction we can definitely "ascribe concrete properties to nonexistent entities" (Frigg 2010b: 261) such as in the modelling of pseudoreferents, and this because we are entitled to do so within the operational parameters of make-believe, yet I find it problematic to carry this analogy over to modelling, as Frigg does, especially in a model system which is intended to simulate an actual one.

In the main, simulation modelling does not involve imagining imaginary properties but imagining that a model has been attributed actual ones and seeing what emerges when these properties are applied and set in motion. Finally, since simulation modelling involves a very particular form of scientific representation, we cannot concede Frigg's claim that "a structure is not about anything in the world, let alone about a particular target system" (2010b: 254) since the very hypothetical system he proposes as the object of study (the simulation itself) needs to be grounded in laws and behaviours of the actual target system. Therefore, in simulation modelling (at least) it would also be imprecise to assert that a "hypothetical system [is] distinct from the target system" (2010b: 254) and while this may be true of the modelling of literary fictions (what is conveyed in fiction may or may not resemble or correspond with an actual state of affairs), it is certainly not the case with simulation modelling. Simulation modelling and fiction modelling part ways in their target outcome since they adopt a different teleology (simulation modelling, for instance has epistemic functions while the modelling of fictional characters and worlds is not necessarily so, and in general, is not). But

[2] Walton in fact claims that in a novel such as *Gulliver's Travels* or the play *Macbeth* the nature of the work itself leads the reader or spectator to specific imaginings. Thus, Walton concludes, "the work is a prop" (1990: 51).

we also need to consider what happens in the case of narrative simulation, which combines aspects of simulation modelling with conditions prevalent in fiction, and one where games of make-believe become structurally complex.

4. World-Building as Simulation Modelling

Following Roman Frigg's proposition that scientific modelling and fictional representation have rules in common, correlations can also be drawn between simulation modelling and narratology. In 1969, Tzvetan Todorov proposed a "narratology" that went beyond the study of text-based discourse to an actual scientific theory that would address the logic and structural properties of narrative as "a universe of representations" (Meister 2014). This would open the study of narratives to new modes and disciplines. Out of necessity, in this argument I adopt a text-based approach to narratology while explaining how specific structures embedded in narrative attribute it the quality of narrative simulation, as opposed to conventional mimesis. If narrative can be conceived of as a "universe", as Todorov has claimed (Meister 2014), then we can theorise about the construction of entire, possible worlds as textual models.[3] However, while the construction of fictional worlds is conventionally based on mimeticism, some fiction ventures beyond conventional mimesis to acquire the status of text-based simulation, with narratives that either simulate cognitive processes in real-time or simulate actual reader behaviours beyond the phenomenological.[4]

We should ask: what makes a fictional world a 'complete' world, one which is sufficiently cross-referential to sustain belief in its constructs? Michael Heim describes a world's "totality" in terms of "a felt totality or whole" (qtd. in Ryan 2001: 91), "not a collection of things but an active usage that relates things together [in a] total environment or surround space" (qtd. in Ryan 2001: 91). While Heim uses this concept of a total world for virtual realism, specifying the interoperability of the fictional world's constituents (X acts on Y) as a form of causality, his concept can be reduced to one phenomenological imperative: affect. This condition is also present in textual worlds. A fictional world, whether a visual and interactive one or one which simply relies on cognitive immersion, must construct its narrative/s in such a way that its user/reader feels or experiences the game/text as possible. This is why apart from the interconnectedness of objects and individuals and their habitable environment, Marie-Laure Ryan has added phenomenological requisites to the structuring of complete fictional worlds, such as the "intelligible totality for external observers" and "field of activity for its members" (2001: 91).

Fictional worlds are more than mimetic narrative constructs; they are approaches to narrative phenomenology. For Ryan, this means experiencing "the text as world", of being "immersed" in the textual world (2001: 90) while for Richard Gerrig this experience is akin to being "transported" (1998: 10) to a secondary

[3] Other obvious narrative modes such as film and digital games also permit this, the latter being the most convincing due to their immersive and interactive nature.

[4] While it is beyond the scope of this essay to engage in narratological analysis, it may still be worth mentioning works by Virginia Woolf such as 'Kew Gardens', James Joyce's *A Portrait of the Artist as a Young Man* and 'The Dead', and Bret Easton Ellis's *American Psycho* as examples of cognitive simulation, and Michael Cunningham's *The Hours*, Ian McEwan's *Atonement*, and Mark Z. Danielewski's *House of Leaves* as examples of affective modelling.

world, making some aspects of the reader's "world of origin [temporarily] inaccessible" (1998: 11). Similarly, what Victor Nell has called "reading entrancement", or being absorbed or "lost in a book" (qtd. in Ryan 2001: 96), implies an almost effortless passage from physical reality to fictive reality, provided that the narrative is structured in such a way that it does not place increasing demands on a reader's consciousness during the largely unconscious decoding of the information presented. These approaches to world-building focus on the reader's experiencing of the fictional world through a very active make-believe process which sufficiently simulates, if not the texture, then at least a mentally intelligible perception of that world. At this point the question moves from the ontological to the phenomenological. As Pimentel and Teixeira have observed, it is not "whether the created world is as real as the physical world, but whether the created world is real enough for [the reader] to suspend [their] disbelief for a period of time" (qtd. in Ryan 2001: 89). Considering that the world-as-text is a linguistic construct requiring near-instantaneous conversion of letters into semantically relevant content, this is no mean feat.

Modelling a textual world goes beyond mimetic representation. If it is meant to elicit behaviour or affect, it requires simulationist strategies which often go unnoticed. Ryan explains that the

> idea of a textual world presupposes that the reader constructs in imagination a set of language-independent objects, using as a guide […] textual declarations, but building this always incomplete image into a more vivid representation through the import of information provided by internalized cognitive models, inferential mechanisms, real-life experience, and cultural knowledge, including knowledge derived from other texts (2001: 91).

The terms Ryan uses for her description of linguistic structures which generate virtual scenarios and characters—"constructs"; "objects"; "declarations"; "representation"; 'import of information"; "internalized models"; "inferential mechanisms"; "real-life"—recalls a discourse of simulation modelling where virtual objects are imputed properties and rules based on external real-life targets. But curiously, while Ryan seems to downplay the idea of the text-as-world by treating it as metaphor (2001: 90-93), the modelling of successful microcosmia out of text—one which is written, but more significantly, one which is read[5]—needs to be discussed in nothing less than simulationist terms. This is rendered more imperative in the light of Frigg's declaration that structures are non-referential, becoming meaningful model systems only when they are used as such. Similar to Walton's make-believe scenarios involving props whose function must be "authorised", the properties of a textual world model must be tacitly agreed upon by a minimum of two parties. A fictional world only comes to 'exist' upon its moment of narration (and consequently, its moment of reception).[6]

How does a fictional world's structure become both referential and meaningful? Ryan argues that a textual world "entails a referential or 'vertical' conception of meaning" where "language is meant to be traversed toward its referents" (2001: 92). This goes against the poststructuralist view that signification exists solely as

[5] Narrative simulation is eventually an end-process that is triggered through the act of reading similar to the execution of computer code.

[6] In what can be compared to a dry run or testing of the writing process, the fictional world can be assumed to be self-narrated at first—the author doubling as a first critical reader in the same manner that the first critical gaze cast upon a work of art is the artist's.

a set of "horizontal relations between the terms of a language system" (2001: 92) and assumes a referential base, a primary world or an actual reality, from which signification emanates.[7] In other words, textual worlds need to primarily subordinate language use from the semiotic to the purely semantic during the reading process, for, as Sven Birkerts has argued, "when we are reading a novel we don't, obviously, recall the preceding sentences and paragraphs. In fact we generally don't remember the language at all, unless it's dialogue" (qtd. in Ryan 2001: 92). A fictional world may be constructed out of text, but it is read into being. The act of reading "is a conversion, a turning of codes into contents" (qtd. in Ryan 2001: 92) claims Birkerts, much like the systematic attribution of representational value to structures in Frigg's model-systems or the rendering of abstract digital inputs into meaningful and complex visual outputs in a computer simulation. In turn, it can be assumed that any linguistic or fictional construct that suppresses or delays the decoding process gradually diminishes the reader's suspension of disbelief so necessary for the reading-simulation to run.

A number of assumptions are being made here. Ryan's assertion that "language is meant to be traversed towards its referents" (2001: 92) holds true only of mimetic texts "devoted to the representation of states of affairs involving individual existents situated in time and space" but not for "universals, abstract ideas, and atemporal categories" (2001: 92). 'Vertical referentiality' is possible for referents which ostensibly exist in the primary world but certainly not possible when abstract ideas are introduced in the fictional world, to which we can add impossible referents or pseudoreferents which owe their ontology to language. We can therefore question whether fantastic other-worlds or surreal representations of this world are less believable models if their description impedes vertical referentiality.

We are faced with two constraints here: the linguistic structure that permits the system model to cohere (the world-as-text) and the source system which it is meant to emulate (the world). Both are unavoidable in textual world-building and are interdependent—the fictional world only exists because of its linguistic composition, as text. We can see how Ryan's concept of 'vertical referentiality' starts breaking down in instances where mimeticism cannot be sustained linguistically or indefinitely, especially in the description of textual worlds which are possible but nonactual, such as in Philip K. Dick's alternate histories, or the downright impossible, as in most of Jorge Luis Borges's fiction.

While it is true that Ryan treats "the text as world [as] only one possible conceptualization among many others" (2001: 90), we must look beyond the metaphor to locate the model and its functional relationship with the real. Simulation is not analogy but surrogacy. If we respond to a fictional text we do so precisely because we "imagine it as a physical, autonomous reality furnished with palpable objects and populated by flesh and blood individuals" (Ryan 2001: 92). "How could a world be imagined otherwise?" (2001: 92) adds Ryan. How indeed. We do not explicitly treat narrative as metaphor, and in cases where it is, we still seek

[7] As narratology shows, it is not just desirable but vital for the process of fictional mimesis—and simulation itself—to preserve an irreducible materialist ontology in the form of connections or indices of accessibility with the actual world. These relations have been extensively discussed in the work on possible worlds theory of Saul Kripke, David Lewis, Thomas Pavel, Marie-Laure Ryan, Umberto Eco, Lubomír Dolezel and Ruth Ronen (among others), to establish what conditions of necessity and accessibility are imposed in the creation of alternative, non-actual possible worlds (APWs).

an irreducible mimetic element that enables us to sound out the fictiveness and solidity of its referents—a principle of minimal reality. In Heim's words again, a fictional world must have "a felt totality" (qtd. in Ryan 2001: 91). Fictional worlds are therefore "existentially centred around a base we call home" (Ryan 2001: 91). The 'homeliness' or familiarity of fictive experience which grounds it to a 'felt' reality, and any reactions it invites, are well-documented, from Viktor Shklovsky's ostraneine (defamiliarisation) to Sigmund Freud's unheimlich (the uncanny; the unhomely). Literature is meant to open a 'window onto the world', allowing us to gain insight into the very world that generated it, thus the baseline for world-building is "home", the familiar, "the node from which we link to other places and other things, [the] thread weaving the multitude of things into a world", according to Heim (qtd. In Ryan 2001: 91). Ryan concurs by stating that "the most immersive texts are [in fact] the most familiar ones" (2001: 96).

5. Reader-Centric Modelling

The notion of 'home' also correlates with what Kathryn Hume refers to as "consensus reality" (1984: 23), that which "immediately refers us both to the world of the author and that of the audience" (1984: 23), in other words the real or actual world which is the basis of all forms of simulation modelling.

Consider the diagram by Hume below and reproduced in various studies on literary realism. For Hume, the work of fiction results from the reciprocal influence

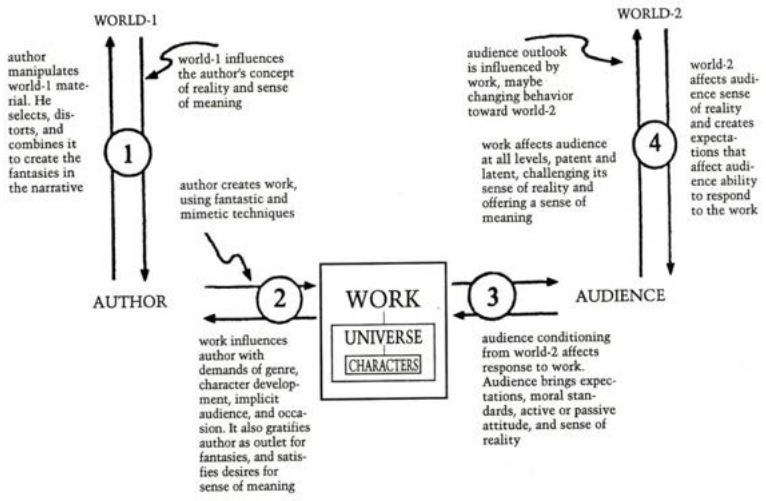


Fig. 1: World-reflection: real world phenomenology giving rise to mimetic fictional world (Hume 1984: 10)

and mediation occurring between "world-1" and "world-2" but although "world-1 is everything outside the author that impinges upon him" (1984: 9) this is not necessarily the world shared by the reader. Says Hume:

> These worlds of experience, world-1 and world-2, differ even if the artist and reader are contemporaries; world-2 indeed differs for each member of the audience. If artist and audience are separated by time, language, religion, culture, or class, the amount of shared reality may be small (1984: 9).

Hume's model of mimetic world-building is based on shared and unshared individual phenomenologies (experiences and sensations of both real and fictive worlds). The model suggests a bi-directional and intersecting process of creation: (i) the writer draws on shared/unshared reality for experience and imagination; (ii) crafts his fictional world complete with life-like or fantastic characters, settings or plots by recoursing to structures, both fantastic and mimetic, that use consensus reality as a referential base; (iii) readers self-identify (or do not) with the fictional world, which has both vestiges of world-1 (the author's) and world-2 (their own); (iv) readers' reactions to the fictional world prompts discussion and critique, and (v) the fictional world influences generic trends in fiction writing, thus opening up the mimetic-reflexive process again.

From Hume's diagram one can infer that what links author and audience is the text, which she calls "work", implying a joint authorship. However, this is inexact. Base reality is missing from the model. This serves both as the writer's point of departure in creating the work in world-1 but also the readers' benchmark for assessing and self-identifying with this work in world-2. Hume's model appears to separate writer and audience by having them occupy, influence and be influenced by their respective worlds, as if the world of the text, or the work itself, were the livable domain of the audience rather than its affective domain. From Hume's annotation to the diagram we read that "world-2 affects audience sense of reality and creates expectations that affect audience ability to respond to the work" (1984: 10). This is not wholly correct. It is the source for the modelled world which is occupied by, and phenomenologically influences, both writer and reader. This is the (mostly) shared reality from which stem both the writer's and reader's knowledge, emotions and expectations of the fictional world. This connection is not displayed in Hume's diagram, leading to the unfortunate conclusion that major divergences seem to exist between worlds-1 and 2, when in reality these only serve as metaphorical labels which have been used by Hume to represent different personal, historical or political realities (or instances of the same world) rather than different worlds.

Hume's concept of world-construction underplays the significance of a dominant and common non-fictional world for the sake of social relativism (what is represented as worlds-1 and 2 in her diagram). This is curious as she still bases her argument that "literature is the product of two impulses" (1984: 20) on "consensus reality" (1984: 20). Mimesis is "vraisemblance to the world we know" (1984: 21) while fantasy "is any departure from consensus reality, an impulse native to literature and manifested in innumerable variations, from monster to metaphor" (1984: 21). Therefore world-construction as a form of simulation modelling must take into account what aspects of the world are to be modelled, but the author must also assume a priori what aspects will diverge—or 'depart', to use Hume's word—from the dominant, and to what extent.[8] But for this to occur, a dominant must be acknowledged. Alan Palmer calls this the "source domain, the real world in which the text is being processed by the reader" (2008: 34), as opposed to the "target domain, the storyworld that constitutes the output of the reader's processing" (2008: 34). This clear distinction between a source domain and a target domain does not imply that features are not shareable or common to

[8] Conventionally, if we regard literature as the product of both mimetic and fantastic impulses, as Hume does, any convergence or divergence from the core of consensus reality is responsible for the various genres and sub-genres that are to be located along the entire spectrum.

both; in fact Palmer explains that access to the fictional storyworld occurs when readers process and negotiate knowledge from both domains (2008: 34). Access to fictional worlds is therefore reader-centric.

6. Simulation-Type Modelling in Literary Fictional Worlds

At this stage, we can synthesise concepts from narratology such as Palmer's concept of source and target domain, Hume's notions of world-1 and world-2 author-audience reciprocity, and Birkerts' assertion that reading is an act of converting code into contents to propose a valid text-as-world model (to borrow Ryan's phrase) which is fully consistent with simulation modelling and which treats it as a fully-fledged system rather than metaphor (see Figure 3 further down).

Hume's mimetic model might have its minor shortcomings however it still bears obvious similarities to simulation modelling in most respects, mostly in situating a reality external to the simulated world as its source (worlds-1 and 2); in devising a medium (the work) for users (the writer and audience) to engage with and manipulate; in suggesting an individual phenomenology (audience affect) and finally in validating personal experience (epistemology). The last two are perhaps the most crucial aspects of this model. For a simulation to matter—how we engage with it, what it can do and what we can learn from it—we demand credibility from the model. This is possible only after we have assessed the model in terms of its functional relations to the source domain.

Naim Kheir's diagram of the simulation process (Figure 2) demonstrates how properties of the physical system (reality) are modelled through a structure (system and computerised models) while the model-designer validates and verifies the system's processes. In this way, the desired match between "observed behaviour" and "predicted behaviour" is obtained in the final or system model (Kheir 1996: 5).

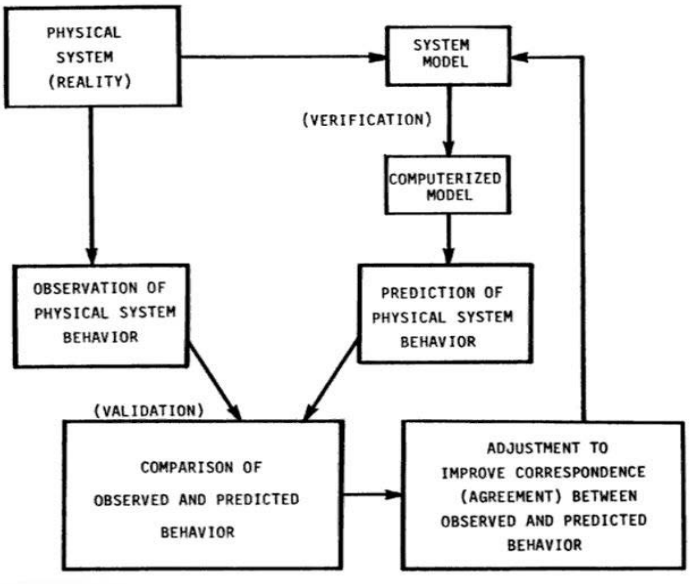


Fig. 2: Cross validation of system and real-world behaviours (Kheir 1996: 5).

The role of the model-designer is to ensure that the data generated by the system model corresponds to that acquired from the physical system to permit accurate replication. (Kheir's model implies a pre-test design and implementation phase where users are absent). Hume's model suggests a similar practice. However, while Kheir's model mostly assumes an epistemological approach to validate system behaviours, Hume's model is mostly phenomenological since it takes audience response and affect into account. Thus, in Kheir we find that an interplay of verification and validation processes is necessary to ensure that "the computerized model represents the system's model within specified limits of accuracy" (1996: 6). Until this is achieved, the model is "modified to reduce the differences between model and system behaviors" (1996: 6). In Hume's mimetic model, this process of verification and validation is implicit in the audience's reception (or rejection) of the work, which might also lead them to changing their behaviour towards world-2 (1984: 10). In the final analysis, both Hume and Kheir's models assume that faithful modelling/simulation of target behaviour or phenomena, whether rendered through text or digital medium, depends on a constant interplay between source-user-target systems, lending more credence to the idea that Ryan's text-as-world can be construed in simulationist rather than figurative terms.

The observation that narrative-as-simulation is different to other fictional narratives since it cannot be based on counter-factual rules would therefore be correct, but only insofar as the distinction with other narratives is made. It is true that while certain liberties may be, and frequently are, exercised by narratives, this cannot absolutely be the case in simulation modelling, where accuracy and credibility are sine qua nons. Thus, the argument might run, total immersion in a fictional world is possible only by removing oneself and one's experiences from the non-fictional world of external reality—a willing suspension of disbelief in the fictional world which is facilitated by readerly transportation from the actual world (Gerrig 1998). In this manner, the fictional and non-fictional world are kept distinct domains with distinct entities and rules of behaviour.

However, as we have seen, this argument is not entirely correct. Even if the reader (or "traveler" in Gerrig's words) "goes some distance from his or her world of origin" (1998: 13) this certainly does not imply that textual distancing warrants complete ontological distancing. According to a "principle of minimal departure" (Ryan 1980: 406) "we reconstrue the world of fiction […] as being the closest possible to the reality we know [making] only those adjustments which we cannot avoid" (1980: 406). Extreme variations and deviations are permissible only in the case of specific narrative genres or works where the internal laws of the fictional world hold sway. Therefore, in the simulation of narrative worlds, suspension of disbelief does not imply a suspension of primary reality but merely the heightening of a secondary one. One does not preclude the other. Indeed, as narratology shows, it is not just desirable but vital for the process of fictional mimesis—and simulation itself—to preserve an awareness of, and an anchorage, to the real.

A simulation-type model for fictional world construction is thereby being offered below (Figure 3) by assimilating some core concepts of narratology explored in this essay.

My proposed model integrates elements from simulation modelling with narratology to demonstrate areas of permeation and overlap between two representational systems:

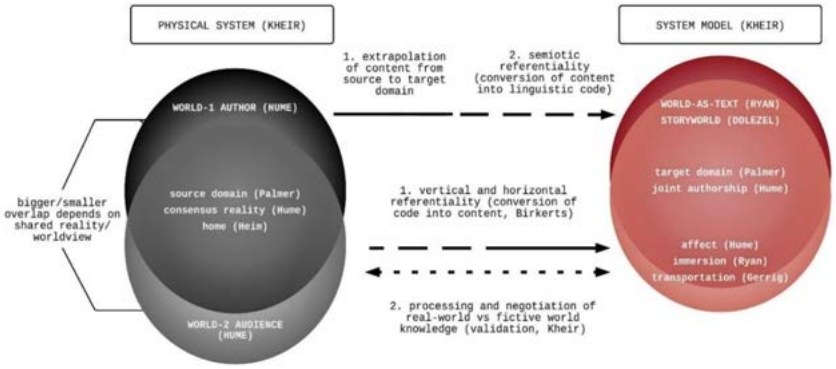


Fig. 3: Simulation-type model for the construction of a fictional world.

1. The physical system and system model are terms used by Kheir to denote the source and target systems in a simulation model. Similar to the construction of a simulation model, the construction of a fictional world entails extrapolation of content (properties, attributes, laws, reference) from the physical system to the system model. A first and irreducible materialist ontology on which behaviours are modelled and compared is therefore also present in fictional world-construction. Palmer's narratological terms for these two distinct domains are the source and target domain, both indistinguishable from any scientific discourse on modelling or simulation.
2. The source domain is essentially distinguished by its referential physicality, although it does encompass idiosyncratic worldviews, personal experience and highly individual realities. Hume treats this domain from the perspective of two worldviews (World-1 the author's, World-2 the audience's) and acknowledges that these views overlap. I have not only preserved this overlap but accentuated it since an irreducible materialist ontology—a principle of minimal reality—which enables us to sound out the fictiveness and solidity of all referents is necessary. For the sake of clarity, the source domain therefore encompasses much more than individual realities but is a (mostly) shared and therefore consensus reality (Hume 1984: 23). Cross-referencing of properties between source and target systems therefore requires a departure from consensus reality or a departure from the familiar, the concept of home according to Heim (qtd. in Ryan 2001: 91).
3. This departure occurs as a parallel and inverse process. The construction of a simulation model entails a process of substitution of content to code which maintains a valid relational status between the source and target referents. Similarly, both author and audience maintain this relational status of referentiality in the construction of a fictional world through the conversion of content to linguistic code, according to Birkerts (qtd. in Ryan 2001: 92). This referential dissolution from referent to sign and from sign to (virtual) referent is denoted by straight and broken lines in the diagram above and occurs as a near-simultaneous and inverse process in the performative act of reading (physical referent in source domain ⇢ linguistic code (system of signs) ⇢ virtual, textual referent in target domain). While this process is assumed to be natural or quasi-instantaneous, this only applies to instances

where reference is vertical and not horizontal (direct from sign to referent rather than indirect, from sign to sign, as distinguished by Ryan).

4. Depending on the complexity of the fictional world, its constituents and its narrative(s), approximation and relatability become conscious processes akin to verification and validation in simulation modelling, as proposed by Kheir (1996: 6). Knowledge, rules and laws pertaining to the fictional world are counter-checked against consensus reality until the audience is sufficiently convinced by the internal consistency of the fictional world.
5. Finally, the audience's active participation in world-(re)construction (transportation for Gerrig; willing suspension of disbelief for Samuel T. Coleridge) solidifies the construction of the storyworld (Dolezel). If the text world can be read into (imaginative) being, then its ontology becomes a shared responsibility. The extent of this joint authorship—how much of the text world is 'written' by the author and how much is 'rewritten' (reimagined) by his audience—is denoted by the overlap where the audience's immersion (Ryan) allows for full phenomenological response (or affect in Hume) to this world. In simulationist terms this effectively means that the user is the final gauge of a system's strength or correctness.

7. Conclusion

Correlations can definitely be drawn between simulation modelling and narratology. This is evident in the way models construct frames of reference for target systems through make-believe mechanisms which also validate their truth as fictions—a mechanism readily seen in narratology as a form of textual modelling. While the rules outlined in my proposed model can be applied to the construction of any type of fictional world, narratives which adopt simulationist strategies require a greater degree of audience participation and a discernible amplification of the reality principle in their construction. In this manner, the "accessibility relations"[9] (Ryan 2001: 100) of the target domain to the source is hardly questioned. Or, put otherwise, narrative simulations can therefore be said to describe possible worlds in fiction in the most possible of terms, even if the target outcome is non-actual. This is achieved purely on the basis of modelling, which finally owes much to simulation theory.

References

Casti, J.L. 1997, *Would-Be Worlds: How Simulation is Changing the Frontiers of Science*, New York: John Wiley & Sons.

Coleridge, S.T. 1817, "Chapter XIV", in Watson, G. (ed.), *Biographia Literaria*, London: J. M. Dent & Sons Ltd., 168-174.

[9] Accessibility relations are extensively discussed in possible world theory to establish what conditions of necessity and accessibility are imposed in the creation of alternative, non-actual possible worlds.

Fonseca i Casas, P. 2014, "Model-Based System Design Using SysML: The Role of the Evaluation Diagram", in *Formal Languages for Computer Simulation: Transdisciplinary Models and Applications*, Hershey: IGI Global, 236-66.

Fraassen Van, B.C. 2010, *Scientific Representation: Paradoxes of Perspective*, Oxford: Clarendon Press.

Frigg, R. 2010a, "Fiction and Scientific Representation", in Frigg, R. and Hunter, M.C. (eds.), *Beyond Mimesis and Convention: Representation in Art and Science*, Dordrecht: Springer, 97-138.

Frigg, R. 2010b, "Models and Fiction", *Synthese*, 172, 251-68.

Frigg, R. 2010c, "Fiction in Science", in Woods, J. (ed.), *Fictions and Models: New Essays*, Munich: Philosophia Verlag, 247-87.

Frigg, R. 2002, "Models and Representation: Why Structures Are Not Enough", in Dietsch, P. (ed.), *Measurement in Physics and Economics Project Discussion Paper Series*, London: The London School of Economics and Political Science, 1-42.

Gerrig, R.J. 1998, "Two Metaphors for the Experience of Narrative Worlds", *Experiencing Narrative Worlds: On the Psychological Activities of Reading*, Boulder, CO: Westview Press, 1-25.

Heim, M. 1998, *Virtual Realism*, Oxford: Oxford University Press.

Hume, K. 1984, "Critical Approaches to Fantasy", *Fantasy and Mimesis: Responses to Reality in Western Literature*, London: Methuen, 5-28.

Kheir, N.A. 1996, "Motivation and Overview", in Kheir, N.A. (ed.), *Systems Modeling and Computer Simulation*, New York: Marcel Dekker, 3-26.

Meister, J.C. 2014, "Narratology", *The Living Handbook of Narratology*, www.lhn.uni-hamburg.de/article/narratology.

Nell, V. 1988, *Lost in a Book: The Psychology of Reading for Pleasure*, New Haven: Yale University Press.

New, C. 1999, *Philosophy of Literature: An Introduction*, London: Routledge.

Palmer, A. 2008, *Fictional Minds*, Lincoln, NB: University of Nebraska Press.

Pimentel, K. and Teixeira, K. 1993, *Virtual Reality: Through the New Looking Glass*, St Louis: Intel/Windcrest McGraw Hill.

Rothenberg, J. 1989, "The Nature of Modeling", in Widman, L. *et al*. (eds.), *Artificial Intelligence, Simulation and Modeling*, London: John Wiley & Sons, 75-92.

Ryan, M.L. 1980, "Fiction, Non-factuals, and the Principle of Minimal Departure", *Poetics*, 9, 403-22.

Ryan, M.L. 2001, *Narrative as Virtual Reality: Immersion and Interactivity in Literature and Electronic Media*, Baltimore, ML: Johns Hopkins University Press.

Walton, K.L. 1990, *Mimesis as Make-Believe: On the Foundations of the Representational Arts*, Cambridge, MA: Harvard University Press.

Unlocking Limits

James Nguyen[*†‡] *and Roman Frigg*[‡]

[*]*University College London*

[†]*University of London*

[‡]*London School of Economics and Political Science*

Abstract

In a series of recent papers we have developed what we call the DEKI account of scientific representation, according to which models represent their targets via keys. These keys provide a systematic way to move from model-features to features to be imputed to their targets. We show how keys allow for accurate representation in the presence of idealisation, and further illustrate how investigating them provides novel ways to approach certain currently debated questions in the philosophy of science. To add specificity, we offer a detailed analysis of a kind of key that that is crucial in many parts of physics, namely what we call *limit keys*. These keys exploit the fact that the features exemplified by these models are limits of the features of the target.

Keywords: Scientific modelling, Representation, Limits, Keys, DEKI.

1. Introduction

Many scientific models are representations of a target system, a selected part or aspect of the world. To understand how these models work we have to understand how representation works. In our (2016, 2018) we formulate the DEKI account of scientific representation which assigns a central role to what we call a *key*: a systematic way for moving from model-features to features to be imputed to the models' targets.[1] To the extent that their targets have those features, the models in question are accurate representations.

So far we have discussed the account at a relatively high level of abstraction and said rather little about how keys work. But to understand how a model represents it is crucial to know the details of the key that accompanies it. The aim of this paper is to start filling this lacuna in the DEKI account by characterising a typical kind of key associated with many models in physics, namely what we call *limit keys*. This kind of key exploits the fact that the features of models are

[1] For a discussion of alternative accounts of representation see our 2017, 2020.

Argumenta 6, 1 (2020): 31-45
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.ngu

the results of taking certain features of the target system to a limit. Appropriately understood, these keys allow for models that radically diverge from their targets—in the sense that they are highly idealised—to nevertheless represent them accurately. As such, by making these keys explicit, the epistemic role of certain kinds of idealisation is clarified. However, as we will see, a limit key can only be invoked under particular conditions. Specifying these conditions forces us to pay careful attention to certain choices scientists make in the construction of their models, and doing so sheds a new light on certain controversies about models. Thus, this paper's contribution is threefold. First, it develops the DEKI account of scientific representation by adding an analysis of limit keys. Second, it illuminates a certain area of scientific practice by scrutinising the epistemic function of taking target-features to a limit in a model. Third, it demonstrates how such models can be accurate despite being idealised, thereby contributing to our understanding of the epistemic value of idealisation.[2]

We proceed as follows. In Section 2 we briefly recapitulate the DEKI account of representation. In Section 3 we introduce limit keys. Section 4 illustrates how, and under which conditions, they work via some simple examples. Section 5 discusses the methodological assumptions that underpin the use of limit keys 'in the wild', where the relevant features that have been taken to a limit, and the nature of these limits themselves, are assumed (as part of scientific practice, rather than rigorously proven) to be relatively well-behaved. Section 6 concludes.

Before we begin, it's worth commenting on how the use of limit keys to underpin cases of scientific representation contributes to our broader philosophical account of scientific modelling. We have argued elsewhere (Frigg 2010a, 2010b; Frigg and Nguyen 2016) that scientific models should be thought of as akin to works of fiction. Now, it's important to note that this claim concerns the *ontological* status of scientific models. As such, it is only a *part* of a complete philosophical account of model-based science. The fiction view of models tells us what models are, but not how they function representationally. The DEKI account of scientific representation is thus designed to supplement the fiction view by providing an account of how a scientific model, thought of as a work of fiction, might represent a target system. For our current purposes, the fictional nature of the models in question is left in the background, since we focus on the nature of a particular kind of key.[3]

2. DEKI

The DEKI account of scientific representation provides a general framework for thinking about the representational relationship between models and their targets. The framework specifies four conditions that must be met for a scientific model M to represent a target system T so that reasoning about the former can

[2] Our account thus avoids regarding idealised models as falsities, or misrepresentations. This comes at the costs of rejecting the notion that models have to be interpreted literally. For a discussion of this point see our 2019 and Nguyen 2019.

[3] But see our 2016 for a discussion of the interplay between DEKI and the fiction view of models more generally.

generate hypotheses about the latter. The conditions, which also give that account its name, are denotation, exemplification, keying up, and imputation.

The first condition is that *M denotes T*. Denotation is a two-place relation. A name denotes its bearer; a map denotes its territory; a portrait denotes its subject; and a model denotes its target. Denotation is necessary but insufficient for scientific representation. It's necessary because it establishes the bare sense in which *M* is about *T*. It's insufficient because it doesn't account for how we can reason about target systems via investigating their models, which is what Swoyer (1991) calls 'surrogative reasoning'. DEKI's other three conditions are designed to explain this.

The second condition is that models *exemplify* certain features.[4] Exemplification is instantiation plus reference: something exemplifies a feature if it at once possesses that feature and refers to it. This can be illustrated with Goodman's (1976: 52-56) example of a tailor's book of fabrics. The swatches both instantiate the particular kind of cloth they are—e.g. herringbone or pin-stripe—and also refer to these cloth-properties themselves.

Now, whilst scientific models may exemplify certain features, these features needn't be carried over to their target directly. A piece of litmus paper dipped into an acidic solution exemplifies redness, but it doesn't represent the solution as being red. Rather, the litmus paper—understood as a representation—comes with a *key* which systematically relates colours to pH values. Similarly, whilst a map exemplifies a certain distance between, say, the marks that are labelled 'Newcastle' and 'London', this distance isn't carried over directly to the cities themselves: rather the map comes with a key specifying a scale with which to systematically relate map-distances to the actual distances that the map represents. The DEKI account insists, and that's the third condition, that models function like litmus paper or maps in that they come with a key that associates model-features with target-features. In general terms, a key is a mapping which takes as arguments the exemplified features $P_1, \ldots, P_n$ of *M* and delivers as values some (possibly, but not necessarily, identical) features $Q_1, \ldots, Q_m$.[5]

The final condition is that the model user *imputes* at least one of $Q_1, \ldots, Q_m$ to *T*. If *T* has the feature imputed, then the representation is accurate in that respect. If it doesn't, then *M* still represents *T* as having such a feature; it's just a *mis*representation in that respect.

Tying these conditions together delivers:

DEKI *M* represents *T* iff

1. *M* denotes *T*;
2. *M* exemplifies features $P_1, \ldots, P_n$;
3. *M* comes with a key *K* which associates exemplified features $P_1, \ldots, P_n$ with features $Q_1, \ldots, Q_m$; and
4. a model user imputes at least one of $Q_1, \ldots, Q_m$ to *T*.

[4] We place no restrictions on what counts as a feature. In the current context, (one-place) properties, *n*-place relations, functions, solutions to equations of motion, and structural relationships, among others, count as features.

[5] We are not claiming that there is an easy way to dissociate different model-features, nor that the key is insensitive to relationships between them. This is just a schematic rendering of how keys work, additional constraints may be required.

DEKI provides a general framework in which to think about the relationship between models and their targets, and the framework needs to be filled in in particular cases. In order to understand a particular instance, or style, of scientific representation, the ways in which the conditions are met need to be further explicated. Our concern in this paper is the third condition. What associations between model-features and features to be imputed to the target are there, and how does a key encode them? Our goal here is to illustrate how the account works, and to illuminate a particular kind of reasoning, namely where the key in question exploits the notion of a limit. As we discuss below, by analysing this kind of reasoning in terms of DEKI, we also gain additional understanding of the role of (at least one kind of) idealisation in science.

3. Limit Keys

Many models exemplify 'extremal' features: model-planes are frictionless, model-gases have an infinity of molecules, and model-planets are perfect geometrical spheres. What do models exemplifying such features tell us about target systems that don't, and never will, have such features? The core idea that we develop here is that (at least some) models of this kind should be interpreted as being equipped with a limit key: a key that exploits the fact that the model-features can be understood as resulting from taking certain features of the target to a limit.

To give a definition of limit keys and analyse them, we must first introduce limits. We restrict our attention to two cases: number sequences and function sequences. A *number sequence* is a list of numbers linked by a rule. The list is usually indexed by an index α and the rule is given by an operation. As an example, consider the sequence $1/\alpha$ for $\alpha = 1, 2, 3, \ldots$. We follow an often-used convention and write such sequence as f_α. In our example we have $f_\alpha = 1/\alpha$. Although intuitive, nothing depends on the index being a natural number (in the next section we will see an example where α is a real number).

We can now ask how f_α behaves if α tends toward infinity. That is, we can consider the *limit* of f_α for $\alpha \to \infty$, where the symbol '∞' denotes infinity. If that limit exists and has value L, we write $\lim_{\alpha\to\infty} f_\alpha = L$. The question now is how a limit can be defined precisely and under what circumstances it exists. The standard definition of a limit is couched in terms of positive real numbers ϵ (where 'positive' means $\epsilon > 0$). These numbers can be arbitrarily small, but never equal to 0. Then, the limit L of the sequence f_α for $\alpha \to \infty$ is defined as follows:[6]

(1) $\lim_{\alpha\to\infty} f_\alpha = L$ iff $\forall \epsilon > 0\ \exists \alpha'$ such that $\forall \alpha$: if $\alpha > \alpha'$, then $|f_\alpha - L| < \epsilon$.

Intuitively this means that we can keep f_α as close to L as we like by making α sufficiently large. Limits with $\alpha \to \infty$ are also referred to as infinite limits. If it is not possible to keep f_α as close to L as we like by making α sufficiently large, then the infinite limit does not exist. If the limit exists, we say that the sequence f_α converges toward L. Consider again the previous example of $f_\alpha = 1/\alpha$. We can now take the limit of this sequence for $\alpha \to \infty$ and it is obvious that $\lim_{\alpha\to\infty} 1/\alpha = 0$.

Infinite limits can be taken irrespective of whether α is a natural number or a real number. When we look at cases where α is a real number, we can also ask

[6] This, and the below definition of a finite limit, are standardly stated in books on calculus. See, e.g., Spivak 2006: Chapter 5.

how the sequence behaves when α tends toward a particular (finite) value a. For instance, we can ask how f_α behaves when α tends toward zero, or toward five. The standard definition of such a limit is couched in terms of two positive real numbers, ϵ and δ (where, as previously, 'positive' means that both $\epsilon > 0$ and $\delta > 0$). The definition then says:

(2) $\lim_{\alpha \to a} f_\alpha = L$ iff $\forall \epsilon > 0\ \exists \delta > 0$ such that $\forall \alpha$: if $0 < |\alpha - a| < \delta$, then $|f_\alpha - L| < \epsilon$.

Intuitively this means that we can keep f_α as close to L as we like by keeping α close to a. If this is not possible, then the limit does not exist.

It's crucial not to conflate the limit of a sequence with the value of the sequence at the limit: L and f_a are not the same mathematical objects. To see this, consider the case where $\alpha \to a$. Since the definition of the limit requires $0 < |\alpha - a| < \delta$ (that is, the limit requires that $|\alpha - a|$ has to be strictly greater than 0), α will never be equal to a in taking the limit. So the limit L reflects how f_α behaves when α comes arbitrarily close *a without reaching it*. It does *not* reflect the value of f_α if $\alpha = a$. The same holds for infinite limits: because α tends towards ∞ without ever reaching it, L is not the same as f_∞. To express this difference clearly, we call L the *limit value* and refer to f_α (or f_∞) as the value at the limit.[7]

That two values are conceptually distinct does not mean that their numerical values must be different. If both the limit value and the value at the limit exist and are equal, then the limit is a *regular limit*; if they are different it's a *singular limit* (Butterfield 2011: 1077).[8]

We will see examples of both cases later. Before discussing examples, we can now say what a limit key is. Let the target system have a feature of interest corresponding to some value in the sequence f_α. To study the target, we construct a model in which the parameter α assumes the extremal value. Let us begin with a finite value a. This means that the feature exemplified by the model is f_a. Now assume (i) that the limit L of f_α exists for $\alpha \to a$; (ii) that the value f_a at the limit exists; and (iii) that the limit is regular (i.e. that $L = f_a$). Under these assumptions it follows that for all ϵ there exist a δ such that for all α, if $|\alpha - a| < \delta$, then $|f_\alpha - f_a| < \epsilon$. This can be exploited. If we consider a limit $\alpha \to a$, the model user can infer that as long as α in the target is not more than δ away from a in the model, the value of f_α in the target is no more than ϵ away from f_a in the

[7] In cases where the extremal value is ∞, the below discussion regarding the value at the limit requires that we specify what this value is. In the case of number sequences we can follow Butterfield (2011: 1075) and consider the sequences as containing elements from $\mathbb{N} \cup \{\infty\}$ (or $\mathbb{R} \cup \{\infty\}$), where '$\mathbb{N}$' denotes the natural numbers and '$\mathbb{R}$' denotes the real numbers. This is standard practice in the physics literature where the idea of a 'natural infinite system' corresponding to a system at an infinite limit is often invoked; see, for example, Ruelle's discussion of phase transitions as only occurring in systems that are 'idealized to be actually infinite' (2004: 2).

[8] Although note that Butterfield recommends caution with respect to the use of the term 'singular limit', given the variety of meanings one finds in the literature (see Butterfield 2011: 1068). It's worth noting here that Butterfield uses the phrase 'non-singular' limit to refer to both cases where the limit exists, and is equal to the value at the limit, and cases where the limit exists and there is no obvious value at the limit. Given our current purposes (where we're investigating models which are 'at the limit' so to speak), our use of 'regular limit' is restricted to the first kind of 'non-singular' limit.

model. Or, more colloquially, if the parameter α in the target is close to the model value, then the feature f_α in the target is close to f_a in the model. In this way knowing the model feature gives information about the target feature. If a model user employs knowledge of limits in this way to infer from a model-feature to target-feature she uses a *limit key*. Such a key works by taking the exemplified feature in the model, f_a, and converting it into a logically weaker property: having a feature in the interval $(f_a - \epsilon, f_a + \epsilon)$. It is this weaker feature that is imputed to the target system. In terms of the symbolic notation introduced in the last section, f_a is P and Q is having a feature in the interval $(f_a - \epsilon, f_a + \epsilon)$.[9]

The argument is *mutatis mutandis* the same if we consider an infinite value. In this case the feature exemplified by the model is f_∞. Assume that the limit for $\alpha \to \infty$ is regular. Then the model user can infer that if α in the target is larger than a threshold α', then the value of f_α in the target is no more than ϵ away from f_∞ in the model.

We can now turn to *function sequences*. The difference between number sequences and function sequences is that a function sequence is not a sequence of numbers but a sequence of functions $f_\alpha(x)$. The functions can be of any kind, but to keep things simple we consider real valued functions: $f_\alpha : \mathbb{R} \to \mathbb{R}$, where, as before, '$\mathbb{R}$' denotes the real numbers. An example of such a sequence is $f_\alpha(x) = x^{-\alpha}$. A function sequence can converge toward a limit function in different ways. One of the simplest is *pointwise convergence*: the function sequence $f_\alpha(x)$ converges pointwise toward the function $L(x)$ iff for every $x \in \mathbb{R}$ the value of $f_\alpha(x)$ converges to $L(x)$. If this is the case, we write $\lim_{\alpha \to a} f_\alpha(x) = L(x)$, and *mutatis mutandis* for $\alpha \to \infty$. We call $L(x)$ the *limit function* and $f_a(x)$ the *function at the limit*. As before, the limit function and the function at the limit can, but need not, be the same. If they both exist and are identical, then the limit is regular; if not, then it's singular.

Function sequence limits can be used to reason with the model about the target in the same way as number sequence limits. If the limit is regular it follows that for all x and for all $\epsilon > 0$ there exists a $\delta > 0$ such that for all α, where $|\alpha - a| < \delta$, we have $|f_\alpha(x) - f_a(x)| < \epsilon$ (and, again, *mutatis mutandis* for $\alpha \to \infty$). This means that as long as (for each value of x) α in the target is not more than δ away from a in the model, the function $f_\alpha(x)$ in the target is no more than ϵ away from $f_a(x)$ in the model.[10] The limit key works by taking the exemplified feature of interest in the model, $f_a(x)$, and converting it into a logically weaker feature of interest, namely that the target's feature is somewhere in the interval $(f_a(x) - \epsilon, f_a(x) + \epsilon)$ for all x, which is imputed to the target system.

4. Toy Examples: Stairs and Slopes

Let's see this kind of reasoning in action with two toy examples: one where it works and one where it breaks down. In order to understand a method, it's often illustrative to see where it fails. So we start with an example, based on a number

[9] We drop the subscripts on the P and Q from here on for ease of notation since we're only dealing with a single exemplified model-feature and connecting it to a single feature to be imputed to the target.

[10] Since we're using the notion of pointwise convergence, the values of δ (and ϵ) can vary across different values of x.

sequence with a singular limit, where the limit reasoning fails. We then turn to an example where it works via a function sequence with a regular limit.

Assume that your target system is a set of stairs that you want to carpet. To buy the right amount of carpet you need to know the stairs' total length. The staircase in which the stairs are located has the shape of a right-angled triangle with both sides having unit length, and with the stairs sitting on the hypotenuse. Further suppose that there are a large number of stairs in the staircase and you somehow cannot work out their total length. You therefore resort to a model.

Let $\alpha = 1, 2, \ldots$ be the index of a number sequence. You start with a staircase with two stairs and every time you progress to the next index you double the number of steps in the staircase: for $\alpha = 1$ the staircase has two steps, for $\alpha = 2$ four steps, for $\alpha = 3$ eight steps, and so on. This is illustrated by the three images to the left in Figure 1. In general, for staircases in our sequence, the staircase with index α has 2^α steps. The dependant feature of interest, f_α, is the length of the stairs with index α; that is, f_α is the length of the set of stairs with 2^α steps. The number of steps seems so large to you that your model is a fictional scenario in which the stairs consist of an infinite number of steps. But a staircase with an infinite number of steps is a line, and so this idealisation results in a model, as shown by the 'staircase' to the right in Figure 1, where the length of the stairs is the length of the hypotenuse of a right-angled triangle whose other sides are of unitary length: $f_\infty = f_{\text{model}} = \sqrt{2}$.

You of course know that the number of steps is not infinite, but you think that this is not a problem because you can use a limit key. The number of steps is large, and you think that it is in fact large enough for the length of the model-stairs to be close enough to the length of the real stairs for all practical purposes, in particular to buy the right amount of carpet.

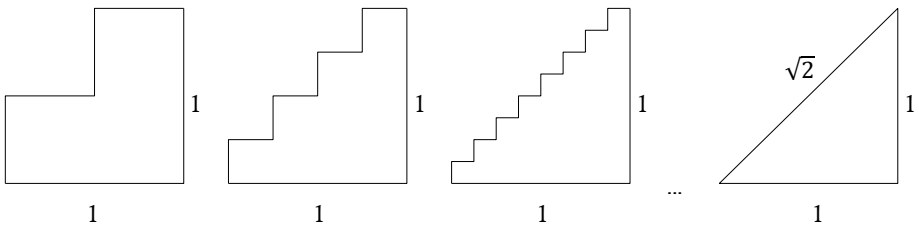


Figure 1: A sequence of staircases, with the value at the limit

This is mistake. Looking at Figure 1, it's easy to see that the total length of the stairs is two *irrespective* of the number of stairs: $f_\alpha = 2$ for all $\alpha = 1, 2, \ldots$. Hence, trivially, $\lim_{\alpha\to\infty} f_\alpha = 2$. So $L \neq f_{\text{model}}$. This shows that the limit is singular and we're now in a position to see how reasoning with a limit key breaks down (we're using definition (1) since we're dealing with an infinite limit). From $\lim_{\alpha\to\infty} f_\alpha = 2$ we know that for every $\epsilon > 0$ there is an α' such that: for all α, if $\alpha > \alpha$', then $|f_\alpha - 2| < \epsilon$. But applying the limit key would amount to mistakenly assuming that for all $\epsilon > 0$ there is an α' such that for all α, if $\alpha > \alpha$', then $|f_\alpha - \sqrt{2}| < \epsilon$. This is false. In fact, for any $\epsilon < 2 - \sqrt{2}$ there is no α' such that for all α, if $\alpha > \alpha$', then $|f_\alpha - \sqrt{2}| < \epsilon$. So no matter how many stairs there are, the length of the stairs doesn't come close to the length of the hypotenuse, not even in the limit for the number of stairs toward infinity! This is why the limit key doesn't

work here, and you would buy the wrong length of carpet if you were to reason in this way. So by using a limit key in a case where the limit in question is singular, the model yields wrong results.

Our second example works with a function sequence and provides an illustration of a case where limit keys work. Suppose your target system is a ski-jumper and you want to know how her position on the slope changes through time. To this end you construct a model, which is a fictional scenario consisting of a rectangular object sliding down a perfect geometrical plane with an inclination of θ. The materials of the object and the plane are such that there is no friction between them, and the only force acting on the object is the linear gravitational force $\vec{F} = mg$, where g is the gravitational constant on the earth's surface. With some simple trigonometry we can calculate the magnitude of the component of the force acting on the object parallel to the surface of the plane: $f^{\parallel}_{\text{model}} = mg \sin(\theta)$, as displayed in Figure 2.

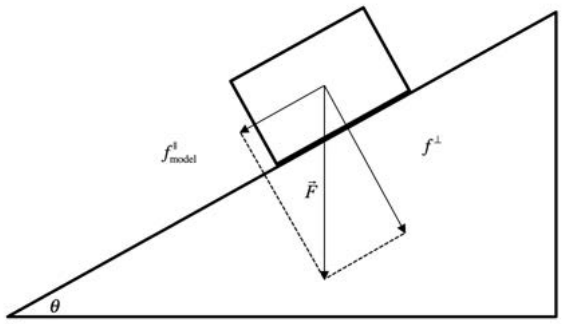


Figure 2: Ski-jumper model

Using Newton's equation, and without loss of generality setting the original position and initial velocity to zero, delivers the following position function along the slope for the object:

(3) $x_{\text{model}}(t) = 1/2t^2 g \sin(\theta)$.

This function is an exemplified feature of the model, and in the idiom of DEKI it is *P*.

We know perfectly well that the real slope isn't a frictionless perfect plane, and that there are forces other than gravity acting on the skier such as air resistance and the Coriolis force. Given this, what does the model tell us about the real-world skier's position? To answer this question we need a key. In keeping with the spirit of our above discussion, we understand the model as a limiting case of the real situation and aim to construct a limit key.

To make a start, let us assume that the only force acting on the skier not taken into account in the model is friction, and that friction is linear. This is a strong assumption and we come back to it later; let's run with it for now to see how the reasoning works. The magnitude of the friction force acting on the skier then is proportional to the magnitude of the force perpendicular to the plane, $f^{\perp} = mg \cos(\theta)$, where the proportionality constant is the friction coefficient μ. Then, the actual force acting on the skier parallel to the slope is given by $f^{\parallel} = mg \sin(\theta) - mg\mu \cos(\theta)$. This means that the actual position function of the skier is:

(4) $x_{\text{friction}}(t) = 1/2t^2g(\sin(\theta) - \mu\cos(\theta))$.

Now regard μ as a freely varying parameter and notice the following relationship between $x_{\text{model}}(t)$ and $x_{\text{friction}}(t)$:

(5) $\lim_{\mu \to 0} x_{friction}(t) = x_{\text{model}}(t)$.

To see this, and to connect it to our above definition of a limit, it suffices to notice that the relevant δ for each ϵ is given by:

(6) $\delta = 2\epsilon/\cos(\theta)t^2$.

It's then easy to see that the condition in definition (2) is satisfied (for all values of t) and that the limit function is equal to the function at the limit. Hence the limit is regular. This allows us to use a limit key: for all times t and for any $\epsilon > 0$, it is the case that as long as $\mu < \delta$, it's guaranteed that $|x_{\text{friction}}(t) - x_{\text{model}}(t)| < \epsilon$. In words: as long as the friction coefficient in the actual system is less than δ, the position function in the model will differ from the actual position function by less than ϵ.

In the terminology of DEKI, the feature exemplified by the model, P, is $x_{\text{model}}(t) = 1/2t^2g\sin(\theta)$. The feature Q is: the position of the skier in the target is in the interval $(x_{\text{model}}(t) - \epsilon, x_{\text{model}}(t) + \epsilon)$ at all times t, where ϵ depends on the lower bound the model user can set on the value of μ in the target. The key then acts to connect feature P to feature Q. We can think about the key as a mapping from the exemplified features to the features to be imputed to the target. So, $K(P) = Q$. The value of the key, i.e. the feature Q, is then imputed to the target. Interpreted in this way, the model is an accurate representation (because the position function of the skier does actually fall within the bound imputed).

It's important to note that this doesn't rely on the idea that the friction acting on the skier is in any sense negligible or makes no difference to her movement. The exact same reasoning can be applied to all skiers irrespective of what the level of friction in the target is. Even if friction plays a significant role in the target system, Equation (6) can be used to say how the real skier moves in exactly the same way in which it is used in situations in which friction is small. We can use the frictionless model to impute Q as above, the only difference being that the interval defining Q is wider. And this will still result in the model being an accurate (albeit logically weak) representation.[11]

5. Limits in the Wild

Let us now return to our assumption that that the only force acting on the skier not taken into account in the model is friction, and that friction is linear. This assumption allowed us to specify the ϵ and the δ explicitly and prove that the limit exists. We made this assumption to illustrate how limiting reasoning works. It is, unfortunately, unrealistic in two ways. First, there are known unknowns: even when further factors are known, it is not always possible to calcu-

[11] Thus, our approach diverges from Strevens' (2008, Chapter 8). According to him, idealisations work by deliberately misrepresenting non-difference makers by taking a parameter representing them to an extremal value. Using limit keys allows distortions to accurately represent systems even where they *do* make a difference. In fact, they allow us to quantify the difference that they make by means of the size of the interval that results after applying the key.

late their effect explicitly. We know that the real slope is uneven in various ways and that this unevenness has an effect on the real skier's motion, but we cannot capture this effect mathematically. Nor can we calculate the effect of air resistance that crucially depends on the skier's shape, which we know not to be a rectangular block! And so on. So we cannot always explicitly specify the difference between a model and the target as we did in the last section; linear friction is a special case in that regard. Second, and worse still, there may be unknown unknowns: we may not know all the factors that influence a situation. For example, the skier may be subject to forces we don't know. Knowing all the relevant factors would require a God's eye perspective that mortal scientists don't have. The consequence of this is that in practice we cannot neatly quantify the differences between model and target, and we cannot rigorously prove that the model is a regular limit of sequence that contains the real-world target.

But it remains that when we reason using a limit key, we're relying on the existence of such a limit. In the abstract, such a key requires the following. We have a model with a particular exemplified feature (P). We assume that the model is the system that would result, were we to take all of the potentially relevant features of the target to a certain limit. As such, by exploiting this, we can reason from the fact that the model exemplifies P, and assuming that the model is the result of taking all of the relevant limits of the target, that the target's feature of interest will be within the interval $(P - \epsilon, P + \epsilon)$ around the feature P exemplified by the model (where ϵ will depend on the limit in question) . In terms of DEKI, Q is 'being in the interval $(P - \epsilon, P + \epsilon)$', and Q is imputed to the target. Now, whether or not the result of this reasoning, i.e. whether target's feature of interest is in this range, is true will depend on whether it is the case that by taking all the limits of features in the target we will in fact arrive at the model in question. And this is usually not the sort of thing that admits mathematical proof.

Does the fact that we cannot prove that the limit exists pull the rug from underneath limiting reasoning? For those who require mathematical proofs, yes. But there are rarely, if ever, mathematical proofs backing the successful application of a model to the world.[12] What scientists will do in this situation is to form a qualitative judgement against their background knowledge. They will take into account everything they know about forces and their effect on bodies, and they will make a qualitative estimate of the magnitude that this effect will have on the skier. This will give them an interval $(x_{\text{model}}(t) - \epsilon^e, x_{\text{model}}(t) + \epsilon^e)$, where the superscript 'e' stands for 'estimate', of which they will be willing to say that the real position of the skier will lie in that interval given everything they know about forces. This defines a feature Q^e that they can then impute to the target.

Limits have not become obsolete. The justification for imputing Q^e rests on the belief that a limit exists and that the model function is only so far away from it. Let us spell this out in more detail. Meet an old friend: Laplace's Demon

[12] And there are good reasons to doubt that we should expect there to be such proofs. Whether or not a model is an accurate representation depends on features beyond the model: it depends on the nature of the target system in question. As such, whilst we may be able to prove that if the target is such that by taking its relevant features to the limit we arrive at the model, then the model will allow us to reason successfully about the target, the antecedent of this conditional isn't the sort of thing that admits mathematical proof.

(Laplace 1814). The Demon knows all the forces and can write down the true position function $x_{skier}(t)$ of the skier. This function will depend on a myriad of parameters. The claim that scientists—mostly implicitly—rely on is that if the Demon now took all of the parameters in $x_{skier}(t)$ to their values in the model, that limit would turn out to exist and to be regular. That is, they assume $\lim x_{skier}(t) = x_{model}(t)$, where we write 'lim' (without subscripts) to indicate that the limit is taken for *all* parameters. Of course, $\lim x_{skier}(t) = x_{model}(t)$ is not provable, not least because human scientists, lacking the powers of the Demon, don't have access to $x_{skier}(t)$. It is a transcendental assumption in the sense that it must be made for it to be possible to apply the model using a limit key even though the assumption cannot be proven. But it is an assumption that scientists must make if they are to assume that the model is informative about the target (through a limit key). If the limit does not exist, or if it is singular, then there is no reason to assume that the target behaves like the model, even if the model's parameter values are close to the target's parameter values.

Carpets and ski jumpers are toy examples. But the same inferential patterns are at work in 'real' applications. Consider the Newtonian model of a planet's orbit. The model involves scientists imagining the following fictional scenario: two perfect spheres, both with a homogeneous mass distribution, are placed in otherwise empty space. One is much more massive than the other, and the only force acting on the spheres is the gravitational attraction between them. Combining these assumptions with Newton's second law, assuming that the heavier sphere is at rest, and letting $\vec{x}$ be the vector pointing from the centre of the heavier sphere to the centre of the lighter sphere, gives an equation of motion for the planet in the model: $\ddot{\vec{x}} = -Gm_s\vec{x}/|\vec{x}|^3$, where m_s is the mass of the heavier sphere, and G is the gravitational constant. The trajectory $\vec{x}_{model}(t)$ of the model planet is the solution of this equation.

This equation of motion isn't the exact equation of motion governing the actual planet: even supposing that Newtonian mechanics were correct, the actual force that determines how a planet moves includes forces beyond its gravitational interaction with the sun. So we have an exemplified feature of a model, $\vec{x}_{model}(t)$, which we know doesn't match any actual feature of the target. What, then, does the motion of model-planet tell us about the motion of a real planet? The answer, we submit, is provided to us by a limit key. We should think of the actual trajectory $\vec{x}_{planet}(t)$, available to the Demon but not to us, as being such that if the Demon took all the parameters in $\vec{x}_{planet}(t)$ to limits corresponding to their value in the model—presumably most of them will be taken to zero given they don't appear in $\vec{x}_{model}(t)$—then the Demon would find that $\lim\vec{x}_{target}(t) = \vec{x}_{model}(t)$. If we combine this result with the assumption that the actual value of these parameters in the real world are not too far away from their values in the model, we can infer that the model trajectory is not too far away from the real trajectory.[13]

[13] Here we state the model-target relationship in terms of the model being 'close' to the real system, as standardly presented in physics. As noted above, limit keys obviously cover such cases, but they're not restricted to situations where the model is 'close' to the target. They just require that there be the right kind of systematic relationship between the parameter values and trajectory.

This kind of reasoning has been incredibly successful throughout the history of physics, and indeed engineering. From planetary motion to rocket launches, it has worked successfully in countless applications. This lends credibility to the use of limit keys in mechanics, and it makes scientists confident that limit keys will also work in future applications. It is important to realise, however, that inductive support for limit reasoning does not 'prove' the method right. In fact, scientists have worried about these limits time and again and delimiting the scope of their successful use has been a scientific endeavour in its own right. As an example, consider Poincaré's study of the role of initial conditions. Among the parameters that $\vec{x}_{\text{planet}}(t)$ contains are the planet's position and momentum at a certain initial time t_0. This is because Newton's equation of motion tells us where a planet is at a later time $t > t_0$ only if we specify the planet's position and momentum at some initial time. This specification is the planet's *initial condition*. In practice scientists can only ever specify an approximate initial condition because it's impossible to measure the condition with absolute precision.

Limit reasoning then would say that if the initial condition in the model is sufficiently close to the initial condition of the real planet, then the model-trajectory is sufficiently close to the real planet's trajectory (the comment in footnote 13 applies again here). Scientists took this assumption for granted until Poincaré showed that it was not true in general. Poincaré studied what is now known as the three-body-system, which is exactly like the Newtonian model except that it has a third sphere in it. If you want an interpretation, you can think of these three spheres as the sun, the earth, and the moon. What Poincaré found was that the three-body-system exhibits what is now known as sensitive dependence on initial conditions: even if two initial conditions are arbitrarily close, their trajectories can diverge. This effect is now also known as *chaos*.[14] This means that the limit does not exist and hence the model cannot be equipped with a limit key. This has far reaching consequences. Specifically, it means that Newton's model cannot be equipped with limit key and be expected to provide true results concerning a planet's trajectory, at least not universally and unrestrictedly. What exactly the restrictions are is a question that is discussed in the discipline of chaos theory. The details are beyond the scope of this paper, but one of the crucial results is that in contexts like the ones that Poincaré considered a limit key can be expected to deliver correct results only for finite time spans. So chaos theory tells us that the transcendental assumption is justified only for finite times.

And questions about limits go beyond initial conditions. What happens if the dynamics of the target system is different from the dynamics of the model in certain respects? This question promoted a study of what is now known as structural stability, which continues to date.[15] So the study of the boundaries of limits

[14] For a discussion of Poincaré's discovery of sensitive dependence on initial conditions see Parker 1998 and for discussion of the implications of chaos for predictability see Werndl 2009. For accessible introductions to chaos see, for instance, P. Smith 1998 and L.A. Smith 2007. For an advanced discussion see, for instance, Lichtenberg and Lieberman 1992.

[15] For technical discussion of results see Pilyugin 1991. Frigg *et al*. 2014 provide an accessible introduction and a discussion of philosophical consequences.

is not only a philosophically interesting issue; it is also a field of active scientific research.

6. Conclusion

Limit keys provide a concrete example of the sort of keys that the DEKI account of scientific representation urges we should focus on when investigating what our scientific models tell us about the world. Understanding how they work contributes to our broader understanding of scientific representation, and indeed the epistemic value of idealisation. Moreover, as demonstrated by the previous discussion, by requiring that we specify the key, thinking about (at least some) models in physics through the lens of DEKI helps us understand what sort of methodological assumptions underpin the use of those models. In order to understand how such models work, we have to pay careful attention to which features of a model are exemplified, and which features of its target are taken to which limit.

This lesson generalises to other, more philosophically contentious, models. For example, the Ising model of ferromagnetism invokes the thermodynamic limit, and is thus set on an infinite lattice (Baxter 1989). Given that its target systems—iron bars for example—do not consist of an infinite number of particles, how should we understand the idealisation present in the model? In this case, the problem is particularly pressing since the model in question underpins much of our current understanding of phase transitions. In the case of (the original interpretation of) the Ising model, the phase transition consists in an iron bar shifting between ferromagnetic and paramagnetic phrases, a transition which is understood as being represented by the occurrence of a non-analyticity in the model's free energy function. Taking the lattice to the infinite limit is *necessary* for the model to exhibit such a transition: for mathematical reasons, a non-analyticity cannot occur in the free energy function of a system with a finite particle number, and hence phase transitions—defined as non-analyticities—cannot occur in systems with finitely many particles. For this reason, physicist David Ruelle says that phase transitions only occur in systems that are 'idealized to be actually infinite' and that this 'idealization is necessary' (2004: 2).

In the DEKI framework, the way of analysing what the model tells us about actual, finite systems, requires specifying a key linking an exemplified feature of the model with a purported feature of the target. As such, we need to specify which feature of the target we're interested in, and how it's related to the relevant feature of the model. There are two available options. The first option is to take the relevant feature of the model to be the non-analyticity of the free energy function; in which case we are in a situation where we have a sequence of systems, each a finite lattice lacking such a feature, and an infinite model at the limit of such a sequence, having such a non-analyticity. Such a position is advocated, for instance, by Batterman (2001, 2011) who argues that the infinite model is different from the finite systems and that phase transitions are therefore emergent phenomena. Under this interpretation we have an example of a singular limit, and, as argued above, we cannot reason about the target based on the limit-key.

An alternative approach is recommended by Butterfield (2011) who argues that the relevant feature is not the non-analyticity of the free-energy function, but rather the free-energy function itself (or more specifically, the magnetisation

of the lattice, which is the partial derivative of the free energy with respect to the external field).[16] In this case, if we again consider a sequence of lattices, we have a sequence of free-energy functions that converges pointwise to the free-energy function of the model (this is despite the fact that each of the free-energy functions on finite lattices is analytic, and the model's free-energy function is not). In which case we can employ the limit key strategy which we discussed in the last section.

Which of these points of view is correct is a deep question in the foundations of physics that we cannot address in this paper. Our aim here is a different one, namely to show that in order to reason using the limit key, the model must exemplify a feature that is the *regular* limit of a target-feature. Where an exemplified feature is like this, the key allows us to export a feature from the model to the target that the latter actually has. Conversely, if the exemplified feature is not like this, using a limit key will make the model an inaccurate representation.

This provides two general morals. First it demonstrates that properly understanding these cases of model-based science requires paying careful attention to which features of the models are exemplified, and which specific features of the target system are taken to which limit. The discussion of the Ising model generalises. Choosing a particular feature of the target system to focus on, and constructing a model that takes it to the limit in the right way, is a significant aspect of scientific modelling. Understood in the way we're urging, it is paramount that any model employing extremal features is evaluated carefully in terms of limits, and of how those limits are constructed. Second, and more generally, it demonstrates that limit keys provide concrete examples of the keys invoked in the DEKI account of scientific representation, thereby illuminating how it is to be explicated in practical applications. As applied to models that are idealised in the sense discussed here, this also demonstrates how idealisation—understood as the, sometimes radical, distortion of a relevant feature of a target—can play a positive epistemic role, despite, or even better, in virtue, of that distortion.[17]

References

Batterman, R.W. 2001, *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*, Oxford: Oxford University Press.

Batterman, R.W. 2011, "Emergence, Singularities, and Symmetry Breaking", *Foundations of Physics*, 41, 6, 1031-1050.

Baxter, R. 1989, *Exactly Solved Models in Statistical Mechanics*, Cambridge, MA: Academic Press.

Butterfield, J. 2011, "Less is Different: Emergence and Reduction Reconciled", *Foundations of Physics*, 41, 6, 1065-1135.

[16] Closely related points are made by Norton 2012. For a discussion see Palacios 2019.

[17] Thanks to Alberto Voltolini for inviting us to be part of this special issue, and to Ashton Green and David Lavis for helpful comments on an earlier draft. Thanks also to the audiences at the Universities of Bordeaux, Cambridge, and Oxford, as well as at the Philosophy of Science Association's Biennial Meeting (PSA18) in Seattle.

Frigg, R. 2010a, "Fiction and Scientific Representation", in Frigg, R. and Hunter, M. (eds.), *Beyond Mimesis and Convention: Representation in Art and Science*, Berlin: Springer, 97-138.

Frigg, R. 2010b, "Models and Fiction", *Synthese*, 172, 2, 251-68.

Frigg, R., Bradley, S., Du, H., and Smith, L.A. 2014, "Laplace's Demon and the Adventures of his Apprentices", *Philosophy of Science*, 81, 1, 31-59.

Frigg, R. and Nguyen, J. 2016, "The Fiction View of Models Reloaded", *The Monist*, 99, 3, 225-42.

Frigg, R. and Nguyen, J. 2017, "Models and Representation", in Magnani, L. and Bertolotti, T. (eds.), *Springer Handbook of Model-based Science*, Berlin: Springer, 49-102.

Frigg, R. and Nguyen, J. 2018, "The Turn of the Valve: Representing with Material Models", *European Journal for Philosophy of Science*, 8, 2, 205-24.

Frigg, R. and Nguyen, J. 2019, "Mirrors Without Warnings", *Synthese*, https://link.springer.com/article/10.1007/s11229-019-02222-9

Frigg, R. and Nguyen, J. 2020, *Modelling Nature: An Opinionated Introduction*, Cham: Springer.

Goodman, N. 1976, *Languages of Art* (2nd ed.), Indianapolis and Cambridge: Hackett Publishing Company.

Laplace, P.-S. 1814, *A Philosophical Essay on Probabilities*, New York: Dover 1995.

Lichtenberg, A.J. and Liebermann, M.A. 1992, *Regular and Chaotic Dynamics*, Berlin: Springer (2nd ed.).

Nguyen, J. 2019, "It's Not a Game: Accurate Representation with Toy Models", *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz010

Norton, J. 2012, "Approximation and Idealization: Why the Difference Matters", *Philosophy of Science*, 79, 2, 207-32.

Palacios, P. 2019, "Phase Transitions: A Challenge for Intertheoretic Reduction?", *Philosophy of Science*, 86, 4, 612-40.

Parker, M.W. 1998, "Did Poincaré Really Discover Chaos?", S*tudies in History and Philosophy of Modern Physics*, 29, 4, 575-88.

Pilyugin, S.Y. 1991, *Shadowing in Dynamical Systems*, Berlin: Springer.

Ruelle, D. 2004, *Thermodynamic Formalism: The Mathematical Structures of Classical Equilibrium Statistical Mechanics*, Cambridge: Cambridge University Press.

Smith, L.A. 2007, *Chaos: A Very Short Introduction*, Oxford: Oxford University Press.

Smith, P. 1998, *Explaining Chaos*, Cambridge: Cambridge University Press.

Spivak, M. 2006, *Calculus*, Cambridge: Cambridge University Press (3rd ed.).

Strevens, M. 2008, *Depth: An Account of Scientific Explanation*, Cambridge, MA: Harvard University Press.

Swoyer, C. 1991, "Structural Representation and Surrogative Reasoning", *Synthese*, 87, 3, 449-508.

Werndl, C. 2009, "What Are the New Implications of Chaos for Unpredictability?", *British Journal for the Philosophy of Science*, 60, 195-220.

Fiction, Models and the Problem of the Gap

Frederick Kroon

The University of Auckland

Abstract

An increasingly popular view holds that scientific modeling involves something akin to the imaginative construction of a fictional story along with its cast of fictional characters, not just the positing of entities (models) that yield a false but useful representation of their targets. The present paper focuses on the following problem for this view of models. If a model is a fiction how can it possibly be said to represent some aspect of the real world? How can the unreal represent the real, and in a way that allows modelers to make predictions about the real, and even explain some of its features? Call this the problem of the gap. The paper begins by motivating the fiction view of models, describing and contrasting the two most popular types of view (both based on Walton's pretense theory of fiction), together with the way they deal with the problem of the gap and some other, related problems. I then sketch a modified version of the fiction view, one that takes on board aspects of both of these approaches by utilizing an important but under-appreciated feature of fiction, and I argue that the view provides natural solutions to this suite of problems.

Keywords: Models, Fiction, Nonexistence, Pretense, Fictional surrogate objects.

1. Introduction

In the early 20th century there was an interesting form of anti-realism to match the anti-realism of logical positivism: Hans Vaihinger and the philosophy of "as if" (Vaihinger 1911). For Vaihinger, the posits of science were to be seen, by and large, as fictions, where fictions were construed as falsehoods: false assumptions that "contradict reality", including false assumption that there exist things of a certain kind (call this the error-theoretic sense of 'fiction'). Such fictions were nonetheless to be retained because they were instrumentally useful. This form of fictionalism is most closely mirrored, perhaps, in the work of van Fraassen, although van Fraassen doesn't regard the theoretical unobservable posits of science as fictions but only as posits whose existence is irrelevant to the development and usefulness—in the sense of empirical adequacy—of science (van Fraassen 1980).

Argumenta 6, 1 (2020): 47-63
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.kro

Not surprisingly, it is hard to find fictionalists in Vaihinger's strong sense. There is widespread agreement among scientists and philosophers of science that we should take current formalisms of this or that theory with a grain of salt, but relatively few would agree that even when we get to the final theory about some domain we should reject its posits as non-existent and the theory itself as only instrumentally useful. In his seminal retrospective assessment of Vaihinger's ideas (Fine 1993), Arthur Fine offers a different view of the applicability of Vaihinger's ideas. He thinks that where they have particular resonance is in the area of *scientific modeling* rather than theoretical science:

> Preeminently, the industry devoted to modeling natural phenomena, in every area of science, involves fictions in Vaihinger's sense. If you want to see what treating something "as if" it were something else amounts to, just look at most of what any scientist does in any hour of any working day (Fine 1993: 16).

Many philosophers of science and scientists have come to accept that models should indeed be classed as fictions of a certain kind, but there is an escalating debate about what this means. As suggested earlier, Vaihinger understood 'fiction' in the error-theoretic sense, and it clear that at least in 'Fictionalism' Fine uses the term in the same way. But an increasingly popular view holds that models are fictions in a somewhat different and arguably richer sense—that modeling involves something akin to the imaginative construction of a fictional story along with its cast of fictional characters, not just the positing of entities (models) that, by dint of the involvement of processes like idealization and abstraction, yield a false but useful representation of their targets. When contemporary theorists talk of *the fiction view of models*, it is this work-of-fiction understanding that they tend to have in mind.

The present paper is mainly about work-of-fiction fictionalism about models but its focus is a problem that also arises for the error-theoretic kind. The problem is this. If a model is a fiction, whether because its posits are akin to fictional characters in a fictional story or because it posits nonexistent items, how can it possibly be said to represent some aspect of the real world? How can the unreal represent the real, and in a way that allows modelers to make predictions about the real, and even explain some of its features? Call this the problem of the gap.

The paper's format is as follows. Section 2 sketches and motivates work-of-fiction fictionalism about models (from here on: *the fiction view of models*), while sections 3 and 4 describe and contrast the two most popular types of view (both based on Walton's pretense theory of fiction), together with the way they deal with the problem of the gap and some other, related problems they face. Section 5 sketches a modified version of the fiction view that takes on board aspects of both of these approaches by utilizing an important but under-appreciated feature of fiction, and describes its own solution to this suite of problems.

2. Models and Fiction

Roughly speaking, in modeling scientists apply prepared descriptions and theoretical laws that they know to be false in order to understand and predict features of target

structures in the world. How to characterize the activity of modeling and its products in more precise terms is of course a difficult and contentious matter, something that is underscored by the sheer variety of "things" (the scare quotes are there to remind us of the problem being tackled) that are called models. On the surface, not only do we have concrete models like wind tunnels or string-and-ball models of the solar system, but there are also models that involve idealization and abstraction with respect to properties of a target (frictionless planes, point masses, etc.), as well as models of a more mathematical kind that are focused on hypothetical structures, such as the Lotka-Volterra model of predation.

What is at any rate clear is that, while models all involve falsehood (broadly speaking, the error-theoretic sense of fiction described earlier), for some theorists there is more to the role of fiction as a way of understanding modeling than falsehood, even imaginatively constructed falsehood. To understand models, so they think, we need to appeal to *literary fiction*. A relatively early proponent of this idea was the philosopher of biology Peter Godfrey-Smith, who noted that theorists tended to talk about their models in concrete terms, and that this was even true in the case of mathematical models (Godfrey-Smith 2007, 2009). As he saw it, these are never *purely* mathematical, since we tend to distinguish models that use the same mathematics (e.g., the harmonic oscillator model whose mathematics can be used to describe both an idealised spring and a chemical bond). A closer look at such models shows that they purport to represent the real causal structure of target phenomena, with the mathematics serving as an essential tool to doing so. A description of the Lotka-Volterra model of predation, for example, doesn't begin with the mathematics, but might be introduced with talk of two imaginary populations that have properties like birth rates, capture rates, and so on, that can then be described mathematically.

For Godfrey-Smith, attending to the way theorists *talk* about models is crucial to understanding models:

> I take at face value the fact that modelers often take themselves to be describing imaginary biological populations, imaginary neural networks, or imaginary economies. An imaginary population is something that, if it was real, would be a concrete flesh-and-blood population, not a mathematical object (Godfrey-Smith 2007: 735).

(Reflecting this stance, Thomson-Jones talks of 'the face-value practice of modeling'; cf. Thomson-Jones 2010.) Godfrey-Smith then asks what the best way is to account for this way of speaking and the uses to which models are put. His answer points to a striking similarity to fiction. Typical models, including mathematical ones, involve imaginary systems that would be concrete if they were real.[1] The Lotka-Volterra

[1] In this paper I am setting aside the tricky status of concrete models (e.g., wind tunnels, string-and-ball model of the solar system, etc.), although there is reason to think that the way these represent their targets involves us in somehow imagining them *as* their target (Toon 2012: Ch.5; cf. also the DEKA model of Frigg and his co-authors; Frigg and Salis 2020: §3).

model of predation, for example, consists of a model system that is an imaginary population of predator animals and prey animals.[2] These have the properties explicitly attributed to them in the act of modeling (growth and death rates, say); others are inferred from what has been stipulated, using mathematics and biological "laws"; others are don't-cares. (See especially Godfrey-Smith 2009.) But this, he notes, is *very* similar to the way fictional worlds are constructed. The world of Sherlock Holmes is partly a matter of stipulation, partly a matter of inference from what is stipulated; much of it, say the number of hairs on Holmes's head, is a don't-care.

Godfrey-Smith thinks that this analogy between model systems and fictional objects is non-accidental and important. Model-based science is in the business of specifying imaginary worlds, although its purpose in doing so is not literary but purely cognitive: it aims at understanding, explaining, and predicting features of the world. To this end, Godfrey-Smith points out a nice feature of the analogy between models and literary fiction: the way modelers often talk of the *similarity* between models and their targets when they apply the model. This kind of talk is tricky if you think of models as mathematical, say, while the idea of relevant similarity between models and targets appears natural and "unintimidating" from the fiction point of view. Thus, just as we can compare two physical systems, we can compare two fictional systems (e.g., Tolkien's "Middle Earth" and the world of Malory's King Arthur in *Morte D'Arthur*). And just as we can compare a model system to its target physical system we can compare a fictional system with a physical system (e.g., events in Orwell's *Animal Farm* are similar to those in Russia in the first part of the 20th century).

Godfrey-Smith doesn't have a great deal more to say about how best to understand the analogy to fiction, but he says enough to indicate a potentially serious problem for the fiction view of models. In the comparison between a model system and its target physical system, what we seem to be doing is comparing properties associated with one system with properties associated with another. But if model systems are imaginary objects such as concrete infinite populations, frictionless planes, and so on, they don't exist and so can scarcely be said to have properties that can be compared to properties of things in the real world. There is, to put it differently, an ontological gap between model systems (when seen as analogous to fictional characters) and their real-world targets, a gap that makes it hard to see how we can learn about the real world from models. Because its aims are different, there is no such problem for literary fiction.

How is the problem best solved? Godfrey-Smith (2009) discusses a number of options without coming down on one side or the other. What is clear, however, is that one's preferred solution will depend on which version of the fiction view of models one chooses. Consider, for example, the view that model systems are abstract entities. In this case there is no ontological gap since model systems so construed *exist*. What this view retains from the face-value picture of modeling is the idea that the model system is an extra entity. What is not preserved, at least not straightforwardly,

[2] Like many authors, I tend to use 'model' and 'model system' interchangeably. Strictly speaking, however, a model system is the (purported) entity that serves to portray or represent some target or other. The model describes how the model system does this, using whatever parts of science and mathematics are needed.

is the idea that model systems and their targets are similar to the extent that the properties of the one broadly correspond to properties of the other. Instead there is a more abstract mapping of some kind, with only the formal structure of the relations between objects on each side being preserved. As Godfrey-Smith points out, on a familiar Platonistic interpretation of structures such a view seems to inherit well-known problems encountered in the literature on the semantic view of theories.

These problems may disappear if the idea of models as abstract structures is tied more closely to the idea that model-building involves pretense. Thus Thomasson (2020) and Thomson-Jones (2020) develop an "artifactualist view" of models, following the contours of Thomasson's view of fictional characters, on which the content of a text that introduces a model should be understood as occurring in pretense, while in producing such descriptions authors create abstract cultural artifacts. On such a view, there is a sense in which it is correct to say 'point masses don't exist' (just as it is correct to say 'Holmes doesn't exist'), even though as artifacts they literally do exist (just as Holmes literally exists). Proponents claim there is a clear reason for positing such entities: we routinely assert truths external to the pretense, such as 'Bohr's quantized shell model of the atom gets more of an atom's structure right than the plum pudding model'. Given, as they argue, that statements of this kind have no plausible paraphrases that eliminate reference to models, that suggests that such models really do exist.

But of course, abstract objects cannot really have such properties as being a biological population or composed of protons and electrons, so the view doesn't conform as straightforwardly as one might like to the "face-value practice of modeling". In addition, the idea inherits other problems to do with the nature of such abstract objects (see, for example Brock 2010). I will here set it aside in order to consider pure pretense theories that try to do without such extra objects.

3. The Fiction View of Models and *De Dicto* Imagining

Consider one such account, due to Roman Frigg (this is the account most fully discussed in Godfrey-Smith 2009). The account adapts Kendall Walton's well-known make-believe account of fiction (Walton 1990) to the case of models. Walton focuses on the way a text can be used as a resource in games of make-believe in which participants pretend, imagine, or make believe that the world is as the text represents it as being. If readers let their imaginings be directed in this way, they are then participating in a game of make-believe that is *authorized* by the work. For Walton, a proposition can be said to be fictional—true in the fiction—just in case participants in such a game of make-believe are supposed to imagine it as true. There are two types of fictional truth: the *primary* fictional truths are evident in the work itself, taking proper account of the linguistic conventions that allow us to understand the work, while the *implied* fictional truths are generated from the primary ones by taking into account what the world is like, or perhaps what the community of origin of the text believed the world was like.[3] It thus turns out that it is true in the Holmes stories that *Sherlock Holmes lived nearer to Paddington Station than to Waterloo Station* (no sentence in the Holmes corpus

[3] For further details, including criticism, see Kroon and Voltolini 2019.

actually says this, so this is an implied truth), while it is false in the stories that Holmes had a wife, for example.

Frigg thinks descriptions of models are structurally much like works of fiction, even if their purpose is very different (they are supposed to provide an understanding of the world, not be a source of entertainment). A model description serves as a prop for a game of make-believe in which participants imagine that the world is as the model description represents it as being. Mimicking the case of fiction, not only are there propositions explicitly authorized by the model description, but in addition there are implied truths:

> What is explicitly stated in a model description (that the model-planets are spherical, for example) are the primary truths of the model, and what follows from them via laws or general principles are the implied truths (Frigg 2010: 260-61).

For another example, take the classical case of Fibonacci's population model, as described in Frigg and Salis 2020. Here the primary truths of the model include such claims as *The rabbits breed in six month intervals*, and the implied truths include claims like *The rabbit population grows monotonically*, which can be derived from the basic assumptions of the model supplemented with some basic facts of arithmetic.

Note the apparent lack of worrisome metaphysical commitments. As it seems, such models are committed only to systems comprising concrete things such as populations of reproducing rabbits, perfectly spherical planets in circular orbits around a massive sun, planes not subject to friction, and so on. The modeler proceeds by imagining a system comprised of such things, and then draws conclusions about their properties using relevant theories and mathematics. There are no further commitments, say to model systems as abstract artifactual entities.

But how do we relate the model system to the target? Unless we are talking of physical models, there is no physical resemblance between model system and target system—the left-hand side of any relation of resemblance is purely in the modeler's head. How can we possibly plug this gap and show how modelers can apply their models to the real world? Here we seem to encounter the problem of the gap in its most pernicious and challenging form.

Frigg's answer is to draw on the way we can use 'transfictional' claims to say what the real world is like. Uttering a sentence like 'Morris Zapp is no more conceited than most academics' allows us to state something about the conceitedness of academics by taking a property that Morris Zapp has in David Lodge's *Changing Places* and affirming that this same property is abundantly instantiated among real world academics. He thinks we can do the same with models. If, for example, I say that some actual rabbit population behaves like a population in the model, certain properties are on the table that can be compared to the properties of a real rabbit population:

> [T]ransfictional statements about models should be read as prefixed with a clause stating what the relevant respects of the comparison are, and this allows us to rephrase comparative sentences as comparisons between properties rather than objects, which makes the original puzzle go away (Frigg 2010: 263).

This is a kind of reductive account: rather than comparing nonexistent things directly with existent things (an impossibility, Frigg thinks, since there are no such things as nonexistents), we compare (existent) properties that imagined things have in a model to (existent) properties of things in the real world. But this account faces a number of objections. Godfrey-Smith points out that many of the properties that are being introduced when dealing with fictional models will be uninstantiated ones and that these may raise special problems of the same kind as those seen with fictional objects.[4] But I suspect that Godfrey-Smith's deeper underlying worry is that models should be seen as *representing* their targets, and a package of allegedly nonexistent entities and properties seems particularly ill-suited to this task. So the wider problem that is not solved by the property-comparison account is this: *how can the model-target gap be closed when there literally are no concrete models to represent the target*?[5] Call this the no-representation problem.

In later work, Frigg is explicit about the need for an account of how models *represent* their target systems. To deal with the problem of representation, Frigg and his co-authors develop the idea, inspired by Goodman and Elgin, of a model M *t-representing* a target system T. Briefly, M *t-represents* T if it denotes T and represents T *as* being a certain kind of thing Z exemplifying Z-style properties, properties that are then related via a key to another set of properties at least some of which M imputes to T (see, e.g., Frigg and Salis 2020: §3). This makes sense if models are existing concrete objects, such as a string-and-ball model of the solar system, but not when they involve such things as immortal rabbits and frictionless planes. Following Salis (see especially Salis 2020), Frigg and Salis respond to this problem by modifying the account of a model. Instead of taken models to include such things as nonexistent immortal rabbits, they associate the model with the content of the fiction together with the text that describes the content, not with the fictional object that is described in the text. So conceived,

> [a] model is a tuple $M = [D, C]$, where D is the description of the model and C is the [full] content of the description … (i.e. the set of propositions that are specified by D together with the principles of generation). … C now takes the place of what one intuitively would call the 'model system' (such as Fibonacci's immortal rabbits). Because model-descriptions and their contents exist, models thus construed are *bona fide* objects (akin to fictional stories) that can enter into relations (Frigg and Salis 2020: 202).

[4] It is not entirely clear what Godfrey-Smith had in mind. Levy takes him to mean that "the model, being merely imaginary, cannot instantiate properties" (Levy 2015: 790). More plausibly, he has in mind such "properties" as *being a non-extended physical object*. Thomasson (2020) responds that the existence of uninstantiated properties can be argued for 'by making pleonastic inferences' such as moving from "The wand is not magical" to "the property of magicalness is not possessed by the wand" to infer that there is a property of magicalness that the wand (indeed everything) lacks. But this move looks question-begging. If someone says, pretending that a gu-gu is a new kind of primitive primate, that you are a "gu-gu", your reasonable protest that you are not a gu-gu (on the grounds that you are a human) doesn't show that there is such a property as being a gu-gu. That there is a genuine property of being a gu-gu requires at the very least some intelligible account of what it is to be a gu-gu.

[5] See also Toon 2012: 58 and Levy 2015: 789-90.

Assuming this notion of a model is even coherent,[6] how do models so construed *denote* and *represent* target systems? Frigg and Salis give only the briefest of indications:

> The model thus defined exists and therefore can stand in the denotation relation with real world systems (Salis 2020: 20),

and

> a look at scientific practice suggests that in many cases the denotation of a model piggybacks on the denotation of denoting linguistic symbols. In our example, Fibonacci's model denotes what it does because we use the denoting expression "the rabbit population in the London Zoo" (Frigg and Salis 2020: 203).

Salis (2020) argues in some detail that we can exploit our knowledge of models so construed to learn about target systems. Perhaps. But the question remains in what sense this is like *denotation*, the relation that plays such a straightforward, pivotal role in their account of t-representation.

4. The Fiction View of Models and *De Re* Imagining

Salis herself sees her account as fixing the failures of a rather different way of understanding the way models denote their targets: what she calls the *direct fiction* view defended in somewhat different ways by Toon (2012) and Levy (2015) and one which is immune to the no-representation objection. One way to motivate their view is to look at the work of a prominent opponent of the fiction view of models like Paul Teller. For Teller a point particle or continuous fluid represents *real* objects such as extended objects and bodies of water:

> A real extended object is fictionally described as having no extension. A real body of water is fictionally described as being a continuous fluid. Such cases constitute fictional descriptions of real objects. So such cases should be thought of, not as object fictions, but as state of affairs fictions, as fictional characterization of states of affairs of real objects (Teller 2009: 244).

(Here 'fictions' and 'fictional' applies to anything that is non-veridical, while *object fictions* are non-veridical, i.e., nonexistent, entities.) Teller, and, following him, Ronald Giere, argue that while scientists might *call* entire models fictional, this may

[6] Note the following prima facie semantic problem: according to the kind of Millian orthodoxy Walton accepts, there may be no propositions expressed by the relevant text. Take model descriptions that contain names of, as it seems, nonexistent entities, for example the silogens referred to in models of fractures in micron-sized pieces of silicon (Giere 2009) or the ether, as modeled by Maxwell's mechanical model. Frigg's original account can maintain that the model descriptions only describe entities from the point of view of the pretense, and so may only express what Kripke (2013) calls *pretend* propositions. The new account cannot afford this luxury, or may need to adopt a more descriptivist, e.g., Ramseyan, view of the content of model descriptions.

be for no other reason than that they contain component object fictions (cf. Giere 2009).

Although Teller does not explicitly put it this way, his account suggests that a model just *is* its target, but a target that is misdescribed through the use of idealization, abstraction, and approximation. That is precisely how Toon (2012) and Levy (2015) see models. But where Teller sees a role for "fictional" (i.e., "non-veridical") characterizations of states of affairs involving real target objects, Toon and Levy see something much closer to the deployment of the literary notion of fiction. Modeling, on their view, aims to provide an imaginative description of real things, with a description of the model prescribing, effectively through the use of Walton's machinery of rules of generation, what we are to imagine about the real system. In the case of the ideal pendulum, for instance, model-users are required to imagine real springs as perfectly elastic and the bob as a point mass, with laws and mathematics needed to supply a stock of inferred truths about the movement of the bob under these conditions. These inferred truths can then be used to make predictions about, and explain features of, real pendula. Levy's presentation of this idea appeals to Walton's notion of prop oriented games of make-believe, for example games in which we imaginatively speak of Italy as a boot or of thunderclouds as faces as a means of thinking and reasoning about them (Walton 1993). Levy's suggestion is that

> we treat models as games of prop oriented make-believe—where the props, as it were, are the real-world target phenomena. To put the idea more plainly: models are special descriptions, which portray a target as simpler (or just different) than it actually is. The goal of this special mode of description is to facilitate reasoning about the target. In this picture, modeling doesn't involve an appeal to an imaginary concrete entity, over and above the target. All we have are targets, imaginatively described (Levy 2015: 791).

This is not the only seemingly significant difference between the way Toon and Levy describe their accounts. For Toon (2012), model descriptions of models with targets (e.g., the simple pendulum) prescribe imaginings about real concrete targets, while model descriptions of models without targets simply prescribe imaginings or, at best, imaginings about purely fictional systems (Toon 2012: §3.3). By contrast, Levy (2015: §4.4) argues that there are no targetless models, appearances notwithstanding. Models that appear targetless may do so because, for example, "the specific range and features of the intended target are not known for sure', or because they are "generalized models [that] work as hubs anchoring specific models" (2015: 796-7). Other apparent models like the Game of Life are genuinely targetless, but Levy thinks that these are little more than "bits of mathematics" rather than models in a full-blooded sense (2015: 797).

Following Salis, I want to highlight two problems for Levy's account in particular: the *no-target* and *indirectness* problems. First, the account seems to have nothing to say about certain familiar kinds of models that, unlike the targetless models Levy mentions, seem *aimed* at a target. Examples include Maxwell's mechanical model of the ether and models of synthetic molecules that will never be created in a lab, perhaps because the models reveal that any such entities would cause great harm. Secondly, the relation between model and target seems typically far less direct than Levy and Toon make it out to be. As Salis (2020) puts it:

> Stating that model descriptions are about real objects does not dispense with fictional entities (and the controversies they generate) because model descriptions always involve apparent reference to some fictional objects. … [Take the simple pendulum]. The model description of the simple pendulum is not about any particular pendulum. It does not start with 'Imagine of this particular pendulum in front of you that it is a point mass suspended by a massless, unstretchable string'. Rather, it apparently refers to an imaginary system consisting of a point mass and a massless string and hence prescribes imagining about a fictional system (Salis 2020: 12).

I think we should, with some qualifications described below, accept Toon's answer to the first objection, which is that some imagining is directed at fictional target systems (and hence that this kind of imagining is effectively embedded in imagining, rather than knowing, that there is a real target system). The second objection strikes me as in some ways more important, although it is not hard to see the beginnings of an answer. As Salis points out, modelers typically do not begin with an instruction to imagine, of some particular thing or of any of a class of particular things, that it has certain properties specified in the model. The relation between model and target is often, even typically, more indirect. But *contra* Salis, we should at the same time note that there are also cases where the relation is described as being much more direct, in much the way emphasized by Toon and Levy. Science texts, for example, often talk about the different idealized ways in which atoms are *described* by Rutherford and Bohr, say, not just about the different ways in which atoms are *modeled*, or about the way a model of the solar system might *describe* the Earth as being a point-like object that doesn't rotate.[7] A rather nice example is given by Levy, who quotes the two ways in which Turing characterizes his mathematical model of the growing embryo: in one version "the cells are idealized into geometrical points" while in the other "the matter of the organism is imagined as continuously distributed" (Levy 2015: 782).

Not only do both the direct and indirect perspectives occur in the literature. It is also clear that the difference in perspective would not strike modelers themselves as particularly significant. They would be unlikely, for example, to reject a presentation of the Fibonacci population model that went as follows:[8] "My kids have two rabbits, one male and one female. Their names or identities don't matter. What matters is that they are ready to mate. Let's describe how their number will grow by making some simplifying assumptions. Assume that rabbits always mate six months after birth, that the female of each pair gives birth to exactly one male-female pair another six months after mating, that they never die, etc. [Now comes the Fibonacci calculation of rabbit pair numbers at all future moments.] That is the Fibonacci population model!" None of us, modelers included, would be nonplussed by such a presentation of the model.

[7] See Matthews *et al.* 2005 for the various ways in which the simple model of the pendulum has been described.

[8] If the presentation was intended for a journal or a text meant for researchers, they would, perhaps, be chided by a referee; but there the reason has to do with the culture of scientific academic writing (which is also the reason why there might not be criticism if the presenter was known to be famous).

What is more important to scientists who use the model is that it can be applied to target populations other than the one featured in the introduction.

5. Learning from Fiction

To see how the fiction view of models is able to cast light on this seemingly odd duality of perspectives, it is time to apply some lessons from the case of literary fiction. Consider real objects that feature in fictional works, for example Napoleon in *War and Peace*. While real individuals can appear in fictional works, in the works they are often very different from the way they actually are. These differences, whether large or small, have given rise to the view that real individuals as they appear in works of fiction should be regarded as distinct fictional characters: fictional *surrogate* objects for short. Meinongians in particular should see the attraction of the view. Just as they think it is true that Andrei Bolkonsky (a purely fictional character in Tolstoy's *War and Peace*) was wounded at the Battle of Austerlitz, they should also think it true that Napoleon rescued Andrei at this battle: for Meinongians, fictional truth suffices for truth. Since in reality Napoleon did not rescue Andrei, it must be a surrogate object that did so—the Napoleon of *War and Peace*. Artifactualists, who think that fictional objects are abstract objects created by authors, have provided other reasons for thinking that there must be a surrogate fictional Napoleon.[9]

But even pretense theorists should admit a sense in which real objects have fictional surrogates. After all, in our pretenses the Napoleon of the story must be distinguished from the real Napoleon. We are pretend-referring to someone who rescued Andrei at Austerlitz. Napoleon wasn't like that! Similarly, when I read *War and Peace* and admire Napoleon for his kindness in rescuing Andrei it is the Napoleon of *War and Peace* I admire; I might detest the real Napoleon on the grounds that he, on the other hand, would never have done such a thing. Note again how we use certain familiar qualifiers to draw the contrast; we talk of *the Napoleon of* War and Peace or of *Napoleon as he was in* War and Peace, and contrast that person with other versions of Napoleon: the *real* Napoleon, say, or *Napoleon as he was in Vincent Benét's* 'The Curfew Tolls'. Our ability to make sense of these distinctions doesn't depend on whether one is a pretense theorist or a fictional realist of some kind (Kroon 1994).[10]

But how do these fictional surrogates of Napoleon relate to (the real) Napoleon? Fictional realist proponents of fictional surrogacy tend to agree that the surrogates in some sense represent their real-world counterparts, even if there is disagreement about the nature of this relation.[11] If one is a pretense theorist, however, there is much simpler account one can give of such relationships: there *is* in fact no relationship between a Napoleon surrogate like the Napoleon of *War and Peace* and Napoleon himself since

[9] See Motoarca 2014 and Voltolini 2013, 2020.

[10] In fact, artifactualist believers in fictional surrogate objects could probably adapt the kind of surrogacy-friendly view of modeling defended in this paper to yield an alternative to the artifactualist account of modeling found in Thomasson 2020 and Thompson 2020. (They may, of course, have independent reasons to resist such an extension of the notion of surrogacy.)

[11] It needn't be representational in any strong sense; Voltolini, for example, considers it a many-many relation of similarity (Voltolini 2020: 815).

the Napoleon of *War and Peace*, an occasionally kind person who rescued Andrei Bolkonsky, doesn't exist. Instead, we should say the following. In writing *War and Peace* Tolstoy wanted his readers to imagine, of Napoleon, that Napoleon did certain things that he did not in fact do. Much else that he wanted his readers to imagine about Napoleon is based on facts about the latter's actual life and deeds. But when we, in response, engage imaginatively with *War and Peace* we do so from the inside: in the scope of our pretending that the world is as reported in *War and Peace* we represent to ourselves someone who is rather different from the real Napoleon (of course in so doing we must import facts about the real Napoleon and his exploits in so far as these don't conflict with the prescription to imagine what the novel tells us—there could be no fiction without such an anchoring in reality). So, while the story is partly *about* Napoleon and his exploits (here 'Napoleon' refers in the standard way to Napoleon), when we engage with the story we are not referring to him. This is because we are not referring at all: we are only *pretend*-referring, referring from inside the scope of the pretense (or, as some prefer, referring at a pretend or fictional context instead of at a real context).[12] In the scope of the pretense, we are referring to someone who has the properties he is ascribed in the novel—and that person, aptly characterized as *the Napoleon of* War and Peace, doesn't exist.

(Quick proof that we are not really referring. If we were, our utterances would be up for evaluation for truth or falsity in the usual way, and so it would be entirely appropriate for listeners to accuse us, over and over again, of uttering falsehoods. But this would not be appropriate—our utterances are not truth-normed in this way. Despite this, we can learn a lot about the world of Napoleon by reading *War and Peace*. Doing so requires some sensitivity, but, roughly speaking, if Napoleon is described in *War and Peace* as having done X and there are no artistic ends that would be served by Tolstoy asking us to imagine this even though he believed that Napoleon did not in fact do X, then, given that Tolstoy is reliable where Napoleon is concerned, it is probably safe to infer that Napoleon *did* do X. In short, it is appropriate to *export* fictional truths under certain circumstances, that is, to interpret them as genuine, non-pretend truths, even though it is admittedly difficult to frame rules about how to do this.)[13]

Return now to the case of models, and consider again the kind of Waltonian pretense accounts discussed in previous sections. Levy and Toon think that in modeling we are imaginatively re-describing real-world systems (but sometimes fictional systems, if Toon is right). By contrast, Frigg and co-authors like Salis argue that model systems are the product of *de dicto* imagining the existence of model systems like point masses, infinite populations of immortal rabbits, and the like. The view I prefer borrows from the lessons we have just learned about Napoleon and his surrogates: to the extent that in modeling we are indeed imaginatively re-describing real-world systems, that idea does not in any way get rid of the idea of nonexistent model systems. It just requires us to rethink their role and nature.

Here, in brief, is the idea. Let X be a real-world system or object—a real pair of rabbits, say, or the solar system or a pendulum, or…—and suppose we imaginatively

[12] For discussion of this use of the notion of context, see Kroon and Voltolini 2019: §2.1.
[13] For useful discussion, see Friend 2014.

represent it as satisfying various assumptions *F*, both idealizing and auxiliary (such as the existential assumption that there exist silogen atoms; Giere 2009), while also abstracting away features of no concern. (For the sake of brevity, I'll simply talk of *assumptions of idealization and abstraction*.) Then the model system is object X conceived of as conforming to these assumptions of idealization and abstraction *F*, say a physical pendulum idealized as a point mass bob suspended from a string of zero or negligible mass, with the only forces acting on the bob being the force of gravity and tension from the string. (Given the principles of generation for the pretense, it is also part of the pretense that this object has properties *P* whose possession by X follows from X satisfying *F*, given mathematics and relevant scientific laws.) That is the surrogate object we encounter from inside our pretense, just as the Napoleon of *War and Peace* is the surrogate figure we encounter as we engage with *War and Peace*. And this object does not in fact exist just as the Napoleon of *War and Peace* doesn't exist.

How do we apply and learn from models once we understand them this way? Here is one suggestion: in the same way as we learn about Napoleon by reading and engaging with *War and Peace*. From in the scope of the pretense we can engage with the model system and work out how it behaves under various conditions. At this point there is only pretend-reference and a pretend-ascription of what properties the system would have under these conditions, or, if you prefer, reference and property ascription at a pretend-context. But as in the case of *War and Peace* there is also a non-pretend way of reading the sentences that record these findings: reference and property ascription at a real-world context. We have been assuming that the pretense was based on (*de re*) imagining that a certain target system conformed to a degree of idealization and abstraction. If so, when we refer *apart* from the pretense (that is, at a real context), we are referring to the target system simpliciter. What it is reasonable to export from the pretend-truths of the model and how to qualify or amend these in light of facts about the target system depends on features and the intended scope of the model. Perhaps the process can be understood using the notion of partial truth that Levy favors (Levy 2015: §4.2).[14] We needn't take a stand on this. All I here want to emphasize is that this is a version of a pretense account that takes due note of the role of surrogate objects, and thereby suggests a prima facie attractive way of closing the gap between model systems and the real-world systems they represent.

Note that this way of describing how we learn from models is misleading in so far as it treats the pretense as focused on the content of models—as content oriented rather than prop oriented make-believe, to use Walton's terms (Walton 1993). What is really going on, of course, is a little different, since models are unlike stories in their orientation: stories are meant to be engaged with from the inside (to treat them as learning tools is not to do them justice), while the purpose of models is to facilitate reasoning about external target-systems. So, it would be more accurate to say that we begin with target systems and an interest in explaining and predicting features of such

[14] Levy describes his use of the notion of partial truth as follows: "The idea, to put it tersely, is that while model descriptions are typically idealized, hence not true of their targets simpliciter, they are nevertheless partly true, at least when successful" (Levy 2015: 792). Although Levy doesn't say this, note that this assumes that applying models involves what I earlier called exportation: to apply a model we must export what is in the first instance merely imagined.

systems and that pretend-reference and pretend-ascription of properties only come in once we try to meet this interest by talking about these systems through assumptions of idealization and abstraction. In short, the pretense involved in modeling should be seen as externally oriented, not content oriented.[15]

In fact, it is tempting to say that the purpose of model systems is to *represent* external target systems. But that way of putting the point hides another problem (the *no-representation* problem) which affects both Frigg's early account of models and the present account—strictly speaking, there are no model systems that can do the representing. Before turning to this problem, let me first deal with a problem that may initially loom even larger: the *indirectness* problem. As presented, the account only works (if at all!) for models based directly on target systems; but as Frigg and Salis emphasize, models are often, perhaps typically, not based on real-world targets this way (set aside models of specific objects like the solar system). But here the very nature of models shows us the way out. I have already suggested that there is something almost incidental about the fact that modelers don't appeal *de re* to real-world targets when they describe their models. They could have, without this affecting their models. That suggests the following solution to the indirectness problem. When devising models, modelers imagine that some *arbitrary* system of the relevant kind is subject to certain assumptions of idealization and abstraction, not (*pace* Salis) some *specific* system. In effect, the modeler in the pendulum case means something like "Consider a pendulum—*any* pendulum—made subject to idealization and abstraction as follows". At the point where the modeler comes to apply her model to a real-world system X, she effectively instantiates to the specific system X. At this point, and without any loss of generality, she simply takes X to be the system that is imaginatively reconfigured to conform to certain assumptions of idealization and abstraction. For purposes of application, the simple pendulum model can thus be thought of as *this* particular physical pendulum imaginatively reconfigured as being *F* as easily as *that* particular physical pendulum. So the fact that models are (typically) introduced without reference to a specific target system is not of any significance on this picture, and is fully consistent with the idea that pretense in the case of modeling has an external orientation.

What about the no-representation problem? Here the pretense account on offer has limited leeway. Absent existent model systems, it is only in the scope of a pretense that model systems (now construed as pretended surrogate objects) denote target systems. Frigg and Salis point out that some "may think that [this] is too feeble a notion to account for how science represents its objects and nothing short of 'real' denotation between models and their targets is good enough" (Frigg and Salis 2015: 201-2). They say they want to keep an open mind about this issue, but as we saw earlier they also propose a new version of the fiction view that, in their view, makes room for real denotation. My suspicion is that our intuitions are simply not clear enough to determine if more than pretense is needed. What certainly is clear is that in saying that models denote real-world targets, we are asserting something about the world and the

[15] I don't use the term 'prop oriented' since, as the indirectness problem reminds us, there need not be any particular prop that is invoked in the imaginative construction of the model.

way modeling activity is directed at the world—we are not merely *pretending* that models and modeling are directed at the world. But there is a familiar way in which this worldly orientation of our utterance can be captured by the kind of pretense view on offer: we often make substantive claims about the world by talking *through* our pretenses, and statements to the effect that a certain model represents a particular target system may be no exception.[16] So I remain unconvinced that more than pretense is needed to make sense of the way we describe models as denoting their targets. If more than pretense is needed, and if, in particular, we want a robust, "non-feeble" account of the relationship between model and target system, then—and only then—do we have to change the way we think of models.

That leaves us with the final problem: the no-target problem, which affects the present view no less than Levy's. In the case of some attempts at modeling there turns out to be no target system, and hence no system that is imagined as subject to certain idealizing assumptions. Such is the fate of Maxwell's mechanical model of the ether and models of synthetic molecules that will never be created in a lab. The solution to this quandary is essentially that proposed by Toon (2012): the target systems are themselves (at best) fictional systems.[17] But more should be said. To the extent that there is no ether there can (of course!) be no model of the ether either. But in Maxwell's case there was the *belief* that the ether exists, and that he had a model of it. To be able to talk of Maxwell's work, commentators simulate this belief: they do *as if*, or pretend that, the ether exists. They do this effortlessly, in much the same way as, picking up a book like *The Hobbit*, we effortlessly engage in the pretense that there are such things as hobbits. And if you are a chemist in the business of hypothetically working out what a certain molecule would be like when no such molecule yet exists, you do *as if* there is such a molecule and (in the scope of your pretense) you describe your attempts to model it. Commentators, including philosophers writing about modeling, will effortlessly join in the pretense, much as friends whom you regale with the exploits of Bilbo in *The Hobbit* join you in your pretense.

[16] We can, for example, extend the original pretense to allow talk of a relationships of denotation by letting the extended pretense include something like the following rule of generation: it is correct to pretend that model system X denotes target system Y just if those who instituted the original pretense mandated that we are to imagine X as a system that conforms to certain assumptions of idealization and abstraction, and that we do so by pretending that Y is subject to these assumptions of idealization and abstraction. (As a consequence, the relevant model description denotes Y at a real context while denoting X at a pretend-context in which it is pretended that Y conforms to certain assumptions of idealization and abstraction.) Assuming that in uttering a statement of the form 'Model system X denotes its target system Y' our interest is in communicating the circumstance that makes this statement true in the extended pretense, the *real* content of our utterance is that this circumstance obtains. Hence in uttering a statement 'Model system X denotes its target system Y' we are making a substantive claim about the world. (See Everett 2013: Ch. 3 on 'Talking Through the Pretense', esp. §3.3.2 which is concerned with modeling fictional characters; see also Armour-Garb and Woodbridge 2015 for a general discussion of how an involvement in pretense allow us to make claims about the world.)

[17] For a very different non-pretense way of dealing with models, including targetless models, see Suárez's influential account of an inferential conception of representation in Suárez (2004).

So are there targetless models? Not on the view proposed. In any such case of an apparent targetless model there is a fiction in which it really is a model with a target—a fiction that we, as commentators, effortlessly engage with. Since on the fiction view of models the usual kind of model systems are to be understood as fictional systems, we should consider these wannabe models *fictional* fictional systems, much as Gonzago, a character within a play enacted in *Hamlet*, should be considered a *fictional* fictional character rather than a fictional character (Kripke 2013: 72-73).[18]

References

Armour-Garb, B. and Woodbridge, J.A. 2015, *Pretense and Pathology: Philosophical Fictionalism and Its Applications*, Cambridge: Cambridge University Press.

Brock, S. 2010, "The Creationist Fiction: The Case against Creationism about Fictional Characters", *The Philosophical Review*, 119, 3, 337-64.

Everett, A. 2013, *The Nonexistent*, Oxford: Oxford University Press.

Fine, A. 1993, "Fictionalism", *Midwest Studies in Philosophy*, XVIII, 1-18.

Friend, S. 2014, "Believing in Stories", in Currie, G., Kieran, M., Meskin, A. and Robson, J. (eds.), *Aesthetics and the Sciences of Mind*, Oxford: Oxford University Press, 228-48.

Frigg, R. 2010, "Models and Fiction", *Synthese*, 172, 2, 251-68.

Frigg, R. and Salis, F. 2020, "Of Rabbits and Men: Fiction and Scientific Modeling", in Armour-Garb, B. and Kroon, F. (eds.), *Fictionalism in Philosophy,* Oxford: Oxford University Press, 187-206.

Giere, R. 2009, "Why Scientific Models Should not be Regarded as Works of Fiction", in Suárez 2009, 248-58.

Godfrey-Smith, P. 2007, "The Strategy of Model-Based Science", *Biology & Philosophy*, 21, 5, 725-40.

Godfrey-Smith, P. 2009, "Models and Fictions in Science", *Philosophical Studies*, 143, 101-16.

Kripke, S. 2013, *Reference and Existence*, Oxford: Oxford University Press.

Kroon, F. 1994, "A Problem about Make-Believe", *Philosophical Studies*, 75, 201-29.

Kroon, F. and Voltolini, A. 2019, "Fiction", *The Stanford Encyclopedia of Philosophy*, Zalta, E.N. (ed.), https://plato.stanford.edu/archives/win2019/entries/fiction/.

Levy, A. 2012, "Models, Fictions, and Realism: Two Packages", *Philosophy of Science*, 79, 5, 738-48.

Levy, A. 2015, "Modeling without Models", *Philosophical Studies*, 172: 781-98.

Levy, A. and Godfrey-Smith, P. (eds.), 2020, *The Scientific Imagination: Philosophical and Psychological Perspectives*, Oxford: Oxford University Press.

[18] Many thanks to Carola Barbero, Matteo Plebani, and Alberto Voltolini for organizing the FINO GC / SIFA MidTerm Conference (June 2019) at the University of Turin where an early version of this paper was presented, and to members of the audience for their stimulating questions and criticisms. Special thanks to Alberto Voltolini for suggesting a number of useful revisions.

Matthews, M., Gauld, C. and Stinner, A. (eds.) 2005, *The Pendulum: Scientific, Historical, Philosophical and Educational Perspectives,* Dordrecht: Springer.

Motoarca, R. 2014, "Fictional Surrogates", *Philosophia*, 42, 1033-1053.

Salis, F. 2020, "The New Fiction View of Models", *British Journal for the Philosophy of Science*, 71, 1-28.

Suárez, M. 2004, "An Inferential Conception of Scientific Representation", *Philosophy of Science*, 71, 5, 767-79.

Suárez, M. (ed.) 2009, *Fictions in Science. Philosophical Essays on Modeling and Idealization*, London & New York: Routledge.

Teller, P. 2009, "Fictions, Fictionalization, and Truth", in Suárez 2009, 235-47.

Thomasson, A.L. 1999, *Fiction and Metaphysics*, New York: Cambridge University Press.

Thomasson, A.L. 2020, "If Models Were Fictions, Then What Would They Be?", in Levy & Godfrey-Smith 2020, 51-74.

Thomson-Jones, M. 2010, "Missing Systems and the Face-Value Practice", *Synthese*, 172, 2, 283-99.

Thomson-Jones, M. 2020, "Realism about Missing Systems", in Levy & Godfrey-Smith 2020, 75-101.

Toon, A. 2012, *Models as Make-Believe: Imagination, Fiction, and Scientific Representation*, London: Palgrave Macmillan.

Vaihinger, H. 1911, *Die Philosophie des Als Ob*, Leipzig: Meiner. Translated by C.K. Ogden as *The Philosophy of 'As If'*, London, 1924: Routledge and Kegan Paul.

van Fraassen, B.C. 1980, *The Scientific Image*, Oxford: Oxford University Press.

Voltolini, A. 2013, "Probably the Charterhouse of Parma Does Not Exist, Possibly Not Even That Parma", *Humana-Mente Journal of Philosophical Studies*, 25, 235-61.

Voltolini, A. 2020, "Real Individuals in Fictions, Fictional Surrogates in Stories", *Philosophia*, 48, 2, 803-20.

Walton, K.L. 1990, *Mimesis as Make-Believe*, Cambridge, MA: Harvard University Press.

Walton, K.L. 1993, "Metaphor and Prop Oriented Make-Believe", *European Journal of Philosophy*, 1, 1, 39-57.

Learning through the Scientific Imagination

Fiora Salis

University of York

Abstract

Theoretical models are widely held as sources of knowledge of reality. Imagination is vital to their development and to the generation of plausible hypotheses about reality. But how can imagination, which is typically held to be completely free, effectively instruct us about reality? In this paper I argue that the key to answering this question is in constrained uses of imagination. More specifically, I identify make-believe as the right notion of imagination at work in modelling. I propose the first overarching taxonomy of types of constraints on scientific imagination that enables knowledge of reality. And I identify two main kinds of knowledge enabled by models, knowledge of the imaginary scenario specified by models and knowledge of reality.

Keywords: Scientific models, Imagination, Make-believe, Counterfactual imagination, Knowledge.

1. Introduction

How do we learn about reality through scientific models? Answering this question requires distinguishing between two main kinds of models, material and theoretical. Material models are physical objects that serve as representations of physical systems. Theoretical models are mathematical models that do not exist as physical objects and for this reason are sometimes called 'fictional'. Morgan (1999) originally argued that learning with models involves two steps, model building and model manipulation. Frigg and Hartmann (2018) notice that material models are used in experimental contexts and do not raise any special problems beyond general questions about learning through experimentation. Fictional models, however, do raise serious concerns. What are the constraints on model building and model manipulation in fictional models? Answering this question requires that we recognise the crucial role of imagination in fictional models. Learning through fictional models requires imagination and the value of scientific imagination depends on its ability to produce valuable (that is, potentially true) hypotheses.

Argumenta 6, 1 (2020): 65-80
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.sal

Consider a simple but paradigmatic example, Maxwell's thermodynamic ideal gas model, which represents a gas as a large number of particles bouncing against each other and against the walls of a closed container. The model is usually identified with the following equation:

$pV = nRT$.

The equation per se is not a model of anything unless it is used under an interpretation. In this case, p is pressure, V is volume, n is the number of moles, R is the gas constant, and T is temperature. To facilitate mathematical treatment, the model makes certain simplifying assumptions. It assumes that the gas is composed of molecules construed as point particles having no volume in and of themselves, exerting no intermolecular forces, and bouncing against each other and against the walls of the container in elastic collisions that do not involve any conversion of kinetic energy into other forms of energy. Of course, there are no gases that are composed of such idealised molecules. Real gases are composed of molecules that have some finite volume, that exert intermolecular forces and collide in non-elastic ways. The ideal gas model describes an imaginary gas composed of imaginary particles interacting under imaginary conditions. Nevertheless, the model provides a useful approximation of the behaviour of many real gases under temperatures that are near room temperature and pressures that are near atmospheric pressure.

Philosophers usually recognise that imagination has an important role in modelling. Cartwright understands modelling as offering "descriptions of imaginary situations or systems" (2010: 22). Godfrey-Smith suggests we "take at face value the fact that modelers often take themselves to be describing imaginary biological populations, imaginary neural networks, or imaginary economies" (2006: 735) and sees modelling as involving an "act of imagination" (2009: 47). Harré sees models as things that are "imagined" (1988: 121). Sugden regards models as "imaginary" worlds (2009: 5). Weisberg discussing the Lotka-Volterra model of predator-prey interaction reports that Volterra "imagined a simple biological system" (2007: 208) and further recognises that "[m]odelers often speak about their work as if they were imagining systems" (2013: 48). Frigg (2010), Levy (2015), Salis (2019; 2020a), Salis and Frigg (2020), and Toon (2012) present analyses that place acts of imagination at the heart of modelling.

When it comes to explaining how models enable knowledge of reality, however, standard explanations dismiss uses of imagination in modelling as ill-suited to scientific reasoning. Notwithstanding their differences, these accounts agree in connecting learning with the representational function of models (Giere 1988; Knuuttila and Voutilainen 2003; Mäki 1992, 2005; Suárez 2004; Swoyer 1991; Weisberg 2007, 2013). On these views, a model description (the mathematical equation and linguistic assumptions of the ideal gas model) specify a simplified surrogate (the idealised gas) of a real system (some real gas). The surrogate is called model system and the real system is called target system. A model system is interpreted as a symbol representing a target in ways that enable the generation of plausible hypotheses based on a relation of similarity with the target, where similarity is usually understood as the sharing of certain properties in some respects and to some degree.

While these standard proposals advance many important ideas, they do not satisfy one key theoretical requirement that Frigg (2010) calls *naturalism*. According to naturalism, any account of how scientists learn with models should be able

to explain scientific practice, namely it should explain how scientists construct models and how they reason with them. This is what Thomson-Jones (2010) calls the face-value practice of modelling. Scientists present model-descriptions that specify model-systems as objects of study. Model descriptions involve the attribution of properties that only concrete objects can have, yet there are no objects instantiating these properties. Scientists think and talk as if there were such concrete systems having such and such properties, yet they are aware that there are none. They merely imagine that there are systems having such and such properties.

So, modelling crucially relies on imagination. Yet, standard accounts do not offer any explanation of how knowledge of reality is obtained through imagination. The result is a poor understanding of the epistemic role of imagination in modelling. Pre-theoretically, imagination is often thought of as completely free and unconstrained. In this vein, many think of imagination as a means to escape reality, as when we engage in daydreams and fantasies that provide diversion and create new things that depart from reality. Pessimists about our ability to gain knowledge through imagination emphasise the freedom of imagination (Descartes 1985; Norton 1991; Spaulding 2016). There is, however, another pre-theoretical notion of imagination as a means to learn about reality, as when we engage in problem solving, mindreading, thought experimenting, counterfactual reasoning and, of course, scientific modelling. The key to this second notion is the idea that imagination can be constrained in ways that effectively enable knowledge of reality.

In this paper I shed new light on this issue by developing a new notion of constrained imagination that is motivated by the face-value practice of scientists and by the recognition of the importance of scientific cognition involving imagination. In Section 2, I start by identifying two main varieties of imagination that are currently deemed crucial to scientific models, counterfactual imagination (Godfrey-Smith 2020) and make-believe (Salis and Frigg 2020). In Section 3, I argue that standard analyses of counterfactual imagination in modelling raise important issues that deserve further theoretical development. In section 4, I identify make-believe as a more suitable option and explain its role in model building and model development. In Section 5, I put forward a taxonomy of types of constraints operating on imagination in modelling based on contemporary literature in cognitive science and philosophy. Finally, in Section 6, I draw some conclusions.

2. Imagination

What sort of imagination is involved in models? Imagination is ordinarily construed as mental imagery, which is an ability to form a sensory-like representation of something (real or non-existent) in any sensory modality (imagining seeing, imagining hearing, imagining smelling, imagining touching, imagining tasting). The most common variety is visual imagery, which is often referred to as seeing in the mind's eye, imagining seeing or visualising. Scientists often appeal to this pre-theoretical notion in introspective reports and descriptions of activities that were key to the generation of new ideas. In the 19th century, Michael Faraday contributed to the foundations of classical electromagnetic theory by imagining invisible lines of force as narrow tubes curving through space (Tyndall 1868).

Starting from this picture, James Clerk Maxwell studied lines of force by producing a series of mechanical models of the ether, which led to his famous set of equations (Maxwell 1965). In the same century, August Kekulé discovered the structure of the molecule of benzene after a daydream in which he saw a snake biting its own tail (Findley 1948). These and similar cases led to the widespread recognition of the key role of imagery in scientific discovery, conceptual change and innovation (Magnani 2009; Nersessian 2008, 2009).

Whether imagery has a key epistemic role in modelling, however, is currently disputed. In particular, Salis and Frigg (2020) emphasise that mental images are neither necessary nor sufficient to scientific modelling. For example, the ideal gas model requires imagining that the gas be composed of point particles having no volume in and of themselves and bouncing against each other in elastic collisions. These imaginings involve certain theoretical concepts (point particle, volume, elastic collision) and relations within the imaginary scenario described by the model. Whether they are accompanied by mental images or not seems to be irrelevant to the epistemic function of models.

In fact, another notion of imagination has gained traction in the contemporary philosophical literature on scientific modelling, that of propositional imagination. This is an ability to entertain a proposition without any commitment to its truth, with or without forming a mental image. This somewhat minimal notion of imagination, which is akin to a notion of acceptance, has been specified in two main varieties that are deemed crucial to the modelling practice, counterfactual imagination and make-believe.[1]

3. Counterfactual Imagination

Godfrey-Smith (2020) recognises the key role of conditional thinking and, in particular, the counterfactual imagination in modelling. Conditionals are statements of the form *if A then C*. A counterfactual conditional is a subjunctive conditional where the antecedent is known or assumed to be false, or $A \,\square\!\!\rightarrow C$. For example, one might imagine that if Hillary Clinton had won the elections in 2016 (counterfactual antecedent), then the US would have led a coordinated effort to combat COVID-19 with allies in Europe, Asia and the Americas. Godfrey-Smith notices that counterfactual conditionals in modelling often involve generalisations such as *if there were a system like this, it would do that*, or $M \,\square\!\!\rightarrow C$, where the antecedent M stands for the model assumptions and the consequent C stands for the consequence that follows from M. For example, if there were a gas having these and these features, then it would behave like this (ideal gas model); or, if there were two celestial bodies having features F, then they would do that (sun-earth model). The antecedents in these conditionals are assumed to be false. Scientists know that they are never realised in the actual world.

Implicit criteria for how imagination is constrained in counterfactual reasoning have been offered by the influential analyses of counterfactuals put forward by Stalnaker (1968) and Lewis (1973). The leading idea of these analyses is that a counterfactual claim is true in the closest possible world where the antecedent is

[1] Another important notion of propositional imagination is that of supposition (Arcangeli 2018; Nichols 2006), which plays an important role in Sorensen's (1992) account of scientific thought experiments. There are, however, no accounts of modelling in terms of supposition.

true and the consequent is also true. By 'closest possible world' we mean closest to the actual world—or reality. Hence, closeness—reality orientation or similarity—is the key constraint on imagination that emerges from these analyses. When engaging in counterfactual reasoning, we select an antecedent A that is contrary to some relevant fact in the actual world and then draw a consequence C in the A-worlds that are closest to the actual world. However, the antecedent A selects a set of possible worlds (the A-worlds), not all of which are relevant for the assessment of the counterfactual conditional. The A-world that is closest to reality is the one that determines its truth. When one ponders what would have happened if Hillary Clinton had won the elections in 2016, one considers how things would have been in a world that is just like the real world apart from the election of Hillary Clinton in 2016.

There is one general challenge for these analyses, and three specific issues concerning their application to modelling. The general challenge concerns the details of the notion of closeness, which remains insufficiently characterised. Stalnaker appeals to the "intuitive idea that the nearest, or least different, world in which antecedent is true is the one that should be selected" (1981: 88), but does not provide any explanation of what 'least different' means. And Lewis (1973) assumes a primitive notion of similarity of worlds, which "leaves the notion of similarity unconstrained and mysterious" (Arlo-Costa 2019: Sect. 6.1).

The three more specific challenges for an application of these analyses to models are posed by completeness, epistemic access, and intersubjective access.

Salis and Frigg (2020: 43) notice that possible worlds are complete, yet scientific models cannot be said to be complete in the same way. What completeness means is open to interpretation. However, they notice that there is an intuitive link between completeness and the principle of Excluded Middle (EM). According to EM, for any proposition p it is the case that *either p* or *not-p* holds. Models are not complete in this sense because there are many propositions that are neither true nor false in models. For example, the proposition that Mont Blanc is the tallest mountain in Europe is neither true nor false in the ideal gas model. However, if possible worlds are complete, the closest possible world in which M is true is one in which this claim is true even though it describes matters of fact that have nothing to do with the model. On this analysis, the counterfactual "if a gas were composed of point particles exerting no intermolecular forces, then Mont Blanc would be the tallest mountain in Europe" would come out true. The truth value of this counterfactual, however, should be indeterminate. The world of the model does not satisfy EM and is not complete in the same way in which possible worlds are supposed to be complete.

Stalnaker's semantic analysis of counterfactuals does not allow for this kind of indeterminacy because it accepts the principle of Conditional Excluded Middle (CEM). According to this principle, either $M \,\square\!\!\rightarrow\, C$ is true or $M \,\square\!\!\rightarrow\, \neg C$ is true. Stalnaker's semantics uses a selection function that picks a unique closest possible world where C is either true or false and, hence, either $M \,\square\!\!\rightarrow\, C$ or $M \,\square\!\!\rightarrow\, \neg C$ holds. In contrast, Lewis's semantics deploys a relation of comparative similarity that defines a weak total ordering of all possible worlds with respect to each possible world (what he calls 'a system of spheres'). On this proposal, when $M \,\square\!\!\rightarrow\, C$ is true, C is true in all the closest M-worlds. However, when C is true only in some of the M-worlds but not in others, CEM fails because neither $M \,\square\!\!\rightarrow\, C$ nor $M \,\square\!\!\rightarrow\, \neg C$ holds. This allows for the possibility of indeterminacy and is therefore an im-

provement with respect to Stalnaker's original proposal. Lewis's analysis, however, poses a different problem. The specific indeterminacy of models seems to be difficult to capture in a way that applies universally to all models. Salis and Frigg suggest that the particular way in which the world of a model is incomplete seems to require "a tailor-made cross-world similarity metric" such that "the counterfactual conditional $M \,\square\!\!\rightarrow C$ has no determinate truth value for all the right *Cs*" (2020: 44).

Williamson (2020) finds the objection unconvincing. He notices that in any conversational context many things that are true are also irrelevant and that a notion of relevance as a standard Gricean conversational implicature could be used to explain the sort of indeterminacy that is characteristic of counterfactual conditionals with irrelevant consequents in models. On this proposal, the world of the model is as complete as any other possible world and the consequents that seem to be indeterminate are determinate yet scientifically irrelevant. Furthermore, a semantic analysis of '*in*' could include a stipulated relevance condition such that *C* is true in the model if and only if two conditions obtain: i) if *M* then *C* holds; and ii) *C* is relevant to *M*. Williamson states that the latter, however, "is hardly worth the trouble, since the irrelevant truth is scientifically harmless" (2020).

While this may be the case, the fact remains that the conditional claim "if a gas were composed of point particles exerting no intermolecular forces, then Mont Blanc would be the tallest mountain in Europe" is intuitively neither true nor false. One may have independent reasons to preserve CEM and the completeness of possible worlds, and hence reject the intuition. Or one may recognise that scientific models pose a serious challenge to the completeness of possible worlds in the context of scientific modelling and go for a different analysis that does not satisfy CEM. This would be coherent with the face value practice of modelling and the theoretical principle of naturalism, and it would provide an opportunity for the development of a potentially more fruitful analysis of the sort of indeterminacy involved in models.

The second issue raised by an interpretation of modelling in terms of counterfactual imagination concerns epistemic access. Salis and Frigg (2020: 44) notice that there is no general agreement on the epistemology of counterfactual conditionals. Kment originally held that our ability to gain counterfactual knowledge "needs to be based on rules that permit us to determine which propositions are cotenable with a given antecedent" (2006: 288). Any epistemology of counterfactual conditionals needs to identify these rules. Currently, however, there is no general agreement on what these rules are. In particular, these rules should rely on a previous understanding of the similarity relation between possible worlds, which (as mentioned above) is still insufficiently characterised. These problems are inherited by a counterfactual epistemology of models. The set of *C*s that are true in a model is different in each case. An epistemology of counterfactual conditionals in models needs to build on a previous understanding of the tailor-made cross-world similarity metric for each case or, as Salis and Frigg tentatively suggest, "perhaps we can identify a series of overarching types of metrics for different types of models" (2020: 44).

The final issue raised by the interpretation of modelling in terms of counterfactual imagination is intersubjective access. Many imaginative activities are solitary and idiosyncratic. This is typically the case in the sort of imaginative activi-

ties involved in dreams and daydreams, and in many cases of counterfactual imagination. Modellers, however, work as members of a scientific community. Their imaginative activities have a social dimension that cannot be explained merely in terms of the ways in which individual modellers think in their own subjective and idiosyncratic ways. Godfrey-Smith himself recognises that model-based science "has sociological and formal features, as well as psychological ones" (2006: 728) and emphasises that he is not interested in providing an account of the psychological mechanisms underlying model-based reasoning. Thus, the analysis of the sort of imagination involved in modelling should build on the social practice of model-based science, and hence on the ways in which scientists think and talk about models as members of specific scientific communities. And while the counterfactual imagination may be compatible with this analysis (once the above problems are solved), a framework that builds merely on this kind of imagination does not have the theoretical resources to explain the social dimension of modelling. This social dimension, as I will argue, is the key feature of a different notion, compatible and yet distinct from the counterfactual imagination. This is the notion of make-believe that I will explore in the next section.

4. Make-believe

Salis and Frigg (2020) argue that make-believe is crucial to theoretical modelling. Walton (1990) originally introduced the notion of make-believe as a social imaginative activity with normative and objective content that is determined by the use of props. Props are ordinary objects that make propositions fictionally true in virtue of a prescription to imagine something. They are material objects that can be perceived and shared by different individuals in a context and thereby provide the physical scaffolding that enables the social, intersubjective dimension of make-believe. Effectively, props afford and constrain the imaginative processes of participants in the make-believe by making manifest the relevant prescriptions to imagine.

What is fictional truth? Naturally, many have spelled out the notion of fictional truth—or fictionality—in terms of fictional worlds. The idea comes from the literature on fiction, where storytelling is often construed as an activity that indicates or creates a fictional world. On this view, Mary Shelley's act of storytelling selects (among the logical space of possibilities) or generates (through her creative imagination) a world where it is true that Dr Frankenstein creates a hideous, intelligent and articulate creature through an unconventional laboratory experiment. This somewhat natural way of thinking about fictional truth as truth in the world of the story is interpreted in two main ways, literal and non-literal—or imaginative. On the literal interpretation, fictional truth is construed as a variety of truth and being fictionally true is being true in a possible (Lewis 1978) or, perhaps, impossible (Berto 2011; Priest 1997) world. This notion of fictional truth as truth in a world fits well with an analysis of modelling in terms of counterfactual imagination, but it also raises similar problems.

On a second, non-literal interpretation, fictional truth is not a variety of truth but a property of the propositions that are among the prescriptions to imagine in force in a fictional story (Eagle 2007; Currie 1990). This alternative notion of fictional truth, which is Walton's (1990) preferred notion, is often paraphrased in terms of correctness with respect to the prescriptions to imagine in force in a par-

ticular game of make-believe and has normative and objective features. It is normative because it depends on the rules that guide the imaginings of participants in the game. It is objective because it is independent of the individual imaginings of participants who may or may not conform with the prescriptions to imagine in force in a certain game of make-believe. Furthermore, on Walton's account, fictional truths divide between primary fictional truths and implied fictional truths of the game. Primary fictional truths are the initial assumptions of an episode of make-believe and they are generated directly from the props. Implied fictional truths are inferences generated indirectly from the primary fictional truths via principles of generation (more on these in the next Section).

As stated above, props are ordinary objects that can be perceived and shared by different individuals in a context. What sort of props are involved in a literary work? It is common to indicate, vaguely, the literary work of fiction as the prop. But we can be more specific and say that the concrete tokens constituting the text of a literary fiction are the props that prescribe to imagine in certain ways. These are concrete marks on paper, a computer screen or a tablet, which can be perceived and shared by different individuals in a context. In some cases, they can also be the concrete sounds produced by someone reading a text aloud, hence enabling an audible rather than visual experience of the text. These visible marks (or audible sounds) are the props that enable and constrain the intersubjective and social dimension of make-believe in literary fictions.

These ideas contribute an explanation of model building and model development. Let us start from model building. A scientist builds a model by specifying a model description—the prop—that prescribes certain imaginings. Like the text of a fictional story, the model description, which involves a linguistic and mathematical description, is constituted by concrete, physical marks that can be perceived and shared in a context. These perceptible marks provide the physical scaffolding that make the social dimension of modelling possible. They can be shared by different scientists in a context, hence enabling intersubjective communication within the scientific community and providing tools for the investigation of particular issues.

Similarly to the text of a story, the model description constrains the model's assumptions, or primary fictional truths, coherently with the model's prescriptions to imagine. These prescriptions to imagine involve the attribution of physical properties that only concrete objects can have, yet there are no such objects. For example, the model description of the ideal gas prescribes imagining that the molecules composing the gas are point particles having no volume of their own and bouncing against each other in elastic collisions. In this way, scientists build an imaginary system wherein imaginary gas molecules interact under imaginary conditions. This imaginary system emerges from the propositions that are among the prescriptions to imagine of the model. Hence, it is natural to interpret model building as a cognitive process that is enabled by a use of imagination that diverts from reality in some respects and to certain degrees for the purpose of building a surrogate, imaginary system.

What sort of object is this imaginary system? Realists about model systems argue that they are abstract created entities (Contessa 2007; Giere 1988). Antirealists hold that there are no model systems (Frigg 2010; Salis 2020a). Imaginings have no ontological commitments. So, for example, imagining a witch or telling

a fictional story about some witch do not commit to the existence of any witch.[2] Similarly, imagining an ideal gas or specifying a linguistic and mathematical description of an ideal gas in the imagination do not commit to the existence of any ideal gas. Walton's theory is compatible with both realism and antirealism about fictional entities. Personally, I have a strong preference for antirealism and I therefore assume that there are no model systems. What follows from this is that model systems are built in the imagination, without any commitment to their existence. Hence, there are no model systems.

Yet, there are models. This much seems undisputable. So, what are they? The term 'model' is often ambiguous between different uses. Sometime it is used to refer to the model system. A realist about model systems can endorse this interpretation of the term 'model' and argue that models are abstract objects. Realism about model systems, however, should be motivated by theoretical considerations that do not depend upon this particular problem. As stated above, I assume antirealism and hold that there are no model systems. This together with the assumption that models are model systems entail the absurd consequence that there are no models. Some other time the term is used to refer to the model description. But a mathematical equation or a string of linguistic symbols are not a model of anything unless they are interpreted in certain ways and according to certain conventions. So, a model description on its own is not a model. However, a model description together with its interpretation (its propositional content) can be identified with the model. On this view, which is the one I favour, a model is akin to a fictional story that the scientist tells by employing certain symbols (linguistic or mathematical) interpreted according to certain conventions.

The propositional content of a model can be analysed according to different accounts depending on one's theoretical stance. Descriptivist accounts will analyse it in terms of general propositions with a uniqueness condition where the description involves apparent reference to a particular (singular) entity, *à la* Russell (1905). Referentialist accounts will analyse it in terms of general propositions and, where certain singular terms such as proper names are involved, singular propositions (realism) or gappy propositions (antirealism), *à la* Braun (2005), or no proposition (antirealism), à la Walton (1990: Ch.10). While philosophers of science have well known descriptivist preferences, choosing over one or the other of these options requires independent theoretical reasons that do not hinge on anything specific to the case of models. For this reason, I will not take a stance on this particular issue.

So, on this proposal, a scientist builds a model (intended in this way) by specifying a model description (the prop) together with its interpretation (the primary fictional truths of the model determined by the model's prescriptions to imagine). These, in turn, specify a model system as the object of study, but only within the make-believe. The model is then developed by eliciting what is implicitly true in

[2] Of course, there were (and in some regions of the world there still are) societies that believed in the existence of witches. These beliefs, however, are rightly rejected in most advanced societies, which find other ways to express their own sexist and misogynistic stances. Fictions can be about real entities. Imaginings, however, do not commit to the existence of the objects they seem to be about. If they are about real entities, they are so in virtue of the existence of these entities.

it—or fictionally true. This requires going beyond the initial assumptions via principles of generation. Specifying what these principles are is no easy feat. I will discuss this problem in the next Section.

5. Constraints on Imagination

Make-believe is a type of imagination that is constrained by the game's prescriptions to imagine and by the principles of generation. In his critical assessment of Salis and Frigg (2020), Williamson (2020) notices that there is no general agreement on the epistemology of make-believe. Understanding our ability to learn through make-believe requires an investigation into the sort of constraints operating on it, including the principles of generation of implicit truths in the model. Salis (2020b) indicates at least three distinct types of such constraints, architectural, context-specific, and epistemic.

Architectural constraints are determined by the cognitive structure of the imagination and operate on all uses of imagination across different contexts. From the contemporary literature in cognitive science emerge two main architectural constraints, mirroring and quarantining.[3] Imagination displays mirroring when imaginings carry inferential commitments that are similar to those carried by isomorphic beliefs—that is, beliefs that have the same propositional content. If I believe that it is raining outside, I also believe that the pavement is wet. Similarly, if I imagine that it is raining outside, I also imagine that the pavement is wet. The inferences we make, however, typically depend on background assumptions and on the specific aims and practical interests that direct our reasoning. Thus, mirroring interacts with context-specific constraints to determine the sort of inferences that are allowed in particular episodes of imagination.

Quarantining is displayed when imaginings do not entail beliefs and do not guide action in the real world. In other words, quarantining guarantees that imaginings have effect only within an imagined episode. For example, if I believe that it is raining outside, and I have a desire not to get wet, I will pick up my umbrella on my way out of the house. But if I merely imagine that it is raining outside, I will not act in the same way. This does not mean that nothing of real-world importance can be learned through imagination. Learning about reality through imagination, however, requires exiting the imagination and exporting what one has learned outside of it and into reality. One can study the ideal gas model without automatically learning anything of real-world importance. Gaining knowledge of empirical truths about real world gases requires exporting what one has learned in the imagination onto reality.

While architectural constraints operate on all uses of imagination through different contexts, context-specific constraints are determined by disciplinary conventions and interpretative practices. Individuals who engage in these practices imagine in ways that are specific to the practices themselves. Context-specific constraints correspond to Walton's principles of generation. They are the constraints that enable the generation of implicit truths in a game of make-believe. Inspired

[3] Salis and Frigg (2020) identify mirroring and quarantining as two key features of propositional imagination (together with a third one, which is the typical freedom of imagination). See also Leslie (1987), Nichols (2004), and Nichols and Stich (2003) for the original discussion of mirroring and quarantining based on experimental and theoretical research in cognitive psychology and philosophy of mind.

by Lewis (1978), Walton (1990) identifies two main principles of generation, the reality principle and the mutual belief principle.[4]

The reality principle keeps the world of the game as close as possible to the real world. Effectively, this principle relies on the notion of closeness that is key to Lewis's and Stalnaker's standard analyses of counterfactual conditionals. This brings some, although not all, of the aforementioned problems into the framework of make-believe. First, the notion of closeness—reality orientation or similarity—is left unconstrained and mysterious because it is insufficiently characterised. Second, within the framework of make-believe, the reality principle could be implemented without commitment to the completeness of possible worlds. This is because Walton's (1990) appeal to reality orientation is quite loose and does not commit to the completeness of fictional worlds. But as Salis and Frigg (2020) emphasise, different models are incomplete in their own specific ways, which raises the issue of how to provide the right cross-world similarity metric for each particular case. Third, there is no general agreement on the rules that enable us to determine which co-inferences are allowed by a given antecedent. These rules should rely on a previous understanding of the notion of similarity of worlds, which is currently unavailable.

The second principle identified by Walton is the mutual-belief principle, which imports the mutual beliefs of the members of the community in which the game originated. Beliefs are of many different kinds. In the context of modelling, theoretical beliefs and experts' opinions are fundamental for drawing certain inferences within particular models. More context-specific constraints are also possible and new research through historical and contemporary case studies may contribute a better understanding of what they are. Among them are mathematical constraints provided by the particular mathematical tools deployed in a model, interpretations of data, and more.

Finally, epistemic constraints are determined by the particular sort of knowledge we want to acquire. In the context of modelling, there are two main types of knowledge gained through imagination, knowledge of the imaginary scenarios described by model descriptions, and knowledge of empirical truths about reality. These different types of knowledge correspond to two different types of claims generated through imagination in modelling, knowledge claims about the imaginary system specified by the model and knowledge claims about reality.

Knowledge claims about imaginary systems are the claims scientists make within a game of make-believe, such as "the ideal gas is composed of point particles" (in the ideal gas model). These claims are produced within the make-believe, not without it. They are internal claims about the ideal gas (the imaginary system), not external claims about the model (the complex entity constituted by model description and model content). Scientists merely imagine the content of these claims (rather than believing it), which are merely fictionally true (rather than genuinely true).

But what sort of justification do scientists have to make these claims? In the traditional theory of knowledge, justification has the special role of ensuring that "a true belief isn't true merely by accident" (Steup 2018). A belief that p is justified if and only if there are some grounds that properly increase the probability that it

[4] See also Evans (1982) for a classical discussion of these two principles within the framework of make-believe, and Friend (2016) on the reality principle and a different take on the aboutness of fictional stories.

is true. When we think about the notion of justification in the context of knowledge of imaginary scenarios, the question we need to ask is: What sort of grounds probabilify knowledge claims about imaginary systems? Most plausibly, the relevant sort of grounds must depend on specific modelling practices. Mathematical constraints operate on uses of imagination in all theoretical models. Theoretical grounds may play an important justificatory role in many types of models, including macroeconomic models, models in cognitive neuroscience and models in physics. However, they may play a more limited role in mechanistic models in chronobiology and models in medicine. In these cases, empirical grounds (broadly construed) may play a more relevant justificatory role. More fine-grained distinctions about the specific constraints at work in different modelling practices and even in specific models could be made through case studies.

In the ideal gas model, the principles of generation are quite straightforward and they are provided by the mathematical constraints imposed by the model equation. The model assumes that the volume V of the imaginary gas is proportional to the number of moles n. So, when one doubles n, keeping pressure and temperature constant, V doubles too. In this way one learns about the properties of an imaginary gas. Learning about real gases, however, requires exporting what one has learned about the imaginary system outside of the make-believe and onto reality via the formulation of theoretical hypotheses. These hypotheses are of two kinds, model-world comparisons and direct attributions.

Model-world comparisons are claims that scientists make about the model system and the real system of interest. Often, they are based on a relation of similarity, which is usually interpreted as the sharing of certain properties in certain respects and to certain degrees. So, one can claim that the ideal gas and some real gas have similar behaviours in certain respects. For example, one can claim that when one doubles the number of moles n of an ideal gas and a real gas, keeping pressure and temperature constant, the volume V of the two gases will double too. The ideal gas, however, is only a fiction, a useful construct of the imagination. So, it cannot have the sort of properties that it supposedly shares with real gases. More generally, model systems are constructs of the imagination that do not exist (they are creatures of the imagination that inhabits a model's fictional scenario) and therefore cannot have any of the properties that they supposedly share with their targets.[5] As a consequence, there cannot be any real similarity between models and reality. But then how can we make sense of the common practice of scientists to compare properties of the model system with properties of real systems?

Answering this question requires that we reconceptualise the notion of similarity in terms of imagined similarity, that is, in terms of the attribution of certain properties to model systems in the imagination, and more specifically within a game of make-believe. According to Walton (1990), games of make-believe can be of two main sorts, authorised and unofficial. A game is authorised when its fictional truths are determined by the model's prescriptions to imagine and the relevant principles of generation. For example, the claim "the ideal gas is composed of point particles" is true in the ideal gas model. A game is unofficial when its fictional truths are determined by some *ad hoc* rules. The claim that "when one doubles the number of moles n of an ideal gas and a real gas, keeping pressure and temperature constant, the volume V of the two gases will double too" is true only in an unofficial game of make-believe constrained by *ad hoc* rules combining

[5] See Hughes 1997 for a similar concern.

the original prescriptions to imagine of the ideal gas model and new prescriptions to imagine determined by the ways in which real gases behave. Real gases and imaginary gases cannot share any properties, so the claim is literally false. But the same claim is fictionally true when assessed from within the unofficial game of make-believe because they share such properties in the imagination. Knowledge claims generated from model-world comparisons can be assessed only within unofficial games of make-believe and therefore involve epistemic constraints that are similar to those involved in knowledge claims about imaginary systems. Their content is the object of imagination rather than belief. And they can only be fictionally true (or false) when assessed within a game of make-believe (even if unofficial).

Typically, however, scientists build and develop models to learn about reality, to gain some better understanding of it and, possibly some new knowledge. This requires stepping out of the imagination through the formulation of theoretical hypotheses that do not involve any reference to imaginary systems. A scientist can claim that "if one doubles the number of moles of a real gas, keeping pressure and temperature constant, the volume doubles too". This is a hypothesis that is exclusively about a real system and that can be assessed and even tested for truth. This second sort of hypotheses is enabled by the development of the model in make-believe. But it is exported outside of it in the form of a direct attribution to real systems of the properties attributed to model systems in the make-believe. The knowledge claims generated from direct attributions export what one has learned about the model system into reality, they are exclusively about reality and can be assessed for truth. The attitude one has towards their content is belief and the sort of justification they require is typically provided by empirical evidence.

6. Conclusion

In this paper I advocated the view that scientific modelling crucially relies on imagination of the make-believe variety and that this must be constrained in certain ways to enable knowledge of reality. I described the first overarching taxonomy of types of constraints on imagination in modelling, architectural, context-specific and epistemic. And I identified two main varieties of knowledge generated through modelling, knowledge of the model imaginary system and knowledge of reality. One aspect of the proposal that should be emphasised is that the above taxonomy is open and does not exhaust the many possible specific constraints on uses of imagination in particular modelling practices. New research through case studies is required to specify the different context-specific and epistemic constraints at work in different modelling practices. This, however, will be the aim of future work.

References

Arlo-Costa, H. 2019, "The Logic of Conditionals", *The Stanford Encyclopedia of Philosophy*, Zalta, E.N. (ed.), https://plato.stanford.edu/archives/sum2019/entries/logic-conditionals/.

Arcangeli, M. 2018, *Supposition and the Imaginative Realm: A Philosophical Inquiry*, New York: Routledge.

Berto, F. 2011, "Modal Meinongianism and Fiction: The Best of Three Worlds", *Philosophical Studies*, 152, 3, 313-34.

Braun, D. 2005, "Empty Names, Fictional Names, Mythical Names", *Noûs*, 39, 4, 596-631.

Cartwright, N. 2010, "Models: Parables v Fables", in Frigg, R. and Hunter, M.C. (eds.), *Beyond Mimesis and Convention. Representation in Art and Science*, Berlin and New York: Springer, 19-32.

Contessa, G. 2007, "Scientific Representation, Interpretation, and Surrogative Reasoning", *Philosophy of Science*, 74, 48-68.

Currie, G. 1990, *The Nature of Fiction*. Cambridge: Cambridge University Press.

Descartes, R. 1985, *Meditations on First Philosophy*, ed. and trans. J. Cottingham, Cambridge: Cambridge University Press.

Eagle, A. 2007, "Telling tales", *Proceedings of the Aristotelian Society*, 107, 125-47.

Evans, G. 1982, *The Varieties of Reference*, Oxford: Oxford University Press.

Findlay, A. 1948, *A Hundred Years of Chemistry*, 2nd ed., London: Duckworth.

Friend, S. 2016, "The Real Foundation of Fictional Worlds", *Australasian Journal of Philosophy*, 95, 1, 1-14.

Frigg, R. 2010, "Models and Fictions", *Synthese*, 172, 251-68.

Frigg, R. and Hartmann, S. 2018, "Models in Science", in Zalta, E.N (ed.), *The Stanford Encyclopedia of Philosophy*, <https://plato.stanford.edu/archives/sum2018/entries/models-science/>.

Giere, R.N. 1988, *Explaining Science: A Cognitive Approach*, Chicago: The University of Chicago Press.

Godfrey-Smith, P. 2006, "The Strategy of Model-Based Science", *Biology and Philosophy*, 21, 725-40.

Godfrey-Smith, P. 2009, "Models and Fictions in Science", *Philosophical Studies*, 143, 101-16.

Godfrey-Smith, P. 2020, "Models, Fictions, and Conditionals", in Godfrey-Smith, P. and Levy, A. (eds.), *The Scientific Imagination*, Oxford: Oxford University Press, 155-77.

Harré, R. 1988, "Where Models and Analogies Really Count", *International Studies in the Philosophy of Science*, 2, 118-33.

Hughes, R.I.G. 1997, "Models and Representation", *Philosophy of Science*, 64, 325-36.

Kment, B. 2006, "Counterfactuals and the Analysis of Necessity", *Philosophical Perspectives*, 20, *Metaphysics*, 237-302.

Knuuttila, T. and Voutilainen, A. 2003, "A Parser as an Epistemic Artefact: A Material View on Models", *Philosophy of Science*, 70, 1484-95.

Leslie, A. 1987, "Pretense and Representation: The Origins of 'Theory of Mind'", *Psychological Review*, 94, 4, 412-26.

Levy, A. 2015, "Modeling Without Models", *Philosophical Studies*, 172, 3, 781-98.

Lewis, D.K. 1973, *Counterfactuals*, Oxford: Basil Blackwell.

Lewis, D.K. 1978, "Truth in Fiction", *American Philosophical Quarterly*, 15, 1, 37-46.

Lewis, D.K. 1979, "Counterfactual Dependence and Time's Arrow", *Noûs*, 13, 455-76.

Magnani, L. 2009, *Abductive Cognition*, Berlin and New York: Springer.

Mäki, U. 1992, "On the Method of Isolation in Economics", *Poznań Studies in the Philosophy of Science and Humanities*, 26, 316-51.

Mäki, U. 2005, "Models are Experiments, Experiments are Models", *Journal of Economic Methodology*, 12, 303-15.

Maxwell, J.C. 1965, *The Scientific Papers of James Clerk Maxwell*, Niven, W.D. (ed.), New York: Dover.

Morgan, M. 1999, "Learning from Models", in Morgan, M. and Morrison, M. (eds.), *Models as Mediators. Perspectives on Natural and Social Science*, Cambridge: Cambridge University Press, 347-88.

Nersessian, N. 2008, *Creating Scientific Concepts*, Cambridge, MA: The MIT Press.

Nersessian, N. 2009, "Conceptual Change: Creativity, Cognition, and Culture", in Meheus, J. and Nickles, T. (eds.), *Models of Discovery and Creativity*, Dordrecht: Springer, 127-66.

Nichols, S. 2004, "Imagining and Believing: The Promise of a Single Code", *Journal of Aesthetics and Art Criticism*, 62, 129-39.

Nichols, S. 2006, "Just the Imagination: Why Imagining Doesn't Behave like Believing", *Mind and Language*, 21, 4, 459-74.

Nichols, S. and Stich, S. 2003, *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*, New York: Oxford University Press.

Norton, J. 1991, "Thought Experiments in Einstein's Work", in Horowitz, T. and Massey, G.J. (eds.), *Thought Experiments in Science and Philosophy*, Lanham (MD): Rowman & Littlefield, 129-48.

Priest, G. 1997, "Sylvan's Box: A Short Story and Ten Morals", *Notre Dame Journal of Formal Logic*, 38, 4, 573-82.

Russell, B. 1905, "On Denoting", *Mind*, 14, 479-93.

Salis, F. 2019, "The New Fiction View of Models", *The British Journal for the Philosophy of Science* (online first), https://doi.org/10.1093/bjps/axz015

Salis, F. 2020a, "Scientific Discovery through Fictionally Modelling Reality", in Ippoliti, E. and Nickles, T. (eds.), *Scientific Discovery and Inferences*, *Topoi*, 39, 927-37.

Salis, F. 2020b, "Of Predators and Prey: Imagination in Scientific Modeling", in Moser, K. and Sukla, A. (eds.), *Imagination and Art: Explorations in Contemporary Theory*, Leiden: Brill.

Salis, F. and Frigg, R. 2020, "Capturing the Scientific Imagination", in Godfrey-Smith, P. and Levy, A. (eds.), *The Scientific Imagination*, Oxford: Oxford University Press, 18-50.

Sorensen, R. 1992, *Thought Experiments*, New York: Oxford University Press.

Spaulding, S. 2016, "Imagination through Knowledge", in Kind, A. and Kung, P. (eds.), *Knowledge through Imagination*, Oxford: Oxford University Press, 208-26.

Stalnaker, R. 1968, "A Theory of Conditionals", in Rescher, N. (ed.), "Studies in Logical Theory", *American Philosophical Quarterly*, Monograph Series, Vol. 2, Oxford: Blackwell, 98-112.

Stalnaker, R. 1981, "A Defence of Conditional Excluded Middle", in Harper, W.L., Stalnaker, R. and Pearce, G. (eds.), *IFS: Conditionals, Belief, Decision, Chance and Time*, Dordrecht: Reidel, 87-105.

Stalnaker, R. 1986, "Possible Worlds and Situations", *Journal of Philosophical Logic*, 15, 1, 109-23.

Steup, M. 2018, "Epistemology", in Zalta, E.N. (ed.), *The Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/archives/sum2018/entries/epistemology.

Suárez, M. 2004, "An Inferential Conception of Scientific Representation", *Philosophy of Science*, Supplement, 71, 767-79.

Sugden, R. 2009, "Credible Worlds, Capacities and Mechanisms", *Erkenntnis*, 70, 3-27.

Swoyer, C. 1991, "Structural Representation and Surrogative Reasoning", *Synthese*, 87, 449-508.

Thomson-Jones, M. 2010, "Missing Systems and the Face Value Practice", *Synthese*, 172, 283-99.

Toon, A. 2012, *Models as Make-Believe: Imagination, Fiction, and Scientific Representation*, Basingstoke: Palgrave Macmillan.

Tyndall, J. 1868, *Faraday as a Discoverer*, London: Longmans, Green &Co.

Walton, K. 1990, *Mimesis as Make-Believe*, Cambridge, MA: Harvard University Press.

Weisberg, M. 2007, "Who is a Modeler?", *British Journal for the Philosophy of Science*, 58, 2, 207-33.

Weisberg, M. 2013, *Simulation and Similarity: Using Models to Understand the World*, New York: Oxford University Press.

Williamson, T. 2005, "Armchair Philosophy, Metaphysical Modality, and Counterfactual Thinking", *Proceedings of the Aristotelian Society*, 105, 1-23.

Williamson, T. 2020, "Review of The Scientific Imagination, by Arnon Levy and Peter Godfrey-Smith", *Notre Dame Philosophical Reviews*, https://ndpr.nd.edu/news/the-scientific-imagination-philosophical-and-psychological-perspectives/

Spoiler Alert! Unveiling the Plot in Thought Experiments and other Fictional Works

Daniele Molinari

University of Parma

Abstract

According to a recent philosophical claim, "works of fiction are thought experiments" (Elgin 2007: 47), though there are relevant differences, as the role of spoilers shows—they can ruin a novel but improve the understanding we can gain through a thought experiment. In the present article I will analyze the role of spoilers and argue for a more differentiated perspective on the relation between literature and thought experiments. I will start with a short discussion of different perspectives on thought experiments and argue that the mental-model view and the conception of games of make-believe are most promising for developing the present analogy. Then I will assess the similarities and differences between thought experiments and other works of fiction. I will focus on the role of spoilers and, more generally, on the foretaste context, of which they are part. This context guides readers of literary works of art to draw their attention to the literary and aesthetic quality of the text. In the case of thought experiments, on the other hand, it (i) prompts them to accept the presence of fictional elements in worldly-cognitive works and (ii) draws their attention towards cognitively relevant elements of the story. A discussion of Borges' Pierre Menard in the last part will show that literary works of art become thought experiments if they are embedded in an appropriate foretaste context. Spoilers, thus, unveil that even works which—due to their length or plenty of detail—usually are not considered thought experiments, can perform similar cognitive functions.

Keywords: Thought experiments, Works of fiction, Spoiler, Imagination, Foretaste context.

1. Introduction

A brand-new mystery story of your favorite author has just been published. As often, several plot anticipations begin to appear online. You cannot keep curiosity at bay and instantly try to find out who the new murderer is—a weak moment that just ruins your enjoyment of the long-awaited work. This love-hate relationship results from two conflicting tendencies: (i) the desire to immediately release

Argumenta 6, 1 (2020): 81-97
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.mol

the tension and find out how the work ends and (ii) the desire to enjoy a good, unspoiled read, to work your way through the text step by step and slowly unveil the solution. This is particularly evident in the case of criminal novels or movies, where suspension builds up, holding the reader in tension, and twist turn their expectations upside down. It is the main purpose of works of this kind to keep the reader in suspense—and we admire them for that. Does this point apply to all works of fiction? If we look at thought experiments as fictional narratives,[1] spoilers seem to lose all their destructive aura and turn into a most helpful tool.

According to a recent claim, at least some works of literature are thought experiments (Carroll 2002; Davies 2007, 2018; Elgin 1996, 2007, 2014, 2017, 2019). This claim sheds light on a much-debated question: can fiction provide knowledge? It may be obvious that Nobel-laureates such as Wisława Szymborska and Gabriel Garcia Márquez wrote literary masterpieces of outstanding cultural and cognitive value. But it is less obvious how exactly these or other works of fiction impart relevant knowledge about the real world to the reader. Catherine Elgin provides a straightforward explanation: "works of fiction are thought experiments" (2007: 47). Thus, if a work of fiction can widen our cognitive horizons, it will do so in the same way a thought experiment does. This suggestion is, however, only the tip of the iceberg that hides, on a more profound level, a series of problems. Its plausibility depends essentially on the conception of thought experiments that one endorses, i.e. which position one opts for in the lively debate on the nature of thought experiments that is immersed in the icy waters, as it were: how can thought experiments—which are never actually carried out and often involve a fictional narrative—add to our empirical knowledge or contribute to our scientific understanding?

I will develop my argument in two steps. First I will briefly recall the main positions in the current debate on thought experiments and suggest that Elgin's claim is best suited to the view that thought experiments are mental model reasoning and works of literature are games of make-believe. Then, I will present some analogies between works of literature and thought experiments and show how the role of spoilers can help us to shed a new light on the differences between them. I will argue that spoilers can be useful in thought experiments and scientific papers, but counter-productive in many genres of fictional works as they ruin the reader's experience. It will be helpful to focus on the context in which spoilers can be embedded. There are works of fiction that can be used in two different ways, as literary works or as thought experiments, depending on context; in the first case, the author of the work will avoid spoilers so as not to diminish the aesthetic quality of the work, while in the second case spoilers can be helpful as they guide the reader's attention and so enhance the cognitive efficiency of the work.

2. Thought Experiments and Games of Make-Believe

2.1 How Do Thought Experiments Work?

Let's start with thought experiments. They give rise to an epistemological puzzle (Kuhn 1977; Davies 2007) that can easily seem paradoxical: they can and often do enhance our understanding of reality even though they do not provide new empirical data. In recent debate at least three main accounts can be distinguished:

[1] Why should not we? Thought experiments are devices with narrative structure in which fictional events dynamically occur (Egan 2016, Nersessian 1991, Willée 2019).

the *Argument View*, the *Platonic Perception View*, and the *Mental-Model Reasoning View*. Let's have a quick look at the three positions.

John Norton claims that, behind their aesthetic and narrative features, thought experiments always ground themselves in deductive or inductive argumentations (1996, 2004). Each thought experiment can, therefore, be reduced to an argument with no epistemic loss: aesthetic and narrative features are only used for illustrative purposes. In an antithetical way, James Brown argues for a Platonic solution, taking thought experiments as "telescopes" directed to the realm of abstract entities (1991, 2004). The way in which we gain knowledge concerning this kind of artifacts has nothing to do with empirical experience. Rather, it is grounded on an *intellectual perception*—a special activity that allows us to grasp independent and outside-from-space-and-time laws of nature.

The attraction of the *Platonic view* lies in the fact that it recognizes thought experiments as a peculiar source of knowledge, different from empirical data-gathering or logical reasoning. Norton's position, on the other hand, has the virtue of parsimony, while Brown can do justice to the point that thought experiments are both essential and peculiar.

While for Norton, thought experiments are only "arguments in disguise" and can easily be replaced by them, the *Platonic Perception View* attributes a central relevance to them, but does invite for serious concerns: first, it does not provide a reliable account of how abstract entities can be grasped in intuition; second, the unexplained use of the metaphor of "intellectual perception" raises suspicion as it comes along with inappropriate empirical commitments. Norton's *Argument View*, on the other hand, fails to account for imagination-based thought experiments, such as cases where you have to imagine a certain shade of blue or those in which you put yourself in someone's else position to assess moral judgement.

Supporters of the *Mental-Model Reasoning view* can combine the strengths of the other two positions and avoid their problems by arguing that thought experiments provide cognitive advancement grounded on experience in virtue of an essential use of imagination rather than collecting new empirical data. In virtue of imagination, simulation, and memory, thought experiments can reconfigure previously obtained empirical data in new ways, prompting new experience-related knowledge (Gendler 2010; Miščević 1992, 2007; Nersessian 1991, 1992, 1999, 2018). This solution takes up central insights from Ernst Mach and Philip Johnson-Laird. The former argues that thought experiments use an instinctive kind of experience stored in memory but not yet propositionally articulated (Mach 1976). According to the latter we use mental models in the understanding of narrative texts, in which we experience some sort of contemplation of a fictional situation (Johnson-Laird 1983).

In conducting a thought experiment, we deal with a hypothetical state of affairs in a "What would happen if…" style. We reflect the consequences a counterfactual, fictional state of affairs would have. In doing so, the reader employs many non-linguistic cognitive resources, such as her familiarity with her own body, spatial intuitions, and forms of tacit knowledges (*Know-how*). In thought experiments, we imagine *seriously*: it is thanks to the meticulousness with which these cognitive resources are used, with respect to the constraints designed by the author, that a thought experiment can play a role that is analogous to that of a real experiment. Successful thought experiments invite the reader to imagine scenarios that allow her to draw true, or at least plausible, conclusions regarding the real world.

Tamar Gendler, who also draws on Mach's account, further develops this position, using the terms "mental representation" (2010: 47). She argues that thought experiments are moments in which one is "contemplating the imaginary case in question" using "a store of unarticulated knowledge of the world which is not organized under any theoretical framework" (Gendler 2010: 39).

It is worth noting that the authors here considered explicitly state that their theories regard scientific thought experiments only. According to Rachel Cooper, this is a "strategy of caution" (2005: 329): there are different types of thought experiments, and scientific ones seem *prima facie* easier to classify. Gendler does state, however, that the only difference between scientific and non-scientific thought experiments lies in the fact that the first concern "features of the physical world" (2010: 45). For these reasons, I believe that a unifying account of thought experiments is to be preferred. In this article I will consider thought experiments in general.

2.2 Games of Make-Believe and Pierre Menard

The role of imagination and its limits have been of interest in the entire history of philosophy. In what follows—to spoil the choice right away—I will rely on Kendall Walton's account, which provides an interesting explanation of the analogy between works of literary fiction and thought experiments. In particular, Walton's notion of *fictional world*—which can be compared to "mental model", a "space" within which the development of a fictional state of affairs is imagined, will be of interest.

For our purposes, one highlight of *Mimesis as Make-Believe* (Walton 1990) is the concept of *principle of generation*—a more or less explicit rule that gives structures to a fictional world. For example, a bunch of bored children decide to play together, and imagine that the apples on the kitchen-table are hand bombs. From the beginning of the game, and until its conclusion, the principle of generation "apples are dangerous hand bombs" is pretended to be true and the children begin to behave accordingly. All children who accept this principle of generation become players attuned to the same fictional world—a micro-world within which apples are fictitious hand-grenades. In light of that, a player who eats an apple (i) either pretends to commit suicide, (ii) proposes a further principle of generation and pretends to defuse or hide the bomb, (iii) does not play correctly, (iv) or is just hungry and decides to "detune" from the fictional world and "re-attune" to the real world.

According to Walton, principles of generation are central also in other contexts: figurative sculptures, paintings, and works of fiction prompt the viewers'/readers' imagination. Principles of generation are, thus, best understood as rules that prescribe what to imagine and which every player has to respect in order to take part in the game. According to this conception, imagination is not an unregulated, creative faculty but "it is a realm in which the play of ideas is bound by constraints the imaginer sets" (Elgin 2014: 227).

However, the mere desire to comply with a rule is not enough to act in accordance with it (Wittgenstein 2009: §202). An external criterion is needed. In Walton's perspective, *props* as warrants of the correctness of a game perform this function. A *prop* is an object that makes it possible to retrieve principles of generation and give coherence to a fictional world. The children involved in the *apple-*

bomb world rely on real apples to generate the fictional truth that there are hand-grenades on the kitchen table. The apples, thus, serve as external criterion in this.

Similarly, also fictional texts could be understood as props in games of make-believe: Jorge Luis Borges' short story (1999) generates the fictional truth that Pierre Menard planned to re-write Cervantes' *Don Quixote*, just like John Searle's article (1980) made it fictionally true that the person in the room can answer questions formulated in Chinese without mastering the language—and so made it possible that all readers take part in the same game of make-believe, the *Chinese-room* thought experiment. All props work independently from particular acts of imagination, but not in isolation: a (more or less) explicit agreement, in the stipulation of the principles of generation that guide the game, is necessary to start it and to make sure that all can enjoy the same fiction work or thought experiment—that might deviate from one another only in minor details (Meynell 2018: 503). This shows that Walton's theory has a well-marked normative and social dimension.

According to Walton, thus, the great works of literature are fine props that prompt the readers' imagination in different ways: by surprising them, by making them reflect on determinate matters or on the psychology of a character, or by making them identify with strange beings, etc. Sometimes, a work of literature can be used to show, to test, or to argue for a certain theory or hypothesis. Thought experiments understood as mental model manipulation seem to work precisely in that way: thanks to a set of prescriptions the author of a thought experiment invites the readers to imagine a certain fictitious state of affairs and so designs a model that shows something meaningful.

A potential critique of the position that I have discussed so far could emerge from a recent suggestion, formulated by Fiora Salis and Roman Frigg (2020), according to which only propositional imagination is necessary for the performance of thought experiments and games of make-believe. The contents of propositional imagination are propositions. It has three main features: (i) one can freely imagine any proposition one desires, (ii) an imagined proposition imposes inferential commitments similar to those imposed by a proposition that is believed, and (iii) to imagine a proposition does not require one to believe it. The authors argue that games of make-believe are cases of propositional imagination which, in addition of the above features, are (iv) social activities structured through (v) normative aspects. Such a classification is inspired by Gregory Currie (1990), who argues that activities of make-believe are propositional attitudes similar to belief and desire. Salis and Frigg, following Currie, suggest that to play a game of make-believe might arouse mental images, but this is not a necessary feature, and therefore it is not a relevant element of make-believe.

According to Salis and Frigg, the five features that identify make-believe listed above are also shared by thought experiments: to imagine a set of propositions that describe a hypothetical scenario, to imagine their possible inferences, and to imagine the conclusion that is to be drawn from such a set of imagined propositions do not require the contemplation of a quasi-visual situation. The conclusions of thought experiments, therefore, presuppose the possession of linguistic competences, rather than pre-theoretical information stored in memory or any kind of "phenomenological" imagination.

Walton is less reductive on this point; he distinguishes different types of imagination: "imagining a proposition, imagining a *thing*, imagining *doing* something" (1990: 13) and states:

> Props prescribe nonpropositional imaginings as well as propositional ones. They do not thereby generate fictional truths, but the mandated nonpropositional imaginings are a distinctive and important part of our games of make-believe (Walton 1990: 43).

Although the term 'make-believe' suggests some resemblance to beliefs, and therefore to propositional attitudes, Walton does not privilege propositional imagination over other forms of imagination: depending on the fictional world that is presented in them, games of make-believe can involve different forms of imagination.

Mental model reasoning seems to provide further evidence for Walton's liberal position, as it underlines the fact that we can gain meaningful insights from different imaginative resources, such as imagine-that-color ("imagining a *thing*") and imagine-that-feeling ("imagining *doing* something", or "imagining *being affected* by something"). Hume's notorious thought experiment of the *Missing Shade of Blue* (1999: 9f.) and Thomson's *Dying Violinist* (1971: 48f.), essentially involve non-propositional imagination: the visual imagination of a particular nuance of blue in the first case and the imagination of feeling empathy for a character, in the latter. Since both types of imagination cannot be reduced to propositional imagination, the attempt to reduce the imaginative activities required by thought experiments to propositional imagination must fail.

We are now in a position to come back to Elgin's claim concerning an analogy between literature and thought experiments. Borges' *Pierre Menard, Author of the Quixote* seems to substantiate this view: it invites the reader to consider a counterfactual scenario, adopting the sober and direct style of a literary review. As a work of literature, the text exemplifies Borgesian humor and an excellent mix of essayistic and narrative genre. As a thought experiment, it develops an argument in favor of the thesis that a work of literature cannot easily be separated from the historical context and that extrinsic features of the work (partially) determine its identity-criteria (Bailey 1990; Danto 1981; Goodman and Elgin 1988; Lamarque 2009).

We should be cautious, however. The example discussed, which is right in the *Goldilocks zone* between works of literature and thought experiments, is likely to be more an exception than a prototypical example. It is true that works of literature and thought experiments have important analogies, but we have to take their differences into account to avoid a *petitio principii*. I will return to *Pierre Menard* later in this paper, when I will discuss the role of spoilers in works of fiction.

3. Exploring the Analogy

There are apparent similarities between literary works and thought experiments. In what follows, I will discuss four points that illustrate the strength of Elgin's analogy: (i) both develop their plot in a narrative, (ii) in indeterminate or incomplete contexts, (iii) are subject to reality constraints, (iv) and both can provide an advancement of the reader's understanding. Elgin's analogy has its limits, however, as I will try to show in section 3.2, where I will come back to the phenomenon of spoilers.

3.1 Similarities

(i) First of all, both works of literature and thought experiments develop a narrative. Both describe series of events that are causally related or groups of

similar but causally unrelated events.[2] It is worth noticing that thought experiments are usually presented *in medias res*, which likens them to standard experiments, that are also conducted *in medias res* (Elgin 2014: 225), while other fictional works may use different narrative devices depending on what type of game of make-believe they are meant to induce.

(ii) Second, narrated events always occur in partially indeterminate or incomplete contexts. Unlike possible worlds, at least in David Lewis' conception,[3] works of literature and thought experiments leave many aspects open. It is neither true nor false that Pierre Menard has a brother who is a musician, or that the *Society of Music Lovers* in Judith Thomson's thought experiment publishes a journal that is dedicated entirely to Chopin's style. These details are not mentioned in the text and, therefore, play no role in the games of make-believe that they prompt. Ignoring details of a fictional world that are not present in the plot helps the author to control the scenario and to highlight the ones that she wants to draw attention to. We find a similar strategy in real experiments conducted in laboratories: also here it is important to suspend all irrelevant elements that could distort results and lead the reader astray (Elgin, 2014: 222)—and in thought experimentation this selective ability is at its best use.

(iii) Moreover, elements of the real world are typically carried over to the fictional worlds that are described in works of fiction or in thought experiments. This is what Stacie Friend calls the *reality assumption* (2017: 31), i.e. the assumption that readers of fictional works usually import aspects of the plot from the real world. It is true in the story that Pierre Menard has got a brain, or that the *Society of Music Lovers* is not composed by a bunch of domesticated monkeys, even if Borges or Thomson did not explicitly mention any of these facts. This point does not conflict with (ii). Although it suggests that we add details not present in the plot, these are details that we take for granted in our world-view and that are not specific of a given state of affairs. Moreover, even though some aspects are imported from the real world, other are still left indeterminate.

(iv) One last analogy is about the advancement of understanding: both thought experiments and works of literature can show, defend or confute a hypothesis. Powerful works of literature can have a more lasting impact and make the reader to reflect on their themes in later moments in time. Engaged novels like Orwell's *Animal Farm*, dystopian ones like Huxley's *Brave New World*, and existential novels such as Camus' *The Stranger* are particularly apt to prompt the reader to think about the meaning of life, moral choices, the responsibility of technology etc. In the case of thought experiments, it is more obvious that they can serve cognitive goals—after all, this is the main purpose for which they have been devised.

Notwithstanding these analogies, however, we can note a relevant difference between literature and thought experiments, which becomes particularly evident

[2] In *Magnolia*, directed by Paul Thomas Anderson in 1999, the plot is developed in an alternating intertwining that made school in contemporary cinematography. The stories narrated are almost totally isolated from each other but connected by various themes, including that of cancer.

[3] Lewis argues that every assertion regarding to a possible world has truth-value (1986).

when we look at the way in which the narrative is presented; i.e. at aspects of aesthetic appreciation and of literary style. I will focus on one of these aspects in the next section.

3.2 A Significant Difference: Spoilers

Let me illustrate how the role of spoilers differs between thought experiments and other works of fiction by discussing two emblematic examples, the movie *The Sixth Sense* (directed by M. Night Shyamalan in 1999) and Judith Thomson's aforementioned thought experiment of the *Dying Violinist*. Attention: spoiler alert! In the following discussion, the plot of *The Sixth Sense* will be revealed. If you do not want to ruin your experience of the movie, you better skip the next two paragraphs.

In *The Sixth Sense*, the child psychologist Malcolm Crowe deals with an apparently common case: a 9-year-old boy called Cole feels strongly anxious in every life context. Dr. Crowe takes the child to heart but Cole confesses that his problem is not psychological: he claims to possess the extraordinary capacity to see dead people. Crowe works hard to give the boy a life-purpose and put him in a condition to accept his special capacity. The film ends with a masterful plot-twist: we realize that Dr. Crowe had already died and was dead throughout the movie. His interaction with Cole were possible only due to the latter's paranormal gift. Throughout the movie, the spectator was longing to find out whether or how Dr. Crowe was able to heal Cole from his anxieties—just to find out, at the end, that Dr. Crowe was dead and it was Cole who comforted the wandering dead all the way long.

Shyamalan's movie illustrates well the destructive effect a spoiler can have on the spectator's experience: one of the central pleasures of the film lies in the fact that the final twist turns upside down the storyline of the entire film and forces the spectator to interpret several scenes of the film in a completely new light. One might even want to see the film a second time, just to see whether the entire plot was consistent with the new interpretation and whether there were hidden clues that could have given the surprising final twist away. This second viewing would afford, if any, a very different kind of pleasure. If a potential viewer has never watched that movie but already knows that Crowe is a wandering dead, both kind of pleasure would be ruined. The first because the surprise element, which makes the movie so interesting, would be lost; the second because who knows about the final twist throughout the movie will interpret all relevant scenes "correctly" and is deprived of the enjoyment of performing a "check-reading", i.e. to go back in memory to the relevant moments and reflect whether they are coherent with the new interpretation.

It might be argued that a viewer who already knows about the final twist could still enjoy *The Sixth Sense*, appreciating different aspects of the movie, such as the photography, the way it is directed, or the performance of the actors. Yet, the enjoyment the movie was intended to arouse would be lost, the viewer would not deal it as a story, because "when you have negotiated the intricacies of the plot—when you have experienced the surprises, made the discoveries, had your expectations verified—you have realized the intentions of the novelist [or the screenwriter] *qua* story-teller" (Kivy 2011: 7f.). Peter Kivy suggests that works of narrative fiction can be enjoyed only once if read by the reader as a story to be

told. Further readings could not provide the same kind of pleasure.[4] Accordingly, even though *The Sixth Sense* can be viewed focusing on scenography or photography, or viewed a second time for a "check-reading", it seems quite obvious that the main quality of the film is related to the enjoyment that results from the final plot-twist. Shyamalan's movie is a particularly effective example for analyzing spoilers in fictions since it masterfully shows the destructive effect it can have on the experience of a work of fiction.

Let me illustrate the role of spoilers in thought experiments with a short discussion of Thomson's *Dying Violinist*. In this famous thought experiment the philosopher asks you to imagine yourself waking up in bed next to a famous violinist who, as you learn right after your awakening, suffers from a kidney disease and risks dying. The *Society of Music Lovers* has kidnapped you because you have the same rare blood-type as the violinist and could, with your circulatory system pumping blood also through the violinist's body, save the life of the violinist. The hospital director concisely states: to save his life, you have to stay connected to his body for nine months. At this point, Thomson asks the reader: "is it morally incumbent on you to accede to this situation?" (1971: 49). What should you do if the situation will last not for months but for years, or for the rest of your life? It would certainly be a kind action to save the violinist, but no one seems to be morally constrained to stay in bed and ruin her existence for saving another person's life. This thought experiment invites the reader to imagine a fictional world designed to conceive, by analogy, the possible relationship between a mother and her fetus, and to understand some moral implications of abortion that could easily be underestimated or neglected.

In this paper, I am not interested in discussing this experiment's moral or political implications, but in another question: if a person, who reads through the text and in doing so conducts the thought experiment in her mind, knows already about the final twist beforehand, would that ruin her experience or the effectiveness of the thought experiment? I do not think so. Moreover, it seems that also Thomson would agree, since the paper in which the Violinist's case is presented is called *A Defense of Abortion*. It just seems that with this title Thomson wants to get the reader "straight to the point", without trying to hide the cognitive instances with which she elaborates her fictional story. Rather, she uses the "spoiler" to attract the readers' attention and arouse their curiosity.

This observation can be explained by the fact that every thought experiment is part of a broader argumentation or theoretical context—which we can call *foretaste context*—in which spoilers about the cognitive instances of the fictional work are presented and which, therefore, makes it easier for the reader to accept that a fictional narrative is embedded in a theoretical, scientific discussion.[5]

It is possible to introduce in this context instructions that guide the interpretation of the thought experiment and to communicate to the reader which aspect

[4] An exception to this point could be due to the viewer's bad memory. If over time she has forgotten the plot, a new experience of the same movie would realize again its story-telling intentions (Kivy 2011: 8).

[5] Note, however, that spoilers are not a necessary element of a foretaste context. It can be generalized as the ever-present contextualization that embeds all texts. It may be more or less complex, contain spoilers or not, be marginalized or significant in order to serve the purposes of the author. Thus, the aim of this paper is to analyze the role that spoilers can play within any foretaste context.

of the narrative is salient for the argument. Even when the thought experiment leads to an unwelcome conclusion,[6] knowing about it in advance does not ruin its effectiveness; on the contrary, spoilers can stimulate and guide a thought experiment's reading[7]—which marks an essential difference to other works of fiction.

The role of spoilers, considered here as a narrative and epistemic device, is useful to unveil a significant difference between the narrative style of works of literature or movies on the one hand, and thought experiments, on the other. Whether or not spoilers can be an effective device depends on how the content of the narrative is presented: showing spoilers concerning the plot—or its extreme consequences—can guide the viewer's attention, or block her emotional involvement and enjoyment, depending on the purposes for which a work is used.[8]

In this paper I suggest that it is in virtue of an effective foretaste context, in which we can read spoilers about claims and conclusions of a fictional state of affairs, that the author of a thought experiment succeeds in two aims: (i) justifying the presence of a fictional element within an essay with worldly-cognitive purposes, and (ii) guiding the reader's attention in order to put her in the condition to better understand which elements of that fiction are salient and which are not. The first claim becomes evident if we imagine the bewilderment of a reader who, while she's struggling with an essay about artificial intelligence, suddenly comes across a paragraph regarding Chinese idioms and people locked up in isolated rooms.[9] An out-of-the-blue employment of fictional stories is not so common in scientific essays, therefore an effective foretaste context is needed to make the reader accept the use of fiction in this kind of context.

The second point highlights the role that the instructions provided by the author play in the reader's interpretation of a thought experiment. Fictional texts often present multiple interpretative layers that result from different aspects of the work; clear-cut instructions can, thus, be helpful to put aside irrelevant interpretations—at least, the ones that are irrelevant for the author's cognitive purposes. If this strategy is applied, even fictional texts that were not designed as thought experiments can be illuminating. This requires us to put the aesthetic qualities aside and focus on those aspects that deepen understanding.

A good example of these two points is, again, Thomson's paper, in which the reader accepts the use of a short fictional story within a theoretical context because it is presented as a counterfactual situation about the main topic, thanks to the title and the style-formula "It sounds plausible [that a fetus could be a person from the moment of conception.] But now let me ask you to imagine this" (1971: 48). The reader, thus, is guided to weigh the elements presented in the fictional narrative: the fictitious fact of having been kidnapped by the *Society of Music Lovers* is hardly relevant, while the fictitious fact of staying in bed for exactly nine

[6] For example, a reader may be skeptical about Thomson's argument but surprisedly conclude, after carrying out her thought experiment, that the author has got a point.

[7] Just briefly think about how much interest the question "do you want to know about that thought experiment which shows how computers cannot think?" can prompt.

[8] An "above board" presentation of what cognitive instances are actually in play could enhance the strength and understanding of an argument better than a plot-twist. This does not mean we always have to ignore emotive responses when our claims are epistemic, nor that emotions play no role in the constitution of beliefs or in the advancement of understanding (see Elgin 2002, 2008).

[9] Fortunately, Searle's *Chinese Room* is well contextualized and does not confuse the reader—at least not for its being fictional.

months immediately stand out as salient. The spoiler in the title suggests that the thought experiment is presented with the aim to advocate women's right to choose whether to carry on a pregnancy or not—this allows the reader to read the story through the interpretative lens intended by Thomson and to neglect, or at least bracket, alternative interpretations.

Complementary to Elgin, who states that "[true descriptions can be] embedded in a work of fiction, a context in which an author is free to take liberties with truth in order to serve his aesthetic ends" (2007: 43), I think it is important to note that the opposite situation can also occur. Just as it is possible that fictional works may contain true statements within them—so it is possible that worldly-cognitive works may contain fictional elements, if fictionality helps the author to pursue her cognitive ends.

In addition, I suggest that, with an effective foretaste context, an author may be able to make a fictional story work like a thought experiment. This does not hold for any story, though: there must be a cognitively relevant and not overly ambiguous content emerging from the narration. In literary works of art of this kind, spoilers guide the reader's attention, and focus it on internal argumentation present in a literary work, on plausible causal chains of fictional events, on a thesis shown in the work, rather than focusing on the aesthetic pleasure that is aroused by it.

Note that it is still possible to perform a thought experiment without any previous spoiler about the plot and the cognitive instances involved. In these cases, if we cannot properly speak of "spoilers", we can still find interpretative instructions that guide the readers' attention. When the focus of the reader's attention is guided "at a later stage", she could—and probably will—do a "check-reading" of the thought experiment, just like in *The Sixth Sense* mentioned above, and check whether the interpretation is consistent with the fictitious state of affairs.

Thus, the use of spoilers is not a necessary condition for carrying out a thought experiment, even though it is an interesting device. Could spoilers nevertheless be a sufficient condition for a text to be a thought experiment? It depends on what they are used for. If their purpose is primarily epistemic, the work is used as a thought experiment. Spoilers, however, can serve different functions: some deep spoilers of *Pierre Menard* can be helpful in a creative writing course as a virtuous example. Thus, spoilers alone are not sufficient; they can fulfill different functions in the foretaste context of a work.

The point here is that a useful condition for using a fictional story with cognitive purposes is a theoretical context that puts aspects like entertainment and suspense aside. The author's instructions for the interpretation can be placed before or after the presentation of the fictional part, but, in order to render it more likely that the reader will accept the presence of a fictional story and avoid bewilderment, some spoilers can be helpful. Granted this, in the next section I will discuss in more detail an example of a short story that can work as thought experiment by virtue of an effective foretaste context.

3.3 Spoilers and Pierre Menard

In section 2.2 I mentioned Jorge Luis Borges' short story *Pierre Menard* as a good example of both a thought experiment and a work of literature. In the present section I will take a closer look at the work, analyzing how it presents itself and how it works in its original context. Finally, I will discuss the question of whether

an efficient foretaste context can "transform" *Pierre Menard*'s short story into a thought experiment.

The cognitive force of *Pierre Menard* lies in its illustration of how two texts that are identical word for word can be considered two distinct works of art, if we consider elements such as the social, historical and cultural contexts, as well as the intentions of the authors. It will be useful to have a look at the way the story is presented and, with all due respect to Derrida's claim,[10] the author's intentions: the text has the form of a short story and it has been published in a collection of short stories entitled *Ficciones*.

The foretaste context in which the work is presented contains no spoiler. It seems that the author did not want to force a single interpretation onto the reader. In fact, we only need to know a little bit about Borges to understand the plausibility of this point: throughout his *oeuvre*, Borges play on ambiguity so effectively that this clearly contributes to its outstanding literary and aesthetic value.

Nonetheless, the meta-narrative level in which *Pierre Menard* is developed makes it particularly versatile. If we focus on certain passages of the work rather than others,[11] we can easily find an ontological theory that is presented in an aesthetically and cognitively successful way. These considerations have led some philosophers to take up *Pierre Menard* as a thought experiment, quoting it with the appropriate foretaste context. Let me give you some examples that illustrate this point as well as the potential of the appropriate foretaste context.

Discussing the identity criteria of texts, Arthur Danto argues for the possibility that two works that are indiscernible are not identical with one another. With reference to Borges' work, he writes:

> The possibility was first recognized, I believe, in connection with literary works, by Borges, who has the glory of having discovered it in his masterpiece, *Pierre Menard, Symbolist Poet*. There he describes two fragments of works, one of which is part of *Don Quixote* by Cervantes, and the other, like it in every graphic respect—like it, indeed, as much as two copies of the fragment by Cervantes could be—which happens to be by Pierre Menard and not by Cervantes. […] the books are written at different times by different authors of different nationalities and literary intentions: these facts are not external one; they serve to characterize the work(s) and of course to individuate them for all their graphic indiscernibility. […] Borges' contribution to the ontology of art is stupendous: you cannot isolate these factors from the work since they penetrate, so to speak, the essence of the work (Danto 1981: 33-36).

Although Danto recalls the work with a different title, in this section we can see how he uses *Pierre Menard* as a thought experiment. He describes the text—without any attempt to avoid spoilers—and uses it as an integral part of his own argument. If someone who has never read Borges' work would come across this passage, probably all these spoilers and the argumentative context, in which it was

[10] The reference here is to the lucky slogan "there is no outside the text" (Derrida 1976: 158).

[11] In the first part of the work, there is a long and ironic list of the visible work left by Pierre Menard, to be contrasted with his invisible fragment, the re-writing of parts of the *Don Quixote*. An effective foretaste context will take these elements to the background and guide the reader's attention to another part of the story, where comparative analyses between Menard's and Cervantes' texts clearly illustrate the thesis at stake.

embedded, would tempt her to read it. It is worth noting that, in this case, the spoilers would not ruin the experience of the short story. On the contrary, they would add to its enjoyment: the ontological thesis shown in the story is the main point of its plot—at least, in this theoretical context—and there is no need for plot twists or unexpected surprises.

Even more explicitly, Peter Lamarque uses *Pierre Menard* as thought experiment in *The Philosophy of Literature*:

> Jorge Luis Borges's witty short story "Pierre Menard, Author of the *Quixote*" has come to epitomize, for philosophers, thought-experiments about works and texts, supposedly offering a powerful fictional exemplification of the view that distinct works can have identical texts. In the story, Menard, a fictional early-twentieth-century Symbolist poet, has the ambition to write *Don Quixote*, not by merely copying the original, but by a fully inspired act of literary creation. Here is a key, and often quoted, passage from the story:
>
> > It is a revelation to compare Menard's *Don Quixote* with Cervantes's. The latter, for example, wrote (part one, chapter nine):
> >
> > > …truth, whose mother is history, rival of time, depository of deeds, witness of the past, exemplar and adviser to the present, and the future's counsellor.
> >
> > Written in the seventeenth century, written by the "lay genius" Cervantes, this enumeration is a mere rhetorical praise of history. Menard, on the other hand, writes:
> >
> > > …truth, whose mother is history, rival of time, depository of deeds, witness of the past, exemplar and adviser to the present, and the future's counsellor.
> >
> > History, the *mother* of truth: the idea is astounding. Menard, a contemporary of William James, does not define history as an inquiry into reality but as its origin. Historical truth, for him, is not what has happened: it is what we judge to have happened. The final phrases—*exemplar and adviser to the present, and the future's counsellor*—are brazenly pragmatic.
> >
> > The contrast in style is also vivid. The archaic style of Menard—quite foreign after all—suffers from a certain affectation. Not so that of his forerunner, who handles with ease the current Spanish of his time.
>
> […] Whether or not Borges's story in itself provides adequate grounds for distinguishing work from text, it shows in effect the way that distinction could be maintained. A single text could be shared by two distinct works if certain conditions are in place: at the least, the works must have different properties and the texts must be produced by independent creative acts (Lamarque 2009: 74ff.).

Unlike Danto, Lamarque adds two elements: he not only discusses the cognitive value of the work, quoting its most significant part; he explicitly refers to Borges' text using the term 'thought experiment'. Lamarque's term is uncontroversial and does not hurt the reader's common sense: the short story fits well within a theoretical and argumentative horizon that justifies its presence.

Not all philosophers agree with the ontological thesis presented by Borges. Nelson Goodman and Catherine Elgin, for example, consider it with critical intent, when they discuss the criteria of identity of a text and its relationship with different interpretations (1988: 60ff.). This clearly indicates that Goodman and Elgin tacitly treat Borges' work as a thought experiment. After all, it is good prac-

tice that a philosopher, who does not agree with the conclusion of a thought experiment, challenges the fictional scenario or its interpretation.[12] For Goodman and Elgin, therefore, *Pierre Menard* is a thought experiment that can raise a serious objection to their own claim and, thus, calls for discussion.

Finally, George Bailey dedicates an entire paper to the ontological debate around the *Pierre Menard*'s case. He recognizes in Goodman and Danto the main argumentative poles, arguing that Borges' story should be considered as a valid contribution to the ontology of artworks (Bailey 1990: 340). This shows the importance of Borges' short story, which has come to stand for a specific philosophical position within a prolific debate in contemporary aesthetics.

It is important to note, however, that Borges does not use a spoiler in the story's foretaste context to guide the reader towards a specific ontological conclusion. It rather seems plausible that he has played with the absence of any interpretative line, with the aim to prompt astonishment in the reader who, reading *Pierre Menard*, might wonder whether what she has just read is nothing but brilliant nonsense or whether, perhaps, Borges has got a point. The philosophers considered above certainly accept the second possibility and use *Pierre Menard* in a context which, at the cost of unveiling the short story's plot, gives more prominence to the thesis it can be taken to prove.

This shows a central point. If we put *Pierre Menard, Author of the Quixote* in a theoretical context which explicitly states the main point of its plot, the short story does not appear flawed. We would consider the argument to be properly developed—and might be surprised by how well it was written, especially when compared to other thought experiments.

Thus, Borges' work can be considered both a literary work of art and a thought experiment, depending on the context in which it is presented and used. If we grant the lesson that Borges seems to illustrate in his short story, perhaps we could argue that there are two distinct works—a literary work of art and a thought experiment—that are identical word for word, but differ in context, purpose, style and have to be interpreted in different ways. The whole thing about anticipation and explanation of a plot's main points therefore seems to be a constancy, and also an important literary device, in guiding the modes of access, interpretation and identification with which a reader approaches a fictional text in cognitive contexts. Although it has previously been argued that *Pierre Menard* is both an outstanding work of literature and an illuminating thought experiment, the "transformation" is not immediately established and we have to add some spoilers, as well as to change the context of presentation of the short story, i.e. we have to change what I have called *foretaste context*, to make it really work as a thought experiment.

It should be noted, however, that Borges' short story is a most effective example: it is particularly apt to become a thought experiment, since it is short enough and the main point of its plot can be easily recognized as a substantial thesis concerning the ontology of artworks. It is more common to find larger and more heterogeneous works of literature that cannot, if considered in their whole, act as thought experiments even though some of them might contain suitable passages that do. Even in these cases, the foretaste context has the task of guiding the

[12] In this context it can be interesting to recall that John Searle presents his thought experiment of the *Chinese Room Argument* in connection with a discussion of several replies that have been raised against it (1980: 419-24).

reader to accept the author's choice of a certain part of the text, and to focus her attention in order to bring out the cognitive instances that are so fundamental in thought experimentation, but that, without explicit instructions, can also serve the function to add to the pleasure of reading.[13]

References

Bailey, G. 1990, "Pierre Menard's Don Quixote", *The Jerusalem Philosophical Quarterly*, 39, 339-57.

Borges, J. 1999, *Collected Fictions*, London: Penguin Books.

Brown, J.R. 1991, "Thought Experiments: A Platonic Account", in Massery, G.J. and Horowitz, T. (eds.), *Thought Experiments in Science and Philosophy*, Maryland: Rowman and Littlefield, 119-28.

Brown, J.R. 2004, "Peeking into Plato's Heaven", *Philosophy of Science*, 71, 1126-38.

Carroll, N. 2002, "The Wheel of Virtue: Art, Literature and Moral Knowledge", *Journal of Aesthetic and Art Criticism*, 60, 3-26.

Cooper, R. 2005, "Thought Experiments", in *Metaphilosophy*, 36, 328-47.

Currie, G. 1990, *The Nature of Fiction*. Cambridge: Cambridge University Press.

Danto, A. 1981, *The Transfiguration of the Commonplace. A philosophy of Art*, Cambridge, MA: Harvard University Press.

Davies, D. 2007, "Thought Experiments and Fictional Narratives", *Croatian Journal of Philosophy*, 7, 29-45.

Davies, D. 2018, "Art and Thought Experiments", in Stuart M.T., Fehige, Y. and Brown, J.R. (eds.), *The Routledge Companion to Thought Experiments*, London and New York: Routledge, 512-25.

Derrida, J. 1976, *On Grammatology*, Baltimore: John Hopkins University Press.

Egan, D. 2016, "Literature and Thought Experiments", *The Journal of Aesthetics and Art Criticism*, 74, 139-50.

Elgin, C. 1996, *Considered Judgement*, Princeton: Princeton University Press.

Elgin, C. 2002, "Art in the Advancement of Understanding", *American Philosophical Quarterly*, 39, 1-12.

Elgin, C. 2007, "The Laboratory of the Mind", in Huemer, W., Gibson, J. and Pocci, L. (eds.), *A Sense of the World. Essays on Fiction, Narrative, and Knowledge*, New York: Routledge, 43-54.

Elgin, C. 2008, "Emotion and Understanding", in Brun, G., Dogouglu, U. and Kunzle, D. (eds.), *Epistemology and Emotions*, London: Ashgate, 33-49.

Elgin, C. 2014, "Fiction as Thought Experiment", *Perspectives on Science*, 22, 221-41.

Elgin, C. 2017, *True Enough*, Cambridge: MIT Press.

Elgin, C. 2019, "Imaginative Investigations: Thought Experiments in Science, Philosophy and Literature", in Bornmüller, F., Franzen, J. and Lessau, M. (eds.), *Literature as Thought Experiment? Perspectives from Philosophy and Literary Studies*, Paderborn: Fink, 1-16.

[13] I am deeply grateful to Wolfgang Huemer for his helpful suggestions on earlier versions of this paper.

Friend, S. 2017, "The Real Foundation of Fictional Worlds", *Australasian Journal of Philosophy*, 95, 29-42.

Gendler, T.S. 2010, *Intuition, Imagination and Philosophical Methodology*, Oxford: Oxford University Press.

Goodman, N. and Elgin, C. 1988, *Reconceptions in Philosophy and other Arts and Sciences*, Cambridge: Hackett.

Hume, D. 1999, *An Enquiry Concerning Human Understanding*, Oxford: Oxford University Press.

Johnson-Laird, P. 1983, *Mental Models*, Cambridge: Cambridge University Press.

Kivy, P. 2011, *Once-Told Tales. An Essay in Literary Aesthetics*, Hoboken: Wiley-Blackwell.

Kuhn, T. 1977, "A Function for Thought Experiments", in *The Essential Tension*, Chicago: University of Chicago Press, 240-65.

Lamarque, P. 2009, *The Philosophy of Literature*, Oxford: Blackwell.

Lewis, D. 1986, *On the Plurality of Worlds*, Oxford: Blackwell.

Mach, E. 1976, "On Thought Experiments", in McGuinness, B. (ed.), *Knowledge and Error*, Dordrecht-Boston: Reidel, 134-47.

Meynell, L. 2018, "Images and Imagination in Thought Experiments", in Stuart, M.T., Fehige, Y. and Brown, J.R. (eds.), *The Routledge Companion to Thought Experiments*, London and New York: Routledge, 498-511.

Miščević, N. 1992, "Mental Models and Thought Experiments", *International Studies in the Philosophy of Science*, 6, 215-26.

Miščević, N. 2007, "Modelling Intuitions and Thought Experiments", *Croatian Journal of Philosophy* , 7, 181-214.

Nersessian, N.J. 1991, "Why Do Thought Experiments work?", *Proceedings of the Cognitive Science Society* ,13, 480-86.

Nersessian, N.J. 1992, "In the Theoretician's Laboratory: Thought Experimenting as Mental Modeling", *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2, 291-301.

Nersessian, N.J. 1999, "Model-Based Reasoning in Conceptual Change", in Magnani, L., Nersessian, N.J. and Thagard P. (eds.), *Model-Based Reasoning in Scientific Discovery*, New York: Kluwer Academic and Plenum, 5-22.

Nersessian, N.J. 2018, "Cognitive Science, Mental Modeling, and Thought Experiments", in Stuart, M.T., Fehige, Y. and Brown, J.R. (eds.), *The Routledge Companion to Thought Experiments*, London and New York: Routledge, 309-26.

Norton, J.D. 1996, "Are Thought Experiments Just What You Thought?", *Canadian Journal of Philosophy*, 26, 333-66.

Norton, J.D. 2004, "Why Thought Experiments Do Not Transcend Empiricism", in Hitchcock, C. (ed.), *Contemporary Debates in the Philosophy of Science*, Oxford: Blackwell, 44-66.

Salis, F. and Frigg, R. 2020, "Capturing the Scientific Imagination", in Levy, A. and Godfrey-Smith, P. (eds.), *The Scientific Imagination*, Oxford: Oxford University Press, 17-50.

Searle, J. 1980, "Minds, Brains and Programs", *Behavioral and Brain Sciences*, 3, 417-57.

Thomson, J. 1971, "A Defense of Abortion", *Philosophy & Public Affairs*, 1, 47-66.

Walton, K. 1990, *Mimesis as Make-Believe: On the Foundations of the Representational Arts*, Cambridge, MA: Harvard University Press.

Willée, A. 2019, "Thought Experiments as a Narrative Genre", in Bornmüller, F., Franzen, J. and Lessau, M. (eds.), *Literature as Thought Experiment? Perspectives from Philosophy and Literary Studies*, Paderborn: Fink, 83-96.

Wittgenstein, L. 2009, *Philosophical Investigations*. Oxford: Blackwell.

From Fictional Disagreements to Thought Experiments

Louis Rouillé

Institut Jean Nicod

Abstract

In this paper, I present a conceptual connection between fictional disagreements and thought experiments. Fictional disagreements happen when two readers disagree about a fictional detail. The "great beetle debate" is a paradigmatic case. Nabokov once argued that Gregor Samsa, in *The Metamorphosis*, metamorphosed into a beetle. Yet many critics and readers imagine Gregor to be a big cockroach. Analysing a fictional disagreement is interesting because it exhibits the informational structure which is common to all fictions. First, it shows the distinction between the fictional foreground (what is expressed by the narrator) and background (what the reader automatically infers from the narration). Second, it shows how the fictional background is filled with the reader's representations of reality and other shared conventional representations. The fictional background is a sophisticated mixture of traceable fictional and non-fictional bits of information. I argue that one can use this complex informational structure to explain how it is possible to extract new information originating in fiction for non-fictional purposes. The possibility of "learning from fiction" has led to a long-standing philosophical debate. However, everyone agrees on the possibility of extracting fictional information: this corresponds to drawing a moral from a given fiction. This possibility is, I argue, analogous to performing a thought experiment. I show that thought experiments and fictional disagreements exploit the same informational structure. Instead of filling the fictional background, one informs one's non-fictional representations using the same informational channels in reverse direction.

Keywords: Truth in fiction, Fictional Disagreement, Learning from fiction, Philosophy of literature

1. Introduction

"Truth in fiction" has become a well-known problem for those who are interested in the semantics of fictional discourse, both literary theorists and philosophers.[1] However, philosophers have become interested in a notion of fictional truth which

[1] In this article, I will focus on fictional texts, though everything I say should apply to other media as well.

Argumenta 6, 1 (2020): 99-116
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.rou

is most of the time uninteresting for literary theorists. Indeed, philosophers have been puzzled by the fact that readers automatically infer trivial fictional truths which are not explicitly stated. In reading, say, Shakespeare's *Hamlet*, one automatically infers that Hamlet is a human being, that he has two lungs and a liver, though nothing is explicitly said about this. Consequently, virtually every philosopher of fiction agrees that the fictional truths well exceeds what is explicitly in the text.[2]

This led philosophers of fiction to distinguish between the fictional foreground and background. To use Walton 1990's terminology, the fictional foreground contains the *primary* fictional truths, which are derived from the fictional text only; while the fictional background contains the *secondary* fictional truths, which are derived using some primary truths. The primary fictional truths are not necessarily the propositions explicitly expressed in the text, for there can be unreliable narrators. When the narrator is reliable, though, the fictional foreground coincides with what is explicitly narrated.

There are several competing frameworks designed to model how the fictional foreground and background are constructed in the mind of the reader. Walton's influential account explains how the reader imagines the foreground using the notion of "props in games of make-believe", and how the background is filled by using general "principles of generation". The two mechanisms are nicely integrated into a single abstract model which is now widely taken as a basis for further investigation on the notion of fictional truth. My present contribution will be within this general framework.

In this paper, I discuss "fictional disagreements". *Fictional disagreements* are controversies about how to fill the background of a story. They typically happen when two readers disagree about some detail of a fiction. One paradigmatic fictional disagreement called the "great beetle debate" was recently unearthed in Friend 2011. On the basis of this case study, I will claim that thought experiments exploit the same information channels as those needed for fictional disagreements.

2. The Great Beetle Debate

2.1 Nabokov's Argument

From his arrival in the United States in 1940 until the success of *Lolita* in 1955, Vladimir Nabokov taught foreign literature at Cornell University. In a lecture posthumously published in Nabokov 1980, Nabokov offered an original literary interpretation of Kafka's *Metamorphosis*. He makes a great deal of what can be thought of as a fictional detail, speculating about the kind of insect Gregor Samsa has turned into.

Everyone agrees that Gregor has turned into an insect. This is not explicitly stated though.[3] Kafka gives only a vague description of Gregor's physical appearance after the metamorphosis. The most precise description of Gregor is to be found in the opening sentence of *The Metamorphosis*:

[2] In D'Alessandro 2016, one can find a defence of "explicitism" about fictional truth. He is, to my knowledge, the only dissonant voice in the philosophical community.

[3] This is true of the original in German, though not always in the various English translations.

> One morning, upon awakening from agitated dreams, Gregor Samsa found himself, in his bed, transformed into a monstrous vermin.[4]

The term "vermin" does not immediately indicate that Gregor is an insect, for this word can be used literally to denote other animals like rodents or metaphorically to denote despicable human beings.[5] However, given the fictional foreground, it is clear that Gregor has turned into an insect, for it is explicitly said that Gregor can walk on walls, and the food he eats also indicates that he is an insect.

Nabokov then considers the question: what insect?

> Commentators say *cockroach*, which of course does not make sense. A cockroach is an insect that is flat in shape with large legs, and Gregor is anything but flat: he is convex on both sides, belly and back, and his legs are small. [...] he has a tremendous convex belly divided into segments and a hard rounded back suggestive of wing cases. [...] In the original German text the old charwoman calls him *Mistkafer*, a "dung beetle." It is obvious that the good woman is adding the epithet only to be friendly. He is not, technically, a dung beetle. He is merely a big beetle (Nabokov 1980: 258-59).

In fact, Nabokov has an entomological argument. The reader, in the opening scene of the story, is required to imagine that Gregor is stuck on his/its back. As it happens, cockroaches do not get stuck on their backs but beetles do. Here is Nabokov's argument made explicit:

- Gregor is stuck on his back in the opening scene of *The Metamorphosis*.
- Cockroaches do not get stuck when they are put on their back (because they are flat and have long legs).
- On the contrary, it is typical of beetles to get stuck on their back.
- Therefore, in *The Metamorphosis*, Gregor is a beetle (and not a cockroach).

Finally, Nabokov famously drew how he imagines Gregor to be (see Figure 1).[6]

2.2 Debating Nabokov's Argument

Despite Nabokov's asserting tone, his argument can be questioned. First, one might wonder whether there is a fact of the matter (so to speak) about the ultimate nature of Gregor's insecthood. After all, Kafka himself did not commit on any

[4] This is Joachim Neugroschel's translation (Kafka 1915: 90). Here is the Kafka's original wording in German: "Als Gregor Samsa eines Morgens aus unruhigen Träumen erwachte, fand er sich in seinem Bett zu einem ungeheueren Ungeziefer verwandelt."

[5] As in: "The vermin who looted houses after the hurricane" (Merriam-Webster online).

[6] Nabokov was a semi-professional lepidopterist. Yet, Nabokov's argument should be made more precise to achieve a natural science standard of rigour. Indeed, entomologists distinguish between about 4,000 species of cockroaches and more that 250,000 beetle species (see Capinera 2008: 437, 938). Many cockroaches species do not match Nabokov's description at all, like the Oriental (*Blatta orientalis*) or the Florida woods (*Eurycotis floridana*) cockroach. Nabokov's argument is plausible only if one he meant to talk about German (*Blatella germanica*) or American cockroach (*Periplaneta americana*). (See figure 70 of Capinera 2008 reproduced at the end of this paper for photographs of these cockroach species.) As for what Nabokov had in mind when he says "beetle", given his drawings, it corresponds to a familiar species of the Scarabaeidae, probably something like the common brown beetle. Thanks to an anonymous referee for pressing me to make this point.

specific kind of insect. And it seems uncontroversial that the answer to such a question has no bearing over the comprehension of the story. So why not leave Gregor's insecthood undetermined? One reader would imagine him/it as a big cockroach, another as a beetle, a third as a bedbug, etc. Nabokov's argument, so the objection goes, is idle.[7]

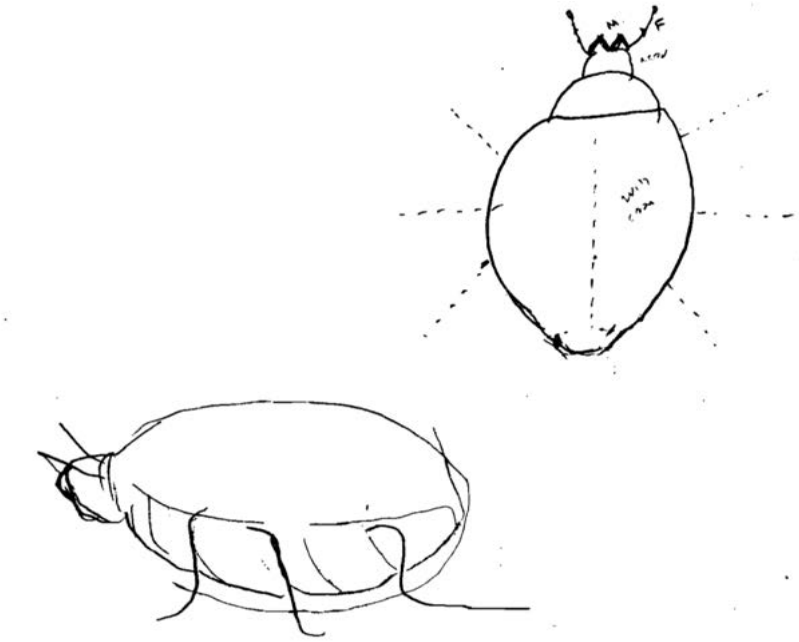


Figure 1: Nabokov's hand-drawing of Gregor (Nabokov 1980: 259)

From a literary viewpoint, this questioning the very relevance of Nabokov's question can be used to reject his literary interpretation of Kafka's story, which is essentially grounded in this detail. However, from a philosophical viewpoint, there is no reason to think that Gregor's insecthood is indeterminate in his world, even though we shall never know. Indeed, when it comes to fictional background, one should distinguish between a descriptive and a normative claim. One thing is what readers do in fact imagine, another is what they should imagine on the basis of a correct reading of the text. Fictional truth, by definition, is on the normative side. A proposition is fictionally true iff there is a *prescription* to imagine it (Walton 1990: 39). So Nabokov argument makes sense even if he was the only reader on earth to care about this bit of fictional truth.

[7] Actually, it can be argued that Kafka remained *purposefully* vague on this point since he wrote to his publishing house, talking about the cover of his book: "The insect itself is not to be drawn. It is not even to be seen from a distance". His wish was fulfilled in the original edition of 1915. Nowadays, however, it is not so rare to see a pictorial representation of Gregor on the front page...

That being said, it is indeed useful to distinguish between several kinds of prescriptions to imagine, so as to have a finer-grained view of fictional truth as in Friend 2017b:

> Although prescriptions to imagine are sometimes associated with mandates, we need not imagine everything that is fictional. If we want to understand a work, some kinds of imagining are required. One could not grasp the basic plot of *Gulliver's Travels* without imagining Gulliver travelling to Lilliput, Brobdingnag, and so forth. A fuller appreciation demands recognizing how mistaken Gulliver is about himself (something children often miss). Still, even a fuller appreciation does not require imagining that Gulliver has internal organs, though it is surely fictional that he does. It is helpful to distinguish these obligations. I will say that a work *mandates* imagining that P if failure to imagine that P would mean falling below a minimum threshold for comprehension. A work *prescribes* imagining that P if we should imagine that P to have a fuller appreciation of the story. Finally, a work *invites* imagining that P on the following condition: if the question arises and we must choose between imagining that P and imagining that not-P, we are required to imagine the former. What is fictional in a work is what the work invites imagining. Although we need never imagine that Gulliver has internal organs, if the question came up it would be absurd to deny that he does (Friend 2017b: 2).

That Gregor fictionally is a beetle (assuming he is) is clearly not a mandate. Nabokov claims it is a prescription but most readers, I think, would rather take it as an invitation. The first objection points to this debate.

A second more radical objection consists in denying that Nabokov's argument is correct. Supposing that it is valid, it must be an enthymeme. The hidden premise is that cockroaches and beetles behave in a similar fashion in the real world and in Gregor's world. In other word, real entomological facts carry over into the background of *The Metamorphosis*. It is not clear that one should accept this premise. Questioning this premise will put us at the center of long-standing debates in the philosophy of fiction about the so-called "reality principle".

The reality principle is a particularly efficient way of filling the fictional background of a story as remarked in Walton 1990:

> The basic strategy which the Reality Principle attempts to codify is that of *making fictional worlds as much like the real one as the core of primary fictional truths permits*. It is because people in the real world have blood in their veins, births, and backsides that fictional characters are presumed to possess these attributes (Walton 1990: 145).

Although virtually every philosopher of fiction agrees with "the basic strategy", the specifics of this principle are much debated.[8] If one wants to apply a reality principle so as to get the premise Nabokov needs, one would say something like: Since it is true that cockroaches do not get stuck on their back (and beetles do), it must be true in *The Metamorphosis* that cockroaches do not get stuck on their back (and beetles do).

[8] See Lewis 1978 for an interpretation of it within a possible-world semantic framework. See Everett 2013: 23 for a discussion of two principles called "Incorporation" and "Reality". See Friend 2017b for a criticism of the reality principle and a defence of her "reality assumption".

However, there are several cases in which the reality principle must give way. First and foremost, what is imported to fill the fictional background must be compatible with the fictional foreground. For instance, that Gregor as turned into a monstrous insect is a primary truth. So any real fact incompatible with such a metamorphosis should not be imported into Gregor's world, on the pain of inconsistency. However, selecting the relevant real facts to be imported in Gregor's world is not an easy task. For instance, how much entomology should one bring in a world where humans can turn into insects the size of a big dog? (Gregor's size can be derived from the fact that he opens the door standing on his/its back legs.) Actually, there are good reasons not to bring too much. Indeed, as shown in (Haldane 1926: 3), an enormous insect would have difficulty breathing if we follow entomology to the letter. Consequently, contrary to what Nabokov thought, it is debatable whether one can use the reality principle to derive the hidden premise.

There are two other cases in which the reality principle must give way (they will be interesting later on). One is when it is not reality but ideology that is used to fill the fictional background.[9] In some cases indeed, what the author and readers commonly believe is more relevant than reality itself. For instance, if a story originates in a community where it is commonly believed that the earth is flat, then it is widely acknowledged that in the fiction the earth is flat, even if the fictional foreground does not require it. The other case is when the reader is expected to fill the fictional background using some prior knowledge of shared conventions. For instance, there is a convention according to which dragons breath fire. So if there is a dragon in a story, one automatically infers that this dragon breaths fire even though it may not be explicitly said so. Since there is no dragon in reality, this fictional truth cannot come from using a reality principle.

2.3 Truth in Fiction and Interpretations

It is important to emphasise that a fictional disagreement is a disagreement about the interpretation of a fiction in a very specific sense of "interpretation". In (Friend 2017a: 388), three kinds of interpretative activities are carefully distinguished. One is *elucidation*: it consists in making explicit what is merely implicit in the fiction. The second is *explication*: it consists in "ascertaining the meanings and connotations of words, or passages" of a fiction. The third is *thematic interpretation*: it consists in "identifying the themes and theses in the work as a whole".[10]

Here is the precise definition of *elucidation* in (Friend 2017a: 388-89):

> To elucidate a work is to determine what is going on in the storyworld, what is "true in the story"—or as I prefer, *storified*—where this is not specified by the explicit text and may even contradict it (as with unreliable narrators).

Given what Friend later says, we can distinguish between "trivial" and "substantive" elucidations. *Trivial* elucidation is a case where what is elucidated has no bearing upon other kinds of interpretative activities. For instance, elucidating the blood type of Hamlet is trivial in this sense. (Literary critics are usually not interested in trivial elucidation, although philosophers love it.) *Substantive* elucidation, by contrast, is a case where what is elucidated has crucial consequences which

[9] Such cases motivate the shift from (Analysis 1) to (Analysis 2) in Lewis 1978.
[10] Friend's terminology is adapted from Beardsley 1958.

feed into the other interpretative activities. For instance, elucidating whether Hamlet is mentally ill is substantial in this sense. (Philosophers tend to avoid these complicated cases, whereas literary critics look for them.)

In practice, substantive elucidations touching upon crucial aspects of a narrative can hardly be distinguished from the other kinds of interpretation, as noted in Friend 2017a:

> Many puzzling works demand efforts at elucidation. Anyone who fails to wonder who Godot is and why Vladimir and Estragon awaits him, or who is unperturbed by Bartleby's intransigence has simply not engaged with the relevant works (Friend 2017a: 389).
> [Footnote: As these cases indicate, elucidation cannot always sharply be distinguished from other dimensions of interpretation or criticism.]

In such cases, the substantive elucidation is very likely to be controversy-ridden. For instance, two literary critics can disagree on whether Godot fictionally is God, and this would surely affect the thematic interpretation of Beckett's play. However, there is no reason to think that disagreements over an elucidation happen only when it is substantive. Some fictional disagreements focus on trivial elucidations. Two readers can passionately disagree on Hamlet's eye-color and this would not affect the other kinds of interpretation.

Nabokov's critical genius consists in grounding a thematic interpretation on a case of elucidation which is *prima facie* not substantial. Despite his ingenious rhetoric, I suggested that the great beetle debate might very well be trivial. If that is the case, then the great beetle debate gets stuck at the level of elucidation.

3. Conditions of Possibility of Fictional Debates

I can now generalise: a *fictional disagreement* is a disagreement about the (possibly trivial) elucidation of a background fictional truth. Note that two readers can also disagree on how to fill the fictional foreground, for example in the case of a subtle unreliable narration. These do not count as fictional disagreements in my sense, though. I will thus set aside such cases to focus on the fictional background.

In this section, I claim that fictional disagreements come with a necessary condition, namely that there are open information channels. I will first explain what I mean by "information channel". I will then show that my claim makes adequate empirical predictions for when there is no available information channel, there can be no fictional disagreement.

3.1 General Picture of the Information Flow in the Fictional Background

In order to fill the fictional background, the reader first needs a fictional foreground, which I will take as a given. Moreover, we have seen that they need some information coming from the outside of the fiction, and a general mechanism to combine this outside information with the fictional foreground.

Let us now focus on the needed outside information. It must be one of two things: either it originates in reality or it does not. When I say that some information originates in reality, I mean the reader's representation of reality, be it knowledge or belief. For expository purposes, I will call this information originating in one's (accurate or not) representation of reality "factual". The fictional

background can thus be filled with facts. I say that the facts are imported into the fictional background through an information channel which links the fictional world and the real world.

If the outside information is not factual, then it is fictional. For instance, when one fills the background of a dragon story with "this dragon breaths fire", one imports information from genre conventions which originate in some seminal fiction (or perhaps in myths). However, one can also use "local conventions", so to speak. For instance, a leitmotiv in movies or opera is like a convention operating at the level of the fiction itself: each time you hear a tune, you are expected to fill the background with "such character is around". So the outside fictional information can either come from a different fiction or from the fiction itself. For expository purposes, I will call both kind of information "conventional". The fictional background can thus be filled with conventions. I say that the conventions are imported into the fictional background through an information channel which links the fictional world to another fictional world or to itself.

My picture is not very controversial, since it is a tidying up of the mainstream view as, for instance, described in Lewis 1978:

> I have said that truth in fiction is the joint product of two sources: the explicit content of the fiction, and a background consisting either of the facts about our world (Analysis 1) or of the beliefs overt in the community of origin (Analysis 2). Perhaps there is a third source which also contributes: carry-over from other truth in fiction. There are two cases: intra-fictional and inter-fictional (Lewis 1978: 45).

(Analysis 1) corresponds to a factual channel linking the fictional background to the reader's knowledge, while (Analysis 2) links it to the reader's beliefs about reality. "Intra-fictional" corresponds to a conventional channel linking the fictional background to the fiction itself, while "inter-fictional" corresponds to a link to a different fiction.

My claim is that fictional disagreements are controversies about how much an information channel should be open, hence they presuppose that the relevant informational channel is open. In the great beetle debate, the relevant information channel is factual. The debate boils down to whether one should or should not open an information channel so as to fill the fictional background with fine-grained entomological facts, as Nabokov suggest we should.

My claim entails that where there is no information channel available, there can be no fictional disagreement. I think this prediction is empirically accurate, as I will presently show.

3.2 Fictional Background of *The Nose*

In 1836, Gogol published *The Nose*. The main character of the short story is the Collegiate Assessor Kovalyov who, one morning, wakes up to find his nose missing. He becomes literally nose-less. Alarmed, looking for his nose all around the city of St Petersburg, Major Kovalyov runs into his nose in the street, dressed in the uniform of a higher-ranking official than himself. Suddenly, the nose enters a church. Stunned Major Kovalyov follows him in. Inside the church, the two characters exchange a few words. Finally, the nose sends Kovalyov packing using his higher-ranking authority. Major Kovalyov then tries to start legal proceedings against his nose without success. The story unfolds with other interesting twists and turns.

Let us focus on a detail of the story. One of the fictional event is the follow ng: nose-less Major Kovalyov is walking in the street and runs into his nose walking in the same street, dressed in the uniform of a high-ranking official. So the reader is explicitly asked to imagine the Nose *walking* down a street. But how would it do that? Does the Nose have legs? There are good reasons to think that it does not have legs, because it is a nose: but at the same time, if it is *walking* down a street, it probably has legs. I suspect there are different natural ways of getting around this difficulty, now I have raised the question: one would imagine the Nose, dressed, levitating ahead; or one would imagine a big nose bumping forward; or even a nose with long thin legs walking like a crane; or a very human-like creature whose face consists only of a nose.[11] There are probably other ways of realising this invitation to imagine the Nose walking down the street.

Gogol's *Nose* and Kafka's *Metamorphosis* are actually very similar but they crucially differ in that there can be no "great nose debate". First, most readers enjoy the works without a clue: if the questions did not arise, these elucidations would have remained in the background where they belong.

Second, both fictional events are just as impossible and implausible as can be. One may have the feeling that the two fictional events differ in their plausibility or possibility. But it is an irrational feeling. Both turning into a monstrous insect or losing one's nose while sleeping are impossible, implausible events. So any theory of plausibility or possibility which makes a significant difference between Gregor's fate and Kovalyov's lot will be deeply counter-intuitive to say the least.

Third, both stories require imagining that a supernatural event occurred in a world where the "laws of reality" are as we experience them otherwise (as such, both are *fantastic* stories as defined, *inter alia*, in Todorov 1970). Indeed, in Kovalyov's world, everything seems to be normal except for a nose on the loose. For instance, when Kovalyov tries to start legal procedures against his nose, the office clerk follows the same procedures as in reality and since there is no possibility of charging one's body part in reality, he cannot follow on from Kovalyov's demand. This prompts an incident in the office, since Kovalyov insists and gets angry at the office clerk.

Consequently, the two stories can systematically be compared, as in Erlich 1956:

> Clearly, Gogol's nonsense narrative lacks the quality of an existential disaster. Yet it shares with the grimmer story of Kafka the discrepancy between its "realistic" mode of presentation and the utterly incredible central event (Erlich 1956: 102).

The relevant natural science for the studying of walks and gaits is a branch of anatomy called functional anatomy. Functional anatomy studies the relationship between anatomy and movement, thus establishing facts about the constraints of the skeleton on possible movements, how the muscles of the body are to cooperate to produce a given movement, how the balance of the body coincides

[11] There are numerous illustrations and adaptations of Gogol's story for visual media. I noticed that this last option is very widespread (see for instance Alexandre Alexeieff's and Claire Parker's animated film from 1963). But the imagination of illustrators knows little limit, as a quick search on the internet reveals!

with movement, etc.[12] Unfortunately, functional anatomy says nothing about individual organs. A nose is not in any straightforward sense a possible object of functional anatomy.

Now, suppose there is a real natural science whose object of study is the walk, or rather the way of locomotion, of organs separated from the body. Let us call this science *schismatic functional anatomy*. This natural science would describe the way livers move on their own when separated from their body of origin, the locomotion of lungs, single arms, brains and pairs of eyes, and so on. Naturally, schismatic functional anatomy would have something to say about the walks of noses; the different walks available for separated noses would probably depend on their size: snub noses do not move in the same way as big noses do; probably it would depend on the species they are separated from: dog noses tend to be quadruped whereas human noses tend to be bipeds (these are only statistical facts); and many other factors. As in all natural sciences, controversies and discoveries are part and parcel of the positive knowledge it delivers; the history of schismatic functional anatomy is also quite a thing, since most of the scientists were born Russian, a coincidence which is still an open area of research.

If schismatic functional anatomy were a natural science, one could open a factual information channel and construct an argument, analogous to Nabokov's, to fill the fictional background of the Nose's walk. So there would be a "great Nose debate" just like there is a great beetle debate. However, schismatic functional anatomy is, as far as I know, merely a pseudo-natural science; hence, this is merely a pseudo-"great Nose debate".

In the same manner, one can find no conventional information channel available to fill the fictional background of the Nose's walk. Indeed, I have never heard of a convention which specifies how noses progress when severed from their body of origin.

This completes what can be thought of as a *reductio* argument. Suppose, there can be a fictional disagreement about the way the Nose walks analogous to the great beetle debate. Then, schismatic functional anatomy would be a natural science. But schismatic functional anatomy is manifestly not a natural science. Hence, there is no fictional disagreement. There can be no fictional disagreement where there is no available informational channel.

4. From Fictional Disagreements to Thought Experiments

It is not clear that the expression "thought experiment" denotes a unified set of phenomena as is shown in Stuart, Fehige and Brown 2017.[13] However, virtually everyone acknowledges that thought experiments and fictions share characteristic features (see Davies 2007 for a detailed analysis of this claim); many even argue that they are essentially similar (in particular, see Elgin 2007).

From the philosophy of fiction viewpoint, the putting together thought experiments and fiction serves a precise purpose in a now longstanding debate between cognitivism and anti-cognitivism. One of the main arguments in favour of the cognitive value of fictions is built on this widely accepted closeness between

[12] This is a very difficult and fascinating natural science which is currently challenged in its results by the advance in robotics. Making robots which can walk is a surprisingly difficult task; especially if one wants to design robots which can walk *like human beings*.
[13] See also Gendler 2016: 25 for an insightful tripartite view on thought experiments.

fictions and thought experiments (Davies 2017: 512, premise (2)). The basic idea is that if thought experiments have cognitive value, then so have fictions. Anti-cognitivists usually try to find principled reasons to distinguish literary fictions from thought experiments when it comes to cognitive value (see for instance Lamarque and Olsen 1994, arguing at length that fictional "truth" is not a kind of truth). Interestingly, both cognitivists and anti-cognitivists agree on the fact that some information originating in fiction can travel out of the fiction. What they disagree about is the cognitive value of such information: cognitivists argue that it can be knowledge under suitable conditions, while anti-cognitivists deny this.

In this section, my aim is to discuss this phenomenon both cognitivists and anti-cognitivists agree upon and I intend to remain neutral on whether the information extracted from a fiction can qualify as knowledge or not.[14] Very often, people justify their claims about "moral, psychological and social" facts by quoting relevant fictions (Carroll 2002: 3). I will leave aside the question whether people *should* avoid doing this if they want to be rational.

I think I should emphasise the scope and the limit of the phenomenon I aim to analyse. It consists in taking up a "crucial question unanswered" raised in Searle 1975, namely that:

> serious (i.e., nonfictional) speech acts can be conveyed by fictional texts, even though the conveyed speech act is not represented in the text. Almost any important work of fiction conveys a "message" or "messages" which are conveyed by the text but are not in the text. [...] Literary critics have explained on an ad hoc and particularistic basis how the author conveys a serious speech act through the performance of the pretended speech acts which constitute the work of fiction, but there is as yet no general theory of the mechanisms by which such serious illocutionary intentions are conveyed by pretended illocutions (Searle 1975: 332).

Drawing a moral is indeed a very familiar phenomenon, and in some cases, like with fables or satirical stories, the reader is actually expected to do so. In such cases, it is widely acknowledged that fictions function as thought experiments.

One might question the analogy between the moral drawing activity and thought experiments, despite a consensus among philosophers of fiction. For instance, thought experiments seem to be conceptually linked to the notion of *possibility* in a way fictions are not.[15] Taking Putnam's Twin Earth thought experiment as a paradigmatic example: one can reject Putnam's defence of semantic externalism using this thought experiment on the ground that Putnam's story is *impossible* for physico-chemical reasons. By contrast, the fact that a fiction describes an impossible situation does not seem to preclude one to draw some moral. In this sense, at least, thought experiments can be thought of as distinct from fictions. However, my aim is to show that the informational structure which is necessary for drawing a moral is also necessary for thought experiments. I will show that the informational channels automatically deployed to fill the fictional background can be used in reverse direction, so to speak, in order to extract some fictional information. Consequently, I hope to shed interesting new light merely on the conditions of possibility of thought experiments, not on their argumentative

[14] For the record, I am attracted to the cognitivism of Novitz 1987. His two-stage model of how one can learn from fiction is somewhat close to what I will present below.
[15] Thanks to an anonymous referee for raising this objection.

efficiency. This can be seen as a limit beyond which the analogy between thought experiments and the moral drawing activity falters.

4.1 Fictional Background as a Mixture

Fictional disagreements show without a doubt that every fictional background is inter-connected with relevant representations of reality as well as other available fictional representations we may have. Developing on a striking geological metaphor, Proust thus talks about the "historical substratum" which constitutes the fictional background of Balzac's novels in his essay entitled *Sur la lecture*.

However, we have also seen that the fictional background cannot consist only of this substratum coming from the outside. In other words, factual and conventional information is combined with what I shall call "free imagination". How much free imagination there is depends on how much and how many informational channels are open. In the case of Gregor's metamorphosis, according to Nabokov, there is not much free imagination at play, for there is an informational channel originating in entomology which fills almost all of the detail of Gregor's physical appearance. In the case of Kovalyov's nose, however, the reader is free to imagine the Nose's physical appearance. A fictional background should thus be thought of as a sophisticated mixture of free imagination, factual and conventional information. The possible mixtures are primarily constrained by the fictional foreground.

Importantly, a fictional detail is "free" only relative to an informational channel. As such, some parts of the fictional background can be free relative to one informational channel and not relative to another. To illustrate this point, let us focus on *The Nose* again. Relative to functional anatomy, the reader should freely imagine the Nose. However, it seems clear that the fictional background of *The Nose* should not be freely imagined relative to, say, gravitational physics. Indeed, physical things are clearly weighty in Gogol's story. Since the Nose is a physical thing, it is subject to gravitation in Kovalyov's world. Consequently, if the Nose was to stumble while walking, it would fall down; if it was to climb up some steep stairs, it would not do it effortlessly; etc. The Nose is thus free relative to functional anatomy, but not relative to gravitational physics.

4.2 Reversal of the Direction of Fit

Using an informational channel to fill the fictional background is using it one way. To borrow Anscombe's famous notion, we can say that filling the fictional background has a direction of fit which goes from outside to inside the fiction. Once the informational channel is open and has been used, I suggest that we can use it the other way by simply reversing the direction of fit. Given that the fictional foreground is dynamic upon reading a story, the fictional background has to be accordingly updated. Consequently, reversing the direction of fit *at the end of the story* can convey some new information. Doing this corresponds to drawing the moral of a story.

By reading a fiction, the reader updates the foreground with fresh semantic information. This new information may or may not force the reader to substantially update the background by opening some new information channels. At the end of this repeated process, when the story ends, the reader has in mind a fictional background and some open informational channel. One can use the already open channels in the other direction by holding fixed the fictional background

and by asking oneself what are the facts or conventions which would have produced the fixed background in the first place. Since the background is a mixture containing free imagination, the information travelling back from the fiction in this manner is necessarily new.

Once the information has travelled outside the fiction, the reader would treat the information as they would some information coming from a non-fictional source. That is, they would first ponder it to decide whether it should be accepted, and then make the modifications to fit this new information into their cognitive system if necessary. Here, the plausibility of the information coming from the fiction, as well as the trust one can place in the author of the fiction would clearly play a crucial role. As such, the nature of the informational channel first deployed is important. Consequently, the more "realistic" or naturalistic the fictional background, the more reliable will the information retrieved be labelled. Here, "realistic" should also be understood relative to an information channel. Some situation will be "realistic" according to an informational channel (i.e. against some factual or conventional background) and not "realistic" according to another one. It is thus a term of art which measures how much an informational channel is open. The more "realistic", the less free imagination is required to fill the background; the less "realistic", the freer the reader is to imagine the fictional background.

To illustrate this mechanism thanks to which one extracts fictional information, it is useful to first give a ludicrous example where some information is retrieved from a fiction but it does not get inserted into the reader's knowledge. In the foreground of *The Nose*, many things are mandated to be imagined. For instance, the story mandates imagining the Nose fully dressed, with a hat, walking and talking. The reader should freely imagine this in the process of reading. Once the reading is finished, the reader has a representation of the Nose in mind. They can retrieve some new information by using the informational channel about walking creatures. Indeed, since the reader had to imagine Kovalyov walking down a street, they must have deployed a relevant informational channel to fill the background with a walking human. Now, the reader can thus update one's "knowledge" of how noses walk as if it was not freely imagined but the result of some factual information. If the reader imagined a biped nose, they would thus have some new information about independent human noses, namely that they are bipeds. Of course, calling this bit of information which originates in *The Nose* "knowledge" is very weird, for it contradicts a very robust facts about walking creatures, namely that they are organised bodies of organs and never individual organs. In other words, the reliability of this information about walking noses is zero, but this bit of information is both new and originating in the fiction.

Doing this is absurd, for the right response to Gogol's story is not that of drawing a moral about a super-natural event. As Roman Jakobson puts it, *The Nose* should be interpreted as a "realised oxymoron" for there is something utterly absurd in the fact that the Nose has nothing to do with a *real* nose. "Such is Gogol's 'Nose' which Kovalyov recognises as a nose even though it shrugs its shoulders, wears full uniform and so on." (Cited in Shukman 1989).

By contrast, take La Fontaine's first fable, inspired by Aesope, *The Grasshopper and the Ant* which is clearly an invitation to draw some moral. The foreground of this fable features two talking insects. Of course, the reader is expected to freely

imagine that insects can talk.[16] As the story goes, the reader is to imagine that a spendthrift singing Grasshopper unsuccessfully begs for food a stingy summer-worker Ant. Let us hold fixed the fictional background as the reader has freely imagined it. The fictional background is such that an informational channel originating in folk-psychology had to be open. Indeed, the two fictional characters have an explicitly human-like psychology. When the reading is done, the reader can use this channel in the other direction to answer the question: what kind of psychological facts would make the resulting fictional background as close to the facts as possible? One moral of the story is thus: stingy people do not lend what they earned. Probably this piece of information matches many of the reader's experiences with stingy people. Moreover, La Fontaine enjoys a very high reputation when it comes to folk psychology. Consequently, if asked whether stingy people tend to share what they earned in reality, the reader would probably feel confident in saying "no", and quoting La Fontaine's fable, even though they know that it is fiction.

Let me emphasise again how fictions differ from thought experiments when it comes to *possibility*. Fables about talking insects invite us to imagine impossible situations in some intuitive sense. This could be thought of as a problem, were one using La Fontaine's fable as a thought experiment in the course of an argument. Indeed, one would simply dismiss the story as a relevant piece of information for any kind of argument. However, this does not affect the drawing a moral from the story, because the fictional information one is expected to extract from the fable is travelling a particular information channel, i.e. a channel linking the fable with the reader's folk psychology representations. The moral drawn is not about ants and grasshoppers. It is about the usual subjects of folk psychology, i.e. ordinary folks. Informational channels are thus used to bring out bits of information, abstracted away from other parts of the story. Just like they are used to fill the fictional background with originally disconnected bits of information.[17]

5. Conclusions

One can see how easy it is to use a fiction as a thought experiment to inform our non-fictional representations. I showed how this extracting a moral exploits the same mechanisms as the filling of the fictional background. In a sense, it is as if fiction was made for it!

According to the present picture, the pressing question is not: how come we can learn from fiction? But: how come we usually do not? The answer to this question, I suggested, should be the same as for non-fictional source of information.[18] Consequently, coherence with already accepted knowledge or beliefs as well as the reliability of the source are expected to be central. How "realistic" the fictional background, or rather how much the reader is expected to freely imagine

[16] For there are no facts nor conventions about the detail of, say, phonatory devices of talking insects.

[17] Thanks to an anonymous referee to make me think of this special role of abstraction which is at play in the moral drawing activity.

[18] This reversal of how the problem is usually framed within the philosophy of fiction is, if I understand well, in keeping with some recent empirical results. I refer to some personal discussions with Stacie Friend who has ongoing work at the interface of philosophy and experimental psychology with Greg Currie and Heather Ferguson. See the details of the research project here.

such and such background element is also predicted to play an important role in my picture, since it greatly determines the kinds of informational channels which should be open to fill the background in the first place.

Interestingly, the results of the case study about fictional disagreements carry over to thought experiments. Informational channels must be in place: they are necessary conditions for such phenomena to happen. One can now see why thought experiments can benefit from being built on little narratives. Indeed, a fiction comes with a background and a fictional background automatically opens informational channels. A thought experiment can thus surreptitiously (or conspicuously) exploit these open channels to convey the information the reader is required to ponder. The rhetorical efficiency of thought experiments thus construed rests on the fact that everything happens in the background, i.e. automatically and probably mainly unconsciously.

I want to end with two side consequences of the claim I made. First, a consequence of my view is that each time some moral can be drawn, one can create a corresponding fictional disagreement. I think it is an empirically adequate prediction. Indeed, often, when one wants to question a moral drawn from a fiction, one starts a fictional disagreement. For instance, suppose one reader takes La Fontaine's fable to have another moral, namely that singers and artists are lazy, inconsequential people.[19] One way is to exhibit a real artist who is neither lazy nor inconsequential. Another is to deny that the story is "realistic" on this fact, i.e. to argue that the reader should freely imagine the Grasshopper's psychology, given a text analysis. One could thus argue that the Grasshopper's psychology in the fiction is really at odds with folk-psychology and it's the story which is "wrong". For instance, one might hold that real artists are proud; as a singer and artist, the Grasshopper should thus be proud; but the Grasshopper fictionally has no pride, for it begs for food. Consequently, one could start a disagreement whether the Grasshopper is a *genuine* artist or not, for if it was, it would be too proud to beg for food and it would happily die when the time comes.[20] One can see that this disagreement has an outside to inside direction of fit and aims at contradicting the moral according to which artists are lazy, inconsequential people.

Finally, I focused my claim on factual informational channels. However, given the general picture I presented, there is no reason to think that one could not export some fictional information so as to update or create conventions in the real world. I think this is a plausible fact, for it is clear that, say, conventions about dragons are ultimately grounded on seminal fictions (or maybe myths). Interestingly, genre conventions seem to be a very sophisticated phenomenon which ends up crystallising information coming from many different fictions into a grossly coherent body of information. In this sense, a genre convention can be thought of as non-fictional for it somehow acquires a sort of independence from its fiction(s) of origin. Borges wittingly emphasised this fact when he set about doing an encyclopedia of mythical creatures. In Borges 1967, he describes the "western dragon" as an entomologist would describe cockroaches and beetles:

[19] There is no explicit moral for *The Grasshopper and the Ant* in La Fontaine's text, which is quite remarkable. So the moral I drew above should not be taken as exhaustive in any way.
[20] See for instance Suits 1978 for a nice interpretation of Aesope's fable along these lines.

> A tall-standing, heavy serpent with claws and wings is perhaps the description that best fits the Dragon. It may be black, but it is essential that it also be shining; equally essential is that it belch forth fire and smoke. The above description refers, of course, to its present image; the Greeks seem to have applied the name Dragon to any considerable reptile (Borges 1967: 152).

References

Beardsley, M.C. 1958, *Aesthetics: Problems in the Philosophy of Criticism*, New York: Harcourt, Brace and World.

Borges, J.L. 1967, *The Book of Imaginary Beings*, 2nd ed. 1974, transl. N.T. di Giovanni, Harmondsworth: Penguin.

Capinera, J.L. 2008, *Encyclopedia of Entomology*, Leipzig: Springer.

Carroll, N. 2002, "The Wheel of Virtue: Art, Literature, and Moral Knowledge", *The Journal of Aesthetics and Art Criticism*, 60, 1, 3-26.

Davies, D. 2007, "Thought Experiments and Fictional Narratives", *Croatian Journal of Philosophy*, 7, 19, 29-45.

Davies, D. 2017, "Art and Thought Experiments", in Stuart, M.T., Fehige, Y. and Brown, J.R. (eds.), *The Routledge Companion to Thought Experiments*, London: Routledge, 512-25.

D'Alessandro, W. 2016, "Explicitism About Truth in Fiction", *British Journal of Aesthetics*, 56, 1, 53-65.

Elgin, C. 2007, "The Laboratory of the Mind", in Gibson, J., Huemer, W. and Pocci, L. (eds.), *A Sense of the World: Essays on Fiction, Narrative, and Knowledge*, New York: Routledge, 43-54.

Erlich, V. 1956, "Gogol and Kafka: A Note on 'Realism' and 'Surrealism'", in Halle, M. *et al.* (eds.), *For Roman Jakobson: Essays on the Occasion of His Sixtieth Birthday*, The Hague: Mouton & Co., 100-109.

Everett, A. 2013, *The Nonexistent*, Oxford: Oxford University Press.

Friend, S. 2011, "The Great Beetle Debate: A Study in Imagining with Names", *Philosophical Studies*, 153, 2, 183-211.

Friend, S. 2017a, "Elucidating the Truth in Criticism", *The Journal of Aesthetics and Art Criticism*, 75, 4, 387-99.

Friend, S. 2017b, "The Real Foundation of Fictional Worlds", *Australasian Journal of Philosophy*, 95, 1, 29-42.

Gendler, T.S. 2016, *Thought Experiment: On the Powers and Limits of Imaginary Cases*, New York: Garland.

Haldane, J.B.S. 1926, "On Being the Right Size", *Harper's Magazine*, 152, 424-27.

Kafka, F. 1915, *Die Verwandlung*, transl. J. Neugroschel, "The Metamorphosis", in *The Penal Colony, and Other Stories*, New York: Scribner 2010.

Lamarque, P. and Stein H.O. 1994, *Truth, Fiction, and Literature: A Philosophical Perspective*, Oxford: Clarendon Press.

Lewis, D. 1978, "Truth in Fiction", *American Philosophical Quarterly*, 15, 1, 37-46.

Nabokov, V. 1980, *Lectures on Literature*, ed. by F. Bowers, Intro by J. Updike, Tampa: Mariner.

Novitz, D. 1987, *Knowledge, Fiction & Imagination*, Philadelphia: Temple University Press.

Searle, J. 1975, "The Logical Status of Fictional Discourse", *New literary history*, 6, 2, 319-32.

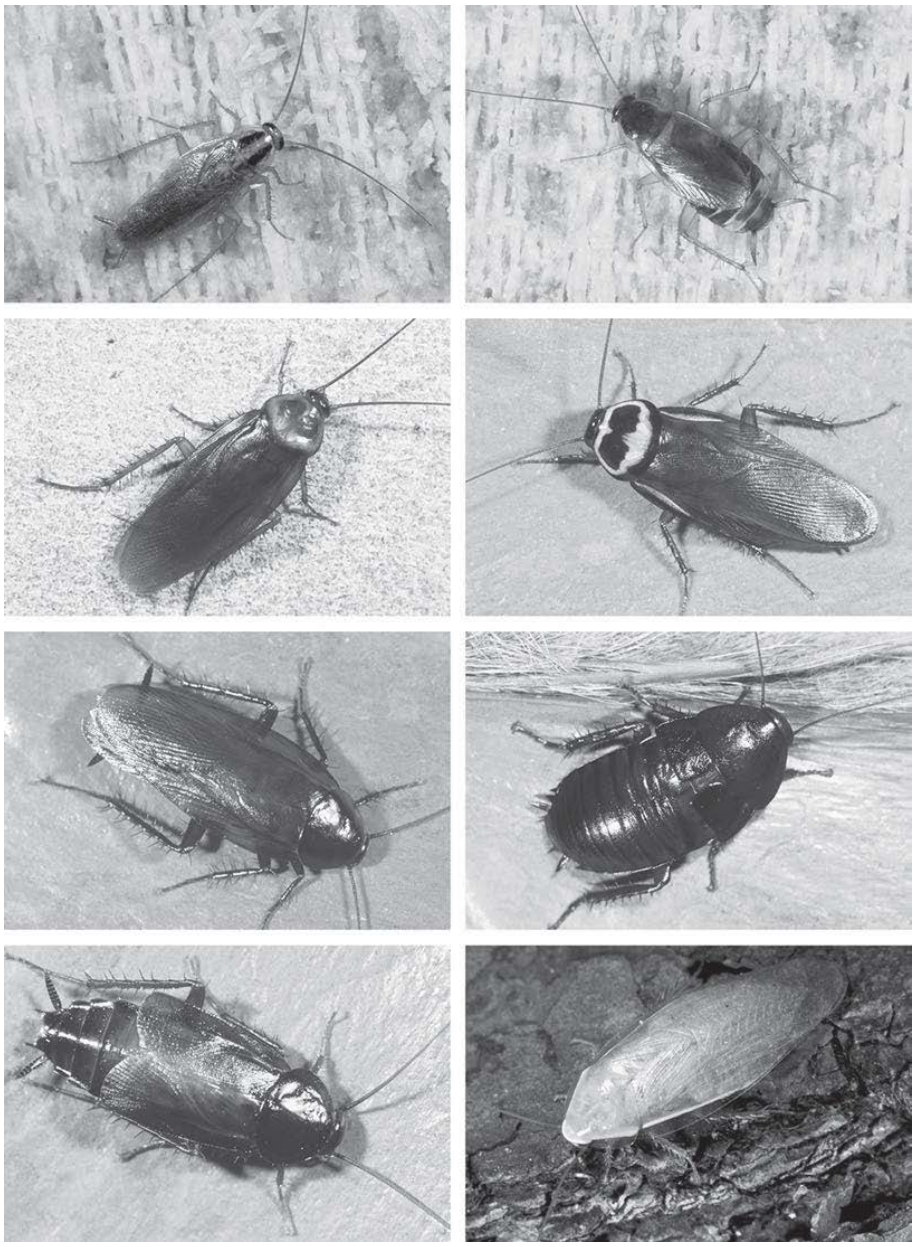
Shukman, A. 1989, "Gogol's The Nose or the Devil in the Works", in Grayson, J. and Wigzell, F. (eds.), *Nikolay Gogol: Text and Context*, London: Macmillan, 64-82.

Stuart, M.T., Fehige, Y. and Brown, J.R. 2017, "Thought Experiments: State of the Art", in Stuart, M.T., Fehige, Y. and Brown, J.R. (eds.), *The Routledge Companion to Thought Experiments*, London: Routledge, 1-28.

Suits, B. 1978, *The Grasshopper: Games, Life and Utopia*, Toronto: University of Toronto Press.

Todorov, T. 1970, *Introduction à la littérature fantastique*, Paris: Seuil; transl. R. Howard, *Fantastic: A Structural Approach to a Literary Genre*, Ithaca: Cornell University Press, 1975.

Walton, K. 1990, *Mimesis as Make-believe: On the Foundations of the Representational Arts*, Cambridge, MA: Harvard University Press.



Cockroaches (Blattodea), Figure 70. Some common cockroaches: top left, German cockroach, *Blatella germanica*; top right, brown-banded cockroach, *Supella longipalpa*; second row left, American cockroach, *Periplaneta americana*; second row right, Australian cockroach, *Periplaneta australasiae*; third row left, smoky-brown cockroach, *Periplaneta fulginosa*; third row right, Florida woods cockroach, *Eurycotis floridana*; bottom left, Oriental cockroach, *Blatta orientalis*; bottom right, Cuban cockroach, *Panchlora nivea* (photos by J.L. Castner, University of Florida).

Game Counterpossibles

Felipe Morales Carbonell

KU Leuven

Abstract

Counterpossibles, counterfactuals conditional with impossible antecedents, are notoriously contested; while the standard view makes them trivially true, some authors argue that they can be non-trivially true. In this paper, I examine the use of counterfactuals in the context of games, and argue that there is a case to be made for their non-triviality in a restricted sense. In particular, I examine the case of retro problems in chess, where it can happen that one is tasked with evaluating counterfactuals about illegal positions. If we understand illegality as a type of restricted impossibility, those counterfactuals are non-trivial counterpossibles. I suggest that their non-triviality stems from their role in practices of rule coordination and revision, and suggest that this model could be generalized to counterpossibles in different domains. I then compare the approach to the accounts of Vetter 2016 and Locke 2019.

Keywords: Counterpossibles, Games, Retro problems, Constitutive and regulative norms, Rule coordination and revision.

1. Introduction

There is an ongoing debate about the status of counterpossibles, counterfactuals with impossible antecedents. There are roughly two camps: one defends the view that counterpossibles are vacuously true, while the other defends the view that counterpossibles can be non-vacuously true or false.[1] One of the main motivations for the non-vacuity position is the defense of a series of metaphysical views (Nolan 2014 gives an overview of the many topics in which counterpossibles might play a crucial role). However, metaphysics is contentious enough that the case for counterpossible non-vacuity has remained inconclusive. Recently, some authors who defend non-vacuity have tried the different strategy to show that there are less contentious independent contexts in which it is necessary to distinguish between the truth value or acceptability of counterpossibles. For example, Baron, Colyvan, and Ripley 2017 argue that there can be genuine math-

[1] Lewis 1973, Williamson 2007, Emery and Hill 2016 and Vetter (2016 defend the orthodoxy. Nolan 1997, Kim and Maslen 2006, Brogaard and Salerno 2013, Kment 2006, Priest 2016, Berto *et al.* 2018, Locke 2019, Tan 2019, Berto and Jago 2019 defend non-vacuity.

Argumenta 6, 1 (2020): 117-133
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.mor

ematical counterpossibles, and Tan 2019 argues that there can be genuine counterpossibles in the natural sciences. Without holding an opinion on whether those applications of the strategy work, the main aim of this paper is to examine whether this strategy can pan out in the context of games and play. I will argue that the strategy does indeed work in this context, albeit with significant restrictions. By observing how counterfactuals behave in the context of games, we can get indirect evidence about whether this strategy can be useful more broadly. If we can only account for counterpossibles with restrictions even in the case of games, there might be restrictions also for the use of counterpossibles in different contexts.[2]

2. Some Generalities about Counterfactuals in Games

Reasoning in the context of games is often *explicitly conditional* reasoning. For concreteness, take chess. Planning a move involves reasoning about the consequences of the move: if we move a pawn to a certain position, the king will be exposed; if we castle, the attacker will have to move their knights to a certain area of the board if we want to push from this side; and so on. The conditionals that are evaluated in the context of play encode information about the outcomes of hypothetical scenarios where strategic choices are made, and they can be highly complex: in multi-player games, they are often not only about the direct effects of actions in the game state, but also about the beliefs of other participants of the game about the game itself: "if I move this piece here, my opponent will think that I plan to do this, so he will…".

While some of the relevant conditionals in game playing are indicative (as in the examples I just gave), it can be equally common to reason using counterfactuals of the form:

(1) If A were to happen, B would happen.

There is a rich literature on counterfactual reasoning in the context of games from the perspective of game theoretical issues (cf. Binmore 1987, Bicchieri 1988, Stalnaker 1996 and Skyrms 1998). For example, we can describe the prisoner's dilemma in counterfactual terms: in that formulation, it is about what would happen if a number or individuals were made to choose between cooperating or defecting against each other, given certain payoffs. The use of counterfactuals instead of future indicative conditionals seems to provide greater flexibility, since it allows the evaluation of situations that are detached from the current circumstances. Think of the different contexts in which we would use the conditionals "If Liam doesn't kick the ball to the right, someone else will" and "If Liam hadn't kicked the ball to the right, someone else would have". Clearly, there are contexts in which the truth conditions of these conditionals diverge.[3]

[2] The topic of counterpossibles in games is interesting also from a different perspective. One could defend the non-vacuity of counterpossibles from an anti-realist perspective by treating counterpossible-talk as a kind of game or fiction (cf. Kim and Maslen 2006). Our discussion here could also bear on the scope of this research direction.

[3] It is important to observe, with Lewis (1973: 4), that there are apparently subjunctive conditionals which have the same truth conditions as indicative conditionals, so that the apparent use of subjunctive conditionals in reasoning does not immediately entail that we are dealing with counterfactual reasoning.

Counterfactual reasoning in the context of games can be also *backwards looking*, in case where what we are talking about isn't the outcome of an action, but what explained the action:

(2) If A had happened, B would have happened.[4]

For example, given a surprising event in a game, we might reason about what explained it, so that we can then evaluate our future strategy. In some contexts we also might want to evaluate backwards looking counterfactuals for reasons that don't bear on future play at all. I will assume that both forward and backward looking counterfactual can be given a unified account.[5]

3. Counterpossibles in Games

Ordinarily, we only consider game counterfactuals with possible antecedents. This is reasonable because we are interested in problems of evaluating courses of action, where the possibility of acting on the information given by those evaluations matters (a different way of putting this point is that when reasoning counterfactually when playing games, we are looking for *guidance*). However, this is not decisive on whether there couldn't be cases where we evaluate counterfactuals with impossible antecedents. In principle, it is possible for judgements about what we ought to do to be independent of judgements about what would happen, even though in many cases it is clear that our judgements about what we ought to do are informed by what we believe would happen.[6] Pushing this line of thought would force us to take a stance on a whole host of difficult issues.[7] Instead, we should examine whether there can be direct counterexamples to the restriction of having possible antecedents.

Consider the chess board on top of the following page. Two questions: a) how can we proceed from this position to a win? b) how did we arrive at the current position? I will leave the former aside. The second question is characteristic of retrograde (or 'retro') analysis.[8] During retro analysis, one can make counterfactual judgments like

(3) If white were in this position, the bishop in g1 would have, at some point in the course of the game, moved from h2 or along the a7-g1 diagonal.[9]

Now, it turns out that it is impossible to arrive at this position during actual play (that is, in a game that begins from the standard starting position). I will call this

[4] In contexts where one could produce this conditional, it might be possible to also produce the conditional expressions 'If A happened, B must have happened', 'given that A happened, B must have happened'.

[5] Cf. Bennett (2003: Ch. 18) for a defense.

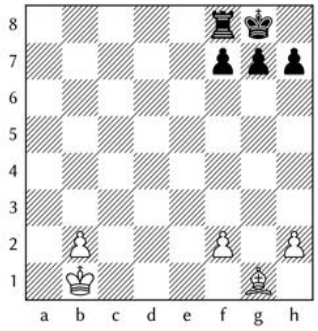
[6] Sinnott-Armstrong 1984 raises this possibility about the 'ought implies can' principle, which he takes to be a mere conversational implicature. In his view, while someone literally could be obligated to do something that they wouldn't be able to do, it would be pointless to say that they ought to do it because such saying wouldn't provide advice.

[7] Cf. the debate between Streumer 2007 and Heuer 2010 on whether there can be reasons to do or try the impossible.

[8] Smullyan 1979 is the most accessible introduction to retrograde analysis.

[9] It is plausible that a seasoned player or problem solver would recognize the impossibility of the board *visually* or *imaginatively* instead of relying on explicit counterfactual reasoning. Cf. the psychological literature on chess cognition tracing back to De Groot 1965 and Chase and Simon 1973.

type of impossibility *chess-impossibility*.[10] Chess-impossible positions are also called illegal (cf. the FIDE's laws of chess 2018: 3.10.3).[11]



The relevant modality pertains to *states* of a game, although there is closely related modality that pertains to *game-plays* (which in the case of chess are sequences of states, but in other cases can be processes in a richer sense). In this paper I will limit myself to the discussion of chess-impossibility in the stative sense only, but I think that much of what I will say here can be applied to the process conception of possibility. In any case, there should be a closely related counterfactual that says:

(4) If the board had come to be in this position, the bishop in g1 would have, at some point in the course of the game, moved from h2 or along the a7-g1 diagonal.

The antecedent of the counterfactual is chess-impossible, so we classify the counterfactual as *chess-counterpossible*.

How to evaluate counterfactuals like (4)? It is usually recognized that counterfactual evaluation is always performed against the background of some body of relevant assumptions. This body of assumptions is fixed by the context, which in turn is fixed by the task at hand. In the case of (4) we should presumably include assumptions about the rules of chess in this background; for example, that the bishop moves diagonally an arbitrary number of free spaces, and

[10] For a more formal treatment of chess-possibility, see the appendix.

[11] Dawson and Hundsdorfer (1915: 9) make an interesting distinction between impossibility and illegality: "We use with forethought the word *illegal* to define any condition which could not arise in actual play. The word *impossible* is often used in the same sense, but it is not satisfactory, and we shall not use it. There is no such thing as an impossible position, provided you have enough chess-men in your box to draw on. The word always provoked Sam Loyd. 'Impossible?' he would say, 'you say these men could not have got into such a position! Why, they *are* in that position; I put them there myself!' To this no answer can be made". Dawson and Hundsdorfer's point is that all chess diagrams are *constructible*, whereas not all of those possible diagrams are legal or could happen in actual play. The size of the possibility spaces is vastly different: roughly speaking, there are 10^{71} possible diagrams, and while the number of possible legal positions is an open question, it is widely believed to be within the 10^{40} to 10^{50} range. Cf. Steinerberger 2015. Sam Loyd (1841-1911) was a well-known chess problem composer and puzzle creator. For a very interesting overview of his position on the significance of chess impossibilities, which is more nuanced than Dawson and Hundsdorfer report, see White 1962: 444-54.

that the starting position of white's bishops is c1 and f1. Since we know that bishops move diagonally and that the bishop in g1 could not have started in its current position (call it P), the last move of the piece could have only initiated at either a position in the diagonals a7-g1 or h2-g1 (this is something that we can deduce or imagine). It is at this point that we could judge that if P were to happen, it would have followed such move (since they are the only apparently possible moves); that is, we could be disposed to accept. However, in both diagonals there are pawns blocking the bishop, which we also know couldn't have moved from their initial positions (pawns do not move backwards). So the bishop couldn't have arrived at g1 from either direction, since they are blocked, and the position is impossible.[12] Here we can suspend judgement on whether this means that we should reject our initial acceptance of the counterfactual; in any case, the standard semantics gives the verdict that the counterfactual is vacuously true.[13]

Once we reach an impossibility like this, we might be interested in evaluating whether there are changes to the setup that would make it possible (for example, we might realize that the type of play that would follow from an illegal position is interesting in a way that we judge should be allowed). Since the impossibility follows from rules about the movement of the chess pieces, and more precisely of a subset of those pieces, one might want to exchange those rules for more suitable ones. This immediately puts us in the position to consider rules that would deliver situations which are strictly speaking impossible in the relevant sense (chess-impossible in the case of (4)). There may be several viable variations of the set of rules that would yield the wanted result.[14] Consider the following counterfactual:

(5) If bishops in chess jumped over pieces of their own color once, the bishop in g1 would have moved from e3.

Again, the antecedent of this counterfactual is chess-impossible since the bishop in chess does not jump over pieces of their own color. Conditional (5) codifies a change to the rules that would allow a bishop in the a7-g1 diagonal to reach g1 (since the diagonal h6-c1 is empty, we can allow free movement for the bishop from its original position to g1). This might suggest that we should accept (5) as

[12] It is plausible that a seasoned player or problem solver would recognize the impossibility of the board *visually* or *imaginatively* instead of relying on explicit counterfactual reasoning. Cf. the psychological literature on chess cognition tracing back to De Groot 1965 and Chase and Simon 1973.

[13] On the supposition that the antecedent is indeed impossible; otherwise we have reason to think that in the closest worlds, whenever the antecedent is true, the consequent is false, so the counterfactual evaluates as false. Suppose that we rejected (4), and moved on to judge that it is false; our options would be either a) to reject the orthodoxy about counterfactuals, or b) to reject the classification of (4) as a counterfactual conditional. Lewis (1973: 24) already considers the possibility that so-called counterpossibles might be *sui generis*, but dismisses it without much comment.

[14] We could come up with rules by transposing and varying the movesets of the relevant pieces (e.g., having the pawns move backwards or sideways) or of other pieces (e.g., having the bishop move like the knight). We could also come up with entirely new move ideas; for example, having the bishop wrap around the board (so that it could continue from the diagonal a3-f8 into the diagonal g1-h2, for example), which no other piece does.

true in a way that doesn't follow automatically from the orthodox vacuity assumption.[15]

4. Defending the Legitimacy of Game Counterpossibles

There are several ways to handle counterfactuals like these. In this section, I will address several of them and argue that there are reasons to think that to handle games counterfactuals we have to be able to account for non-vacuous counterpossibles.

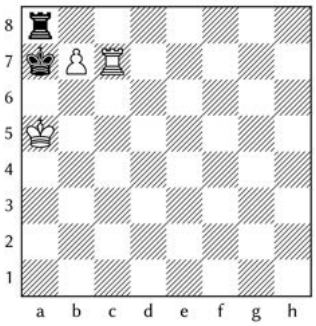
Perhaps the non-vacuity intuition could be explained by an ambiguity in the description of the evaluation of the counterfactual. Call the variation of chess that has the modification to the rules that we just described, chess*. Since P is a possible state of chess*, we can say that the chess-impossibilities of the antecedents of (4) and (5) are chess*-possibilities, so that the judgements about the chess-impossibilities' non-vacuity is simply a reflection of the judgements about the chess*-possibilities' non-vacuity (this sort of strategy is used often by defenders of orthodoxy). But then, there should be a worry that the reasonableness of counterpossible-talk relies on changing the subject, and thus on a form of modal illusion. The idea would be that in cases like these, our acceptance of the counterfactuals would rely on our acceptance of counterpart counterfactuals about similar things which are nonetheless strictly speaking different from the ones we are taking the counterfactuals to be about. If bishops in chess jumped over pieces of their own color once, it wouldn't be chess anymore.[16] While this might work to dismiss counterpossible-talk as misguided in a range of cases, it seems that this strategy of ambiguity elimination cannot be applied so clearly in many game cases.

Take for an example the following chess problem.[17] Suppose that the board is as follows, and it is white's turn:

[15] Does this count against the orthodoxy? We can see the orthodox view as giving explanations for why counterfactuals are true. Does the view have to be committed to those explanations being the only possible explanations? Perhaps some true counterpossibles are overdetermined as true: vacuously and non-vacuously (this could be spelled out in terms of counterpossibles possibly having multiple truthmakers, cf. Armstrong 2004: 21). For a view that went in this direction, it would be more important to establish the possibility of false counterpossibles, since the orthodoxy does not have resources for handling them.

[16] Cf. Kripke 1981: 113-14: "Could *this* table have been made from a completely *different* block of wood, or even of water hardened into ice […]? […] thought we can imagine making a table out of another block of wood or even from ice, identical in appearance to this one, and though we could have put it in this very position in the room, it seems to me that this is *not* to imagine *this* table as made of wood or ice, but rather is to imagine another table, *resembling* this one in all external details, made of another block of wood, or even of ice". Cf. Yablo 1993 for a different account of modal error, and Yablo 2000 for criticism of so-called "textbook Kripkeanism". Cf. also Byrne 2007 and Stoljar 2006 on "proposition confusion".

[17] I take this example from Smullyan 1979: 77. He also comments on a position where it is not obvious whether it is possible to castle, that raises similar worries.



Can white win in one move? According to the current rules of chess, this is impossible (white cannot capture the king or put it in check in one move). However, consider:

(6) If white were to promote the pawn in b7 to a black knight in b8, white would win in one move.

This counterfactual seems true for non-vacuous reasons (if the white pawn is promoted to a black knight in b8, the black king is in check from the white rook, and can only move to a6 and b6 where it can be captured by the white king). However, according to the current rules for chess, the move described by the antecedent of (6) is disallowed because one can only promote a pawn to a piece of its same color. So the antecedent of (6) seems to be chess-impossible, and we should treat (6) as a chess-counterpossible. As we sketched above, one could say that the move is possible for the game (call it chess**) with the less restrictive rule for promotion where there is no restriction about the color of the promoted pieces, and then explain the non-vacuity intuition by reference to the intuition about the chess**-possibility. However, it is not obvious that chess** is *not* chess. It would seem odd to say that after that restriction was put in place, the original game ceased to be and it was replaced by a different game (from the perspective of people endorsing the unrestricted rule, chess would become something else, but from the perspective of people endorsing the restricted rule, something turned into chess); rather, it is more natural to say that chess *itself* changed. The problem is not only theoretical, since the historical rules of chess actually changed in order to prevent this sort of situation.[18] The difficulties here lie in the identity conditions for the referent of the term 'chess', and these difficulties ramify in various directions.

One immediate response might be to notice that the term 'chess' can be equivocal between a broad sense and a narrow sense. In the broad sense, when we talk about 'chess', we talk about a class of games that share similarities (in the structure of the board used, the type of pieces, the rules, the goals). In this

[18] A late 19th century rulebook states the following promotion rule: "A Pawn is 'queened' when it has reached the last square of a file on which it is advancing, or when it captures a hostile piece on the eight row. It may then be exchanged for a Queen or Rook or a Bishop or Knight. Thus, a player may have two or more Queens, Rooks, Bishops or Knights on the board at the same time, or he may refuse promotion to his Pawn" (Steinitz 1889: xxiv). Note also that the pawn is not obliged to promote.

case we also talk of chess *variants*.[19] In the narrow sense, when we talk about 'chess' we refer to a specific instance of chess in the broad sense. However, what precisely, is that type? The precise reference of the term 'chess' when used in the narrow sense will vary from context to context: a person talking about chess in a narrow sense now could be talking about a different thing than what a person talking about chess in a narrow sense a hundred years ago would be talking about. We can expect the issues involved in fixing the reference of the term in a given context to be similar to those that solving our main problem requires, so the distinction between broad and narrow senses of the term is not sufficient.[20]

A more promising observation is that not all rules for chess will have the same role in fixing the reference of the term. While the game supervenes on the rules, the identity of the game might not supervene on the set of all the rules that apply to it. If so, then varying certain rules will not yield chess impossible situations, and consequently counterfactuals that involve variations to those rules will not be counterpossibles. One way to implement this strategy is to deploy Searle (1969)'s distinction between *constitutive* and *regulative* rules. The former "create or define new forms of behavior", while the latter "regulate antecedently or independently existing forms of behavior" (ibid.: 33-34). The type of rule will determine the modal character of facts about the bindingness of the rules (that is, whether they are necessarily or contingently binding). Someone who adopted this strategy could rely on something like the following plausible sounding principles:

Constitutive Necessity

> For some A regulated by a set of rules R, if some r in R is a constitutive rule for A, it is necessary that instances of A must obey r (where the inner necessity, which is deontic, is different from the outer necessity).

Regulative Contingency

> For some A regulated by a set of rules R, if some r in R is a regulative rule for A, it is contingent that instances of A must obey r (that is, it is possible that instances of A must obey r and it is possible that instances of A are not obliged to obey r; again, the inner necessity is not the dual of the outer possibilities).

Given these, we could say that counterfactuals about the application of different constitutive rules in the context of a practice A are A-counterpossibles, while counterfactuals about the application of different regulative rules have possible antecedents. Rules about the starting positions and basic movement of chess pieces seem like good examples of constitutive rules; then, we should count (4) and (5) as chess-counterpossibles. Whether we should count (6) as chess-counterpossible depends on whether we count the promotion rule as regulative

[19] Pritchard 2007 gives a compendium of chess variants, counting more than 1600 games. To those we should add variants that have been created only for the construction of problems. In his introduction, John Beasley counts as a variant "any game […] related to, derived from, or inspired by chess" (ibid: 13), which probably includes too much, but he also holds the opinion that strictly speaking 'true' chess games keep the goal of the game to capture the 'king' piece, and distinguishes these from other games that change the goal but keep the pieces, and from games that call themselves 'chess' but hold no resemblance from it whatsoever.

[20] There can be a range of senses between the broadest and narrowest. When I talk about the narrow sense, because of the contextual sensitivity I already mentioned, I mean the variably narrow sense that is sufficient to determine legality for positions.

or constitutive rule. If we count the rule as regulative, we should say that (6) is an ordinary counterfactual. However, there is a problem with treating the promotion rule as merely regulative. If we didn't have the promotion rule, a whole class of possible chess games would be excluded.[21] While the rule was adopted independently of the basic rules about the movement of the pieces (so it obviously didn't contribute to the *creation* of chess playing), it nevertheless *defines* what chess games are possible, and how they will pan out. This suggests that we should treat any rule that affects the possibility-space of chess (defined in this case as the set of possible positions) as a constitutive rule.[22] But if so, we cannot rely on the distinction to dismiss the legitimacy of chess-counterpossibles; on the contrary, the problem itself might turn out to be about what are the constitutive rules of the game, or what rules can play a constitutive role for chess. I should make it clear that my point here isn't that the application of the distinction couldn't work in any case; in effect, it might be useful to handle counterfactuals about rules like those of refereeing and tournament play (if applicable), since rules like those seem to be correctly characterized as regulative. It is not correct to say that by refereeing being done in one way or another, the game that is being played is different in one case or another. It is also incorrect to say that amateurs and professional players play different games. For those cases, the distinction between regulative and constitutive rules might be useful, with the appropriate revisions.[23]

Perhaps, then, we should treat *some* game counterfactuals as genuine counterpossibles (those we cannot rule out by the simple application of the broadness and constitutivity criteria), and try to account for their non-vacuity in a less indirect way.[24] For this, I think we should consider the functions that these counterpossibles could play in their contexts of use. As we pointed out above, ordinary

[21] It might also mean that the game rules would give no direction about what to do when pawns moved to the opposite end of the board, which would make the pawns unique. However, historically the promotion rule was not universal.

[22] In turn, this seems to indicate that Constitutive Necessity is incorrect at least for games, since for at least some rules that could play a constitutive role, there is a possibility where instances of the game must obey the rule, without it being necessary that instances of the game must obey it. This is compatible with it being necessary that the game obeys some of the rules that can constitute it. Alternatively, but at a greater theoretical cost, we could keep Constitutive Necessity but allow for incompatible constitutive rules to apply to instances of games, in which case we would also need to add a pragmatic story about why contextually certain rules are salient instead of others (cf. footnote 5 above). To make the incompatibility more palatable, we could adopt a logic where it can happen that "A is true", "B is true", but "A and B is not true" (cf. Lewis' observations on the "method of union" for truth in fiction, Lewis 1983: 277).

[23] How much weight should we give to these intuitions about what counts or not as the same game? Couldn't it be that the ordinary conception of games is incoherent, or that alternative conceptions are at least equally good? While these are definite possibilities, in the case of games any potential mismatch between their nature and ordinary talk about them must be treated with care, because the constitution of games is given by the practices of people who engage in them, including our talk about them. So while these intuitions are not infallible, the objection has less bite than usual. However, this line of defense of intuitions doesn't necessarily generalize to cases unlike games.

[24] There might be other ways to dismiss game counterpossibles as non-genuine that I haven't considered. Here I am only claiming that the lines of attack above are not sufficient to dismiss them.

counterfactuals often arise in play because they are needed for planning and strategic thinking. On the contrary, we don't expect counterfactuals like those in our examples to arise during normal play (and to be taken as true or false), except in cases where our reasoning about play is faulty (for example, to evaluate as true and to move according to the antecedent because one wants to arrive at the consequent position would be a misplay). Rather, we expect these counterfactuals to be evaluated in contexts where play is not the point. In the case of chess counterpossibles there seem to be two contexts where counterpossibles might arise.

The first case is that of retro problems. While they can be interesting from the perspective of endgame analysis (and thus implicitly from the perspective of chess possibility), they exist independently of play.[25] Chess problem solving exists outside the institution of play that extends to tournament play, and consequently has entirely different criteria of fairness, and depending on the setup, of what counts as an admissible solution.[26] This might suggest that the notion of impossibility in use here differs systematically from the notion of impossibility in use during play. However, in the case of retro problems with impossible setups, the relevant notion of impossibility is often the regular one: the point of the problems is to explain the illegality of the positions, which is not always obvious. Backtracking to a move and position that couldn't have happened, we reason about intermediate steps that also couldn't have happened. It seems to me that the more flexible way to do this is by allowing counterpossible reasoning.

The second case is the evaluation of rules; for example, when faced with issues that require a decision on how to implement a rule (due to ambiguity in the rule, or because the rule doesn't handle corner cases). This could be observed above in the case of the restricted and unrestricted promotion rules. In regular play, finding oneself in an illegal position indicates that someone made a mistake or cheated; finding oneself in an ambiguous situation, on the other hand, forces an examination of the rules, and of the consequences of possible changes to the rules. In those cases we want to distinguish between game-impossible scenarios, so we need a way to hold the relevant counterfactuals as true or false. Counterpossible reasoning could be used here.

In games like chess the practices that can allow for counterpossible reasoning and playing are relatively independent. However, this is a contingent feature of these practices. Peter Suber's 'nomic' game illustrates how both practices can be fully integrated.[27] In nomic, each 'move' can consist in the modification of the game's rules. A nomic game starts with a minimal set of rules about how the players should proceed, and specifies how rule changes can be incorporated (by default there is a 'democratic' mechanism where a player proposes rules and the other players either accept or reject the proposal). Given these facts about the game, what can be nomic-possible and nomic-impossible is much less clear than

[25] The problem literature precedes the existence of modern chess, with many medieval examples. It is worth mentioning that in some cases problems were embedded in games of gambling (cf. Murray 1913, II: Ch. VII).

[26] Cf. White (1962: 449) on the construction of problems with illegal positions: "If you want to use an extra officer or two, why not do so? There is nothing morally wrong about it. Your result will be distasteful to many solvers; but it will do them no harm". White, of course, assumes the modern practice of treating problems as intellectual exercises, while historically this was not always the case (see footnote 24).

[27] Suber 1990. For a multi-player chess variant of nomic, see Howe 2000.

in the case of chess possibility and impossibility (with suitable changes, the sphere of possibilities at any stage can grow and shrink widely). While one could say that everything is nomic-possible and nothing is nomic-impossible, these are *not* the notions of possibility and impossibility that would be used in counterfactual strategic reasoning during actual nomic play, which would be the proper counterparts of the notion of chess-possibility and chess-impossibility that we examined earlier. Thus, there might be a need for the evaluation of counterfactuals about genuine nomic-impossibilities. Admittedly, one could adopt the possibilist view according to which everything whatsoever is nomic possible, and supplement it with a pragmatic account that filters out irrelevancies. However, given the context sensitivity of counterfactuals, this might underutilize the resources that the context provides to determine their semantic content.[28] While we still get a liberal account of entertainable 'situations' or 'worlds' (that includes impossibilities *stricto sensu*), we have an 'inner' notion of possibility that we can then use to pragmatically rule out irrelevancies in context.

My proposal to understand counterpossible talk in the context of games (and perhaps more generally) can be sketched as follows. Games of the type we have discussed here supervene on rules.[29] If you change the rules too much, you start playing a different or divergent game. But before that happens, you will have potential variations that still count as the same game as we have been playing all along. What counts as merely a variation and what counts as a divergent game depends on criteria which are given in the context, and which are themselves subject to revision. In practice, surrounding or embedded in the game proper there is always a meta-game (or a collection of meta-games) that deals with managing revisions of this sort. It seems like counterpossible-talk can play a crucial role here, because it offers a way to express and discuss the consequences of adopting variant rules while keeping the distinction between variants and divergencies using a constant modal conceptual framework. Chess-impossibility stands in a relation to chess-possibility that chess*-possibility does not stand in relation to chess-possibility. While counterpossibles are context sensitive, they don't shift the modal framework in use implicitly.[30] If they did, they would be pointless in many cases, since they would change the subject too radically. Even when they don't change the subject, they can still be pointless in cases where the task at hand is to evaluate courses of actions, since it is doubtful

[28] The case of nomic is important because it puts pressure on the idea that we could understand the possibility of non-vacuous counterpossible-talk in terms of a sharp distinction between object languages and meta-languages for games (where counterpossibles are vacuous at the object level and possibly non-vacuous at the meta-level).

[29] Cf. Kreider 2011 for discussion of the relation between rules and games, and Ridge 2019 for an overview of the philosophical literature on the nature of games.

[30] This assumes a more or less traditional contextualist view. Ludlow 2014 offers a more dynamic view along similar lines, where the meaning of terms can change between and within conversations (cf. his chapter 5 specially on how he addresses troubles for his account). Like in the current proposal, Ludlow emphasizes the practices of negotiation of meaning and concepts. Unlike in the current proposal, in Ludlow's proposal the negotiation is purely metalinguistic, while I think it might have to do with the open-endedness of the referents of terms as well (in the case of games, the open-endedness of our conceptualizations is grounded on the open-endedness nature of games).

that they could be of direct use for guidance.[31] This is why we don't find them in play. Instead of giving a pragmatic account of the acceptability of counterpossibles, we should also be able to give a pragmatic account of the restrictions that we make in ordinary contexts to counterfactuals with possible antecedents (in which case instead of having a restricted default semantics which is pragmatically extended, we have a liberal default semantics which is pragmatically restricted).[32]

5. Divide and Conquer, or Normativist Subsumption?

It can be useful to contrast the current proposal in its general form with two recent views: Vetter's (2016) 'divide and conquer' strategy, and Locke's (2019) normativist account.

Vetter's (2016) aim is to defend the orthodoxy about counterpossibles using what she calls a 'divide and conquer' strategy, by distinguishing between cases where counterpossibles should be vacuous, and cases where they might not be. Arguably, the current proposal shares this 'divide and conquer' structure, although it draws the division between admissible and inadmissible cases differently.

The crux of Vetter's argument lies on the distinction she makes between epistemic and circumstantial modality, which she contrasts as follows:

> circumstantial modality concerns the objects, properties, and relations that a given claim is *about*, not [like in the epistemic case] any representational or cognitive features of the terms we use to refer to them (Vetter 2016: 2698).

With this distinction in hand, she proceeds to argue that non-vacuous seeming counterfactuals are always epistemic. The reason for this is that they would give rise to referential opacity, which gives evidence for an epistemic reading. This suggests that in the cases of seemingly non-vacuous game counterpossibles we have considered, the non-vacuity intuition can be explained away by proposing epistemic readings for the counterfactuals. Note that Vetter's view is not that counterpossibles are always vacuous, but that circumstantial counterpossibles always are.

However, this does not seem plausible in the case of the counterfactuals that we have considered. They are explicitly not about the representational features of games, or of our epistemic situation relative to them. They are about what would happen or would have happened in the context of games. This is a circumstantial subject matter, and the corresponding modalities should be correspondingly circumstantial.[33]

[31] However, see Heuer 2010. In any case, counterpossible talk in the uses I describe here could be indirectly of use because in some cases guidance requires changes to the operative modal framework.

[32] We don't need to choose: my point is that we have both strategies available instead of just the first.

[33] Locke 2019 raises the same criticism about the scope of Vetter's strategy, giving as a counterexample the counterfactual 'if a steel Penrose triangle were placed in a 4000 deg. F oven, it would melt.'

Locke 2019 offers a theory of counterpossibles that applies a more general modal normativist framework to the case of counterpossibles. Modal normativism is the view that the primary function of modal claims is expressive or normative.[34] The basic idea is that modal claims, in Brandom's turn of phrase, 'make explicit' the rules of use of our terms. Thomasson describes modal normativism about metaphysical necessity as the view that modal claims about necessity "serve the *prescriptive* function of expressing semantic rules for the terms used in them, or their consequences, while remaining in the object language" (Thomasson 2007: 136).

The last point is an important similarity between the modal normativist view and the current proposal. As I said before, if we are to accept seemingly non-vacuous counterpossibles, we should be careful not to change the subject. The goal of having modal language belong to the object language is precisely to avoid this issue. Consequently, modal normativist views do not have the problem that I raised for Vetter's account concerning the subject matter of counterpossibles.

Locke states normativism about counterpossibles as follows:

> metaphysical counterpossibles function to illustrate or express changes, or consequences of changes, to the actual constitutive rules that govern language use while remaining in the object language where terms are used rather than mentioned (Locke 2019: 8).

This follows the constraint we raised before that if there are genuine game counterpossibles, at least some (if not all) of those should relate to constitutive rules. A further similarity between Locke's view and the current proposal is the way Locke deals with the problem of changing the subject:

> I claim that, since object language claims about metaphysical necessities and possibilities illustrate the actual rules or permissions that govern the use of modal vocabulary, object language claims about non-trivial metaphysical impossibilities illustrate non-trivial changes in those rules and permissions. In the right context, claims about non-trivial metaphysical impossibilities are an important object language resource for "mis-using" language without being subject to rebuke or interpreted as incompetent, e.g. in the case of a charitable philosophical dispute. This is because small, relevant changes in the actual rules that govern the use of some expression neither result in a radically different expression nor do they result in a complete change of subject (Locke 2019: 11).

The current proposal manages to tell roughly the same story without having modal language as a whole play a normative or expressive function. Perhaps representational language is normative or expressive, but that is an even greater departure from orthodoxy that we are not forced to make just for the sake of being able to handle counterpossibles. This aspect of the normativist proposal is underplayed by Locke because of his underlying commitment to normativism about modality in general, but in the present context the issue is more pronounced. Furthermore, modal normativism depends on having an account of the adequacy of the constitutive rules of language use (thus, Thomasson 2007: 138 says that normativism "requires that we first accept that our terms *have* rules

[34] Cf. Brandom 2008 and Thomasson 2007, 2013.

of use"). That makes the possibility of contexts where counterpossibles are used to discuss potential revisions to those very same rules a bit awkward; this seems to be the reason why, in the 2007 paper, she claims that under normativism there are substantive limitations about what kind of revisionary projects can be undertaken. In recent work Thomasson (2017) introduces the idea that metalinguistic negotiations might have non-semantic consequences, which allows for more revisionary projects; Locke 2019 adopts this solution. The solution in the current proposal is that the appropriateness of counterpossibles depends on the features of the local context, not of global standards of use (of course, the local context might in turn refer back to broader standards). This means that disputes about counterpossibles might not necessarily be resolved definitely through conceptual analysis, like Locke (2019: 20) suggests; indeed, they might only be resolved temporarily or not at all.[35]

6. Appendix: Chess Possibility

Semi-formally, a board *b* is chess-possible iff it can be reached in any number of steps by the application of chess-rules *R*, from a starting board *s*.

A *diagram* is a sentence describing the complete state of a board (essentially the information encoded in a FEN string). We will work in a language with variables for diagrams ($p_1 \ldots p_n$), two constants: *i* for the current diagram and *s* for the starting diagram, and three modal operators: $\Diamond_{\rightarrow}$, $\Diamond_{\leftarrow}$, and $\Diamond_s$ that build sentences out of sentences. The informal interpretation of these operators is "it is possible to advance to position...", "it is possible to have come from position..." and "it is chess possible that...", respectively. We also have the usual negation and the connectives for conjunction, disjunction, and material implication.

A *chess-frame* is a 4-tuple $<W, s, R_{\rightarrow}, R_{\leftarrow}>$, where *W* is a set of possible (constructible) boards, s is a selected member of *W* that represents the starting position, $R_{\rightarrow}$ is a binary relation over *W*, and $R_{\leftarrow}$ is another binary relation over *W*. We use two binary relations instead of one because we want to track more perspicuously (1) what moves can be made legally from a position (this is what $R_{\rightarrow}$ tracks) and (2) what moves could have been made legally to arrive at a position (this is what $R_{\leftarrow}$ tracks), and some moves in chess are not reversible (the pawns can only move forward). $R_{\rightarrow}$ and $R_{\leftarrow}$ can be understood as the converse of each other, so that $R_{\rightarrow}\, ab \leftrightarrow R_{\leftarrow}\, ba$.[36] It is worth noting that neither relation is reflexive (it is not possible to make a move that doesn't change the state of the board), but both relations are transitive (if it is possible to arrive from one direction at a position *A* from a position *B*, and it is possible to arrive from the same direction at a position *B* from a position *C*, it is possible to arrive from the same direction to A from C). We extend frames with a function I that assigns a unique diagram to every board in W to obtain a *chess-model* (a different way to present this would

[35] I would like to thank the reviewers for their suggestions, and Jan Heylen, Lars Tump and Kristine Grigoryan for their feedback on earlier versions of the paper.

[36] We implicitly assume that we track information about the players and the turns (for example, to prevent white to move twice in a row, or—in some variants—to allow for such things). A different approach would be to have one the frames be a triple <W, s, R> where R is a set of binary relations over W where each represents a possible move according to a rule.

be to make boards themselves diagrams, and to let diagrams represent themselves).

We define a valuation V_M for a model M as a function that assigns truth values (0 or 1) to each well-formed-formula to each member of W as follows, where δ is any diagram, φ and ψ are any *wffs*, and w is any board:

$V_M(\delta, w) = 1$ iff $\delta = I(w)$
$V_M(\neg\varphi, w) = 1$ iff $V_M(\varphi, w) = 0$
$V_M(\varphi\rightarrow\psi, w) = 1$ iff $V_M(\varphi, w) = 0$ or $V_M(\psi, w) = 1$
$V_M(\Diamond_{\rightarrow}\varphi, w) = 1$ iff for some $w' \in W$ with $R_{\rightarrow} ww'$, $V_M(\varphi,w') = 1$
$V_M(\Diamond_{\leftarrow}\varphi, w) = 1$ iff for some $w' \in W$ with $R_{\leftarrow} ww'$, $V_M(\varphi,w') = 1$
$V_M(\Diamond_s\varphi, w) \quad = 1$ iff $\varphi = I(s)$ or $V_M(\Diamond_{\rightarrow}\varphi, s) = 1$

For the three modal operators, there is a derived notion of necessity that is their dual. There are four types of possibility in the model: a) the combinatorial possibility of diagrams, which is assumed for W in the frames, b) the forward looking possibility $\Diamond_{\rightarrow}$, c) the backwards looking possibility $\Diamond_{\leftarrow}$, and c) the composite $\Diamond_s$, which is what we call chess-possibility properly speaking. Because of this, the model includes worlds which are constructible and sharply distinguishable, but impossible in a definite sense, without a need to mark those explicitly.

$(\Diamond_s\varphi \,\&\, \Diamond_{\rightarrow}\varphi) \rightarrow \Diamond_s$ is a non-theorem: there can be positions that can move towards chess-possible positions that couldn't have come from the standard position. On the other hand, if we can advance to an impossible position, the current position is impossible: $(\neg\Diamond_s\varphi\&\Diamond_{\rightarrow}\varphi) \rightarrow \neg\Diamond_s i$. In this model, some impossible positions share with the starting position the property of being terminal nodes: there is no position that they could have come from. But it is clear that in many cases we want to reason about illegal positions that derive from legal positions through misplay. To model this, we should introduce additional accessibility relations that models transitions from positions through mistakes (forwards and backwards, like above). In the system extended in this way we can reason backwards from impossible positions to positions that caused the illegality.

References

Armstrong, D.M. 2004, *Truth and Truthmakers*, Cambridge: Cambridge University Press.

Baron, S., Colyvan, M. and Ripley, D. 2017, "How Mathematics Can Make a Difference", *Philosopher's Imprint*, 17, 3, 1-19.

Bennett, J. 2003, *A Philosophical Guide to Conditionals*, Oxford: Oxford University Press.

Berto, F., French, R., Priest, G. and Ripley, D. 2018, "Williamson on Counterpossibles", *Journal of Philosophical Logic*, 47, 4, 693-713.

Berto, F. and Jago, M. 2019, *Impossible Worlds*, Oxford: Oxford University Press.

Bicchieri, C. 1988, "Strategic Behavior and Counterfactuals", *Synthese*, 76, 135-69.

Binmore, K. 1987, "Modeling Rational Players I", *Economics & Philosophy*, 3, 2, 179-214.

Brandom, R. 2008, *Between Saying & Doing: Towards an Analytic Pragmatism*, Oxford: Oxford University Press, https://doi.org/10.1093/acprof:oso/9780199542871.001.0001.

Brogaard, B. and Salerno, J. 2013, "Remarks on Counterpossibles", *Synthese*, 190, 639-60.

Byrne, A. 2007, "Possibility and Imagination", *Philosophical Perspectives*, 21, 1, 125-44.

Chase, W.G. and Simon, H.A. 1973, "Perception in Chess", *Cognitive Psychology*, 4, 55-81.

Dawson, T.R. and Hundsdorfer, W. 1915, *Retrograde Analysis: A Study*, White, A.C. (ed.), Leeds: Whitehead and Miller.

De Groot, A.D. 1965, *Thought and Choice in Chess*, The Hague: Mouton.

Emery, N. and Hill, C. 2016, "Impossible Worlds and Metaphysical Explanation: Comments on Kment's 'Modality and Explanatory Reasoning'", *Analysis*, 77, 1, 134-48.

FIDE 2018, "FIDE Handbook", https://handbook.fide.com/chapter/E012018.

Heuer, U. 2010, "Reasons and Impossibility", *Philosophical Studies*, 147, 2, 235-46.

Howe, D. 2000, "Nomic Chess", https://www.chessvariants.com/multiplayer.dir/nomicchess.html.

Kim, S. and Maslen, C. 2006, "Counterfactuals as Short Stories", *Philosophical Studies*, 129, 81-117.

Kment, B. 2006, "Counterfactuals and the Analysis of Necessity", *Philosophical Perspectives*, 20, 237-302.

Kreider, A.J. 2011, "Game-Playing Without Rule-Following", *Journal of the Philosophy of Sport*, 38, 1, 55-73.

Kripke, S. 1981, *Naming and Necessity*, Oxford: Blackwell.

Lewis, D.K. 1983, "Postscripts to 'Truth in Fiction'", in Id., *Philosophical Papers*, I, Oxford: Oxford University Press, 276-80.

Lewis, D.K. 1973, *Counterfactuals*, Oxford: Blackwell.

Locke, T.D. 2019, "Counterpossibles for Modal Normativists", *Synthese*, https://doi.org/10.1007/s11229-019-02103-1.

Ludlow, P. 2014, *Living Words: Meaning Underdetermination and the Dynamic Lexicon*, Oxford: Oxford University Press.

Murray, H.J.R. 1913, *A History of Chess*, London: Oxford University Press.

Nolan, D. 1997, "Impossible Worlds: A Modest Approach", *Notre Dame Journal of Formal Logic*, 38, 535-72.

Nolan, D. 2014, "Hyperintensional Metaphysics", *Philosophical Studies*, 171, 1, 149-60, https://doi.org/10.1007/s11098-013-0251-2.

Priest, G. 2016, "Thinking the Impossible", *Philosophical Studies*, 173, 10, 2649-62.

Pritchard, D.B. 2007, *The Classified Encyclopedia of Chess Variants*, Beasley, J.D. (ed.), 2nd edition, Harpenden: John Beasley.

Ridge, M. 2019, "Play and Games: An Opinionated Introduction", *Philosophy Compass*, https://doi.org/10.1111/phc3.12573.

Searle, J. 1969, *Speech Acts*, Cambridge: Cambridge University Press.

Sinnott-Armstrong, W. 1984, "'Ought' Conversationally Implies 'Can'", *The Philosophical Review*, XCIII, 2, 249-61.

Skyrms, B. 1998, "Subjunctive Conditionals and Revealed Preference", *Philosophy of Science*, 65, 4, 545-74.

Smullyan, R. 1979, *The Chess Mysteries of Sherlock Holmes*, New York: Knopf.

Stalnaker, R. 1996, "Knowledge, Belief and Counterfactual Reasoning in Games", *Economics and Philosophy*, 2, 133-63.

Steinerberger, S. 2015, "On the Number of Positions in Chess Without Promotion", *International Journal of Game Theory*, 44, 761-67, https://doi.org/10.1007/s00182-014-0453-7.

Steinitz, W. 1889, *The Modern Chess Instructor*, New York: Putnam & Sons.

Stoljar, D. 2006, *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*, Oxford: Oxford University Press.

Streumer, B. 2007, "Reasons and Impossibility", *Philosophical Studies*, 136, 3, 351-84.

Suber, Peter. 1990. *The Paradox of Self-Amendment: A Study of Law, Logic, Omnipotence, and Change*, Berne: Peter Lang.

Tan, P. 2019, "Counterpossible Non-Vacuity in Scientific Practice", *Journal of Philosophy*, 116, 1, 32-60, https://doi.org/10.5840/jphil201911612.

Thomasson, A. 2007, "Modal Normativism and the Methods of Metaphysics", *Philosophical Topics*, 35, 1/2, 135-60.

Thomasson, A. 2013, "Norms and Necessity", *The Southern Journal of Philosophy*, 51, 2, 143-60.

Thomasson, A. 2017, "Metaphysical Disputes and Metalinguistic Negotiation", *Analytic Philosophy*, 58, 1, 1-28.

Vetter, B. 2016, "Counterpossibles (Not Only) for Dispositionalists", *Philosophical Studies*, 173, 10, 2681-2700.

White, A.C. 1962, *Sam Loyd and His Chess Problems*, 1st edition 1913, New York: Dover.

Williamson, T. 2007, *The Philosophy of Philosophy*, Oxford: Blackwell.

Yablo, S. 1993, "Is Conceivability a Guide to Possibility?", *Philosophy and Phenomenological Research*, 53, 1-42.

Yablo, S. 2000, "Texbook Kripkeanism and the Open Texture of Concepts", *Pacific Philosophical Quarterly*, 81, 1, 98-122.

Fiction, Imagination, and Normative Rationality

Malvina Ongaro

University of Piemonte Orientale

Abstract

Rationality is a cornerstone of economics. The properties defining rationality are embodied by the Rational Agent, whose actions are prescriptive for economic agents. However, the Rational Agent is a fictional character: so why should real agents act like it? The Rational Agent takes its normative force from the arguments in support of the properties it embodies. In this paper, I explore the grounds for the normative force of the Rational Agent by looking at one of them. I explain the compelling pull of the famous Dutch Book argument using tools from narratology. I contend that the argument presents a branching narrative structure that allows the comparison of outcomes. Thus, the agent can see that one option serves her economic desires better than the other, and this is the specific way in which it provides normative support to the rational agent. Since the comparison of outcomes requires the use of imagination, I conclude the paper drawing some implications of my analysis for a connection between imagination and action.

Keywords: Rationality, Normativity, Narratives, Imagination, Dutch Book.

1. Introduction

There is a fictional character haunting economics: the Rational Agent.[1] Microeconomic models base their conclusions on assumptions about what constitutes rational economic agency. These assumptions are embodied by the Rational Agent but, as features of real, human agents, they would be highly unrealistic. However, it is often contended that, since they are taken as features of *rationality*, their goal is not to provide an accurate description of actual economic agency, but rather to prescribe a blueprint for rational behaviour. In short, the fictional character behaves how we *should* behave, i.e. how we would behave if we were

[1] The subject of models of economic agency goes often under the name *homo economicus*. Here, I refer to it as *rational agent* mainly for historical reasons. Genealogically, the two notions are distinct, and talks about normative rationality are more tightly connected with the latter (Morgan 2006).

Argumenta 6, 1 (2020): 135-146
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.ong

rational. It only exists in microeconomic models, where it makes choices that are supposed to be prescriptive for real agents: since the model shows that the Rational Agent would do *a*, then real agents have a reason to do *a*.

But how can the actions of a fictional character be normative for real people? What reasons do we have to act like someone who does not even exist? The apparent contrast between fictional world and normative arguments motivates this paper. Consequently, the *broad* research question from which we move is the following: how can fictional constructs act normatively?

I plan to address this question in the following way. I will start introducing the Rational Agent as the personification of a bundle of normative requirements on rational agency. I will then present the well-known 'Dutch Book', a normative argument in support of one such requirement. With this background in place, I will formulate my *narrow* research question: what is it that makes arguments like the Dutch Book normative? I will construct the answer to this question in three steps. First, I will clarify the sense in which I take such arguments to be normative. Second, I will propose that the argument displays a structure similar to the branching structure theorised by Beatty (2017) for explanatory narratives. Third, I will claim that this structure allows real agents to compare different outcomes and see that the one delivered by complying with the assumptions of rationality is the one that better serves their economic motivations. Thus, fictional constructs can provide normative grounds for human agency. Finally, I will expand on my argument and propose that the mechanism of outcome comparison is based on the cognitive capacity of imagination. If this is so, then my account of normative arguments illuminates some interesting connections between imagination and action.

2. The Rational Agent

The rationality of agents is one of the central assumptions in neoclassical microeconomic models. This puts the Rational Agent at the foundations of microeconomics, since it embodies the properties that economists have taken to constitute economic rationality. Traditionally, the central properties defining economic rationality, and hence the Rational Agent, are the following:

(1) *Logic*: The Rational Agent reasons according to classical logic.
(2) *Probabilism*: The Rational Agent has credences that respect the probability calculus.[2]
(3) *Rational Preferences*: The Rational Agent has preferences that are complete, transitive, and independent.[3]
(4) *Maximisation*: The Rational Agent chooses what is ranked highest in its order of preferences.

[2] Of course, *Probabilism* is not necessary in all those contexts where there is no uncertainty.
[3] While the name and specification of the requirement of Independence vary across different formal systems (e.g. von Neumann and Morgenstern 1944; Savage 1954; Jeffrey 1990), the core idea is that the preference between any two alternatives should be independent of both the state of the world in which they obtain and all irrelevant alternatives. Typically, full theories will include other (technical) requirements, which however are not strictly requirements of rationality.

Even under such a rough presentation, these properties strike as remarkably inaccurate descriptions of actual human agents. Our reasoning often violates classical logic, and nobody has truly probabilistic credences. Indeed, the Rational Agent is a very unrealistic character. Clearly we are not talking of someone *real*—there is no risk of meeting it in line at the post office, for instance. The Rational Agent is nowhere to be found: it is as fictional as Sherlock Holmes. In a truly Meinongian spirit, as a fictional character all we know about it are the properties ascribed to it by economists, and nothing else. It is, so to say, a *thin* character, since we do not assume anything about it beyond what we are explicitly told.

As properties of rationality, (1)-(4) are commonly taken to have a normative character: however unrealistic, they are not meant to constitute a descriptively accurate representation of human agency. They describe not how an economic agent is, but how an economic agent *should* be, or has reasons to be. The Rational Agent acts in microeconomic models and makes choices that are supposed to be prescriptive for real agents, because they are the choices real agents would make if they were rational.

It is worth noting that—although common—this normative interpretation is by no means uncontroversial. Since it is still supposed to apply to human agent, claims about rationality are still open to empirical enquiry. Indeed, the normative interpretation has sometimes been taken as a response to defuse potential empirical counterexamples (Hands 2015).

However, in this paper I will stick to the normative interpretation of properties of rationality for several reasons. First, however contested, the normative interpretation of (1)-(4) is still the standard view n economic methodology. Second, these properties are generally justified with normative arguments showing that it is rational to follow them, rather than with empirical observations. And finally, as long as there is a normative interpretation and there are normative arguments, then it is sensible to investigate the source of this normativity, independently of its adequacy.

So we have a fictional character that makes choices within models, and human agents that are expected to comply with such choices. But what reasons do we have to act like somebody who does not even exist, and that is entirely defined by a bunch of unrealistic properties? How can an unrealistic fictional character have any normative power towards the behaviour of a real person?

To be clear, the Rational Agent is not normative in itself. As we have seen, its role is to flesh out a bundle of properties that are normative. The actions of the Rational Agent have normative power only as representations of the normative implications of (1)-(4).

Of course, each of the properties listed above is supported by arguments justifying it as a property of rationality, and therefore justifying the legitimacy of its inclusion in the set. For instance, the requirement of transitivity in *Rational Preferences* is grounded on arguments that show how intransitive preferences would expose the agent to the possibility of exploitation.

However, the existence of such arguments does not answer our question. Just as there are arguments supporting this view of rationality, there are others opposing it. The debate on what is rightfully rational and on the notion of rationality that economists should care about, if they should care about one at all, is open and heated. But this debate should not concern us. I am not trying to argue for this specific list of properties, or for any such list for that matters: I am

interested in the sources of normativity, not in its objects. Any other property would be interesting, as long as it had a normative status supported by arguments claiming to justify it. It is sufficient for our purposes that there is some property, the legitimacy of which as a feature of rationality is justified by some typical arguments. Even though the validity of these arguments is debated, they have a compelling pull explaining their normative role.

Then, it seems that one could answer the question of what makes the Rational Agent normative by listing the arguments in favour of each property. But this move simply shifts the question. What is it that makes the arguments compelling? Through which mechanisms do they provide normative force to some property? This is the *narrow* question that I will try to address. In order to answer this question, I will focus on one such argument, and try to enlighten the normativity generating mechanisms behind it. We will later see that this mechanism is not specific to the argument I discuss.

One of the most influential arguments in favour of *Probabilism* is the so-called 'Dutch Book' argument. The argument is often presented in a very narrative fashion, constructed as a story in which some character displays non-probabilistic credences in some gambling scenario, and ends up losing money in consequence of her credences. This is an example of a standard presentation of the argument in an introductory text to Decision Theory:

> Suppose, for instance, that you believe to degree 0.55 that at least one person from India will win a gold medal in the next Olympic Games [and to degree] 0.52 that no Indian will win a gold medal in the next Olympic Games […]. Also suppose that a cunning bookie offers you to bet on both these events. […] However, by now you have paid $1.07 for taking on two bets that are certain to give you a payoff of $1 *no matter what happens*. […] Certainly, this must be irrational. (Peterson 2017: 154; emphasis in original).

The normative force of the argument cannot come from some feature of the storytelling. It is not relevant to the final judgement of irrationality that you are betting on the Olympic Games, or that the bookie you meet is cunning. The storytelling may have rhetoric force that is useful to get the message through and make the reader understand the gist of the argument. But it cannot suffice to establish something as a legitimate property of rationality, or it could be sufficient to present the argument under a different storytelling to dispel its legitimising power. Instead, the normative force must reside in the core of the argument, i.e. in that part that remains constant under different clothings. Let us then have a look at the argument in its minimal form, devoid of narrative constructions:

> *Dutch Book*: If an agent has non-probabilistic credences, then there is a theorem that proves that there is a combination of betting contracts[4] (called a Dutch Book) such that the agent faces sure losses.

With this argument in place, the question that remains to be addressed for the rest of the paper is the following: What is the mechanism that makes *Dutch Book* compelling as an argument for the normative validity of *Probabilism*? In order to

[4] A *betting contract* is "a contract to settle a bet or a group of bets at certain agreed betting rates" (Hacking 2001: 164). It is neutral with respect to the role played by the agent, i.e. whether she is the bettor or the bookie in the contract.

attempt an answer to this question, we need first to clarify the notion of normativity at stake. Thus, I will now move to an analysis of the sort of normativity that I take the Dutch Book to confer to *Probabilism*.

3. Normativity

A complete account of what normativity is would be vastly outside of the scope of the present paper. What is interesting for our purposes is not normativity *per se*, but rather the identification of the way in which the Dutch Book argument can be normative. Whether there are other ways for something to be normative, or how powerful or frequent this specific way is, are interesting questions that do not concern us. Instead, I will limit the discussion to two claims that I will try to make as little controversial as possible. Let us start with the first one:

(1) An argument provides normative support for a certain (option)[5] *o* IF it provides a reason for *o*.

Some clarifications on (1). First, we are talking about normative, not motivating reasons (Dancy 2000, Scanlon 1998). We are not looking for the motivation behind some actions, but for a consideration in favour of a certain option. Second, (1) is not meant to be a definition of a normative argument. It is not a biconditional, as it merely provides a sufficient condition for an argument to be normative. This means that there may be many other ways to attain normativity. But as long as (1) is at least one of the possible ways in which an argument can be normative, then there is no obstacle to our discussion. Third, being normative does not imply that the argument is conclusive. Each normative reason provides *pro tanto* justification for a certain option; there may be different normative reasons pulling in the opposite direction, so that the evaluation of an option would require an all-things-considered assessment. Thus, it is not the case that once someone has an argument that provides a reason for option *o*, then *o* is justified once and for all.

With these due clarifications of (1) in place, we need to take a further step, and understand what it means for an argument to provide a (normative) reason for something. Of course, an evaluation of the debate on normative reasons would, again, be far out of the scope of our present inquiry. As before, I will content myself with the following claim, broadly Humean in spirit:

(2) The agent has a reason for a certain option *o* IF she has a desire *d* and *o* serves *d* better than the alternative options.

Again, a few qualifications are necessary. First, (2) is not a definition either. Contrary to other authors supporting a desire-based view of reasons (e.g. Williams 1979, Schroeder 2008, Goldman 2009), I do not claim that this is a requirement of reasons. (2) does not provide necessary and sufficient conditions for the agent to have a reason for something: it merely states two jointly suffi-

[5] The claim is expressed in terms of options. Since it is assumed to be possible to have either probabilistic or non-probabilistic credences, or else there would be no need for an argument, then I take the Dutch Book argument to support an option. Others may prefer to look at *o* as a choice or an action. But nothing in our discussion hinges on the ontological category to which we ascribe the target of the argument. This does not imply that (1) applies equally to all sorts of categories: as long as it applies to your favourite ontological account of the target of the Dutch Book, we are good to continue.

cient conditions, without making any claim about other potential ways in which an agent could have a reason for something. As long as one concedes that these conditions do indeed provide one with a reason, then the discussion can proceed. Second, once again having a reason does not imply that the agent should act according to that reason. She may have other reasons supporting different courses of action, and any justification of her behaviour should come after an all-things-considered assessment of the different reasons she has.

Since both our claims provide sufficient conditions, we have identified a plausible route to attain normativity, one that does not pretend to exhaust the discourse on what constitutes normativity. Putting (1) and (2) together, we obtain (3):

(3) An argument provides normative support to option *o* IF it shows to the agent that *o* serves her desire *d* better than the alternative options.

If my reasoning is correct, this is one way in which an argument gets normative force. In what follows, I will try to show that it is the way in which the Ducth Book argument gets its normative force in support of *Probabilism*, and that it does so in virtue of the specific structure it displays. To introduce this structure, our next step requires a little detour into narratives.

4. Branching Structures

Narratives are often taken to be appropriate tools to describe what happened, but not to explain *why* it happened. Beatty (2017) argues against this position, claiming that certain narratives manage to explain some present outcome by putting it against the background of what could have happened instead. To illustrate his idea, Beatty introduces the example of Mlle Amélie, the protagonist of Kate Chopin's story "Regret". At the age of fifty, Mlle Amélie comes to regret declining an old marriage proposal, as she realises that it meant missing the possibility of having children of her own. According to Beatty, this story has a structure that can be represented as in Fig. 1:

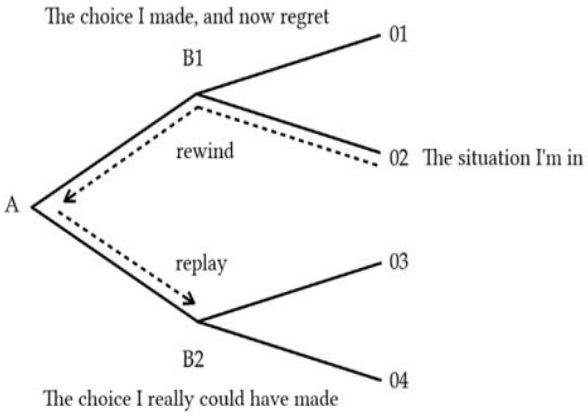


Fig. 1 – From Beatty (2017: 32).

In order to properly explain the situation 02 at which Mlle Amélie ended up being, 02 has to be put against the background of its alternative possibilities. To understand why Mlle Amélie feels regret at 02, we need to entertain the

thought that things really could have gone differently, and she really could have ended up at 03 or 04. The feeling of regret comes from a comparison between the present situation, which is the result of the choices made at some crucial node in the past at which other possibilities were really open, and the alternative outcome that could have resulted from following one of these other possibilities, and that is imagined to be better. In this way, narratives can create a branching structure that develops around the crucial nodes in the past that correspond to some difference-making events. Thus, they allow the reader to consider the ramifying possibilities in the past, and to explain the present as the path identified by what happened at the crucial branching nodes.

And here we arrive at the central point of my proposal. My suggestion is that the Dutch Book argument presents a structure very similar to Mlle Amélie's story, and that it is precisely this structure that provides the mechanism by which the argument gets normative force in the sense identified by (3).

Since this claim brings together two fields as foreign as narratology and decision theory, I will try to make it more plausible by proposing to look at the Dutch Book argument itself as a narrative, even in its barest version. After all, there is no need to be a fictional story to be a narrative. Many current accounts see narrativity as a spectrum, a property that different things can have more or less of (e.g. Ryan 2007, Currie 2010). Rather than by a specific definition, narratives are characterised by a set of typical features, none of which is neither necessary nor sufficient to identify a narrative. And the Dutch Book argument displays a remarkable set of such characteristic features: it presents an ordered series of events, some of which are purposeful actions carried out by intelligent agents, forming a chain and leading to a closure. Hence, even though it may not strike as stereotypically narrative, the Dutch Book still seems to present an interesting degree of narrativity. If this is so, then narratology may provide fruitful tools to investigate the mechanisms behind the Dutch Book.

In the Dutch Book, the agent is at an initial node, at which two different possibilities open: she can comply with *Probabilism* and have probabilistic credences, or she can violate it and have non-probabilistic credences. What the argument does is to show the outcomes of these possibilities, just as narratives do when employing Beatty's branching structure. Therefore, similarly to Beatty's reconstruction of Chopin's "Regret", the structure behind the Dutch Book argument can be schematised as in Fig. 2:

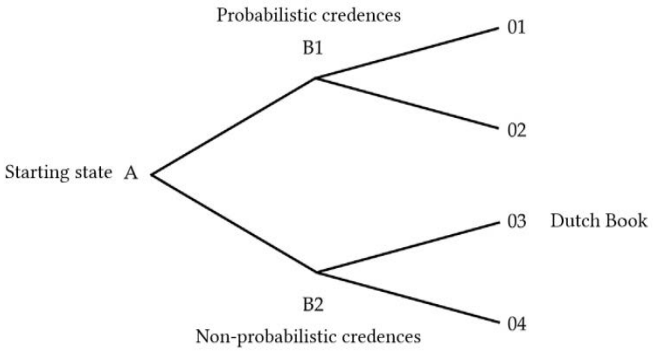


Fig. 2

There are, however, some important differences between the narrative structure theorised by Beatty and the one I propose to see behind the Dutch Book. First, in Mlle Amélie's story the crucial node at which the possibilities branch is situated in the past, while in the Dutch Book it is the starting point of the narrative. Consequently, Mlle Amélie compares the result of the actual course of events with a counterfactual outcome, the possibility of which is ruled out by the choice she made. On the other hand, the Dutch Book compares two equally open possibilities, neither of which has become the actual one yet. And finally, while the first narrative identifies the crucial nodes in the past to explain and understand Mlle Amélie current situation in light of what could have happened instead, the second one employs the branching structure to fulfil a normative function. It is now time to tackle more directly the question of how it manages to do so.

5. Comparing Outcomes

In Mlle Amélie's story, the explanation of the current state of regret is completed by the comparison of that current state with the counterfactual alternative outcomes. The branching structure contributes to the explanatory function of the narrative by allowing such comparison, thanks to the identification of a crucial node in the story from which different outcomes follow. The regret comes from the comparison, and thus the comparison is needed to explain it. Narratives can fulfil an explanatory function thanks to branching structures (Beatty 2017).

As I have argued, the Dutch Book argument presents a similar branching structure. However, if the structure is similar, the function is different: the Dutch Book argument is not intended to explain some current state of affairs. Instead, it is intended to provide normative support to *Probabilism.* But even though it intends to achieve a different goal, the Dutch Book argument exploits the branching structure for the same reason as Chopin's 'Regret': such structure permits the comparison between alternative outcomes of a single node. Thanks to the branching structure, the agent who finds herself at the starting position can see the outcomes of the options in front of her. Through their comparison, the agent can see that only compliance with *Probabilism* guarantees that she is safe from combinations of betting contracts where she would certainly lose money.

Let us now assume that the agent has the desire not to lose money, an assumption that should not strike as particularly controversial—especially on the background of the economic context in which the argument appears. Then, the Dutch Book argument effectively shows that one of the options in front of the agent serves that desire better than the alternative one, since she can see that having non-probabilistic credences would expose her to the risk of Dutch Book contracts. Therefore, the argument provides support for *Probabilism* precisely on the lines of (3):

(3) An argument provides normative support to option *o* IF it shows the agent that *o* serves her desire *d* better than the alternative options.

Applying (3) to our case, the Dutch Book argument provides normative support to *Probabilism* because it shows to the agent that *Probabilism* serves her desire not to lose money better than the alternative option, which would expose her to sure loss. If, as I have argued above, (3) is a viable way to attain normative force, then the Dutch Book argument attains normative force. The comparison of outcomes is the mechanism that provides it with its compelling appeal.

Just as a branching structure provides an explanation of Mlle Amélie's regret by putting the current state against its counterfactual alternatives, so another branching structure provides normative support to *Probabilism* by putting its outcomes against those of the alternative option. The two narratives share the same structure (with the due differences noted), use it to implement the same mechanism of outcome comparison, but exploit that mechanism to fulfil two different functions.

Even though the Rational Agent is a fictional character, its actions can carry normative force because the rationality of its properties is supported by normative arguments like the Dutch Book. The Dutch Book provides normative grounds because it provides reasons for the option it supports. It provides reasons because it shows that that option serves the desire not to lose money better than the alternative, and it does so thanks to a branching structure that links the different options to their outcomes.

Thus, we have seen how one argument provides normative grounds to a certain requirement of rationality on the lines of (3) thanks to its branching structure. But this mechanism is not idiosyncratic to the Dutch Book. Indeed, branching structures like the one illustrated convey normativity along (3) in many other arguments in the debate on normative rationality. *Money Pump* arguments in favour of Transitivity (Davidson et al. 1955) follow the Dutch Book structure quite closely, and a similar analysis can be applied there. Even arguments *against* a certain requirement can employ similar mechanisms. One interpretation of the famous *Allais Paradox* (Allais 1953), for instance, identifies precisely in the possibility of feeling regret the justification for the violation of the requirement of Independence (Loomes and Sugden 1982). Tracing (3), this interpretation contends that the Paradox provides normative support to the violation of Independence because it shows that the violation serves the agent's desire to be safe from regret better than the alternative option.[6]

If this is so, then outcome comparison seems to enjoy some degree of robustness as a mechanism for arguments in normative rationality. More specifically, it applies to both supporting and opposing arguments.

But in order to be able to compare different outcomes, the reader must be able to represent things as they are not. This requires the appeal to the representation capacity of imagination. In the next section, I will explore some interesting implications of the role played by imagination in the comparison of outcomes.

6. Imagining Outcomes

Liao and Gendler (2019) characterise the act of imagination as representation "without aiming at things as they actually, presently, and subjectively are". To explain Mlle Amélie's regret, the reader of the story must be able to represent things not only as they *actually* are (in the world of the story): she must also be able to represent things as they could have been. To grasp the normative stance of the Dutch Book argument, the reader must be able to represent things not only as they *presently* are (at the starting state): she must also be able to represent things as they would be, conditional on the direction taken at the starting node. The cognitive act of comparison involves representations coming from what

[6] I am grateful to an anonymous reviewer for suggesting this application.

Weinberg and Meskin (2006) call the "imagination box", i.e. the cognitive system responsible of the generation of imaginings.

Therefore, it seems that the mechanism by which the Dutch Book argument gets its normative force is grounded in the cognitive capacity of imagination. This opens new perspectives on the relevance of imagination for action.

Typically, one of the features that philosophers use to distinguish imagination from belief is that the former is somewhat disconnected from the action-guiding system (Currie and Ravencroft 2002, Kind 2013): your imagining a venomous snake in front of you will not cause the same reaction as your believing that there is a venomous snake. While beliefs guide your actions, imaginings do not (typically) do so. Nonetheless, the role that imagination is called to play in the comparison of outcomes points to two routes by which imagination can indeed contribute to action guiding.

First, since imagination is needed to root the normative force of arguments like the Dutch Book, then imagination is needed to provide legitimacy to prescriptions based on such arguments. If a certain course of action is advised on the basis of arguments grounded on the type of normative support described above, then the force of that indication requires the imaginative comparison of different outcomes. Thus, by providing the mechanism from which normative arguments draw their force, imagination generates compelling action-guiding prescriptions. In this normative dimension there is a connection between what is generated in the "imagination box" and action.

Second, the comparison of the outcomes of different options goes beyond normative purposes. According to the classical schema of decision-making, belief and desire are the only components mediating between sensory inputs and action outputs. This schema finds its counterparts in decision theory in terms of probabilities over possible states and utilities over possible outcomes. In this classical binary view, there is no obvious room for imagination. However, some authors claim that imagination does play a role, and that therefore this schema is inadequate. Van Leeuwen (2016) sees a role for imagination in the representation of possible states of the world and possible actions to take, which are needed to build the decision matrix required by decision theories. He does not, however, consider outcomes, which nonetheless need to be represented and inserted in a matrix. Nanay (2016) addresses this dimension more directly, as he sees a crucial component of decision-making in the agent imagining her future self in the imagined outcome.

However, it is important to note that the role of imagination is substantially different from that of belief and desire. While these motivate the agent's choices, imagination provides the mechanism that allows the agent to evaluate the situation and represent all its relevant dimensions. Imagination provides the background against which the agent can represent and compare different outcomes, and thus decide on one of them according to her beliefs and desires. If this is so, then the standard picture is preserved at the level of action motivation. But Van Leeewen and Nanay are right in claiming a role for imagination in decision-making. This role is to act as the cognitive mechanism allowing the representation of the decision problem and the comparison of the different outcomes yielded by the alternative options at hand, not unlike what happens in Mlle Amélie's story and in the Dutch Book argument. Thus, in allowing outcomes comparison in decision-making, imagination finds a further way to connect to action.

7. Conclusions

As any model, microeconomic models involve an array of assumptions. Among these, the rationality of agents plays an undoubtedly central role. In that context, rationality is a technically defined concept consisting of a list of properties that are embodied in the Rational Agent. As these properties are very unrealistic, the Rational Agent is a fictional character that cannot describe real human agents. However, the choices it makes within the models are typically supposed to be normative. But how can the actions of a fictional character bear any normative pull for real agents?

In order to answer this question, I have introduced the Rational Agent and its defining properties. The legitimacy of each property as a feature of rationality is supported by some arguments. But the mere existence of these arguments is not an answer to the question of the roots of normativity. And this is not because the arguments are debated, but because it only shifts the broader question to the narrower question of what makes such arguments normative. Contested as they may be, they have an undeniable compelling pull, or they would not even be discussed. The famous Dutch Book argument in support of probabilistic credences presents a good case: what makes it so compelling that it can function as a normative argument?

The search for the answer has consisted in three steps. First, I have proposed a way in which an argument can provide normative support, i.e. by showing that the option it supports serves some desire of the agent better than its alternatives. While I do not claim that this is the only one, I do claim that this is the type of normative support that the Dutch Book offers. Second, I have identified in the Dutch Book the same branching structure that Beatty (2017) identifies in some narratives. In both cases, the structure points to some crucial nodes at which different routes depart. But while in Beatty's examples the structure fulfils an explanatory function, in the Dutch Book case its function is normative. Third, I have argued that this branching structure permits the comparison of the outcomes resulting from the different options. In doing so, it makes it clear to the agent which option serves her desire best, thus providing normative support in the sense proposed. Interestingly, the Dutch Book is not a special case: similar mechanisms support other arguments in normative rationality. Finally, since the comparison of outcomes requires imagined representations, then this mechanism shows some interesting connections between imagination and action: not only does imagination root normative action guidance, but it also provides a necessary background for decision-making, thus enriching the standard binary belief-desire schema.

References

Allais, M. 1953, "Le comportement de l'hommerationneldevant le risque: critique des postulats et axiomes de l'écoleaméricaine", *Econometrica: Journal of the Econometric Society*, 21, 4, 503-56.

Allais, M. and Hagen, G.M. (eds.) 2013, *Expected Utility Hypotheses and the Allais Paradox: Contemporary Discussions of The Decisions Under Uncertainty with Allais' Rejoinder,* Dordrecht: Springer Science & Business Media.

Beatty, J. 2017, "Narrative Possibility and Narrative Explanation", *Studies in History and Philosophy of Science*, *Part A*, 62, 31-41.

Currie, G. 2010, *Narratives and Narrators: A Philosophy of Stories,* Oxford: Oxford University Press.

Currie, G. and Ravenscroft, I. 2002, *Recreative Minds: Imagination in Philosophy and Psychology*, Oxford: Oxford University Press.

Dancy, J. 2000, *Practical Reality,* Oxford: Oxford University Press.

Davidson, D., McKinsey, J.C.C. and Suppes, P. 1955, "Outlines of a Formal Theory of Value, I", *Philosophy of Science*, 22, 2, 140-60.

Finlay, S. and Schroeder, M. 2017, "Reasons for Action: Internal vs. External", in Zalta, E.N. (ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2017 Edition.

Goldman, A.H. 2009, *Reasons from Within: Desires and Values*, Oxford: Oxford University Press.

Hacking, I. 2001, *An Introduction to Probability and Inductive Logic*, Cambridge: Cambridge University Press.

Hands, D.W. 2015, "Normative Rational Choice Theory: Past, Present, and Future", https://ssrn.com/abstract=1738671 (July 2020).

Jeffrey, R.C. 1990, *The Logic of Decision*, Chicago: University of Chicago Press.

Liao, S. and Gendler, T. 2019, "Imagination", in Zalta, E.N. (ed.), *The Stanford Encyclopedia of Philosophy,* Spring 2019 Edition.

Loomes, G. and Sugden, R. 1982, "Regret Theory: An Alternative Theory of Rational Choice under Uncertainty", *The Economic Journal*, 92, 368, 805-24.

Kind, A. 2013, "The Heterogeneity of the Imagination", *Erkenntnis*, 78, 1, 141-59.

Morgan, M.S. 2006, "Economic Man as Model Man: Ideal Types, Idealization and Caricatures", *Journal of the History of Economic Thought*, 28, 1, 1-27.

Nanay, B. 2016, "The Role of Imagination in Decision-making", *Mind & Language*, 31, 1, 127-43.

Peterson, M. 2017, *An Introduction to Decision Theory*, Cambridge: Cambridge University Press.

Ryan, M.L. 2007, "Toward a Definition of Narrative", in Herman, D. (ed.), *The Cambridge Companion to Narrative,* Cambridge: Cambridge University Press, 22-35.

Savage, L.J. 1954, *The Foundations of Statistics,* Hoboken, NJ: John Wiley and Sons.

Scanlon, T.M. 1998, *What We Owe to Each Other,* Cambridge, MA: Belknap Press of Harvard University Press.

Schroeder, M. 2008, "Having Reasons", *Philosophical Studies*, 139, 57-71.

Van Leeuwen, N. 2016, "Imagination and Action", in Kind, A. (ed.), *The Routledge Handbook of Philosophy of Imagination,* London and New York: Routledge, 306-19.

von Neumann, J. and Morgenstern, O. 1944, *Theory of Games and Economic Behaviour,* Princeton: Princeton University Press.

Weinberg, J. and Meskin, A. 2006, "Puzzling over the Imagination: Philosophical Problems, Architectural Solutions", in Nichols, S. (ed.), *The Architecture of the Imagination: New Essays on Pretence, Possibility, and Fiction*, Oxford: Oxford University Press, 175-202.

Williams, B.A.O. 1979, "Internal and External Reasons", reprinted in his *Moral Luck*, Cambridge: Cambridge University Press, 1981, 101-13.

Advisory Board

Book Reviews

De Florio, Ciro & Frigerio, Aldo, *Divine Omniscience and Human Free Will: A Logical and Metaphysical Analysis*.
London: Palgrave Frontiers in Philosophy of Religion, 2019, pp. x + 268.

The logical tension between the doctrines of divine omniscience and human freedom has been studied and discussed for centuries. Is there a logical conflict between these two Christian doctrines, or can we somehow maintain both in a logically consistent manner? In other words, is it meaningful to claim that an agent can choose freely between alternative options if we assume that God already now knows what the agent is going to choose? Over the years, an enormous number of books and papers have been published on this topic. Some would probably assume that it is unlikely that anybody could add anything new of value to this story. However, this is exactly what Ciro De Florio and Aldo Frigerio (both staff members in the Department of Philosophy, Università Cattolica del Sacro Cuore, Milan, Italy) have done. In their book on the tension between the two doctrines, they offer not only a very interesting logical and metaphysical analysis of the classical problems but also important new insights on the topics. There can be no doubt that this book will be extremely helpful to anyone who wants to study these topics in a systematic manner.

Although the book deals with problems that are relevant in theology, the authors make it clear that this is not a book of theology and that they make no presuppositions of faith in it. The book should be seen as "a book of philosophy of religion, which is the rational investigation on the content of religious beliefs" (viii). The book is "dedicated to the logic-metaphysical analysis of the problem of theological fatalism" (2). The basic concepts and ideas of this classical discussion are presented in chapter 1 of the book, "The Battle for Free Will".

In the book, the authors make use of modern tense-logic, which was first introduced by A.N. Prior (1914–69) and further developed by several writers working in the Priorean tradition. This means that the authors formalize the claims in question in terms of Prior's propositional operators, *P* ("it has been the case that …"), *F* ("it will be the case that …"), *H* ("it has always been the case that …"), and *G* ("it will always be the case that …"). Furthermore, they make use of branching time models and basic ideas of formal semantics. In chapter 2, "Metaphysics and Logic of Time", the authors carefully present the formal concepts used in the current analysis of the topics in question. With their work, the authors offer a very strong case for the use of tense-logic as a powerful formal tool in the analysis of the problem of theological fatalism. In fact, the authors claim that the metaphysics of time characterized through systems of temporal logic "is not merely tangent to the foreknowledge dilemma but, quite the opposite, it is an essential part" (261). The authors are clearly right. Having the tense-logical formalism available makes it possible to formulate important distinctions that would be very hard to present without this formal tool. In this way, the use of temporal logic (and, in particular, tense-logic) defines an approach or perhaps even a paradigm for the study of the topics related to the problem of theological fatalism.

In chapters 3-6, the authors examine the responses to the problem of theological fatalism that are currently the most important. The authors carry out this

Argumenta 6, 1 (2020): 149-163
ISSN 2465-2334

DOI 10.14275/2465-2334/202011.boo

task very carefully, making use of conceptual analysis and the methods of temporal logic and formal semantics.

In chapter 3, "Extreme Measures", the authors consider two types of response to the logical tension or apparent conflict between the doctrines of divine omniscience and human freedom. In each of the two cases, the response depends on a reinterpretation or redefinition of one of the two key concepts involved in the problem, the concept of divine omniscience and the concept of free will.

Open Theism is a response that goes back to Prior, which he termed the Peircean solution. According to this view, future contingents cannot be true now. This means that there is no true statement about what a person is going to do freely tomorrow. In consequence, God cannot know today what a person is going to choose freely tomorrow. If this view is accepted, there is no conflict between the doctrines of divine omniscience and human freedom. Critics of this response point out that this is a very weak and rather unusual understanding of divine omniscience. However, the authors find that Open Theism is formally consistent. In fact, they point out that the difficulties of the view are "more theological than philosophical". They ask, "Is the concept of God advocated by open theists really in accordance with the God of the Bible?" (92).

Theological Determinism involves a redefinition of free will that denies what the authors call the Principle of Alternate Possibilities (19): "If you cannot do otherwise when you do an act, you do not do it freely". In this way, the theological determinist proposes a concept of free will compatible with God's full sovereignty over the universe. The authors argue that this is indeed a rather weak notion of freedom and is far from the idea of libertarian freedom (95) that most people refer to when they speak about free choice.

In chapter 4, "God Knows the True Future: Ockhamism", the authors deal with the other famous response presented by Prior. This is a solution inspired by scholastic logician and philosopher William of Ockham (1285–1347). Like Prior, the authors use a formalization of Ockham's position in terms of tense-logic and branching time models. Like in Prior's first formalization of Ockham's ideas, they include the notion of the true future corresponding to the detailed divine omniscience.

Ockham held that the combination of the doctrines of divine foreknowledge and human freedom does not lead to any contradiction. His way out of the problem of theological fatalism was to deny, at least in the most general sense, the principle called the necessity of the past: "If an event e occurred in the past, then it is accidentally necessary that e occurred then" (121). This means that pastness does not generally imply necessary pastness. In symbols: $Pq \supset \Box Pq$, where the operator $\Box$ stands for necessity (or as Prior would put it, "now-upreventability").

It is well-known that the Ockhamist has to specify the cases in which $Pq \supset \Box Pq$ does not hold. In the book, the authors use a number of illustrative examples referring to Emma and Thomas (the children of Ciro De Florio). For instance, let's assume that Emma has been invited to a party that is going to take place tomorrow (and only once). Emma is considering going to the party but decides not to. If p stands for "Emma is at the party", then $\sim p$ will be the case tomorrow and in fact at any other time as well. This means that in the past (e.g. yesterday), it was the case that she would never go to this specific party, so $PG\sim p$ also has to be true now. If $Pq \supset \Box Pq$ is accepted in general, no matter what

q stands for, then it follows from standard tense- and modal logic by a little deduction that □~*Fp*. If so, it would not only be the case that Emma is not going to the party, but it would be necessary for her to stay away from the party (and impossible for her to go to it). This is clearly a conclusion that we want to avoid (given that we want to insist on indeterminism). The only way out is to make sure that *PG~p* does not become necessary just because it is true. In fact, *PG~p* is what the authors call "a semantic soft fact", since the truth of the proposition fully "depends on what the agents will choose at a later time" (126). The proposition *PG~p* is not really about a past event, and this means that on the Ockhamistic view the necessity of the past does not apply here. Consequently, there is no reason to hold that this proposition is now necessary.

Whereas semantic soft facts may be seen as "innocuous", the authors hold that "things become more complex when one passes from semantic soft facts to epistemic soft facts" (127), i.e. when we consider a modification of the above proposition, namely *PKG~p*, where *K* is an operator that stands for "God knows that". The key question here is, of course, whether there is a proper difference between (a) "yesterday, it was true that Emma would never go to the party" and (b) "yesterday, God knew that Emma would never go to the party". If the answer is no, (b) will be just as "innocuous" as (a), which means that on the Ockhamistic view necessity of the past does not apply here. If the answer is yes, we have to account for the logical properties of the operator *K* in order to deal with the problem in a satisfactory manner. In this case, it is an open question whether the necessity of the past should apply. No matter what, it is obvious that the Ockhamistic denial of the principle of the necessity of the past (121) leads to a number of conceptual challenges and, in this sense, to some considerable costs. Clearly, in the example used here, *KG~p* would be true yesterday, but if Emma had in fact attended the party, *K~G~p* (equivalent with *KFp*) would have been true yesterday. For this reason, it appears that Emma can influence the past, at least in sense that her going to the party would have made God know yesterday that she was going to be at the party. It could perhaps be maintained that this should be seen as a kind of backwards causation. The authors have nicely illustrated this problem using their so-called "butterfly schema" (129).

In chapter 5, "Molinism", the authors consider another response to the main problem of the doctrines of divine omniscience and human freedom. This solution was formulated under the inspiration of the works of Luis de Molina (1535–1600). According to Molina's view, God knows not only what any agent is going to do freely at any future time but also what any agent in any counterfactual situation would freely choose. Unlike the Peircean solution (and Open Theism) discussed in chapter 3 and unlike the Ockhamism discussed in chapter 4, this is not a solution that Prior studied. The first attempts at formalizing Molina's approach in terms of temporal logic were carried out in the late 1990s, mainly in response to an analysis published in the important paper, "Indeterminism and the Thin Red Line" by Nuel Belnap and Michael Green.[1] In their paper, Belnap and Green introduced the term "the thin red line" (abbreviated TRL) as a name of the chronicle in a branching time diagram corresponding to the Ockhamistic true future. In fact, Belnap and Green tried to show that the acceptance of "the thin red line" in a system would make the system deterministic and make the representation of time linear instead of branching. In order to es-

[1] *Philosophical Perspectives*, 8, 1994, 365-88.

tablish their conclusion, Belnap and Green argued rather convincingly that any defender of "the thin red line" would also have to accept "a thin red line" through any counterfactual moment in the branching time diagram. Belnap and Green argued that this additional property would make the whole branching time system collapse into a linear structure. This was later shown to be wrong, and Belnap and Green have admitted their mistake. It is in fact possible to construct a consistent model, TRL+, in which the property in question holds and which can be seen as a nice formalization of Molinism. In their book, De Florio and Frigerio discuss the properties and problems of the TRL+ model (162 ff.). They point out that the system should be seen as an enriched form of the Ockhamistic framework. Molinism, however, has to pay some rather high theoretical costs. One problem seems to be that given the obvious semantics of the TRL+ model, $p \supset HFp$ (so-called retrogradation) will not be a valid thesis. The Molinist can, of course, choose to accept this invalidity and argue that for some reason, retrogradation will not be reasonable in all cases. However, it will probably be even more interesting to Molinists to find that De Florio and Frigerio have offered a modified semantical model for TRL+ that should be satisfactory for Molinists and that validates the principle of retrogradation (see 64 ff. & 244 ff.).

One other problem regarding Molinism and the TRL+ model in particular has to do with grounding. How can a claim regarding what an agent would freely choose in some counterfactual situation ever be true? What could make such a claim true? The authors are quite right that the Molinists have to be ready to pay a remarkable theoretical cost if they insist that certain aspects of reality would make such counterfactuals true. However, William Lane Craig has proposed "a complete liberalization of grounding" according to which "any proposition p is grounded on the fact that p" (see 185 ff.). This solution can of course be further discussed, but at least formally it solves the problem and may in fact be the best way out for the Molinist.

In chapter 6, "The Timeless Solution", the authors consider the classical solutions to the dilemma of omniscience based on the Timeless Eternalist view and the B-theory of time. A rather complex discussion for and against this view has been ongoing for years. The authors offer a very informed discussion of this philosophical and theological debate, taking the views of the key debaters like Stump, Kretzmann, Plantinga, Craig, and Rogers into account.

The authors do not claim to have solved the problems related to the logical tension between the two doctrines. However, they do offer an original and very interesting contribution, so-called Perspectival Fragmentalism, which is partly inspired by the works of Kit Fine, a former student of Prior. With their perspectival semantics, the authors want to extend the notion of "truth at a moment" to "truth at a moment from a given perspective". Although some aspects and details of this original contribution ought to be discussed and developed further, there is obviously much inspiration to find in this suggestion. This interesting idea is likely to generate further analysis and deeper investigation of the problem of theological fatalism.

I would highly recommend this book to anyone interested in the logical analysis of the problems of divine omniscience and human freedom.

Aalborg University, Denmark PETER ØHRSTRØM

Giombini, Lisa, *Musical Ontology: A Guide for the Perplexed*.
Milano: Mimesis International, 2017, pp. 374.

Music is probably the most common artistic experience in our everyday lives. It impacts our daily reality in so many different ways, that it is rare to find a person who has never dedicated some thoughts to it. As Kania[1] noticed, it is really unlikely that even people without a specific theoretic and philosophical background do not have personal views or intuitions about music. It is natural, then, that music has generated intense philosophical discussions about its features, elements, and its nature. It has also led to the creation of dense but enlightening books such as the one I am going to analyse here.

The philosophy of music is currently an interesting and vast field of philosophical speculation, within which a particularly broad debate has flourished around questions concerning the metaphysical nature of musical pieces.[2] The identification of essential features of musical works, the existing relations between performances and scores, and the ways in which music occupies space and time, are just some of the core queries that have risen philosophical interest.

Imagine having in front of you a score of Beethoven's 5th symphony. Some questions might come naturally to mind: do you perceive this piece of music visually? What is the relation between that piece of paper and a performance of the same symphony? In virtue of what kind of properties do we consider a certain entity *that* specific symphony?

Broadly speaking, two reactions are possible. On one side, we might take these questions as genuine ontological questions, and proceed to explore them further. Indeed, many philosophers[3] are attracted by the idea of explaining what musical pieces are, aiming to individuate the essential properties of musical works, and to understand the type of relations existing between performances, recordings and music transcriptions. On the other side, as the reader can probably imagine, a certain degree of skepticism arises about the meaningfulness of this metaphysical debate. Indeed, what is the impact that answers to those metaphysical questions can have on our understanding and appreciation of listening to music? Does the metaphysical debate really tell us something about music or our experience of it? Is it interesting for artistic reasons? These queries have led some philosophers to think about second-order questions concerning the goals and methods appropriate to the philosophy of music.

In her book, Lisa Giombini shows how, in order to have a better guide to choose among the different first-order ontological positions, it is necessary to have a clear idea about the second-order debate on meta-philosophical questions. The author proposes an original and articulated meta-philosophical view, comparing and contrasting her position with the main objections developed

[1] Kania, A. 2017, "Philosophy of Music", in E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2017 Edition.

[2] Here, I follow the author in the intention to not draw a distinction between ontology and metaphysics in this specific context.

[3] See Kania 2017 (mentioned in note 1), chapter 2.2., for a general overview or part 1 of Giombini's book for a broader analysis of this debate. See also Davies, S. 2003, "Ontologies of Musical Works", in Davies, S., *Themes in the Philosophy of Music*, Oxford, Oxford University Press; Kivy, P. 1983, "Platonism in Music: A Kind of Defence", *Grazer Philosophische Studien*, 19, 109-29; Levinson, J. 1980, "What a Musical Work is", *Journal of Philosophy*, 77, 1, 5-28.

against musical ontology. Specifically, she tries to dismiss part of the original debate assuming an "*halfway weak* ontological position" (201), to argue for an "*historical ontology*" (237) and to defend the use of a method of "*reflective equilibrium*" (278) to choose the relevant intuitions about music that have to be preserved in the debate.

The substantial second part of her book is exactly intended to show the criticisms that are moved against musical ontology and to reply to them highlighting her personal position. In chapter 4, after pointing out the difficulties generated by the interaction between metaphysics and aesthetics, Giombini introduces an overview of skeptical positions about music ontology, providing a schema (150) to present the four main adversary views that she calls: *Eliminativism, Aestheticism, Historicism* and *Semanticism*. The common idea shared by all these positions is that the domains of aesthetics and metaphysics have to be considered separately. In the following chapters (from 5 to 8), Giombini critically explores the four approaches.

In the metaphysical debate, *Eliminativists* suggest that there is no reason to create an ontology of everyday life objects, artifacts and human creations generally speaking. Specifically, some philosophers, such as van Inwagen[4] and Unger,[5] support the idea that ordinary objects do not exist. As an example of a similar move, Giombini mentions Cameron, who applies this same eliminativist strategy to the case of musical works.[6]

Then, Giombini argues that the principle of simplicity (Ockham's Razor) is not enough to justify a complete elimination of ordinary objects and art works in our metaphysics, and to reduce everything to a more fundamental level of reality. Indeed, such a move, she argues, would have led us to a misrepresentation of the world forcing us to embrace a type of physical reductionism or a form of semanticism. In this way, Giombini refuses to completely abandon musical ontology as promising field of investigations.

Aestheticism suggests that ontological investigations cannot cast any lights on the understanding and appreciation of arts. Indeed, according to this view, what has to be explained in artistic contexts is merely the aesthetic value of art works and experiences. Giombini identifies Ridley's position as the paradigmatic version of this view, and goes on presenting his major argument against musical ontology.[7] Ridley's main idea is that musical ontology does not provide any interesting outcome for musical appreciation or action, so it has to be considered useless and has to be abandoned.

It is in reaction to this argument that the author begins to build her own view. Giombini argues that the ontological debate on music seems pointless just when philosophers forget "real musical activities" (190) and do not consider them in the construction of their theoretic frameworks. She suggests that the uninteresting part of the debate on musical ontology concerns what, following

[4] Van Inwagen, P. 1990, *Material Beings*, London: Cornell University Press.
[5] Unger, P. 1979, "There Are No Ordinary Things", *Synthese*, 41, 2, 117-54.
[6] Cameron, R.P. 2008, "There Are No Things That Are Musical Works", *British Journal of Aesthetics,* 48, 3, 295-314.
[7] Ridley, A. 2003, "Against Musical Ontology", *The Journal of Philosophy*, 100, 4, 203-20; Ridley, A. 2004, *The Philosophy of Music*, Edinburgh: Edinburgh University Press.

Dodd (2008),[8] she identifies in the first part of her book as the "*categorical question*", namely the tentative to recognize what kind of entities musical works are. Indeed, Giombini says that considering a piece of music as a tridimensional or quadridimensional entity fails to provide any information on music as an artistic experience or as a practice, even if it can give us some insights on the general metaphysical debate. On the contrary, answers to the question called by the author the "*identity question*" seem to be extremely relevant for aesthetic purposes: being able to identify a certain piece of music as *that* specific piece says something about the notion of authenticity. Following Giombini's argument, we end up no longer dealing with an evaluation of a performance of a piece of music (good or bad), as Ridely wanted. Instead, we are investigating what can be considered as a proper recording/performance/transcription of a specific piece. With this move, Giombini shows how a "*halfway weak* ontological position" (201) can help in addressing issues related to art works in a sense that is relevant for art criticism and artistic practice. Furthermore, she demonstrates that philosophy can provide more than just a guide for evaluative judgments.

Giombini defines as *Historicists* theorists who want to analyse music considering it as the result of a social, cultural and historical context. These authors are generally unsatisfied by the philosophical approach because, following a scientific methodology, it tends to consider musical works as independent entities not related to the historical and cultural frameworks where they were developed. Goehr's and Bourdieu's works stand as the main contributions supporting this view.[9] Goehr suggests that historical analysis is necessary in the context of studies about musical phenomena. In her analysis, Giombini firstly notes how favourably Goehr considers works by continental philosophers such as Nietzsche and Foucault on the notion of genealogy. Secondly, Giombini explains how Goehr employs this philosophical notion, constructing a genealogical theory of the concept of musical work. Furthermore, Goehr wants to demonstrate that the analytic debate has overgeneralised one specific conception of the art work, namely, the one that appears in the 18th century and through which we can describe the paradigmatic case of Beethoven's 5th symphony. On the other side, Bourdieu, following a different approach, points out how both music creations and also their appreciation are the result of social processes that are, with their dynamics, the generators of history. Giombini highlights how both these views reject the analytic approach because they claim that it applies a form of "*scientism*" to art (216).

In the middle of chapter 7, Giombini shows three different versions of Essentialism, namely theories that try to individuate the essential properties that works of art possess in virtue of being works of art. The essentialist approach is clearly antithetical to historicism, because it tries to define properties and features regardless of historical, cultural and social factors. However, Giombini rejects this set of positions because the nature of the answers that they can provide

[8] Dodd, J. 2008, "Musical Works: Ontology and Meta-Ontology", *Philosophy Compass*, 3, 6, 1113-14.

[9] Bourdieu, P. 1984, *Distinction: A Social Critique of the Judgment of Taste*, Translated by Richard Nice, Cambridge, MA: Harvard University Press; Bourdieu, P. 1989, "The Historical Genesis of a Pure Aesthetic", in Shusterman, R. (ed.), *Analytic Aesthetics*, New York: Blackwell, 147-60; Goehr, L. 2007, *The Imaginary Museum of Musical Works*, Oxford: Oxford University Press.

is too general. In her opinion, essentialism fails to give useful information for philosophical investigation, due to its tendency to create "universal generalization[s]" (232).

At the end of this section, Giombini tries to propose a personal synthesis between ontology and *historicism* that she calls *Historical Ontology*. In her words:

> an ontological-historical approach to music would address musical phenomena […] to explore their appearing and disappearing as object of theoretical and critical inquiry (238-39).

The main proposal expressed here is to study musical objects as objects of intellectual and artistic investigations. This approach, she argues, will allow philosophers to avoid putting in the same category pieces that come from different traditions (e.g. pieces from Western tonal tradition and jazz improvisations) and it will also provide a more nuanced picture of music.

Finally, the last category of adversary views analysed is called *Semanticism*. Approaches that fall in this category consider the questions that music ontology tries to solve as issues originated by language and the meaning of words, rather than genuine enquiries related to what exists in the world. Thus, metaphysical questions have to be addressed trying to clarify the terminology and concepts pertaining to specific artistic terms. Thomasson's work exemplifies this approach.[10] On one side, she suggests that musical ontology has to follow our commonsensical understanding of works of art. On the other, in her view, the investigation of ontological issues has to be carried out through the *conceptual analysis* of the linguistic practice that involves the vocabulary related to music entities. However, as Giombini notices in the section of the chapter dedicated to criticism to semanticism, linguistic practices are not constant, and neither are the beliefs related to them. There is a risk to fall into cultural relativism, where entities are influenced by historical and spatial contexts. Furthermore, a worse problem arises from the role of intuitions in being the relevant ground for beliefs and practices. Indeed, intuitions are usually conflicting and contradictory, so if we consider them the warranty of a certain practice and consequently of a certain reality, then we end up with a theory constructed over an inconsistent basis.

To tackle this issue, Giombini proposes to rely on a strategy to individuate consistent and relevant intuitions to take into account just the "right" ones. Thus, she suggests to employ Rawls' *reflective equilibrium* methodology, namely the practice of considering just those intuitions that constitute a rational and coherent framework.[11]

What emerges from these five chapters is Giombini's personal meta-ontological view, where she partially absorbs some of the critics against music ontology. Overall, she argues that an adequate ontology should 1) address the identity question, 2) take into account the historical dimension of changes of the relevant concepts, and 3) describe the parallel between the concepts and objects in the domain.

[10] Thomasson, A.L. 2007, "Artifacts and Human Concepts", in Laurence, S., Margolis, E. (eds.), *Creation of the Mind: Theories of Artifacts and Their Representation,* Oxford: Oxford University Press; Thomasson, A.L. 2005, "The Ontology of Art and Knowledge in Aesthetic", *Journal of Aesthetics and Art Criticism*, 63, 3, 221-29.

[11] Rawls, J. 1971, *A Theory of Justice*, Cambridge, MA: Harvard University Press.

In chapter 9, Giombini rejects the debate between realist and antirealist approaches on meta-ontology of art and music. Here, she basically enlarges her position, showing how her view can dismiss the dichotomy between realist and antirealist theories, switching the focus of the debate to a "*deontological*" (304) point.

> What is worthwhile is how artistic phenomena, events and products are transformed into objects of aesthetic appreciation and philosophical consideration and the way in which they take the form of ontological entities (304).

The book ends with a reflection on the concept of an art work, applying the methodological approach developed in the previous chapters. In the conclusion, Giombini makes a general point about the difference among works of art and art phenomena, highlighting how *historical ontology* provides interesting information about art reality.

This review was mainly intended to present in some detail the second section of this *Guide for the perplexed*. My purpose was to stress the interesting original position drawn by Giombini on the meta-ontological issues discussed. Indeed, her nuanced view, on what she defines as the *second order* of ontological queries generated by music ontology, sounds appealing and able to raise the curiosity of the reader.

The first part of the book should also be recommended. Indeed, the first three chapters, preceded by a detailed introduction with an enlightening musical example that guides the reader throughout the whole book, constitutes a clear and systematic presentation of the main positions in the complex ontological debate on music. The first part of the book is basically a short and clear handbook, useful both for someone who is approaching this debate for the first time and for whoever is familiar with the vast literature and is looking for an overall picture of the controversies discussed.

University of Warwick GIULIA LORENZI

McGowan, Mary Kate, *Just Words: On Speech and Hidden Harm*.
Oxford: Oxford University Press, 2019, pp. xi + 209.

There has been a joint effort lately among philosophers, political theorists, and legal scholars to show that speech plays a major role in enacting and bolstering unjust social hierarchies, and that we should pay more attention to linguistic considerations in our attempts to disentangle and resist identity-based disadvantage. Mary Kate McGowan's *Just Words: On Speech and Hidden Harm* is a pivotal contribution to this area.[1] McGowan's central claim is that offhand racist, sexist, or otherwise bigoted remarks impact on the normative landscape in ways that are detrimental to the social standing of certain groups of people (e.g. black people, women), and thus *constitute*, as opposed to merely *cause*, harm. The

[1] Supporters of (what McGowan calls) the "linguistic approach to group-based injustice" (4) include, e.g., Rae Langton, Catharine MacKinnon, and Lynne Tirrell. See Langton, R. 1993, "Speech Acts and Unspeakable Acts", *Philosophy and Public Affairs*, 22, 4, 292-330; MacKinnon, C. 1993, *Only Words*, Cambridge, MA: Harvard University Press; Tirrell, L. 2012, "Genocidal Language Games", in I. Maitra & M.K. McGowan (eds.), *Speech and Harm: Controversies over Free Speech*, Oxford: Oxford University Press, 174-221.

book is divided into two parts. The first part (Chs. 1-4) identifies and argues for a distinctive, covert mechanism by which speech enacts norms that shift the boundaries of what is locally permitted. The second part applies this theoretical apparatus to a series of examples—sexist remarks (Ch. 5), pornography-involving actions (Ch. 6), and public racist speech (Ch. 7)—to demonstrate that everyday verbal bigotry enacts norms that harm people along group lines, and that it does so even when the speaker has no intention of doing so and no special authority. The book closes with a glimmer of hope: the norm-enacting role of speech can be put to use to enact beneficial, rather than harmful, norms and promote egalitarian behaviors and habits (Conclusion).

Just Words forms part of a broader project in contemporary philosophy of language aimed to reinterpret and adjust conceptual tools to incorporate in the discipline the necessary resources to understand speech in a non-ideal, messy world.[2] While traditional accounts of linguistic interactions tend to abstract away from many aspects of a communicative situation to get simple and formalizable models, McGowan's contribution admirably deals with the complexities of real-life conversations, thus offering a more faithful picture of how language concretely works. Because of this, her proposal is quite detailed and difficult to summarize in a few lines. In this review, we first provide a sketch of McGowan's account of covert norm enactment, and then critically focus on her notion of harm constitution.

Speech, says McGowan, enacts norms in two different ways (Chs. 2 to 4). Suppose that, in the context of enacting a new city policy, the mayor of Milan declares, "Smoking is no longer permitted in any city building". This is a 'Standard Exercitive' (20)—a speech act that changes what is permissible in a given context via an exercise of speaker authority. Standard Exercitives enact norms *overtly*: their locutionary content precisely matches the content of the norm(s) they enact. Now suppose that Juan and Stella are discussing their respective cars when Stella says, "My car is so run-down that it's just not worth fixing. I'm afraid I have no choice but to get rid of the car". By bringing up her car, Stella makes it the most salient car in that context, thus enacting a norm about how the phrase 'the car' is to be used in the ensuing conversation. From then on, and until salience facts change again, it will be appropriate for both parties to use 'the car' to refer to Stella's car only. Such a norm is enacted *covertly*: the content of the locution does *not* match the content of the norm (roughly, "Currently, the only referent for the expression 'the car' is Stella's car"). Stella's utterance is a 'Conversational Exercitive' (27)—a non-authoritative act that changes what is permissible in a given conversation solely in virtue of adjusting the conversational 'score'.[3] Since the score tracks all those elements that together determine what counts as correct or otherwise acceptable in a given conversation, adjusting the score therewith changes how conversational participants may or may not act. Since every conversational contribution adjusts the score in multiple ways, adding to a conversation enacts norms for that conversation. Going back to our example, one way in which Stella's move adjusts the score is by intervening on its salience component. Her contribution raises the salience of her

[2] Beaver, D., Stanley, J. 2019, "Toward a Non-Ideal Philosophy of Language", *Graduate Faculty Philosophy Journal: The New School for Social Research*, 39, 2, 503-47.

[3] The notion of score is borrowed from David Lewis. See Lewis, D. 1979, "Scorekeeping in a Language Game", *Journal of Philosophical Logic*, 8, 3, 339-59.

car, and hence makes it the proper referent of the definite description 'the car'. The salience shift, and the consequent adjustment of what is conversationally permissible, is disclosed by the fact that if Juan went on using 'the car' to refer to *his* car without signaling that salience facts have changed again, this would result in confusion. Stella might step in with something like, "Wait a minute. Which car are we talking about?", flagging Juan's breach of a conversational norm.

Like any conversational move, everyday bigoted remarks covertly shift the normative context they occur in. McGowan's central example is a telling case of ordinary sexism (Ch. 5). The case goes like this: Steve and John are co-workers at a workplace in the US. The following exchange takes place in the employee lounge:

> JOHN: So, Steve, how did it go last night?
> STEVE: I banged the bitch.
> JOHN: [smiling] She got a sister? (110).

Steve's utterance enacts a number of norms, e.g. it makes a certain woman the most salient and thus the proper referent of the pronoun 'she'. Crucially, it also enacts norms that make it permissible, in that immediate environment, to degrade women—for instance, to verbally derogate or sexually objectify them. By doing so, it "makes women count as second-class citizens (locally and for the time being)" (112). Somewhat surprisingly, however, McGowan goes on to claim that the enactment of such norms is not enough for Steve's utterance to *constitute* harm. A further requirement is needed—namely, people must exploit the permission they are given. *If* those norms are actually followed, and women are actually discriminated against, *then* (and only then) Steve's utterance constitutes the harm of gender discrimination.

McGowan's notion of harm constitution has a built-in causal element. For an utterance to constitute harm, three conditions must be met: (i) the utterance enacts a norm that prescribes some harmful behaviors; (ii) that norm is followed; and (iii) harm results from following it (24). Constituting harm is, in this view, a special, norm-driven way of causing it. Thus, to say that Steve's utterance constitutes gender discrimination is not to say that his utterance is contemporaneous with the discriminatory harm or that it is sufficient for that harm. Rather, the harm is *causally downstream* from his utterance: for the harm to obtain, others must follow the norms the utterance has enacted.

In the remainder of this review, we question the tenability of McGowan's causal account of harm constitution and tentatively suggest an alternative. Before getting to that, it is useful to illustrate McGowan's way of couching the constitution-causation divide. Consider the following examples.

> *Bigoted CEO*
> Julia, the CEO of a shoe company in Hawkins, is in a meeting with her HR team when she says, "From now on, we no longer hire Italians". Julia's utterance enacts a 'No Italian' hiring policy for her company. In adherence with it, her HR team starts to trash incoming job applications from Italian candidates.

> *Bigoted Employee*
> Jeff is a low-level employee at Julia's company and a very good friend of Mark's, the HR manager. Jeff keeps telling Mark how Italians are slackers and a blight

> on the company's business. As a result of coming to believe these things, Mark starts to trash incoming job applications from Italian candidates.

Jeff, the bigoted employee, manages to (verbally) persuade Mark that it is in the company's best interest not to hire Italians, and because of this, Mark and his team stop hiring Italians. The connection between Jeff's words and the ensuing discriminatory hiring practice is *merely causal*. Julia's case is importantly different. Her utterance causes the same discriminatory conduct on the part of the HR team as Jeff's utterance, but Julia's does so via the enacting of a norm (or policy) prescribing that conduct. As such, Julia's utterance *constitutes* harm (precisely, the harm of anti-Italian discrimination). So, in McGowan's view, the difference between constituting and (merely) causing harm lies in the *means* by which the harm is brought about. Speech constitutes harm if it causes harm via the enacting of a norm that prescribes that harm; speech merely causes harm if it brings that harm about in some other way (e.g. via persuasion) (23).

With this in mind, we can now turn to the controversial aspects of McGowan's causal understanding of constitution. Consider a few alternative endings to *Bigoted CEO*.

> *No Italian Around*
> The Italian community in Hawkins moves out of town for unrelated reasons right after the enacting of the 'No Italian' policy. No Italian ever applies for a job position at the company.

> *Company Bankruptcy*
> Shortly after the enacting of the 'No Italian' policy, the Internal Revenue Service shuts down the company for insolvency. No Italian had happened to apply for a job position there in the meantime.

> *Disobedient HR*
> Mark, the company's HR manager, finds the 'No Italian' policy outrageous. He therefore continues to consider Italian candidates' applications, and since he is authorized to sign job contracts on behalf of the company, he continues to hire Italians if they deserve it.

In *No Italian Around* and *Company Bankruptcy*, the 'No Italian' policy has no applications; *a fortiori*, it cannot be followed and no discriminatory hiring practice ensues. In *Disobedient HR*, the 'No Italian' policy is breached and no actual discriminatory hiring practice follows. Although Julia's utterance successfully enacts a 'No Italian' policy, by McGowan's line of thought, it would *not* constitute discrimination in any of the three ending scenarios. That is, it would not be discriminatory—which strongly runs counter to our intuitions. The same line of reasoning applies to another, perhaps more vivid, example. Imagine that a 'Whites Only' sign is hung on a pub's front door. In a scenario in which, for purely idiosyncratic reasons, no black person ever happens to walk past the pub or to try to get a seat there, the 'Whites Only' sign would *not* constitute discrimination—which, again, seems just wrong.

To avoid such problematic results, we suggest that McGowan's causal account of constitution be shifted in a counterfactual direction, so that for an utterance to constitute harm, only two conditions are required:

(i) the utterance enacts a norm that makes harmful behaviors permitted (or even mandatory);
(ii) if the norm were followed, then harm would result from following it.[4]

A counterfactual account of constitution like the one we have just sketched has advantages over a causal account. First, it has bigger explanatory powers: those utterances that the causal approach unsatisfactorily leaves out are properly numbered among harmful norm enactments. Second, under a counterfactual view, whether an utterance constitutes harm does not depend upon whether some specific individuals actually happen to suffer concrete disadvantages. This is, we think, the right result: constituting harm doesn't seem to be (and perhaps, shouldn't be) dependent upon mere chance. Third, the counterfactual account is compatible with the idea that changes in people's deontic statuses (i.e. in their packages of rights, duties, entitlements, etc.) may be harmful per se, regardless of their concrete causal upshot. This is highly desirable, at least insofar as we want to stay true to the idea that depriving people of certain rights is to harm them—and this is so even if they had not exercised those rights in the past and would not have done so in the future.

One might worry that the counterfactual view makes harm constitution empirically undetectable. We can give this concern its due. Under a causal account, given a certain norm-enacting utterance and a certain *actual* harm, a causal connection is hypothesized between the utterance and the harm. Under a counterfactual account, given a certain norm-enacting utterance and a certain *actual or potential* harm, a causal connection is hypothesized between the former and the latter. If proving the hypothesized causal connection in McGowan's approach is already hard, proving it within a counterfactual framework might be even harder—for it would require us to "go and see" (as it were) not only how things are but also how they could be.

Let us stress, however, that the notion of constitution has been introduced in the debate on speech and harm precisely with the aim of capturing harms, and subtle forms of injustice, which may not be immediately empirically visible.[5] In taking into account both actual and potential harms, the counterfactual view aligns with that aim. Notice, moreover, that in employing counterfactual reasoning to determine what constitutes harm, we are following the very same practice adopted in many legal systems to determine whether newly enacted laws are discriminatory or otherwise unconstitutional. Western legal systems currently rely on two basic models of constitutional review of statutes: the 'con-

[4] We treat (ii) as being compatible with the truth of the antecedent. Our definition thus broadens the range of cases captured by McGowan's, while keeping track of everything her definition does. It takes into account cases where the norm is followed and harm actually results from following it, as well as cases where the norm is not followed, but had it been followed, harm would have resulted from following it.

[5] The notion dates back to MacKinnon's writings on the harms of pornography. See, esp., MacKinnon, C. 1987, *Feminism Unmodified: Discourses on Life and Law*, Cambridge, MA: Harvard University Press; and MacKinnon, C. 1993, *Only Words*, cit. at fn. 1. Harm constitution claims against pornography have been taken to have a dialectical advantage over harm causation claims, for they sidestep questions about the lack of conclusive evidence in support of a causal link between pornography consumption and sexual violence. See Mikkola, M. 2019, *Pornography: A Philosophical Introduction*, Oxford: Oxford University Press, esp. ch. 2.

crete' model and the 'abstract' model. In the *concrete* model, mainly adopted in the US, the review is activated by a claim that the enforcement of an (allegedly) unconstitutional law caused a real person—one of the litigants—actual injury. By contrast, in the *abstract* model—adopted in European countries such as Germany, Austria, Spain, and others—the review can be carried out in the absence of litigation, regardless of, and even prior to, the application of the statute in question. Under the abstract model, certain political actors (usually including opposition legislators) can challenge a statute—e.g. on discrimination grounds—right after its enactment in Parliament and prior to its application. When this happens, in order to ascertain whether the challenged statute is indeed discriminatory, the constitutional court cannot look at whether it has caused any actual discrimination against real people (since the statute has never been applied), but will look at whether it *would* do so, if applied. Abstract review

> proceeds in the absence of litigation: the judge reads the legislative text against the constitutional law and then decides. There is no storyline or, if there is, the story is an imaginary or hypothetical one told to highlight the constitutional moral that comes at the end.[6]

That is to say that the (constitutional court) judge will engage in counterfactual reasoning to determine whether or not the statute in question constitutes harm—e.g. the harm of discrimination.

As one can see, McGowan's causal account and the counterfactual account of (harm) constitution reflect the competing intuitions at the roots of the concrete model and the abstract model of constitutional review. We do not aim to settle which model is to be preferred (we leave this question to legal scholars). What we want to emphasize is that the abstract model of review faces the same empirical difficulties as a counterfactual account of constitution; such difficulties, however, do not stall the legal process, nor are they generally considered sufficient to abandon the model in favor of concrete review.

Before concluding, note that McGowan grants in a footnote that

> One might be tempted to say that certain norms are such that the mere enacting of them is harmful. Consider, for example, the employer's verbal enacting of the discriminatory hiring policy. Even if a discriminatory hiring practice does not result from the enacting of this policy […], that policy in place might be harmful in a counterfactual way. […] Although I here concentrate on cases where actual harm ensues, I leave this possibility open (24, fn. 42).

Our point has been to show that we should not just leave that possibility open, but opt for a counterfactual view on constitution, for it gives us better tools to capture what we intuitively consider as harmful—and perhaps want to consider as such for our legitimate political purposes. The direction in which we suggest to shift McGowan's account retains the core tenets of her framework: we entirely agree that the normative environment we navigate is continuously, and often implicitly, adjusted by the things we say. We also agree that offhand bigoted remarks may (and often do) contribute to structural injustice. McGowan's book provides an exceptionally rich and powerful machinery to unpack the mecha-

[6] Stone Sweet, A. 2003, "Why Europe Rejected American Judicial Review—And Why It May Not Matter", *Michigan Law Review*, 108, 2771.

nisms by which this happens. Unlike McGowan, however, we do not think that actual disadvantages must follow for a norm-enacting utterance to constitute harm.

Vita-Salute San Raffaele University LAURA CAPONETTO
Vita-Salute San Raffaele University BIANCA CEPOLLARO

Advisory Board